RATIONAL AGENTS RESIST ENTROPY: BOUNDED RATIONALITY AS THE PRINCIPLED OBJECTIVE FOR RL UNDER FINITE SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning optimises expected return without explicitly constraining representation complexity, which can make learning statistically difficult under finite samples. We propose bounded-rational RL: augment the return with a penalty on the differential entropy of the agent's noisy representations to control statistical complexity. We justify this objective by proving, under explicit conditions, an entropy-dependent bound on the samples needed to evaluate nearby policy changes. Analysing the policy gradient of this objective provides a behavioral characterisation: the complexity penalty supplies an intrinsic reward for low-surprisal representations and can drive the agent to maintain recurring, predictable situations—entropy resistance. This behavioral characterisation distinguishes RL from passive supervised learning and connects bounded-rational RL to active inference and related frameworks that also produce entropy-resisting behavior from different theoretical foundations. Experiments across partially observable control tasks show both: bounded rationality improves mean returns over unconstrained learning, and agents resist entropy, with larger gains in the more demanding task variants studied.

1 INTRODUCTION

Reinforcement learning maximizes expected return without explicitly controlling representation complexity. In history-dependent tasks, the space of possible observation histories grows exponentially with the horizon. Finite experience may have to support different decisions across this space, making learning statistically difficult. Even when a simpler controller achieves nearly the same return, the return objective provides no explicit reason to prefer it.

In supervised learning, structural risk minimisation (SRM) (Vapnik, 2013) addresses a related trade-off between performance and complexity. Bounded rationality studies decision-making under limited resources (Simon, 1955), with information-theoretic formulations trading utility against information-processing costs (Ortega et al., 2015). These principles suggest that an RL objective should account for the complexity of the behavior an agent must learn. We apply this principle to the distribution of learned representations visited by a controller.

We propose *bounded-rational RL*: augment expected return with a penalty on the entropy of the representations encountered during a task. The policy acts on these representations, so reducing their complexity affects control. This creates a trade-off familiar from SRM: a small loss in attainable return can bring a substantial reduction in statistical difficulty. Our theoretical analysis justifies this objective by connecting representation entropy to statistical generalization. Under stated conditions, we prove an entropy-dependent bound on the samples needed to evaluate small policy changes. This supports favoring nearly optimal policies with simpler representations, when available.

Representation entropy can fall when the encoder maps different histories to similar representations. Actions provide a second route: they can make task-relevant situations recur. Because actions shape future observations, the complexity penalty can favor actively maintaining a predictable regime, even without an explicit reward for stabilization. Maintaining that regime may require continual corrective actions. We call this behavior *entropy resistance*. Our policy-gradient theorem explains this behavior: the penalty supplies an intrinsic reward favoring representations common in the agent's experience,

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

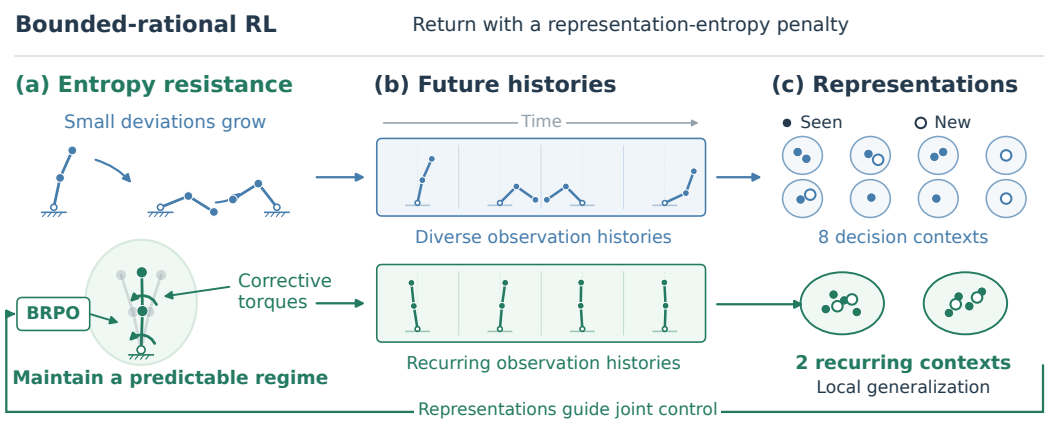

Figure 1: **Entropy resistance connects control and learning.** Conceptual illustration. (a) One controller jointly produces both corrective torques to maintain an upright regime. (b) Actions shape the observation histories encountered next. (c) Those histories map to learned representations: filled markers denote collected experience and open markers fresh visits. Recurring contexts allow experience to be reused.

allowing future representation costs to guide earlier actions. Figure 1 illustrates the connection between entropy resistance, representation complexity, and experience reuse.

In passive supervised learning, the learner receives fixed inputs; an RL agent can act to change the situations it will face. Related behavior appears in active inference, where agents use an internal model to select actions that shape their sensory experience (Friston, 2010; Friston et al., 2015). Here this behavior follows from the finite-sample trade-off between return and representation complexity.

Our contributions are threefold. First, we introduce a bounded-rational RL objective that balances return against representation entropy, following the principle of SRM. Second, we establish its statistical justification and explain entropy resistance through a policy-gradient theorem. Third, we develop Bounded-Rational Policy Optimization (BRPO), with experiments across partially observable control tasks showing improved mean returns, larger gains in the more demanding variants studied, and agents acting to sustain a predictable regime in the double-pendulum task.

2 Related Work

Regularized control and bounded rationality. Regularized control in Markov decision processes (MDPs) analyzes policy entropy and divergence penalties through Bellman operators (Neu et al., 2017; Geist et al., 2019). Information-theoretic bounded rationality trades utility against processing costs (Ortega et al., 2015); mutual-information regularized actor–critic penalizes state–action information (Leibfried & Grau-Moya, 2020). Our cost measures the distribution of noisy representations encountered throughout an episode. Both the encoder and actions shape this distribution. We connect this distribution to learning from limited data and the objective's gradient.

Representation and learning-target compression. Rate-distortion RL balances learning-target or model complexity against performance loss (Arumugam & Van Roy, 2021; 2022). Information bottlenecks compress inputs while preserving relevant information (Tishby et al., 1999; Alemi et al., 2017). Our objective measures complexity in learned representations; return determines which distinctions are worth preserving. Predictive-state representations summarize histories for future prediction (Littman et al., 2001). Approximate information states (AIS) learn summaries that preserve reward and predictive information (Subramanian et al., 2022). Finite-sample RL theory studies learning under assumptions on states and observations (Strehl et al., 2009; Jin et al., 2020). Our bounds relate the evaluation of nearby policy changes to the distribution of representations encountered, without assuming that the encoder preserves all information needed for optimal control.

Predictable behavior. Maximum-entropy RL rewards conditional action entropy (Haarnoja et al., 2018); our objective penalizes pooled noisy representation entropy. SMiRL rewards low-surprise observations and produces stable, repeatable behavior (Berseth et al., 2021), providing a direct precedent for entropy resistance. We connect a learned representation cost to local statistical complexity and distinguish changes to the encoder from changes to the experience produced by actions.

Active inference. Active inference connects inference about hidden states to action that shapes sensory experience. Variational free energy bounds sensory surprise under a generative model (Friston, 2010). Expected free energy evaluates policies through preferred outcomes and information gain (Friston et al., 2015), supporting both maintaining desirable conditions and seeking informative observations. Bounded-rational RL derives entropy resistance from balancing return against representation complexity under finite samples. This gives agents a statistical reason to maintain conditions that simplify learning.

3 PRELIMINARIES

We consider partially observed episodes with T decisions and $T+1$ observations. At decision $t = 0, \dots, T-1$, the agent observes X_t and chooses A_t from the action space $\mathcal{A}$, using the action-observation history $H_t = (X_0, A_0, \dots, A_{t-1}, X_t)$. The environment then generates X_{t+1}, whose distribution may depend on H_t and A_t. Each reward is a known function of the resulting history: $R_t = r_t(H_{t+1})$.

A history-dependent policy specifies $\pi_t(\cdot \mid H_t)$. Let $\tau = (X_0, A_0, \dots, A_{T-1}, X_T)$ denote a complete episode and p^π its distribution. Standard RL maximizes the expected return $J(\pi) = \mathbb{E}_{p^\pi} \sum_{t=0}^{T-1} R_t$.

4 THE BOUNDED-RATIONAL OBJECTIVE

We use a recurrent encoder with parameters ϕ to represent the history: $Z_t = f_\phi(Z_{t-1}, A_{t-1}, X_t) = \phi_t(H_t)$, with $Z_{-1} = 0$ and a fixed start symbol A_{-1}. The deterministic recurrence is called *unifilar* (Shalizi & Crutchfield, 2001); it makes the representation a function of the observed history.

Fixed Gaussian input noise limits the policy head's ability to distinguish nearby representations:

$$Z_{\sigma,t} = Z_t + \sigma\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I_{d_Z}), \qquad A_t \sim \pi_{\theta,t}(\cdot \mid Z_{\sigma,t}).$$

With representation dimension d_Z and head parameters θ, fix $\sigma > 0$ and sample ξ_t independently at every step. The head may use t, or apply the same π_θ at every step. Together, the encoder and head form a controller $w = (\theta, \phi)$, whose history-dependent action probabilities are $w_t(a \mid H_t) = \mathbb{E}_{\xi_t} \pi_{\theta,t}(a \mid \phi_t(H_t) + \sigma\xi_t)$. Let $J(w)$ denote its return.

To measure complexity across the whole episode, sample U uniformly from $\{0, \dots, T-1\}$, independently of the trajectory. Write $\widetilde{Z} = Z_U$ and $\widetilde{Z}_\sigma = \widetilde{Z} + \sigma\xi$, where $\xi \sim \mathcal{N}(0, I_{d_Z})$ is fresh and independent. This pools representations across decision times. We call the resulting law the representation occupancy distribution: it describes how frequently each representation is encountered. Using natural logarithms, differential entropy is $h(W) = -\mathbb{E} \log p_W(W)$, where p_W is the density of W.

Following SRM, the *bounded-rational objective* penalizes the entropy of this overall distribution:

$$\max_{w=(\theta,\phi)} \underbrace{J(w)}_{\text{return}} - \lambda \cdot \underbrace{h(\widetilde{Z}_\sigma)}_{\text{representation entropy}}. \tag{1}$$

The weight $\lambda \geq 0$ sets the trade-off between return and representation entropy. We call its excess over the fixed-noise entropy *normalized representation complexity*. At fixed noise scale and representation dimension, $C(w) = h(\widetilde{Z}_\sigma) - h(\sigma\xi) = I(\widetilde{Z}; \widetilde{Z}_\sigma) \geq 0$, where I is mutual information (Cover & Thomas, 2006). Subtracting this constant noise entropy leaves the maximizers unchanged and gives the normalized objective $F_\lambda(w) = J(w) - \lambda C(w)$ used below.

Assumption 4.1 (Bounded encoder output)**.** For a fixed radius $M > 0$, the encoder output satisfies $\|Z_t\| \leq M$ almost surely at every decision step.

The bound applies to Z_t before Gaussian noise is added.

5 STATISTICAL JUSTIFICATION

With limited experience, a policy update can look promising because it fits the particular histories collected so far. We ask how the entropy penalty helps control this statistical difficulty. We evaluate nearby policy changes using current-policy episodes, reweighted by their relative probabilities under the candidate and current policies (Appendix B). We use the normalized representation complexity $C(w)$ and objective $F_\lambda(w)$ defined in Section 4. A local update makes small changes to the current controller's action probabilities and encoder outputs.

Why the objective favors low complexity. A nearly optimal policy may need far fewer distinctions between histories than an unconstrained optimum. The entropy penalty makes that policy attractive when its reduction in complexity outweighs its loss in return.

Proposition 5.1 (Preference for a nearly optimal, low-complexity policy)**.** *Let J^* be the best return in the controller class and $\lambda > 0$. Suppose a policy w_s loses at most $\beta_s \geq 0$ in return, $J^* - J(w_s) \leq \beta_s$, and has complexity $C(w_s) \leq c_s$, where $c_s \geq 0$. Any policy w with $F_\lambda(w) \geq F_\lambda(w_s)$ satisfies*

$$C(w) \leq c_s + \frac{\beta_s}{\lambda}, \qquad J^* - J(w) \leq \beta_s + \lambda c_s.$$

Thus matching the simpler policy's objective limits both representation complexity and return loss. Appendix B.1 also covers approximate optimization and subsequent local updates.

How entropy controls statistical complexity. When many histories lead to similar representations, their experience can be reused to evaluate decisions. The following result formalizes this connection.

Theorem 5.1 (Learning controlled by representation entropy)**.** *Consider a policy with normalized representation complexity $C(w) \leq c$, where $c \geq 0$, in a finite-action task with a bounded complete-episode return. Suppose rarely visited representations do not dominate estimation error, policy and encoder changes are small enough over the complete episode, and the entropy penalty is estimated accurately enough, as specified in Appendix B. For a candidate set fixed before sampling, the empirical evaluation rule selects an update whose F_λ is within $\varepsilon > 0$ of the best candidate's, with probability at least $1 - \delta$, where $\delta \in (0, 1)$. With other parameters fixed, a sufficient number of independent episodes is*

$$N = O\left(\frac{T^3 e^{O(c)} + \log(1/\delta)}{\varepsilon^2}\right).$$

The horizon factor accounts for evaluating changes over a complete trajectory. The entropy term controls the complexity of the representations encountered along it. Low entropy can therefore reduce the sample requirement even in a long task. If entropy grows linearly with the horizon, the same bound can become exponential in the horizon. The benefit depends on maintaining low entropy and the stated conditions. Appendix A.1 isolates the entropy-dependent bound for evaluating bounded functions of representations; Appendix B.2 gives a sharper complete-objective bound under stronger update and distribution conditions.

Comparing policies with different entropy. The following construction shows how the objective can favor a policy whose nearby updates are easier to evaluate. Both policies are compared in the same environment using the same evaluation rule.

Theorem 5.2 (Contrasting local generalization in one task)**.** *In a common controller class, there are partially observed tasks with a return-optimal policy w_h of complexity $c_h = C(w_h)$ and a policy w_s with zero complexity whose return can be arbitrarily close to optimal. The bounded-rational objective with $\lambda = 1$ favors w_s. For matched small changes to the final action rule, evaluating all such changes to a fixed small error tolerance and confidence requires $e^{\Omega(c_h)}$ episodes around w_h under the estimator in Appendix C. Around w_s, a sufficient episode count is horizon-independent.*

These results justify preferring nearly optimal policies with simple representation distributions.

6 ENTROPY RESISTANCE

The bounded-rational objective controls complexity through both representation and action. Actions shape this experience: they determine which situations occur next and can reduce the complexity

of later decisions by keeping future experience predictable. SMiRL previously demonstrated that minimizing observation surprise can produce stable, repeatable behavior (Berseth et al., 2021). The objective penalizes representation entropy; actions that maintain recurring, task-relevant situations can reduce it. The gradient with respect to policy and encoder explains this effect.

Write p_w^{occ} for the density of $\widetilde{Z}_\sigma$ under the current controller. The cost assigned to a representation z is $u_w(z) = -\mathbb{E}_\xi \log p_w^{\mathrm{occ}}(z + \sigma\xi)$. Subtract this cost from each reward: $\widetilde{R}_t = R_t - \lambda u_w(Z_t)/T$, and let $G_{\lambda,t} = \sum_{s=t}^{T-1} \widetilde{R}_s$ be the resulting return-to-go, summing augmented rewards from t onward.

Theorem 6.1 (Policy gradient of the bounded-rational objective)**.** *Under the differentiability and integrability conditions in Appendix D, the gradient with respect to all controller parameters $w = (\theta, \phi)$ is*

$$\nabla_w F_\lambda(w) = \mathbb{E}_w \sum_{t=0}^{T-1} \left[G_{\lambda,t} \nabla_w \log w_t(A_t \mid H_t) - \frac{\lambda}{T} (D_w Z_t)^\top \nabla_z u_w(Z_t) \right]. \tag{2}$$

Here $D_w Z_t$ is the Jacobian of the encoder output with respect to controller parameters, holding the observed history fixed. The spatial derivative $\nabla_z u_w$ holds the current density fixed; the effect of changing that density is accounted for in the derivation.

The first term updates action probabilities using future rewards and representation costs; both head and encoder can affect these probabilities. The second directly changes how the encoder represents recorded histories. The terms separate action selection from direct representation changes.

These gradients can favor continual corrective actions that keep the agent in a predictable regime, even without an explicit stabilization reward. This is *entropy resistance*. The agent thus shapes the situations it must learn to handle, linking behavior to the statistical benefit in Section 5.

7 Bounded-Rational Policy Optimization

The exact gradient requires the current controller's representation density. BRPO estimates these costs with a diffusion model, alternating between fitting the model and updating the controller. Each iteration collects episodes under the old controller w_0, fits a diffusion model to their pooled representations, and uses its estimated costs in a Proximal Policy Optimization (PPO) update (Schulman et al., 2017). BRPO subtracts representation costs from rewards and directly penalizes the encoder.

Diffusion models and the ELBO. A diffusion model gradually adds Gaussian noise to data $B \in \mathbb{R}^m$. For a variance-preserving (VP) diffusion, $D_\tau = \alpha_\tau B + \sigma_\tau \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_m)$ and $\alpha_\tau^2 + \sigma_\tau^2 = 1$. The noise rate $\beta(\tau) > 0$ sets $\alpha_\tau = \exp[-\frac{1}{2}\int_0^\tau \beta(v)dv]$. Diffusion time τ indexes the artificial noising process and is distinct from the environment's decision time t. A score network $s_\psi(x, \tau)$ learns $\nabla_x \log p_\tau(x)$, where p_τ is the density of D_τ. It is trained with the Gaussian conditional score $-\epsilon/\sigma_\tau$ as its target (Song et al., 2021b). The learned reverse process starts from a standard Gaussian and provides a density model $m_\psi(y)$. Its evidence lower bound (ELBO) is $\mathcal{E}_\psi(y) = \mathbb{E}_{Q(\mathbf{d}|y)} \log[m_\psi(y, \mathbf{d})/Q(\mathbf{d} \mid y)] \le \log m_\psi(y)$, where $\mathbf{d}$ is the noising path, Q its forward law, and $m_\psi(y, \mathbf{d})$ the joint model density. The negative ELBO combines a weighted denoising loss and a terminal contribution (Song et al., 2021a); Appendix E.1 gives the continuous-time formulas.

From the diffusion ELBO to representation entropy. Denoising score matching estimates density scores from noisy samples (Vincent, 2011). VP diffusion supplies the noising process (Song et al., 2021b), and variational diffusion models connect denoising objectives to likelihood bounds (Kingma et al., 2021). We apply these tools to controller representations, accounting for residual error at the final diffusion time and error in the learned score. We apply the VP diffusion to the pooled clean representation $\widetilde{Z}$. Choose a starting diffusion time τ_0 with $\sigma_{\tau_0}/\alpha_{\tau_0} = \sigma$, so the modeled data are $\widetilde{Z}_\sigma = D_{\tau_0}/\alpha_{\tau_0}$. The score network receives diffusion time, with no decision-time input. Write $\ell_\psi(z) = -\mathbb{E}_\xi \mathcal{E}_\psi(z + \sigma\xi)$ for the noise-averaged negative ELBO. For terminal diffusion time $\mathcal{T}$ and $m = d_Z$, BRPO uses

$$u_\psi(z) = \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta(\tau)\, \mathbb{E}_\epsilon \left[\left\| s_\psi(\alpha_\tau z + \sigma_\tau \epsilon, \tau) + \frac{\epsilon}{\sigma_\tau} \right\|^2 - \frac{m}{\sigma_\tau^2} \right] d\tau. \tag{3}$$

Minimizing its batch average trains the score network; sampling τ and ϵ estimates the integral and each representation's cost. The subtracted term is constant across representations.

Proposition 7.1 (Representation entropy and the diffusion ELBO)**.** *Under Appendix E.1's regularity conditions,*

$$\mathbb{E}\ell_\psi(\widetilde{Z}) - h(\widetilde{Z}_\sigma) = e_{\text{diff}} \geq 0,$$
$$\mathbb{E}u_\psi(\widetilde{Z}) + K_{\text{VP}} - h(\widetilde{Z}_\sigma) = g_\psi(w) = e_{\text{diff}} + \frac{\alpha_{\mathcal{T}}^2}{2}(M^2 - \mathbb{E}\|\widetilde{Z}\|^2) \geq 0. \tag{4}$$

Here e_{diff} measures terminal-distribution mismatch and integrated score error, and K_{VP} depends only on the fixed noise, schedule, dimension, and encoder bound.

Thus the average denoising cost, up to a fixed constant, upper bounds entropy. The gap $g_\psi(w)$ measures this overestimate at controller w. Fitting the score model reduces the score-error part of the gap; Appendix E.1 gives each term explicitly.

Updating policy and encoder. With the score network fixed, form augmented rewards $\widetilde{R}_t = R_t - \lambda u_\psi(Z_t)/T$. A critic V_ω predicts the remaining augmented return; advantages $\widetilde{A}_t^0$ estimate an action's gain over this prediction. Write $\phi_{0,t}$ for the old encoder and $\phi_{q,t}$ for a candidate's encoder. For a candidate controller $q = (\theta_q, \phi_q)$, use the action ratio $\rho_t = q_t(A_t \mid H_t)/w_{0,t}(A_t \mid H_t)$. For a batch $\mathcal{B}$, let $\widehat{U}^{\text{dir}}_{\psi,\mathcal{B}}(\phi)$ be the mean of $u_\psi(\phi_t(H_t))$ over episodes and decisions, differentiating with recorded histories fixed. The BRPO update maximizes

$$\widehat{L}_{\text{BRPO}}(q) = \widehat{\mathbb{E}}_{\mathcal{B}} \min\{\rho_t \widetilde{A}_t^0, \text{clip}(\rho_t, 1 \pm \epsilon_{\text{clip}})\widetilde{A}_t^0\} - \lambda \widehat{U}^{\text{dir}}_{\psi,\mathcal{B}}(\phi_q). \tag{5}$$

Here $\text{clip}(\rho, 1 \pm \epsilon_{\text{clip}})$ truncates ρ to $[1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}]$, $\epsilon_{\text{clip}} > 0$ is the clipping threshold, and $\widehat{\mathbb{E}}_{\mathcal{B}}$ averages episode sums. The advantages carry future representation costs to earlier actions; the direct term updates the encoder. At w_0, with exact expectations, the surrogate has the gradient of the objective with fixed diffusion costs, including both effects in Theorem 6.1. Algorithm 1 in Appendix F summarizes the training loop. Appendix F gives the batch normalizations and action-noise implementation. In the ablations, λ_1 weights the direct cost and λ_2 the cost subtracted from rewards. We refer to these as the direct encoder penalty and the reward-side penalty, respectively.

Theorem 7.1 (BRPO improvement with controlled approximation error)**.** *Fix the old controller and score model. Suppose a candidate q keeps the same actions possible, changes action probabilities by at most α in total variation at every history, and satisfies $|u_\psi(\phi_{q,t}(H)) - u_\psi(\phi_{0,t}(H))| \leq e_u$ at every decision and history. Total variation measures the largest probability difference assigned to an action event. Let A_{aug} bound the absolute true augmented advantage. If the empirical surrogate gain $\widehat{\Delta}L = \widehat{L}_{\text{BRPO}}(q) - \widehat{L}_{\text{BRPO}}(w_0)$ differs by at most e_{alg} from the true expected surrogate gain, then*

$$F_\lambda(q) - F_\lambda(w_0) \geq \widehat{\Delta}L - \lambda g_\psi(w_0) - e_{\text{alg}} - 2A_{\text{aug}}T(T-1)\alpha^2 - \lambda e_u(T-1)\alpha. \tag{6}$$

A surrogate gain exceeding these errors improves the bounded-rational objective. The bound covers the ELBO gap, sampled costs and advantages, and changes in visited histories (proof in Appendix F).

8 Experiments

With finite experience, an agent may encounter more situations than it can learn to handle reliably. BRPO seeks to ease this burden by simplifying how it represents experience and acting to keep future experience simple. Across the history-dependent benchmarks below, this objective yields higher mean returns than PPO, with larger gains in the more demanding settings. The double pendulum tests the predicted behavioral effect: BRPO maintains a predictable regime after the explicit stabilization reward is removed. The ablations show how encoder and action-selection updates simplify experience.

8.1 Performance with Limited Experience

We evaluate RockSample (Smith & Simmons, 2004), NoisyTV, LaserTag (Somani et al., 2013), and stacked LightDark (Platt et al., 2010). Our NoisyTV variant uses stochastic distractors inspired by

Burda et al. (2019). We compare BRPO with PPO–RNN (Schulman et al., 2017), AIS (Subramanian et al., 2022), and the planners AdaOPS (Wu et al., 2021), POMCP (Silver & Veness, 2010), and POMCPOW (Sunberg & Kochenderfer, 2018). AIS is trained with PPO and learns predictive history summaries; PPO–RNN sets both BRPO coefficients to zero.

These tasks require different uses of memory: RockSample tracks hidden rock qualities, NoisyTV retains an initial goal cue amid random distractors, LaserTag tracks an unseen moving opponent, and LightDark balances reaching the goal with position-dependent sensing. The planning baselines use the exact environment generative model; learning methods learn from sampled interactions.

Table 1: **Greedy return** (mean $\pm$ std over three seeds). Bold marks the highest mean per task.

Method	RockSample 15×15	NoisyTV	LaserTag	LightDark 2D
PPO-RNN	13.48 ± 2.47	0.2960 ± 0.0419	-10.568 ± 0.247	9.32 ± 0.09
AdaOPS	15.50 ± 0.22	$\mathbf{0.4394 \pm 0.0000}$	-19.249 ± 0.086	-4.40 ± 0.08
POMCP	21.27 ± 1.39	0.3990 ± 0.0031	-15.744 ± 0.412	6.57 ± 0.34
POMCPOW	21.47 ± 0.17	0.4013 ± 0.0048	-16.032 ± 0.569	6.13 ± 0.33
AIS	5.13 ± 3.44	-0.2862 ± 0.0000	-19.882 ± 0.000	-1.73 ± 0.65
BRPO (ours)	$\mathbf{43.13 \pm 1.55}$	$\mathbf{0.4394 \pm 0.0000}$	$\mathbf{-10.367 \pm 0.289}$	$\mathbf{9.38 \pm 0.10}$

Across these different memory demands, BRPO has the highest mean return on all four tasks, tying AdaOPS on NoisyTV (Table 1). The largest gains over PPO occur in RockSample and NoisyTV; gains in LaserTag and LightDark are smaller. Entropy regularization can therefore improve performance within a finite interaction budget, with its benefit varying across tasks. The small differences in LaserTag and LightDark are descriptive comparisons of means; no significance test is reported. Appendix M.1 gives budgets and aggregation rules; environment appendices give task and training details.

8.2 Scaling with Environment Complexity

The need to use limited experience effectively becomes more pressing as the task grows. Figure 2 compares BRPO and PPO–RNN as RockSample maps grow from 7×7 to 19×19 and LightDark stacks $N \in \{2, 4, 8\}$ coordinates, with horizon $12N$. At fixed training budgets, the mean RockSample return gap grows from 11.7 on the smallest map to 37.4 on the largest. For LightDark, the per-coordinate gap grows from 0.027 at two coordinates to 1.338 at eight. Normalizing by coordinate count separates the observed gain from simply summing rewards over more coordinates. The advantage increases with task size in these two families. Task size and horizon indicate difficulty here; they do not directly measure the representation entropy in our bound. Appendix M.2 gives evaluation details.

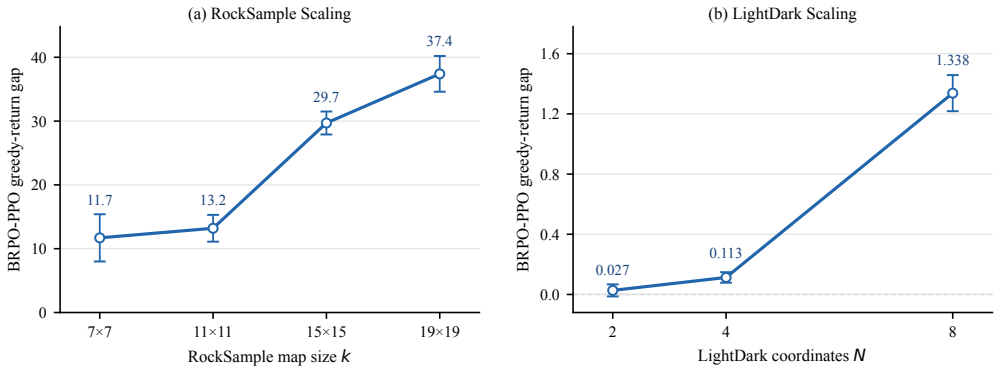

Figure 2: **Scaling with environment complexity.** BRPO–PPO greedy-return gaps: (a) RockSample after 10,000 training iterations; (b) LightDark undiscounted return per coordinate after 1,000 iterations. Points show means over three seeds.

8.3 Entropy Resistance

The double pendulum tests whether continual corrective actions sustain an upright regime. PPO and BRPO share a stabilizing controller adapted to observation noise, then train only on negative full-state prediction error; upright, survival, and fall-related rewards are removed. The policy observes noisy angles with angular velocities hidden, chooses two joint torques, and predicts the four-dimensional state. Reward is the negative squared prediction error; upright fraction never enters it. Appendix K gives the protocol.

Upright fraction counts evaluation steps with both joint angles within 0.1 radians of upright. In Figure 3, both methods start near 0.53; PPO ends at 0.438 and BRPO at 0.496. With initialization and prediction objectives shared, BRPO spends more time upright, supporting the behavioral route to simpler experience (Section 6): corrective actions preserve the regime the agent must predict and control.

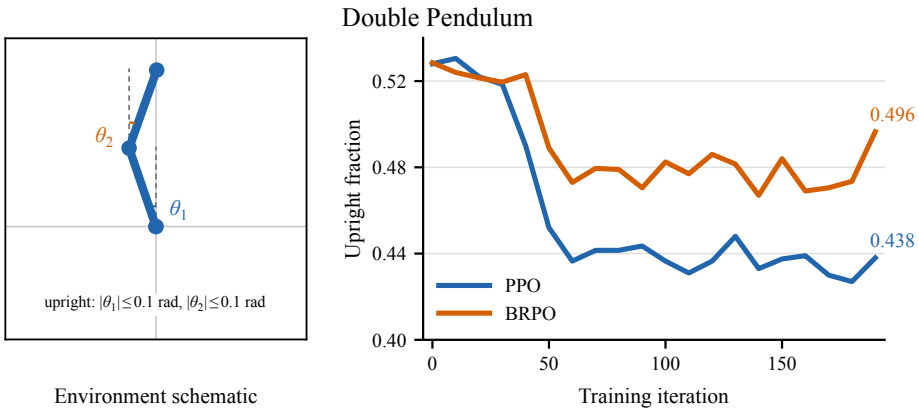

Figure 3: **Entropy resistance in the double pendulum.** Left: partially observed environment and upright region, $|\theta_1|, |\theta_2| \leq 0.1$ radians, with angles measured from upright. Right: upright fraction during prediction-only training from a shared checkpoint; BRPO spends more of the episode upright.

8.4 Mechanism Ablations

Simplifying experience involves both the distinctions retained by the encoder and the situations reached through action. The 2×2 ablations in Table 2 separate the direct encoder penalty (λ_1) from the reward-side penalty (λ_2). Stochastic Drift Arena has a hidden, changing drift and a low-noise region. Reward favors opposing the drift; occupying that region has no direct reward. Here λ_1 alone slightly reduces representation entropy with little behavioral change, while λ_2 substantially increases low-noise-region occupancy. NoisyTVHard requires remembering a one-step goal cue amid stochastic visual distractors; only configurations with λ_1 use the cue to select the appropriate goal. Thus actions maintain predictable surroundings and the encoder retains useful information, with task-dependent effects. Appendix M.3 gives full tables and protocols.

Representation entropy versus standard penalties. The form of the penalty matters: Kullback–Leibler (KL) divergence to a prior encourages matching a chosen distribution, while weight decay constrains parameter magnitude. On RockSample 15×15, representation entropy yields the highest return in this comparison (Table 3). All methods use 8,000 iterations and the final five greedy evaluations. Here q_Z denotes the visited representation density and p_Z the prior; see Appendix M.6.

Stability during optimization. The theory concerns local changes to policies and encoders. To connect this setting to training, on RockSample 15×15 we record policy change and representation drift before and after each PPO update on the same histories. Table 4 summarizes post-burn-in updates across three seeds. In these runs, 96.5% of updates have policy KL below 0.02, alongside small measured encoder changes and changes in score outputs caused by encoder drift. These sampled

Table 2: **Direct encoder and reward-side penalties.** Greedy means ± std: Drift uses five matched seeds; NoisyTVHard uses three seeds and a discounted return. p_B is the fraction of post-burn-in evaluation steps in the low-noise region B; $h(Z_\sigma)$ is the reported noisy representation entropy and may be negative. Active coefficients are 0.5 for Drift and 3×10^{-5} for NoisyTVHard.

	Stochastic Drift Arena			NoisyTVHard
Configuration	Task return	p_B	$h(Z_\sigma)$	Discounted return
Off	5.63 ± 0.24	0.203 ± 0.140	1.03 ± 0.79	-0.1407 ± 0.0393
λ_1 only	5.61 ± 0.24	0.199 ± 0.140	0.87 ± 0.79	0.3340 ± 0.0000
λ_2 only	5.94 ± 0.16	0.657 ± 0.192	-4.27 ± 2.67	-0.1732 ± 0.0365
Both	5.94 ± 0.16	0.659 ± 0.192	-4.41 ± 2.67	0.3340 ± 0.0000

Table 3: **Entropy and alternative penalties on RockSample** 15×15**.**

Method	Penalty	Greedy return
PPO-RNN	None	15.20
KL-to-prior	$D_{\mathrm{KL}}(q_Z \| p_Z)$	10.00
Weight decay	$\|\theta\|_2^2$	10.00
BRPO (ours)	$\lambda h(\widetilde{Z}_\sigma)$	**44.00**

diagnostics document local stability in these runs; they do not verify the uniform conditions of the statistical and improvement theorems. Appendix M.4 reports these diagnostics.

Table 4: **Policy displacement and latent representation drift on RockSample** 15×15. Statistics are computed over post-burn-in PPO updates pooled across three seeds.

Metric	Median
Realised policy KL	8.25×10^{-3}
Latent RMS drift	4.42×10^{-3}
Diffusion-score RMS drift	2.45×10^{-3}

An additional recurrent–Transformer comparison across horizons is reported in Appendix M.5, with the complete results and its single-seed evaluation protocol.

9 Discussion and Conclusion

Bounded-rational RL balances return against representation entropy, giving agents a reason to simplify the situations they must learn to handle. Under the stated conditions, representation entropy controls local statistical complexity. BRPO implements encoder and action updates; experiments show improved returns and agents sustaining predictable conditions.

The encoder and the policy act on the same source of complexity from two directions. The encoder can map different histories to similar representations; actions can make future histories more regular in the first place. Entropy resistance provides this behavioral route: maintaining a predictable regime gives the agent recurring situations in which experience can be reused. The return term determines which distinctions remain worth preserving and which regimes remain useful.

Limitations. Our statistical theory concerns learning near a policy with simple representations. Whether BRPO reliably discovers such a policy with high return remains open. Establishing these statistical conditions throughout training requires further work. Evidence for entropy resistance comes from selected environments and needs broader validation.

AI USE STATEMENT

The core research idea and its formalization, the choice and framing of the research problem, the theorem framework, the narrative structure, and the experimental design were developed entirely by the human authors. Generative AI tools did not contribute to research ideation or determine the scientific direction of this work.

Under the authors' direction, generative AI tools were used extensively during research execution and manuscript preparation. Specifically, they assisted with drafting and revising the paper, developing proofs within the human-designed theorem framework, implementing experimental code according to the human-designed methodology, and surveying relevant literature.

The authors critically reviewed all AI-assisted outputs, independently verified the mathematical arguments and proofs, inspected and tested the experimental code, validated the reported results, and manually checked the existence, relevance, and accuracy of cited references. The authors take full responsibility for the originality, correctness, claims, citations, and final content of this work.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning (ICML)*, pp. 373–382, 2021.

Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Glen Berseth, Daniel Geng, Coline Manon Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. SMiRL: Surprise minimizing reinforcement learning in unstable environments. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 2nd edition, 2006.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 2160–2169, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, volume 80, pp. 1861–1870, 2018.

Chi Jin, Sham M. Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18530–18539, 2020.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 21696–21707, 2021.

Felix Leibfried and Jordi Grau-Moya. Mutual-information regularization in Markov decision processes and actor-critic learning. In *Conference on Robot Learning (CoRL)*, volume 100, pp. 360–373, 2020.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, pp. 1555–1561, 2001.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Pedro A. Ortega, Daniel A. Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

Robert Platt, Russ Tedrake, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Belief space planning assuming maximum likelihood observations. In *Robotics: Science and Systems (RSS)*, 2010.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

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

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3–4):817–879, 2001.

David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, pp. 2164–2172, 2010.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955.

Trey Smith and Reid G. Simmons. Heuristic search value iteration for POMDPs. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 520–527, 2004.

Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. DESPOT: Online POMDP planning with regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1772–1780, 2013.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

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.

Zachary N. Sunberg and Mykel J. Kochenderfer. Online algorithms for POMDPs with continuous state, action, and observation spaces. In *International Conference on Automated Planning and Scheduling (ICAPS)*, volume 28, pp. 259–263, 2018.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer, 2nd edition, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Chenyang Wu, Guoyu Yang, Zongzhang Zhang, Yang Yu, Dong Li, Wulong Liu, and Jianye Hao. Adaptive online packing-guided search for POMDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 28419–28430, 2021.

A Statistical effect of representation entropy

Normalizing the entropy penalty. For the pooled representations in Section 4, independent Gaussian noise gives

$$C(w) = I(\widetilde{Z}; \widetilde{Z}_\sigma) = h(\widetilde{Z}_\sigma) - h(\sigma\xi) \geq 0.$$

Here I denotes mutual information (Cover & Thomas, 2006). Since the noise scale and representation dimension are fixed, $h(\sigma\xi)$ is constant. Thus $F_\lambda(w) = J(w) - \lambda C(w)$ differs from equation 1 only by a controller-independent constant, and the two objectives have the same maximizers and gradients. Pooling decision steps also includes distinctions between times: $h(\widetilde{Z}_\sigma) = T^{-1}\sum_{t=0}^{T-1} h(Z_{\sigma,t}) + I(U; \widetilde{Z}_\sigma)$.

We give the distribution-dependent bounds used in the main text. Throughout this appendix $k_z(y)$ is the $\mathcal{N}(z, \sigma^2 I)$ density, $p(y) = \int k_z(y)\,\mu(dz)$, $Z \sim \mu$, and $Y \mid Z = z \sim k_z$. Let $G \sim \mathcal{N}(0, I_{d_Z})$ be independent Gaussian noise. Thus $C = I(Z;Y) = h(Y) - h(\sigma G)$. Write μ_w^{occ} for the law of the clean pooled representation $\widetilde{Z}$ under controller w.

Why the fixed noise scale matters. The readout noise is part of the controller, rather than an auxiliary perturbation used only to define the penalty. If the clean encoder output is replaced by cZ while σ stays fixed, rescaling the head input by $1/c$ gives $Z + (\sigma/c)G$. For $0 < c < 1$ this increases the effective noise seen by the head. Shrinking the representation can therefore remove distinctions needed for reward; it is not a cost-free change of coordinates. This observation explains why the return term and the representation-entropy term can compete.

Lemma A.1 (A bound using only representation entropy)**.** *For independent $Z_1, \ldots, Z_n \sim \mu$, put $g_f(z) = \int f(y)k_z(y)\,dy$, $|f| \leq 1$, and $D_n = \sup_{|f|\leq 1} |n^{-1}\sum_i g_f(Z_i) - \mathbb{E}_\mu g_f(Z)|$. For every $u > 0$, with probability at least $1 - \delta$,*

$$D_n \leq \frac{e^{u/2}}{\sqrt{n}} + \frac{2(C + 1/e)}{u} + \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Proof. Let $P(dz, dy) = \mu(dz)k_z(y)dy$, $Q(dz, dy) = \mu(dz)p(y)dy$, $r = k_z(y)/p(y)$ and $\iota = \log r$. Since $x\log(1/x) \leq 1/e$ for $0 < x < 1$, $\mathbb{E}_P \iota_- \leq 1/e$, so $\Pr_P\{\iota > u\} \leq (C + 1/e)/u$. Truncate the kernel at $e^u p(y)$: $k_z^u(y) = \min\{k_z(y), e^u p(y)\}$. Its squared expectation satisfies $\mathbb{E}_\mu(k_Z^u(y))^2 \leq e^u p(y)^2$. Consequently the expected L^1 error of its empirical mixture is at most $e^{u/2}/\sqrt{n}$. The omitted nonnegative kernel has expected mass $\mathbb{E}_P(1 - e^{u-\iota})_+ \leq \Pr_P\{\iota > u\}$. The empirical and population omitted masses together contribute at most $2(C + 1/e)/u$. Finally, $D_n = \int |n^{-1}\sum_i k_{Z_i}(y) - p(y)|dy$; changing one Z_i changes it by at most $2/n$. McDiarmid's inequality gives the last term. □

Choosing $u = 4(c + 1/e)/\varepsilon$ when $C \leq c$ proves the entropy-only bound referenced in Section 5. The bound concerns all bounded Gaussian-averaged heads at once. In particular, an error at most ε with probability at least $1 - \delta$ is obtained with

$$n = O\Big(\varepsilon^{-2} e^{O((c+1)/\varepsilon)} + \varepsilon^{-2}\log(1/\delta)\Big)$$

independent representation inputs.

Lemma A.2 (Information-density tail and channel complexity)**.** *Define the Gaussian-channel factor*

$$\mathcal{K}_\sigma(\mu) = \int \frac{\mathbb{E}_\mu k_Z(y)^2}{p(y)}\, dy = \mathbb{E}_P e^\iota.$$

If $\Pr_P\{\iota > aC + s\} \le Ae^{-\beta s}$ *for every* $s \ge 0$*, with* $\beta > 1$*, then*

$$\mathcal{K}_\sigma(\mu) \le \left(1 + \frac{A}{\beta - 1}\right) e^{aC}.$$

For the class $\{g_f : |f| \le 1\}$ *its expected Rademacher complexity on* n *independent inputs is at most* $\sqrt{\mathcal{K}_\sigma(\mu)/n}$*.*

Proof. The identity for $\mathcal{K}_\sigma$ follows by changing measure from P to Q. Apply the tail-integral formula to e^ι and split at e^{aC}:

$$\mathbb{E}_P e^\iota \le e^{aC} + Ae^{aC} \int_0^\infty e^{(1-\beta)s}\, ds.$$

For the Rademacher statement, interchange the supremum over $|f| \le 1$ with the integral and then apply Khintchine and Cauchy–Schwarz:

$$\mathbb{E} \sup_{|f|\le 1} \frac{1}{n} \sum_i \epsilon_i g_f(Z_i) = \mathbb{E} \int \left| \frac{1}{n} \sum_i \epsilon_i k_{Z_i}(y) \right| dy \le \frac{1}{\sqrt{n}} \int \sqrt{\mathbb{E} k_Z(y)^2}\, dy \le \sqrt{\frac{\mathcal{K}_\sigma(\mu)}{n}}.$$

□

A.1 Generalization from independent episodes

We first bound representation-level evaluations, allowing arbitrary dependence between decisions within an episode and time-dependent heads. Fix a controller w and encoder ϕ before sampling independent episodes $\tau_1, \dots, \tau_N$ from w. Let μ_t be the law of $Z_t = \phi_t(H_t)$, $\mu = T^{-1} \sum_t \mu_t$, and $p = T^{-1} \sum_t p_t$, where $p_t(y) = \int k_z(y) \mu_t(dz)$. Thus $C = I(\widetilde{Z}; \widetilde{Z}_\sigma)$. For any encoder ϕ' satisfying $\sup_{t,H} \|\phi'_t(H) - \phi_t(H)\| \le \rho$ and any measurable functions $f_0, \dots, f_{T-1}$ with $|f_t| \le 1$, define

$$L_{\phi',f}(\tau) = \frac{1}{T} \sum_{t=0}^{T-1} \int f_t(y) k_{\phi'_t(H_t)}(y)\, dy, \qquad \mathcal{E}_N^\rho = \sup_{\phi',f} \left| \frac{1}{N} \sum_i L_{\phi',f}(\tau_i) - \mathbb{E}_w L_{\phi',f}(\tau) \right|.$$

The population expectation uses the same sampling policy w for every candidate. These are time-averaged representation evaluations.

Proposition A.1 (Entropy bounds without a horizon factor)**.** *For every* $u > 0$*, with probability at least* $1 - \delta$*,*

$$\mathcal{E}_N^\rho \le \frac{e^{u/2}}{\sqrt{N}} + \frac{2(C + 1/e)}{u} + \sqrt{\frac{2\log(1/\delta)}{N}} + \frac{2\rho}{\sigma}. \tag{7}$$

If the information-density tail satisfies $\Pr\{\iota > aC + s\} \le Ae^{-\beta s}$ *for all* $s \ge 0$*, where* $a, A \ge 1$ *and* $\beta > 1$*, put* $V = (1 + A/(\beta - 1))e^{aC}$*. Then, with the same confidence,*

$$\mathcal{E}_N^\rho \le \sqrt{\frac{V}{N}} + \sqrt{\frac{2\log(1/\delta)}{N}} + \frac{2\rho}{\sigma}. \tag{8}$$

Neither bound has a separate horizon factor.

Proof. First take $\rho = 0$ and write $\widehat{p}_t = N^{-1} \sum_i k_{Z_{it}}$. The supremum over the independently chosen bounded functions f_t is

$$D_N = \frac{1}{T} \sum_t \int |\widehat{p}_t(y) - p_t(y)|\, dy.$$

For the tail bound define $V_t = \int \mathbb{E}_{\mu_t} k_Z(y)^2 / p(y)\, dy$. Independence across episodes and Cauchy–Schwarz give

$$\mathbb{E} D_N \le \frac{1}{T\sqrt{N}} \sum_t \int \sqrt{\mathbb{E}_{\mu_t} k_Z(y)^2}\, dy \le \frac{1}{T\sqrt{N}} \sum_t \sqrt{V_t} \le \sqrt{\frac{\mathcal{K}_\sigma(\mu)}{N}}.$$

Here $T^{-1}\sum_t V_t = \mathcal{K}_\sigma(\mu)$; Lemma A.2 bounds it by V.

For the entropy-only bound, truncate at the overall density: $k_z^u(y) = \min\{k_z(y), e^u p(y)\}$. Since $\mathbb{E}_{\mu_t}(k_Z^u(y))^2 \le e^u p(y) p_t(y)$, the time-averaged expected empirical error of the truncated kernels is at most

$$\frac{e^{u/2}}{T\sqrt{N}} \sum_t \int \sqrt{p(y)p_t(y)}\, dy \le \frac{e^{u/2}}{\sqrt{N}}.$$

The empirical and population omitted masses have total expectation at most $2\Pr\{\iota > u\} \le 2(C + 1/e)/u$, by the argument in Lemma A.1. Replacing one complete episode changes D_N by at most $2/N$. McDiarmid's inequality therefore adds $\sqrt{2\log(1/\delta)/N}$ to either expectation bound. No independence between times in an episode was used.

Finally, Gaussian KL and Pinsker's inequality give $\int |k_{z'} - k_z| \le \|z' - z\|/\sigma$. Changing the encoder moves each empirical and population evaluation by at most ρ/σ, uniformly over all f_t. Thus $\mathcal{E}_N^\rho \le D_N + 2\rho/\sigma$ pathwise, which proves both inequalities. The event is defined by the measurable D_N, so the conclusion holds simultaneously for all admissible encoder changes. □

For $C \le c$ and target error ε, taking $\rho \le \sigma\varepsilon/8$, $u = 8(c + 1/e)/\varepsilon$, and

$$N \ge \frac{8}{\varepsilon^2}\left[e^u + 2\log(1/\delta)\right]$$

makes equation 7 at most ε. Consequently a sufficient episode count is

$$N = O\Big(\varepsilon^{-2} e^{O((c+1)/\varepsilon)} + \varepsilon^{-2}\log(1/\delta)\Big).$$

At fixed accuracy and confidence this is $e^{O(c+1)}$ for the representation-level evaluation class. Under the stated tail condition, it suffices to take $\rho \le \sigma\varepsilon/4$ and $N \ge 8[V + 2\log(1/\delta)]/\varepsilon^2$. For controlled tail constants this improves the accuracy dependence to $O(\varepsilon^{-2}[e^{ac} + \log(1/\delta)])$. These counts measure independent episodes. Collecting N complete episodes uses NT environment decisions. The comparison concerns the statistical class evaluated on the sampling policy's visited histories; the next section treats changed trajectory laws and the complete unnormalized objective separately.

B Complete-episode local learning

For this analysis, assume a finite action space of size $K = |\mathcal{A}|$ and a complete return $G(\tau) = \sum_{t=0}^{T-1} R_t \in [0, R_{\max}]$. The reference controller $w = (\theta, \phi)$ and candidate updates $\mathcal{Q}(w)$ are fixed before collecting N new, independent complete episodes. Action probabilities are averaged over the input noise, as in Section 4. Write p_w^{occ} for the density of $\widetilde{Z}_\sigma$ under w, and, using KL divergence D_{KL}, define the representation cost $c_w(z) = D_{\mathrm{KL}}(\mathcal{N}(z, \sigma^2 I)\|p_w^{\mathrm{occ}})$. The reference density stays fixed when evaluating candidates. To compare them, we weight each sampled episode by the ratio of its probabilities under the candidate and reference controllers, and subtract the estimated representation cost. Equation 9 in Appendix B defines this optimization problem; Section 7 describes BRPO.

Theorem B.1 (Local learning of the complete bounded-rational objective)**.** *Fix accuracy $\varepsilon > 0$ and failure probability $\delta \in (0, 1)$. Suppose the encoder outputs of w and every $q \in \mathcal{Q}(w)$ have norm at most M. Every positive action probability under w is at least $b > 0$. Candidates assign positive probability to exactly the same actions and, for these actions, satisfy $e^{-d_t} \le q_t(a \mid H)/w_t(a \mid H) \le e^{d_t}$, where $d_t \ge 0$ and $\sum_t d_t \le d_{\mathrm{tot}}$. Candidate encoder outputs differ from the reference outputs by at most $\rho \ge 0$, uniformly over histories. Let $Z = \widetilde{Z}$ and $Y = \widetilde{Z}_\sigma$ under w and put $\iota_w(Z, Y) = \log[k_Z(Y)/p_w^{\mathrm{occ}}(Y)]$, where k_z is the $\mathcal{N}(z, \sigma^2 I)$ density. For a complexity budget $c \ge 0$, assume $C(w) \le c$ and the tail bound*

$$\Pr\{\iota_w > aC(w) + s\} \le Ae^{-\beta s} \quad (s \ge 0), \qquad a, A \ge 1,\ \beta > 1.$$

Let $\epsilon_{\mathrm{cost}} \ge 0$ be the cost-estimation tolerance. Suppose the cost estimate $\widehat{c}_w$ satisfies $\sup_{\|z\| \le M} |\widehat{c}_w(z) - c_w(z)| \le \epsilon_{\mathrm{cost}}$ and $\widehat{q}$ maximizes equation 9 over $\mathcal{Q}(w)$ to within optimization error $\eta_{\mathrm{opt}} \ge 0$. If N and the approximation errors satisfy equation 10, and the cost-estimation bound holds with probability one, then, with probability at least $1 - \delta$,

$$F_\lambda(\widehat{q}) \ge \sup_{q \in \mathcal{Q}(w)} F_\lambda(q) - \varepsilon.$$

Write $p_0 = p_w^{\mathrm{occ}}$ and $c_0 = c_w$. For episodes $\tau_1, \dots, \tau_N$ from w, the program is

$$\widehat{\mathcal{L}}_N(q) = \frac{1}{N}\sum_{i=1}^{N} W_q(\tau_i)\left[G(\tau_i) - \frac{\lambda}{T}\sum_{t=0}^{T-1}\widehat{c}_w(\phi_q(H_{it}))\right], \qquad W_q(\tau) = \prod_{t=0}^{T-1}\frac{q_t(A_t \mid H_t)}{w_t(A_t \mid H_t)}. \quad (9)$$

For the theorem's conditions put $\kappa = e^{d_{\mathrm{tot}}}$, $r = M^2/\sigma^2$, $R_\lambda = R_{\max} + 2\lambda r$, $V = (1 + A/(\beta - 1))e^{ac}$, $D_\phi = R_\lambda \kappa T \rho/(b\sigma)$, $e_c = 2M\rho/\sigma^2$, and $\Delta_{\mathrm{KL}} = d_{\mathrm{tot}}\tanh(d_{\mathrm{tot}}/2) + \rho^2/(2\sigma^2)$. The explicit episode and approximation budgets are

$$\begin{aligned} N &\geq \frac{64R_\lambda^2\kappa^2}{\varepsilon^2}\left[\frac{4KT^{3/2}}{b}\sqrt{V} + \sqrt{2\log(2/\delta)}\right]^2, \\ 2D_\phi + 4\kappa\lambda e_c + 2\kappa\lambda\epsilon_{\mathrm{cost}} + \eta_{\mathrm{opt}} + \lambda\Delta_{\mathrm{KL}} &\leq \frac{3\varepsilon}{4}. \end{aligned} \quad (10)$$

With the other constants controlled, the episode budget has the form $O(\varepsilon^{-2}[T^3 e^{ac} + \log(1/\delta)])$. The factor T^3 comes from evaluating changes along the trajectory; e^{ac} captures the statistical complexity of its representations. A bounded entropy budget keeps this latter factor bounded. If entropy is proportional to the horizon, it can instead give an exponential horizon dependence. These statements also require control of the remaining constants and satisfaction of the approximation budget. The environment factors cancel, so $\mathbb{E}_w W_q B(\tau) = \mathbb{E}_q B(\tau)$ for any integrable function of the complete observed trajectory. The likelihood-ratio assumptions give $e^{-d_{\mathrm{tot}}} \leq W_q \leq e^{d_{\mathrm{tot}}} = \kappa$.

Lemma B.1 (Uniform episode evaluation)**.** *Let $\mathcal{Q}(w)$ satisfy the support, ratio and encoder drift assumptions of Theorem B.1. Suppose $B(\tau)$ is fixed across candidates and has absolute value and range length at most R. If the reference occupancy has $\mathcal{K}_\sigma(\mu_w^{\mathrm{occ}}) \leq V$, then, with probability at least $1 - \delta$, the complete-episode importance-weighted empirical mean of B deviates uniformly from its candidate expectation by at most $A_N(R, V, \delta) + D_\phi(R)$, where*

$$\begin{aligned} A_N(R, V, \delta) &= \frac{R\kappa}{\sqrt{N}}\left[\frac{4KT^{3/2}}{b}\sqrt{V} + \sqrt{2\log(2/\delta)}\right], \\ D_\phi(R) &= \frac{R\kappa T\rho}{b\sigma}. \end{aligned}$$

Proof. First freeze the encoder at ϕ_w. For each episode, regard the selected-action probabilities at the T steps as a vector v. Extend its importance-weighted loss to the whole probability cube by clipping each ratio $v_t/w_t(A_t \mid H_t)$ to $[e^{-d_t}, e^{d_t}]$. The extension equals the actual loss on $\mathcal{Q}(w)$. Its coordinate Lipschitz constants are at most $R\kappa/b$, and its Euclidean Lipschitz constant is at most $R\kappa\sqrt{T}/b$. Vector contraction, followed by splitting by time and action, bounds the expected episode-level Rademacher complexity by

$$\frac{\sqrt{2}R\kappa\sqrt{T}}{b}\sum_{t=0}^{T-1}\sum_{a\in\mathcal{A}}\sqrt{\frac{V_t}{N}}, \qquad V_t = \int \frac{\mathbb{E}_{\mu_{w,t}} k_Z(y)^2}{p_0(y)}\, dy.$$

This contraction applies to independent episode vectors; it does not require their coordinates to be independent. Because $p_0 = T^{-1}\sum_t p_{w,t}$, $T^{-1}\sum_t V_t = \mathcal{K}_\sigma(\mu_w^{\mathrm{occ}}) \leq V$. Cauchy–Schwarz yields $\sum_t \sqrt{V_t} \leq T\sqrt{V}$. Two-sided symmetrization and a bounded-difference inequality give A_N; its factor 4 exceeds the $2\sqrt{2}$ resulting from the contraction.

For an encoder-changing candidate, keep its head and replace its encoder output by the old output on the same history. Pinsker's inequality between equal-covariance Gaussians changes every action probability by at most $\rho/(2\sigma)$. The clipped extension above then changes each episode loss by at most $R\kappa T\rho/(2b\sigma)$. Apply this to both empirical and population means to obtain $D_\phi(R)$. The auxiliary head need not itself satisfy the ratio bounds because the extension is globally clipped. □

Lemma B.2 (Frozen occupancy cost)**.** *For $c_0(z) = D_{\mathrm{KL}}(k_z \| p_0)$ and clean outputs bounded by M,*

$$0 \leq c_0(z) \leq 2M^2/\sigma^2, \qquad \|\nabla c_0(z)\| \leq 2M/\sigma^2 \quad (\|z\| \leq M).$$

For every candidate q,

$$F_\lambda(q) = \mathbb{E}_q\left[G - \frac{\lambda}{T}\sum_t c_0(Z_t)\right] + \lambda D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0). \quad (11)$$

If the candidate-to-old trajectory likelihood ratio lies in $[e^{-d_{\mathrm{tot}}}, e^{d_{\mathrm{tot}}}]$ *and encoder drift is at most* ρ*, then*

$$D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0) \le d_{\mathrm{tot}} \tanh(d_{\mathrm{tot}}/2) + \rho^2/(2\sigma^2).$$

Proof. Convexity of KL in its second argument gives $c_0(z) \le \mathbb{E}_{Z' \sim \mu_w^{\mathrm{occ}}} D_{\mathrm{KL}}(k_z \| k_{Z'}) = \mathbb{E}\|z - Z'\|^2/(2\sigma^2) \le 2M^2/\sigma^2$. Differentiating the Gaussian cross entropy gives

$$\nabla c_0(z) = \frac{z - \mathbb{E}_{Y \sim k_z} \mathbb{E}[Z' \mid Y]}{\sigma^2},$$

which proves the gradient bound. Moreover $T^{-1} \sum_t \mathbb{E}_q c_0(Z_t) = \mathbb{E}_{p_q^{\mathrm{occ}}}[-\log p_0(Y)] - h(\sigma G) = C(q) + D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0)$, giving equation 11.

For the final inequality, the trajectory ratio W_q satisfies $\mathbb{E}_w W_q = 1$ and is bounded by the stated endpoints. The chord bound for $x \log x$ on this interval gives $D_{\mathrm{KL}}(P_q^{\mathrm{traj}} \| P_w^{\mathrm{traj}}) \le d_{\mathrm{tot}} \tanh(d_{\mathrm{tot}}/2)$. Add an independent uniform time U to both trajectory laws, and generate a Gaussian readout from $\phi_q(H_U)$ on one side and $\phi_w(H_U)$ on the other. The conditional Gaussian KL adds at most $\rho^2/(2\sigma^2)$. Marginalizing to the readout and applying data processing proves the result. □

Proof of Theorem B.1. Let $\mathcal{L}_0(q) = \mathbb{E}_q[G - \lambda T^{-1} \sum_t c_0(\phi_q(H_t))]$. Lemma A.2 gives $\mathcal{K}_\sigma(\mu_w^{\mathrm{occ}}) \le V$. By Lemmas B.1 and B.2, the exact episode integrand has absolute value and range length at most $R_\lambda = R_{\max} + 2\lambda M^2/\sigma^2$. The encoder change in its costs is at most $e_c = 2M\rho/\sigma^2$ at each history. The population and empirical changes together cost $2\kappa\lambda e_c$. Replacing c_0 by $\widehat{c}_w$ in the empirical program costs at most $\kappa\lambda\epsilon_{\mathrm{cost}}$. Thus, on the event in the theorem,

$$\sup_{q \in \mathcal{Q}(w)} |\widehat{\mathcal{L}}_N(q) - \mathcal{L}_0(q)| \le A_N(R_\lambda, V, \delta) + D_\phi(R_\lambda) + 2\kappa\lambda e_c + \kappa\lambda\epsilon_{\mathrm{cost}} =: E_N.$$

Theorem's sample inequality makes $A_N \le \varepsilon/8$. Empirical η_{opt}-optimality yields $\sup_q \mathcal{L}_0(q) - \mathcal{L}_0(\widehat{q}) \le 2E_N + \eta_{\mathrm{opt}}$. Lemma B.2 then gives

$$\sup_q F_\lambda(q) - F_\lambda(\widehat{q}) \le 2E_N + \eta_{\mathrm{opt}} + \lambda\Delta_{\mathrm{KL}} \le \varepsilon.$$

The cost-estimation event may be conditioned upon or assigned its own failure probability δ_C; in the latter case the joint confidence is at least $1 - \delta - \delta_C$. □

B.1 Preference for a nearly optimal, low-complexity policy

Full statement of Proposition 5.1.

> Let J^* be the largest return in the controller class. Let $\beta_s, c_s, \eta_0 \ge 0$. For $\lambda > 0$, suppose a feasible controller w_s satisfies $J^* - J(w_s) \le \beta_s$ and $C(w_s) \le c_s$. Any w satisfying $F_\lambda(w) \ge F_\lambda(w_s) - \eta_0$ obeys
>
> $$C(w) \le c_s + \frac{\beta_s + \eta_0}{\lambda}, \qquad J^* - J(w) \le \beta_s + \lambda c_s + \eta_0.$$
>
> If w also meets Theorem B.1's conditions and $w \in \mathcal{Q}(w)$, the update $\widehat{q}$ satisfies $J^* - J(\widehat{q}) \le \beta_s + \lambda c_s + \eta_0 + \varepsilon$.

Proof of Proposition 5.1. Use $J(w) - \lambda C(w) \ge J(w_s) - \lambda C(w_s) - \eta_0$, $J(w) \le J^*$, and $C(w) \ge 0$. Rearrangement gives both displayed claims. If $w \in \mathcal{Q}(w)$, Theorem B.1 gives $F_\lambda(\widehat{q}) \ge F_\lambda(w) - \varepsilon$. Since $J(\widehat{q}) \ge F_\lambda(\widehat{q})$, the final return bound follows. □

B.2 Complete-objective learning with cumulative update control

The next result sharpens the main-text sample bound by strengthening the update and temporal distribution conditions of Theorem B.1. Both results use the same empirical program equation 9.

Theorem B.2 (Complete-objective learning without a separate horizon factor)**.** *Fix the reference controller w, candidate set $\mathcal{Q}(w)$, and constants before collecting N independent episodes from w. Suppose $G \in [0, R_{\max}]$, there are K actions, and all reference and candidate encoder outputs have norm at most M on histories in the common trajectory support. Every positive reference action probability is at least $b > 0$, and candidates preserve its action support. For every candidate q, suppose*

$$\sup_{a,y} |\pi_{\theta_q,t}(a \mid y) - \pi_{\theta_w,t}(a \mid y)| \le \nu_t \le b/4, \quad \sum_t \nu_t \le \nu_{\mathrm{tot}},$$
$$\sup_{t,H} \|\phi_{q,t}(H) - \phi_{w,t}(H)\| \le \rho \le b\sigma/2.$$

Let $p_0 = p_w^{\mathrm{occ}}$, assume $C(w) \le c$, and at every decision t set $Y_t = Z_t + \sigma\xi$ using independent Gaussian noise. Assume the uniform temporal tail condition

$$\Pr_w\left\{\log \frac{k_{Z_t}(Y_t)}{p_0(Y_t)} > ac + s\right\} \le Ae^{-\beta s} \quad (t = 0, \dots, T-1,\ s \ge 0), \qquad a, A \ge 1, \quad \beta > 1.$$

Put

$$\mathsf{H} = 2\nu_{\mathrm{tot}}/b, \quad \mathsf{E} = T\rho/(b\sigma), \quad D = \mathsf{H} + \mathsf{E}, \quad \kappa = e^D, \quad V = (1 + A/(\beta - 1))e^{ac},$$

and $R_\lambda = R_{\max} + 2\lambda M^2/\sigma^2$, $e_c = 2M\rho/\sigma^2$, $\Delta = D\tanh(D/2) + \rho^2/(2\sigma^2)$. Suppose the cost estimate has uniform error at most ϵ_{cost} with probability one, and $\widehat{q}$ maximizes equation 9 within η_{opt}. For $\varepsilon > 0$ and $\delta \in (0, 1)$, it suffices that

$$N \ge \frac{16R_\lambda^2\kappa^2}{\varepsilon^2}\left[2 + 8\mathsf{H}\sqrt{KV} + \sqrt{2\log(1/\delta)}\right]^2, \tag{12}$$
$$4R_\lambda\kappa\mathsf{E} + 4\kappa\lambda e_c + 2\kappa\lambda\epsilon_{\mathrm{cost}} + \eta_{\mathrm{opt}} + \lambda\Delta \le \varepsilon/2. \tag{13}$$

Then, with probability at least $1 - \delta$, $F_\lambda(\widehat{q}) \ge \sup_{q\in\mathcal{Q}(w)} F_\lambda(q) - \varepsilon$.

Proof. Freeze the candidate encoder at ϕ_w and write the resulting action probabilities and trajectory weight as q_t^0 and W_q^0. For selected actions in the old support,

$$q_t^0 - w_t = \nu_t g_{q,t}, \qquad g_{q,t}(H, a) = \int f_{q,t}(a, y)k_{\phi_{w,t}(H)}(y)\,dy, \qquad |f_{q,t}| \le 1.$$

For $\nu_t = 0$ take $g_{q,t} = 0$. The condition $\nu_t \le b/4$ gives $q_t^0 \ge 3b/4$ and $|\log(q_t^0/w_t)| \le 2\nu_t/b$. Thus $|\log W_q^0| \le \mathsf{H}$.

Define $V_t = \int \mathbb{E}_w k_{Z_t}(y)^2/p_0(y)\,dy$. The uniform temporal tail assumption and the tail-integral argument of Lemma A.2 give $V_t \le V$ for every t. For the selected-action function class, independent Rademacher signs ϵ_i give

$$\mathbb{E}\sup_{|f|\le 1}\left|\frac{1}{N}\sum_i \epsilon_i \int f(A_{it}, y)k_{Z_{it}}(y)\,dy\right| \le \frac{1}{\sqrt{N}}\sum_{a=1}^K \int \sqrt{\mathbb{E}_w[\mathbf{1}\{A_t = a\}k_{Z_t}(y)^2]}\,dy$$
$$\le \sqrt{KV_t/N} \le \sqrt{KV/N}.$$

The action and representation may be dependent; summing the action indicators recovers $\mathbb{E}_w k_{Z_t}^2$. The scalar map $x \mapsto \log(1 + \nu_t x/w_t)$ is zero at zero and has Lipschitz constant at most $2\nu_t/b$ on $[-1, 1]$. Scalar contraction for the absolute Rademacher supremum, with its factor two, and the triangle inequality therefore bound the absolute Rademacher complexity of $\{\log W_q^0\}$ by $2\mathsf{H}\sqrt{KV/N}$.

Let $B(\tau) = G(\tau) - \lambda T^{-1}\sum_t c_0(\phi_{w,t}(H_t))$; Lemma B.2 gives $|B| \le R_\lambda$. Apply scalar contraction to $x \mapsto B(\tau_i)(e^x - 1)$, with Lipschitz constant $R_\lambda\kappa$ on $[-\mathsf{H}, \mathsf{H}]$. Adding back the fixed function B contributes at most $R_\lambda/\sqrt{N}$ to the absolute Rademacher complexity. Symmetrization and bounded differences now give, with confidence $1 - \delta$,

$$\sup_q |\widehat{\mathbb{E}}_w[W_q^0 B] - \mathbb{E}_w[W_q^0 B]| \le A_N := \frac{R_\lambda\kappa}{\sqrt{N}}[2 + 8\mathsf{H}\sqrt{KV} + \sqrt{2\log(1/\delta)}].$$

Here changing one episode changes the supremum by at most $2R_\lambda\kappa/N$.

Gaussian TV bounds give $|q_t - q_t^0| \le \rho/(2\sigma)$. The actual probabilities remain at least $b/2$, and hence $|\log W_q - \log W_q^0| \le \mathsf{E}$. It follows that $W_q \le \kappa$, $|W_q - W_q^0| \le \kappa\mathsf{E}$, and the empirical and population replacement of $W_q^0 B$ by $W_q B$ costs at most $2R_\lambda \kappa \mathsf{E}$. Replacing the reference encoder cost by the candidate encoder cost costs at most $2\kappa\lambda e_c$ in total, and replacing the exact empirical cost by its estimate costs at most $\kappa\lambda\epsilon_{\mathrm{cost}}$. Thus the empirical program and its population frozen-cost objective are uniformly within

$$E_N = A_N + 2R_\lambda\kappa\mathsf{E} + 2\kappa\lambda e_c + \kappa\lambda\epsilon_{\mathrm{cost}}.$$

Actual trajectory ratios lie in $[e^{-D}, e^D]$. Lemma B.2 therefore bounds the occupancy KL remainder by Δ. Empirical optimization and the frozen-cost identity give

$$\sup_q F_\lambda(q) - F_\lambda(\widehat{q}) \le 2E_N + \eta_{\mathrm{opt}} + \lambda\Delta.$$

The sample budget makes $A_N \le \varepsilon/4$; the error budget bounds all remaining terms by $\varepsilon/2$. This proves the conclusion. □

If $R_\lambda, b^{-1}, K, \mathsf{H}, \mathsf{E}, a, A$ and $(\beta-1)^{-1}$ remain bounded independently of T, and the error budget is met, the sufficient episode count is

$$N = O\big(\varepsilon^{-2}[e^{ac} + \log(1/\delta)]\big).$$

This sharpens the main-text complete-objective bound under stronger conditions. The conditions include cumulative control of the head changes, encoder drift small enough that $T\rho$ stays controlled, a fixed complete-return scale, and tail regularity at each time. They describe a conditional local learning guarantee; occupancy entropy by itself does not enforce these conditions.

B.3 Scope of the statistical guarantees

Proposition A.1 controls bounded, time-averaged functions on visited representations, with a fixed or locally perturbed encoder. Its independent-episode count has no separate horizon factor. This does not change the normalization of the episodic return or control the product of action likelihood ratios under a candidate policy. The complete-objective budget equation 10 retains its trajectory factors; a fixed entropy budget alone does not remove them. The stronger uniform temporal and cumulative update conditions of Theorem B.2 give a separate complete-objective bound without a horizon factor. Its return scale and error budgets must remain controlled; the entropy penalty alone does not ensure this. The complete-episode guarantee assumes a suitable reference policy, a bound on its information-density tail, accurate cost estimates, and small changes in action probabilities and encoder outputs. The preference proposition bounds the complexity and return loss of a controller whose objective value is sufficiently good. Finding that controller and establishing its tail condition remain open search and distribution-control questions. The finite-sample theorem analyzes the optimization problem in equation 9. Applying it to BRPO would additionally require control of the score proxy, sampled advantages, clipping, and optimization errors. Exact score estimation can still leave a gap due to the finite diffusion endpoint, as derived in Appendix E. PPO clipping alone does not ensure the uniform action and encoder bounds used here. The exponential lower bound in Theorem 5.2 concerns uniform error of the specified importance-weighted estimator in the constructed task. Other estimators may behave differently; executing a known optimum is a separate question. This construction's representation dimension and radius also grow with horizon.

C One task with contrasting local generalization

Full statement of Theorem 5.2.

> For every $T \ge 3$ and $0 < \beta_s \le 1/8$, there is a T-decision partially observable task with fixed Gaussian readout noise, a fixed unifilar encoder, and a finite-action controller class $\mathcal{W}_m$ containing a return maximizer w_h and a controller w_s with $0 < J^* - J(w_s) \le \beta_s$. The policies have $C(w_h) = \Theta(T)$ and $C(w_s) = 0$; w_s globally maximizes F_1 in this same class. For $0 < \alpha \le 1/16$, let $\mathcal{Q}_\alpha(w) \subseteq \mathcal{W}_m$

consist of all candidates that keep the initial mode probability and intermediate heads fixed and change only the last head, subject to

$$\sup_{H_m} \mathrm{TV}\big(q_m(\cdot \mid H_m), w_m(\cdot \mid H_m)\big) \le \alpha.$$

The supremum is over legal task histories; action probabilities integrate the Gaussian readout noise. Both centers and their candidates use the same encoder and last-head class defined below. Their occupancy cost is unchanged, so their complete-objective improvement $\Delta F_w(q) = F_1(q) - F_1(w)$ equals their return improvement. Use around either policy the same estimate in equation 14. For every realized batch of $N \le 2^{T-3}$ old-policy episodes,

$$\sup_{q \in \mathcal{Q}_\alpha(w_h)} |\widehat{\Delta F}_{w_h,N}(q) - \Delta F_{w_h}(q)| \ge \frac{153\alpha}{256}.$$

At w_s, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the same uniform error is at most

$$\frac{\gamma_s \alpha}{1-b}\sqrt{\frac{\log(2/\delta)}{2N}},$$

where $b = 1/8$ and $\gamma_s \in (0,1]$ scales the simple-mode reward. Consequently, for $0 < \varepsilon < 153\alpha/256$, uniform error at most ε requires $N > 2^{T-3}$ around w_h. Around w_s, it suffices to take

$$N \ge \max\left\{1, \left\lceil \frac{\gamma_s^2 \alpha^2}{2(1-b)^2 \varepsilon^2} \log \frac{2}{\delta} \right\rceil \right\},$$

which is bounded independently of T since $\gamma_s \le 1$.

For a center w and N independent episodes sampled from it, define

$$\widehat{\Delta F}_{w,N}(q) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{q_{T-1}(A_{i,T-1} \mid H_{i,T-1})}{w_{T-1}(A_{i,T-1} \mid H_{i,T-1})} - 1\right] R_{i,T-1}. \tag{14}$$

Proof of Theorem 5.2. **Task and recurrent encoder.** Fix $T \ge 3$ and put $m = T-1 \ge 2$, $\mathcal{B} = \{0,1\}^m$, $b = 1/8$, $\Delta = b + \beta_s$, and $\gamma_s = (1-\Delta)/(1-b) \in (0,1]$. At time zero an action chooses high mode with probability p or simple mode with probability $1-p$. In high mode the environment reveals independent fair bits $B_1, \dots, B_m$ at times $1, \dots, m$; intermediate actions have no effect. The last action $A_m \in \{0,1\}$ earns $R_m = \mathbf{1}\{A_m = \theta_B\}$, where $B = (B_1, \dots, B_m)$ and the fixed table $\theta = (\theta_j : j \in \mathcal{B})$ is part of the environment. Only the current bit is observed at each time. In simple mode the last action earns $\gamma_s \mathbf{1}\{A_m = 0\}$. All earlier rewards are zero. Let X_0 be a start symbol and use a distinct symbol $*$ for simple-mode observations; high-mode observations are the bits 0 or 1. The terminal observation X_T reports the earned reward.

Use one fixed m-dimensional encoder for both modes. It outputs zero throughout simple mode and at time zero. In high mode it outputs

$$Z_t = a(2B_1 - 1, \dots, 2B_t - 1, 0, \dots, 0), \qquad 1 \le t \le m.$$

One shared deterministic recurrent update sets the state to zero on the start symbol or $*$ and, on a bit, appends $a(2X_t - 1)$ in the next zero coordinate. The nonzero-coordinate count identifies that coordinate. Thus the encoder is unifilar without access to a hidden mode variable. Its clean-output radius is $M_m = a\sqrt{m}$. Both modes use independent $\mathcal{N}(0, \sigma^2 I_m)$ policy-input noise at every decision, with fixed $\sigma > 0$. The initial and last heads may depend on decision time, as allowed in Section 4. Choose finite a/σ so that

$$e_m := m e^{-a^2/(2\sigma^2)} \le \min\left\{\frac{1}{64}, \frac{\beta_s}{2(1-2b)}\right\}.$$

One common noisy-input head class. For $y \in \mathbb{R}^m$, let $j(y) = (\mathbf{1}\{y_k \ge 0\} : 1 \le k \le m)$ be its sign label. Define

$$\mathcal{W}_m = \{w_{p,r} : p \in [0,1], \quad r = (r_j : j \in \mathcal{B}) \in [b, 1-b]^{2^m}\}.$$

The initial head uses the constant mode probability p; intermediate heads choose a fixed action. The last head is

$$\pi_r(0 \mid y) = r_{j(y)}, \qquad \pi_r(1 \mid y) = 1 - r_{j(y)}.$$

This same function is used in both modes and receives only y. In particular its noise-marginalized action-zero probability is

$$w_{p,r,m}(0 \mid H_m) = \sum_{j \in \mathcal{B}} r_j \Pr\{j(Z_m + \sigma G) = j \mid H_m\}.$$

All last-action probabilities lie in $[b, 1-b]$. In simple mode $Z_m = 0$, so the sign label is uniform and this probability is $v(r) = 2^{-m} \sum_j r_j$. It is determined by the same vector r that defines high-mode decisions, with no independent mode parameter.

In high mode write $\widehat{B} = j(Z_m + \sigma G)$. A Gaussian tail bound and union bound give $\Pr\{\widehat{B} \neq B \mid B\} \leq e_m$. The sign channel flips coordinates independently with the same probability, so its transition matrix is symmetric. Since B is uniform, $\widehat{B}$ is uniform too, and Bayes' rule gives

$$\Pr\{B = j \mid \widehat{B} = j\} = \Pr\{\widehat{B} = j \mid B = j\} \geq 1 - e_m > 1/2.$$

Return optimality in the common class. Let $w_h = w_{1,r^h}$, where $r^h_j = 1 - b$ if $\theta_j = 0$ and $r^h_j = b$ if $\theta_j = 1$. Conditionally on $\widehat{B} = j$, the correct action is θ_j with probability at least $1 - e_m > 1/2$, regardless of the other entries in θ. Expected reward is affine in r_j and is therefore maximized by r^h_j. This proves high-mode optimality over the entire last-head class. Write this maximum as J^*_h. Correct decoding yields success probability $1 - b$, and on an incorrect decoding the success probability is at least b. Hence

$$J^*_h = J(w_h) \geq 1 - b - (1 - 2b)e_m, \qquad J^*_h \leq 1 - b.$$

Let $w_s = w_{0,r^s}$, where $r^s_j = 1 - b$ for every j. It uses a constant head on all noisy inputs. For any r, the simple-mode return is $\gamma_s v(r) \leq \gamma_s(1 - b)$. Thus w_s attains the simple-mode maximum,

$$J_s = J(w_s) = 1 - \Delta = 1 - b - \beta_s.$$

The chosen e_m gives $J^*_h \geq 1 - b - \beta_s/2 > J_s$. For any common head r and mode probability p,

$$J(w_{p,r}) \leq pJ^*_h + (1 - p)J_s \leq J^*_h.$$

Hence $J^* = J^*_h$ is the return optimum of $\mathcal{W}_m$, attained by w_h, and $0 < J^* - J(w_s) \leq \beta_s$.

Normalized representation complexity and objective preference. The normalized representation complexity of w_s is zero because its clean representation is constantly zero. For $w_{p,r}$, let $Z = Z_U$ with independent uniform $U \in \{0, \ldots, m\}$ and $Y = Z + \sigma G$. This law depends on p and the fixed encoder, and is independent of r because the last action occurs after all decision representations. The event $E = \{\text{high mode}, U = m\}$ is determined by Z: exactly on this event all coordinates are nonzero. Since $\Pr(E) = p/T$, the chain rule and Fano's inequality give

$$C_m(p) = I(Z;Y) = I(E;Y) + I(Z;Y \mid E) \geq \frac{p}{T}\left[(1 - e_m)m \log 2 - h_2(e_m)\right] = pc_m.$$

Here h_2 is binary entropy, and the inequality at $p = 0$ follows from nonnegativity. For $m \geq 2$ and $e_m \leq 1/8$,

$$c_m \geq \frac{(7m/8)\log 2 - h_2(1/8)}{m + 1} \geq \frac{(7/4)\log 2 - h_2(1/8)}{3} > \frac{1}{4}.$$

The second expression is increasing with m. For pure high mode, decoding the first t bits gives, for $1 \leq t \leq m$, $I(Z_t; Y_t) \geq (1 - e_m)t \log 2 - h_2(e_m)$. The Markov chain $U \to Z \to Y$ gives

$$C(w_h) = I(U;Y) + \frac{1}{T}\sum_{t=0}^{m} I(Z_t;Y_t) \geq \frac{(1 - e_m)m \log 2}{2} - \frac{m}{T}h_2(e_m) = \Omega(T).$$

The clean high-mode representation determines its time and revealed bits. Its discrete entropy therefore equals $H(Z) = \log T + m \log 2/2$, so $C(w_h) \leq H(Z) = O(T)$ and $C(w_h) = \Theta(T)$. Any mode mixture satisfies $J(w_{p,r}) \leq J_s + p\beta_s$. At $\lambda = 1$ and $p > 0$,

$$F_1(w_{p,r}) \leq J_s + p(\beta_s - c_m) < J_s = F_1(w_s).$$

For $p = 0$, $C_m(0) = 0$ and $J(w_{0,r}) \le J_s$. Thus w_s globally maximizes F_1 over $\mathcal{W}_m$, and all maximizers use simple mode.

One estimator for the complete objective at both centers. Because candidates change only the last head, the trajectory law through H_m and all decision representations is unchanged. Their occupancy costs cancel exactly in $\Delta F_w(q)$. Writing $R_m(H, a)$ for the terminal reward, deterministic given history and action in this task, gives

$$\Delta F_w(q) = \mathbb{E}_w \sum_{a \in \{0,1\}} [q_m(a \mid H_m) - w_m(a \mid H_m)] R_m(H_m, a).$$

Old last-action probabilities are at least b, so conditioning equation 14 on H_m proves its unbiasedness for every fixed candidate. Both centers use exactly this rule.

Uniform lower bound at the high-complexity optimum. For any $S \subseteq \mathcal{B}$, define a candidate q_S that retains the preferred action θ_j but lowers its probability from $1 - b$ to $1 - b - \alpha$ on $j \in S$. Equivalently its head parameters are

$$r_j^S = r_j^h - \alpha(1 - 2\theta_j)\mathbf{1}\{j \in S\}.$$

These parameters stay in $[b, 1 - b]$. Either action probability changes by at most α on every noisy input. Integrating noise preserves the bound, so $q_S \in \mathcal{Q}_\alpha(w_h)$ on all legal histories in both modes. For any realized batch of $N \le 2^{m-2}$ high-mode episodes, choose S as the clean histories B absent from the batch. Then $|S|/2^m \ge 3/4$. On a fresh episode, correct decoding of $B \in S$ loses α in expected reward; an opposite contribution requires a decoding error. Consequently

$$\Delta F_{w_h}(q_S) \le -\alpha[(1 - e_m)|S|/2^m - e_m] \le -\frac{185\alpha}{256}.$$

For each sampled history $B_i \notin S$, a changed head entry can be selected only through a decoding error. Thus the noise-marginalized probability of either sampled action changes by at most αe_m. Using the old probability lower bound b and $R_{im} \in [0, 1]$ gives

$$|\widehat{\Delta F}_{w_h,N}(q_S)| \le \frac{\alpha e_m}{b} \le \frac{\alpha}{8}.$$

The candidate may depend on the batch because the claim is a uniform supremum over the fixed candidate class. Subtracting these two bounds yields $153\alpha/256$ for every such batch.

Uniform upper bound at the simple center. Every $q \in \mathcal{Q}_\alpha(w_s)$ remains in simple mode and has one noise-marginalized probability $v = v(r) \in [b, 1 - b]$ of action zero, independent of history, with $|v - (1 - b)| \le \alpha$. The head may vary across sign labels; only its average $v(r)$ enters the policy on simple-mode histories. Let P_N be the sampled frequency of action zero. Then

$$\widehat{\Delta F}_{w_s,N}(q) = \gamma_s \frac{v - (1 - b)}{1 - b} P_N, \qquad \Delta F_{w_s}(q) = \gamma_s[v - (1 - b)].$$

The old action-zero probability is $1 - b$, so Hoeffding's inequality gives $|P_N - (1 - b)| \le \sqrt{\log(2/\delta)/(2N)}$ with probability at least $1 - \delta$. This single event controls every candidate simultaneously, proving the simple-mode bound and its stated sufficient episode count.

The high-mode bound holds for every batch and therefore precludes any confidence guarantee at accuracy $\varepsilon < 153\alpha/256$ when $N \le 2^{T-3}$. Since $m + 1 \le 2^m$ for $m \ge 2$,

$$C(w_h) \le \log(m + 1) + \frac{m \log 2}{2} \le \frac{3m \log 2}{2}.$$

The necessary episode count thus also satisfies $N > \frac{1}{4} \exp[2C(w_h)/3]$. Both sides use the same estimator, action radius, encoder, noise scale and bounded reward scale. □

The comparison concerns uniform evaluation by equation 14 near specified centers in $\mathcal{W}_m$. The head restriction to sign labels is part of this construction. Across the task family, $d_Z = m$ and $M_m = a\sqrt{m}$ can grow; the two centers within each task use the same encoder and noise. Each episode has T decisions, so the interaction count is NT.

D THE EXACT GRADIENT AND ENTROPY RESISTANCE

Regularity conditions. Work at an interior parameter value w where the trajectory laws have locally common support and densities P_w with respect to a common parameter-independent measure. Assume that the action densities, encoder outputs, and noisy occupancy density are differentiable, and that the derivatives of all integrands below admit integrable local dominating functions. These conditions justify differentiation under expectation and integration. Rewards are the fixed history functions in Section 3; the Gaussian noise scale and dimension are fixed.

Proof of Theorem 6.1. Let $p = p_w^{\mathrm{occ}}$ and $u_w(z) = -\int k_z(y) \log p_w^{\mathrm{occ}}(y)\, dy$. Sampling time uniformly gives

$$h(p) = \mathbb{E}_w \frac{1}{T} \sum_{t=0}^{T-1} u_w(Z_t).$$

There are three contributions to its derivative: changing the trajectory law, moving the encoded inputs at fixed histories, and changing the density inside the logarithm. The last contribution is

$$-\mathbb{E}_{Y \sim p_w^{\mathrm{occ}}} \nabla_w \log p_w^{\mathrm{occ}}(Y) = -\int \nabla_w p_w^{\mathrm{occ}}(y)\, dy = 0,$$

where the derivative holds Y fixed. It vanishes because the current density integrates to one. No auxiliary reference density is needed. Thus, with $G(\tau) = \sum_t R_t$ and $G_\lambda(\tau) = \sum_t \widetilde{R}_t$,

$$\nabla_w F_\lambda(w) = \mathbb{E}_w[G_\lambda(\tau) \nabla_w \log P_w(\tau)] - \frac{\lambda}{T} \sum_t \mathbb{E}_w[(D_w Z_t)^\top \nabla_z u_w(Z_t)].$$

The derivative $\nabla_z u_w$ changes its input at the current density; variation of the density itself was already accounted for above. The constant noise entropy in F_λ has zero derivative.

Writing P_{env} for the observation kernels, the complete trajectory density factors as

$$P_w(\tau) = p(X_0) \prod_{t=0}^{T-1} [w_t(A_t \mid H_t) P_{\mathrm{env}}(X_{t+1} \mid H_t, A_t)].$$

Hence $\nabla_w \log P_w = \sum_t \nabla_w \log w_t(A_t \mid H_t)$. For every t, earlier augmented rewards are measurable with respect to H_t, and $\mathbb{E}_w[\nabla_w \log w_t(A_t \mid H_t) \mid H_t] = 0$. Their contribution to the corresponding score term is zero. Replacing G_λ by $G_{\lambda,t}$ in that term therefore gives equation 2. Finally $D_\theta Z_t = 0$ at fixed histories, so the direct term acts only on the encoder parameters. □

For completeness, the separate parameter gradients are

$$\nabla_\theta F_\lambda(w) = \mathbb{E}_w \sum_t G_{\lambda,t} \nabla_\theta \log w_t(A_t \mid H_t),$$

$$\nabla_\phi F_\lambda(w) = \mathbb{E}_w \sum_t \left[G_{\lambda,t} \nabla_\phi \log w_t(A_t \mid H_t) - \frac{\lambda}{T} (D_\phi Z_t)^\top \nabla_z u_w(Z_t) \right].$$

A frozen-density identity for local comparisons. The statistical comparison and PPO-style update use data from an old controller w_0. For that separate calculation, put $p_0 = p_{w_0}^{\mathrm{occ}}$ and $c_0(z) = D_{\mathrm{KL}}(k_z \| p_0) = u_{w_0}(z) - h(\sigma\xi)$. For any candidate q,

$$\frac{1}{T} \sum_t \mathbb{E}_q c_0(Z_t) = \mathbb{E}_{Y \sim p_q^{\mathrm{occ}}}[-\log p_0(Y)] - h(\sigma\xi) = C(q) + D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0).$$

Consequently

$$F_\lambda(q) = J(q) - \frac{\lambda}{T} \sum_t \mathbb{E}_q c_0(Z_t) + \lambda D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0),$$

which is also equation 11. The KL term and its first derivative vanish at $q = w_0$ under the regularity conditions above. This auxiliary identity explains why a cost estimated at the old policy can give the exact gradient there. The direct policy-gradient theorem itself uses the current density.

E THE SCORE PROXY AND ITS APPROXIMATION GAP

E.1 DIFFUSION LIKELIHOOD, ELBO AND REPRESENTATION ENTROPY

The diffusion identities use likelihood weighting $\beta(\tau)$. Assume $\tau_0 > 0$, $\alpha_{\tau_0} > 0$, a finite endpoint $\mathcal{T}$, and a positive integrable schedule on this interval. Assume the score field is differentiable in space, all integrals below are finite, the integrations by parts have zero boundary terms, and the forward and learned reverse path laws are nonexplosive and admit the absolute continuity and finite relative entropy needed for Girsanov's identity. These are diffusion regularity assumptions in addition to the bounded encoder outputs.

This calculation uses the score network with diffusion-time input τ and no decision-time input t. Write $m = d_Z$ for the representation dimension in this calculation. Let $Z \sim \mu_w^{\mathrm{occ}}$, and set $D_\tau = \alpha_\tau Z + \sigma_\tau \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_m)$ is independent of Z and of the input noise used to choose actions, $\alpha_\tau^2 + \sigma_\tau^2 = 1$ and $\sigma_{\tau_0}/\alpha_{\tau_0} = \sigma$. The VP diffusion has drift $-\beta(\tau)D_\tau/2$ and diffusion coefficient $\sqrt{\beta(\tau)}$. Put $r_\tau^z = \mathcal{N}(\alpha_\tau z, \sigma_\tau^2 I_m)$, $q_\tau = \int r_\tau^z \mu_w^{\mathrm{occ}}(dz)$, $a_\tau^z = \nabla \log r_\tau^z$, and $\gamma_m = \mathcal{N}(0, I_m)$. Assume a finite terminal time $\mathcal{T}$, integrable β, and finite integrated score loss. The centered cost u_ψ in equation 3 subtracts a conditional-score norm independent of encoder and score-network parameters.

The continuous-time ELBO. The model input is $Y = D_{\tau_0}/\alpha_{\tau_0} = Z + \sigma\xi$. After that starting time, the forward path follows the VP SDE. With increasing reverse time $v = \mathcal{T} - \tau$, the learned reverse process starts from γ_m and has drift $\beta(\tau)x/2 + \beta(\tau)s_\psi(x, \tau)$ and diffusion coefficient $\sqrt{\beta(\tau)}$. Let m_ψ be the density of its terminal output rescaled by $1/\alpha_{\tau_0}$. For a discretized forward noising path $\mathbf{d}$,

$$\mathcal{E}_\psi(y) = \mathbb{E}_{Q(\mathbf{d}|y)} \log \frac{m_\psi(y, \mathbf{d})}{Q(\mathbf{d} \mid y)}, \qquad \log m_\psi(y) - \mathcal{E}_\psi(y) = D_{\mathrm{KL}}(Q(\mathbf{d} \mid y) \| m_\psi(\mathbf{d} \mid y)) \geq 0.$$

The continuous-time ELBO is the corresponding path-space limit, with relative densities interpreted as Radon–Nikodym derivatives. The regularity assumptions above ensure its finite value. For any noisy representation law p, its expected negative ELBO therefore satisfies

$$-\mathbb{E}_{Y\sim p}\mathcal{E}_\psi(Y) - h(p) = D_{\mathrm{KL}}(p \| m_\psi) + \mathbb{E}_{Y\sim p} D_{\mathrm{KL}}(Q(\mathbf{d} \mid Y) \| m_\psi(\mathbf{d} \mid Y)) \geq 0.$$

This entropy–ELBO connection bounds the entropy of the data distribution used for averaging.

For the VP process the same negative ELBO has the score form (Song et al., 2021a)

$$\begin{aligned}\ell_\psi(z) &\equiv -\mathbb{E}_\xi \mathcal{E}_\psi(z + \sigma\xi) \\ &= -\mathbb{E}_{r_\mathcal{T}^z} \log \gamma_m(D_\mathcal{T}) - m \log \alpha_{\tau_0} \\ &\quad + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta(\tau) \mathbb{E}_{r_\tau^z}[\|s_\psi(D_\tau, \tau)\|^2 + 2\nabla \cdot s_\psi(D_\tau, \tau) + m] d\tau.\end{aligned}$$

The rescaling term $-m \log \alpha_{\tau_0}$ accounts for modeling Y rather than D_{τ_0}. Gaussian integration by parts and the conditional entropy identity

$$\frac{d}{d\tau} h(r_\tau^z) = \frac{\beta(\tau)}{2} \left(\mathbb{E}_{r_\tau^z} \|a_\tau^z\|^2 - m\right)$$

give the equivalent denoising expression

$$\ell_\psi(z) = h(\sigma\xi) + D_{\mathrm{KL}}(r_\mathcal{T}^z \| \gamma_m) + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta(\tau) \mathbb{E}_{r_\tau^z} \|s_\psi - a_\tau^z\|^2 d\tau. \tag{15}$$

In particular, $\mathbb{E} s_\psi \cdot a_\tau^z = -\mathbb{E}\nabla \cdot s_\psi$ and $h(r_{\tau_0}^z) - m \log \alpha_{\tau_0} = h(\sigma\xi)$. These identities supply all constants when converting the likelihood bound to the denoising loss.

Proposition E.1 (Population proxy gap)**.** *With u_ψ from equation 3 and the common encoder radius M, define the controller-independent constant*

$$K_{\mathrm{VP}} = h(\sigma G) + \frac{m}{2} \int_{\tau_0}^{\mathcal{T}} \frac{\beta(\tau)}{\sigma_\tau^2} \, d\tau + \frac{m}{2}(\sigma_\mathcal{T}^2 - 1 - \log \sigma_\mathcal{T}^2) + \frac{\alpha_\mathcal{T}^2 M^2}{2}.$$

Then

$$\begin{aligned}\mathbb{E}_{\mu_w^{\mathrm{occ}}} u_\psi(Z) + K_{\mathrm{VP}} - h(\widetilde{Z}_\sigma) &= D_{\mathrm{KL}}(q_\mathcal{T} \| \gamma_m) + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta(\tau) \mathbb{E}_{q_\tau} \|s_\psi - \nabla \log q_\tau\|^2 d\tau \\ &\quad + \frac{\alpha_\mathcal{T}^2}{2}(M^2 - \mathbb{E}\|Z\|^2) \geq 0.\end{aligned} \tag{16}$$

Proof. The two conditional laws r_τ^z and q_τ follow the same VP diffusion. Differentiating their relative entropy through the Fokker–Planck equations and integrating by parts gives

$$\frac{d}{d\tau} D_{\mathrm{KL}}(r_\tau^z \| q_\tau) = -\frac{\beta(\tau)}{2} \mathbb{E}_{r_\tau^z} \| a_\tau^z - \nabla \log q_\tau \|^2.$$

The Gaussian score identity $\mathbb{E}[a_\tau^Z(D_\tau) \mid D_\tau] = \nabla \log q_\tau(D_\tau)$ makes the cross term vanish when the denoising loss is expanded and averaged over Z. At the initial time, $I(Z; D_{\tau_0}) = I(Z; Z + \sigma G) = C(w)$ because scaling by $\alpha_{\tau_0} > 0$ is invertible. At the terminal time, the KL chain rule gives

$$\mathbb{E}_Z D_{\mathrm{KL}}(r_{\mathcal{T}}^Z \| \gamma_m) = I(Z; D_{\mathcal{T}}) + D_{\mathrm{KL}}(q_{\mathcal{T}} \| \gamma_m).$$

Integrating the entropy dissipation identity therefore yields

$$\begin{aligned} &\mathbb{E}_Z \left[D_{\mathrm{KL}}(r_{\mathcal{T}}^Z \| \gamma_m) + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta \mathbb{E}_{r_\tau^Z} \| s_\psi - a_\tau^Z \|^2 d\tau \right] \\ &= C(w) + D_{\mathrm{KL}}(q_{\mathcal{T}} \| \gamma_m) \\ &\quad + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta \mathbb{E}_{q_\tau} \| s_\psi - \nabla \log q_\tau \|^2 d\tau. \end{aligned}$$

Finally, $D_{\mathrm{KL}}(r_{\mathcal{T}}^z \| \gamma_m) = \frac{m}{2}(\sigma_{\mathcal{T}}^2 - 1 - \log \sigma_{\mathcal{T}}^2) + \alpha_{\mathcal{T}}^2 \|z\|^2/2$. Subtract the centering term in equation 3, add K_{VP}, and use $h(\widetilde{Z}_\sigma) = C(w) + h(\sigma G)$. This proves equation 16. □

Proof of Proposition 7.1. Average equation 15 over $Z \sim \mu_w^{\mathrm{occ}}$. The conditional-score identity and relative-entropy dissipation calculation above give

$$\mathbb{E}\ell_\psi(Z) = h(p_w^{\mathrm{occ}}) + e_{\mathrm{diff}}, \qquad e_{\mathrm{diff}} = D_{\mathrm{KL}}(q_{\mathcal{T}} \| \gamma_m) + \frac{1}{2} \int_{\tau_0}^{\mathcal{T}} \beta(\tau) \mathbb{E}_{q_\tau} \| s_\psi - \nabla \log q_\tau \|^2 d\tau. \quad (17)$$

The terminal Gaussian KL formula and equation 3 also give, for every $\|z\| \le M$,

$$u_\psi(z) + K_{\mathrm{VP}} = \ell_\psi(z) + \frac{\alpha_{\mathcal{T}}^2}{2}(M^2 - \|z\|^2).$$

Averaging establishes equation 4. □

The bound is on the *occupancy expectation* of u_ψ, not on its value for each clean state. For fixed ψ, write $U_\psi(w) = \mathbb{E}_{\mu_w^{\mathrm{occ}}} u_\psi(Z) + K_{\mathrm{VP}}$ and $g_\psi(w) = U_\psi(w) - h(p_w^{\mathrm{occ}}) \ge 0$. Then for any w, w_0,

$$\begin{aligned} &[J(w) - \lambda h(p_w^{\mathrm{occ}})] - [J(w_0) - \lambda h(p_{w_0}^{\mathrm{occ}})] \\ &= [J(w) - \lambda U_\psi(w)] - [J(w_0) - \lambda U_\psi(w_0)] \\ &\quad + \lambda [g_\psi(w) - g_\psi(w_0)]. \end{aligned}$$

Thus improvement in the proxy alone guarantees improvement in the true objective only after accounting for at least the old-point gap and the empirical and optimization errors.

F A CONDITIONAL BRPO IMPROVEMENT STATEMENT

F.1 IMPROVEMENT OF THE ACTUAL DIFFUSION-COST UPDATE

Theorem 7.1 uses the fixed score cost u_ψ, including its finite diffusion endpoint. Define

$$F_\psi(w) = J(w) - \lambda \left[\frac{1}{T} \sum_t \mathbb{E}_w u_\psi(\phi_{w,t}(H_t)) + K_{\mathrm{VP}} \right].$$

For all candidate encoders obeying the same radius bound, Proposition 7.1 gives $F_\lambda(w) = F_\psi(w) + \lambda g_\psi(w) + \lambda h(\sigma\xi)$. Thus

$$F_\lambda(q) - F_\lambda(w_0) \ge F_\psi(q) - F_\psi(w_0) - \lambda g_\psi(w_0). \quad (18)$$

Population surrogate and empirical error. Let $A^0_{\psi,t}(H,a)$ be the exact advantage under w_0 for rewards $R_t - \lambda u_\psi(\phi_{0,t}(H_t))/T$. Assume $|A^0_{\psi,t}| \le A_{\mathrm{aug}}$ at every relevant history and action. With the same clipping threshold as BRPO, put

$$D_\psi(q) = \sum_t \mathbb{E}_{w_0} \min\{\rho_t A^0_{\psi,t}, \mathrm{clip}(\rho_t, 1\pm\epsilon_{\mathrm{clip}})A^0_{\psi,t}\} - \frac{\lambda}{T}\sum_t \mathbb{E}_{w_0}[u_\psi(\phi_{q,t}(H_t)) - u_\psi(\phi_{0,t}(H_t))].$$

At $q = w_0$, $D_\psi(w_0) = 0$, since the old-action mean of each true advantage is zero. The empirical-error condition in the main theorem is precisely $|\widehat{\Delta}L - D_\psi(q)| \le e_{\mathrm{alg}}$. It covers sampled advantages, cost quadrature, finite-batch averages and any action-ratio approximation. When this holds uniformly over candidates on an event of probability $1-\delta$, the theorem holds with that probability for a candidate selected using the batch. The score model may be fit on the same batch; this requires the stated error event to cover that fitted model. The theorem is conditional on that event and does not assume independence of a fitted model and its training data.

Why the surrogate has the right two gradient terms. Assume the controller differentiability and domination conditions in Appendix D also hold for the fixed diffusion cost. At $q = w_0$ all action ratios equal one, so clipping is inactive for the pointwise derivative. Differentiating $D_\psi(q)$ there gives

$$\nabla_q D_\psi(q)|_{q=w_0} = \mathbb{E}_{w_0}\sum_t A^0_{\psi,t}\nabla_w \log w_{0,t}(A_t \mid H_t) - \frac{\lambda}{T}\sum_t \mathbb{E}_{w_0}(D_w Z_t)^\top \nabla_z u_\psi(Z_t) = \nabla_w F_\psi(w_0).$$

The last equality is the policy-gradient identity for the fixed cost u_ψ, with a history-dependent value baseline. It holds for both head and encoder parameters. The separate direct term is needed because differentiating action probabilities alone omits the encoder's effect on costs at fixed histories. The true-objective derivative differs by $\lambda\nabla_w g_\psi(w_0)$. An upper bound on the value of g_ψ does not on its own bound that derivative; the improvement theorem below controls objective values directly.

Proof of Theorem 7.1. First retain the old encoder cost at every history and change only the action kernel. The finite-episode performance-difference identity gives the proxy return difference $\sum_t \mathbb{E}_{H_t\sim q} g_t(H_t)$, where $g_t(H) = \sum_a q_t(a \mid H)A^0_{\psi,t}(H,a)$. The old-action mean is zero. The action TV condition therefore implies $|g_t(H)| \le 2A_{\mathrm{aug}}\alpha$. Coupling the trajectories step by step through the same environment bounds the time-t history-law TV distance by $t\alpha$. Consequently, replacing candidate history laws by old history laws costs at most $4A_{\mathrm{aug}}\alpha^2\sum_t t = 2A_{\mathrm{aug}}T(T-1)\alpha^2$. Importance weighting yields the unclipped old-history action surrogate, which pointwise clipping can only reduce.

Next include the changed encoder cost $\Delta u_t(H) = u_\psi(\phi_{q,t}(H)) - u_\psi(\phi_{0,t}(H))$. Its candidate-history expectation subtracts $\lambda T^{-1}\sum_t \mathbb{E}_q \Delta u_t$. The bound $|\Delta u_t| \le e_u$ and the same coupling show that replacing it by the old-history expectation costs at most $2\lambda e_u \alpha T^{-1}\sum_t t = \lambda e_u(T-1)\alpha$. Together these steps give

$$F_\psi(q) - F_\psi(w_0) \ge D_\psi(q) - 2A_{\mathrm{aug}}T(T-1)\alpha^2 - \lambda e_u(T-1)\alpha.$$

Apply equation 18 and the empirical-error condition to obtain equation 6. □

The theorem uses the realized surrogate gain. Imperfect optimization can reduce this gain without invalidating the bound. The explicit old-point gap is

$$g_\psi(w_0) = D_{\mathrm{KL}}(q_{0,\mathcal{T}}\|\gamma_m) + \frac{1}{2}\int_{\tau_0}^{\mathcal{T}} \beta(\tau)\mathbb{E}_{q_{0,\tau}}\|s_\psi - \nabla \log q_{0,\tau}\|^2 d\tau + \frac{\alpha_{\mathcal{T}}^2}{2}(M^2 - \mathbb{E}_{w_0}\|\widetilde{Z}\|^2),$$

where $q_{0,\tau}$ is the old controller's diffused occupancy. This separates diffusion approximation from the empirical error. PPO clipping alone does not enforce the theorem's action-TV or encoder-cost bounds. Small average score error alone does not establish a uniform gradient approximation; the theorem instead uses the population ELBO gap.

F.2 THE EXACT FROZEN-COST COMPARISON

The next statement isolates the changes made by clipping, visitation drift, encoder drift, and approximation error. It uses the exact frozen cost; all differences between that cost and BRPO's score estimate belong in the stated error.

Proposition F.1 (Local improvement under controlled drift)**.** *Fix an old controller w_0 and its cost c_0. Let $A_t^0(H, a)$ be the true history-conditioned advantage for rewards $R_t - \lambda c_0(\phi_0(H_t))/T$, with $|A_t^0| \le A_{\mathrm{aug}}$. Suppose a candidate q has the old action support and $\mathrm{TV}(q_t(\cdot \mid H), w_{0,t}(\cdot \mid H)) \le \alpha$ at every relevant history. Let $\Delta c_t(H) = c_0(\phi_q(H)) - c_0(\phi_0(H))$ and $\sup_{t,H} |\Delta c_t(H)| \le e_c$. For the population clipped surrogate*

$$S_0^{\mathrm{clip}}(q) = \sum_{t=0}^{T-1} \mathbb{E}_{w_0} \min\{\rho_t A_t^0, \mathrm{clip}(\rho_t, 1 \pm \epsilon_{\mathrm{clip}}) A_t^0\}, \quad \rho_t = \frac{q_t(A_t \mid H_t)}{w_{0,t}(A_t \mid H_t)},$$

one has

$$F_\lambda(q) - F_\lambda(w_0) \ge S_0^{\mathrm{clip}}(q) - \frac{\lambda}{T} \sum_t \mathbb{E}_{w_0} \Delta c_t(H_t) - 2A_{\mathrm{aug}} T(T-1)\alpha^2 - \lambda e_c (T-1)\alpha. \tag{19}$$

If an empirical update expression differs from the first two terms on the right by at most e_{alg}, the right side may be reduced by e_{alg} to cover that error.

Proof. The frozen-cost decomposition equation 11 implies $F_\lambda(q) - F_\lambda(w_0) \ge \mathcal{L}_0(q) - \mathcal{L}_0(w_0)$. First hold the old encoder cost fixed while changing the action kernel. The complete-history performance-difference identity yields $\sum_t \mathbb{E}_{H_t \sim q} g_q(H_t)$, where $g_q(H) = \sum_a q_t(a \mid H) A_t^0(H, a)$. The old-action mean of the advantage is zero, so $|g_q(H)| \le 2A_{\mathrm{aug}}\alpha$. Stepwise coupling of trajectories under the same environment gives at time t a history-law TV distance at most $t\alpha$. Replacing each new history law by the old one costs at most $4A_{\mathrm{aug}} t\alpha^2$; summing $t = 0, \dots, T-1$ gives the third term in equation 19. Importance weighting on old support identifies the old-history expectation with the unclipped advantage surrogate. Pointwise clipping can only lower that surrogate.

Now replace the old encoder's cost at each history by the candidate encoder's cost. This subtracts $\lambda T^{-1} \sum_t \mathbb{E}_q \Delta c_t(H_t)$. Changing its history law to the old one costs at most $2\lambda e_c T^{-1} \sum_t t\alpha = \lambda e_c (T-1)\alpha$. Combining these estimates proves the claim. □

Lemma B.2 gives $e_c \le 2M\rho/\sigma^2$ when the encoder output drift is at most ρ. The error e_{alg} must cover the score proxy, sampled advantages, finite-batch averages, numerical integration and optimization. Neither a low average denoising loss nor PPO clipping alone supplies the proposition's uniform cost and action-drift conditions.

BRPO training loop. For a batch $\mathcal{B}$ of N complete episodes, the notation in equation 5 means

$$\widehat{\mathbb{E}}_{\mathcal{B}} v_t = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} v_{it}, \qquad \widehat{U}^{\mathrm{dir}}_{\psi,\mathcal{B}}(\phi) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} u_\psi(\phi_t(H_{it})).$$

The first quantity estimates an episode sum; the second estimates the time-averaged representation cost. These normalizations match the undiscounted return and representation-entropy penalty in the objective. Each iteration collects complete on-policy episodes using fresh policy-input noise, trains s_ψ on their pooled clean encoder states with the diffusion loss in equation 3, and builds augmented rewards and critic targets. It then computes advantages from those rewards and optimizes equation 5 over the recorded batch, including the direct encoder score term. The next iteration collects new episodes under the updated controller. Discounted generalized advantage estimation (GAE) (Schulman et al., 2016), if used in an implementation, introduces an additional approximation to the undiscounted complete-episode objective analyzed here.

Algorithm 1 summarizes this training loop.

For a marginal action ratio, both numerator and denominator integrate the readout noise. A sampled-noise ratio instead conditions on the recorded noise draw in both likelihoods. Implementations using

Algorithm 1 Bounded-Rational Policy Optimization

Require: Controller $w = (\theta, \phi)$, critic V_ω, score network s_ψ, fixed noise σ, penalty weight λ.

1: **for** each training iteration **do**
2: Snapshot the old controller $w_0 \leftarrow w$.
3: Collect complete episodes under w_0, recording histories, actions, rewards, and representations Z_t before noise is added.
4: Fit s_ψ with denoising score loss on these representations pooled uniformly over episodes and decisions.
5: Estimate $u_\psi(Z_t)$ by equation 3 and form $\widetilde{R}_t = R_t - \lambda u_\psi(Z_t)/T$.
6: Fit V_ω to augmented-return targets and compute old-controller advantages $\widetilde{A}_t^0$.
7: With the score network fixed, update (θ, ϕ) on the recorded batch using equation 5; evaluate its direct encoder term on the recorded histories.
8: **end for**

this joint likelihood must record the sampled readout noise together with each action. The advantage and critic estimates, diffusion integral, and finite-batch optimization introduce the errors listed after Proposition F.1.

Proposition F.2 (Clean occupancy drift and the frozen-cost remainder)**.** *Let μ_0, μ_q be the complete-episode, uniformly time-averaged laws of the clean representations under two controllers, expressed in the same latent coordinates. If $D_{\mathrm{KL}}(\mu_q \| \mu_0) \le \epsilon_{\mathrm{kl}}$, then*

$$D_{\mathrm{KL}}(p_q^{\mathrm{occ}} \| p_0^{\mathrm{occ}}) \le \epsilon_{\mathrm{kl}}.$$

Consequently, the KL remainder in equation 11 is at most $\lambda\epsilon_{\mathrm{kl}}$.

Proof. Both noisy occupancy laws are obtained by passing their clean laws through the same Gaussian channel. Apply the data-processing inequality for KL divergence. □

G ROCKSAMPLE AND MAP-SIZE SCALING

Following RockSample (Smith & Simmons, 2004), we consider finite-horizon tasks on $k \times k$ grids, with $k \in \{7, 11, 15, 19\}$. Each instance contains k rocks at fixed locations; their binary qualities are sampled independently from a uniform prior at the beginning of each episode. For $k \in \{7, 11, 15\}$, the rover starts uniformly in a non-exit column. The 19×19 PathRocks variant instead uses a fixed start on the western boundary and a fixed, path-oriented rock layout. An episode ends when the rover reaches the eastern exit or after 1000 transitions.

Observations and actions. Each observation concatenates one-hot encodings of the rover row and column with one measurement coordinate per rock. The resulting observation dimensions are 21, 33, 45, and 57. A sensing action produces a signed reading for the selected rock, with accuracy decreasing with rover–rock distance; the remaining rock coordinates are zero. The action space contains four cardinal moves, one sampling action, and one sensing action per rock, for 12, 16, 20, and 24 actions, respectively.

Transitions and rewards. Rover motion is deterministic, whereas sensing is noisy and does not alter the state. Sampling a good rock yields $+10$ and permanently changes that rock to bad; sampling a bad rock yields -10. Reaching the eastern exit yields $+10$, and all other transitions have zero reward. We report undiscounted episodic return for every method.

Training configuration. BRPO and PPO–RNN use a 256-dimensional gated recurrent unit (GRU) encoder, batches of 2048 transitions, four PPO epochs per update, minibatches of 512, and a clipping threshold of 0.2. The policy and value learning rates are 3×10^{-4}, and the encoder learning rate is 10^{-4}. At every scale, the discount factor is 0.99 and the generalized advantage estimation (GAE) coefficient is 0.95. The PPO action-entropy bonus coefficient is held at 0.1 for 2000 updates, linearly annealed to 0.01 over the next 5000 updates, and then held fixed. We train for 10,000 updates and evaluate on 100 episodes every 100 updates. BRPO uses $\lambda = 0.1$, diffusion scale $\sigma = 0.1$,

and a VP diffusion score-network learning rate of 5×10^{-4}. PPO–RNN sets $\lambda = 0$ and omits the representation-entropy estimator. On the 15×15 main task, AIS recurrently constructs an approximate information state from the observation, previous action, and previous reward, and trains it through immediate-reward and next-observation prediction; for a controlled comparison, we update its policy with PPO. We additionally evaluate three online POMDP planners on this main task: POMCP performs Monte Carlo tree search from a particle belief, POMCPOW augments this search with observation progressive widening, and AdaOPS uses bound-guided search with adaptive belief packing. All three planners use the exact generative model and undergo 100-episode evaluation.

H NOISYTV

Inspired by stochastic distractors in curiosity-driven learning (Burda et al., 2019), NoisyTV is a finite-horizon partially observed navigation task on an 11×11 grid. At the beginning of each episode, one of two terminal goals is selected uniformly as the correct goal, but its identity is hidden after the initial time step.

Observations and actions. Each observation is an 11×11 RGB image containing the grid, agent, goals, and a 5×5 stochastic "TV" region. A colored cue identifies the correct goal only in the initial observation, requiring the agent to retain this information throughout the episode. The action space contains four cardinal movement actions.

Transitions and rewards. Agent motion is deterministic and walls block invalid moves; TV pixels are independently resampled at every step. Reaching the correct goal yields $+1$, reaching the other goal yields -1, and each nonterminal transition yields -0.015. An episode ends at either goal or after 60 transitions. For all methods, we report undiscounted episode return.

Training configuration. BRPO and PPO–RNN use a 256-dimensional GRU encoder, batches of 4000 transitions, one PPO epoch per update, minibatches of 512, and a clipping threshold of 0.2. The policy and value learning rates are 10^{-4}, the encoder learning rate is 5×10^{-5}, the discount factor is 0.95, and the GAE coefficient is 0.95. The PPO action-entropy bonus coefficient is held at 0.02 for 3000 updates and then linearly annealed to 0.005 over 2000 updates. We train for 5000 updates and evaluate on 100 episodes every 100 updates. BRPO uses $\lambda = 0.001$, diffusion scale $\sigma = 0.1$, and a VP diffusion score-network learning rate of 10^{-3}. PPO–RNN sets $\lambda = 0$ and omits the representation-entropy estimator. AIS uses an LSTM to construct an approximate information state from the observation, previous action, and previous reward, trains the state through immediate-reward and next-observation prediction, and updates the policy with PPO rather than REINFORCE. For the online-planning baselines, POMCP performs Monte Carlo tree search over particle beliefs, POMCPOW adds observation progressive widening, and AdaOPS performs bound-guided search with adaptive belief packing. These planners use the exact environment generative model and are evaluated on 100 episodes.

H.1 NOISYTVHARD

NoisyTVHard is a more challenging variant of NoisyTV on a 13×13 grid. At the beginning of each episode, one of two terminal goals is selected uniformly as the correct goal, but its identity is revealed only at the initial time step.

Observations and actions. Each observation is a 13×13 RGB image of the grid, agent, goals, and a 7×7 stochastic "TV" region. A one-step color cue identifies the correct goal and must be retained while navigating. Actions move in the four cardinal directions.

Transitions and rewards. Agent motion is deterministic and walls block invalid moves; TV pixels are independently resampled at every step. Reaching the correct goal yields $+1$, reaching the other goal yields -1, and each nonterminal transition yields -0.015. An episode ends at either goal or after 75 transitions.

Training configuration. The observation encoder consists of two 3×3 convolutional layers with 32 channels and ReLU activations, followed by a linear projection, layer normalization, and a 128-dimensional GRU. All four ablation settings use batches of 1000 transitions, two PPO epochs per update, minibatches of 512, a clipping threshold of 0.2, discount 0.95, GAE coefficient 0.95, and a constant PPO action-entropy bonus coefficient of 0.02. The policy and value learning rates are 10^{-4}, the encoder learning rate is 5×10^{-5}, and the score network learning rate is 10^{-3}. Training lasts 3000 updates. The four settings independently choose the direct encoder and reward-side penalty coefficients from $\{0, 3 \times 10^{-5}\}$, yielding both off, encoder only, reward only, and both on.

I LaserTag

We use the known-robot-position LaserTag benchmark (Somani et al., 2013) on a 7×11 grid, with the fixed obstacle layout from the standard `cpp_emu_lasertag(4)` configuration. The opponent position is hidden.

Observations and actions. Each observation contains eight normalized noisy range readings along the cardinal and diagonal directions, followed by one-hot encodings of the robot row and column, for 26 features in total. The range-noise standard deviation is 2.5. The action space contains four cardinal movement actions and one tagging action.

Transitions and rewards. Robot motion is deterministic and blocked by obstacles, whereas the opponent moves stochastically with a bias away from the robot. A movement action yields -1; a successful tag yields $+10$ and terminates the episode, whereas an unsuccessful tag yields -10. Episodes are truncated after 100 transitions. Evaluation always uses the original, unshaped reward, and we report undiscounted episodic return for every method.

Training configuration. BRPO and PPO–RNN use a 256-dimensional GRU encoder, batches of 2048 transitions, two PPO epochs per update, minibatches of 512, and a clipping threshold of 0.2. The policy and value learning rates are 3×10^{-4}, the encoder learning rate is 10^{-4}, the discount factor is 0.95, and the GAE coefficient is 0.95. The PPO action-entropy bonus coefficient is held at 0.06 for 2000 updates and then annealed toward 0.015 over 6000 updates. We train for 7000 updates and evaluate on 50 episodes every 100 updates. During training only, we add a clipped positive distance-reduction reward with coefficient 0.25 and noise standard deviation 0.25. BRPO uses $\lambda = 0.001$, diffusion scale $\sigma = 0.1$, and a VP diffusion score-network learning rate of 10^{-3}. PPO–RNN sets $\lambda = 0$ and omits the representation-entropy estimator. AIS recurrently constructs an approximate information state from the observation, previous action, and previous reward, trains it through immediate-reward and next-observation prediction, and updates its policy with PPO; it uses the same training-only distance shaping and unshaped evaluation as BRPO and PPO–RNN. For the online-planning baselines, POMCP searches from a particle belief, POMCPOW adds observation progressive widening, and AdaOPS uses bound-guided search with adaptive belief packing. They use the exact generative model and original reward, with evaluation over 100 episodes.

J LightDark and dimension scaling

LightDark2D. Following the position-dependent sensing idea of LightDark (Platt et al., 2010), the base task comprises two independent one-dimensional coordinates controlled by a single policy. At the start of an episode, each hidden position is drawn independently from $\mathcal{N}(2, 3^2)$. The goal is at zero and the light source is at two for each coordinate; a tag is correct when the corresponding position lies within one unit of the goal.

Observations and actions. The observation contains one noisy position measurement and one exact completion indicator for each coordinate. At position x_i, the measurement noise has standard deviation $|x_i - 2|/\sqrt{2} + 0.01$, and measured positions are divided by 10 before entering the agent. LightDark2D therefore has four observation features and six discrete actions: move left, move right, or tag either coordinate. Tagging a coordinate removes its three actions from the policy support.

Table 5: **LightDark dimension-scaling environments.** Each task contains N independent coordinates and permits one coordinate-level action per time step.

Coordinates	Obs. dim.	Actions	Horizon
2	4	6	24
4	8	12	48
8	16	24	96

Transitions and rewards. A movement action changes the selected position by one and carries no reward. A correct tag yields $+10$, whereas an incorrect tag yields -10; tags are irrevocable. An episode ends after both coordinates have been tagged or after 24 transitions. At the time limit, each untagged coordinate is scored as an incorrect tag.

Dimension scaling. The 4D and 8D tasks stack four and eight independent copies of the same coordinate process. One action still operates on one coordinate per time step, and the horizon grows linearly with N. Table 5 summarises the resulting spaces. To make scores comparable across dimensions, we divide total episodic return by N.

Training configuration. BRPO and PPO–RNN use a 256-dimensional GRU encoder, batches of 2048 transitions, three PPO epochs per update, minibatches of 512, and a clipping threshold of 0.2. The policy and value learning rates are 3×10^{-4}, the encoder learning rate is 10^{-4}, and both the discount factor and GAE coefficient are one. We train for 1,020 updates and evaluate on 100 episodes every 100 updates. The initial and final PPO action-entropy bonus coefficients are $(0.060, 0.0030)$, $(0.043, 0.0022)$, and $(0.034, 0.0017)$ for 2D, 4D, and 8D, respectively; each schedule is held for 750 updates and then annealed. BRPO uses $\lambda = 3 \times 10^{-4}$, diffusion scale $\sigma = 0.1$, and a VP–SDE score-network learning rate of 10^{-3}, while PPO–RNN uses $\lambda = 0$ without entropy estimation.

AIS and the online-planning baselines are evaluated only on LightDark2D. AIS learns its recurrent state through immediate-reward and next-observation prediction and updates its policy with PPO. POMCP, POMCPOW, and AdaOPS use the exact Gaussian observation model; POMCP discretises observations for tree reuse, POMCPOW uses observation progressive widening, and AdaOPS performs bound-guided search with adaptive belief packing.

K Double-Pendulum Entropy-Resistance Experiment

Double-Pendulum We use the MuJoCo environment `DoublePendulumPrediction-v0`. Its physical state is $s_t = (\theta_{1,t}, \theta_{2,t}, \dot{\theta}_{1,t}, \dot{\theta}_{2,t})$, where $\theta_{1,t}$ and $\theta_{2,t}$ are the joint angles. The upright reference state is $s^\star = (0, 0, 0, 0)$. We report upright fraction, the proportion of evaluation timesteps satisfying $|\theta_{1,t}| \leq 0.1$ and $|\theta_{2,t}| \leq 0.1$. This metric is computed from the true simulator state and has no effect on training.

Partial observations. The agent observes the two joint angles with independent Gaussian noise but does not observe angular velocity. In the formal comparison, $\eta_{i,t} \sim \mathcal{N}(0, 0.07^2)$ and the angle channels are $(\theta_{1,t} + \eta_{1,t}, \theta_{2,t} + \eta_{2,t})$. The noise affects only the observation, not the MuJoCo state or the evaluation metric. The first observation of an episode additionally provides the noise level; this context channel is zero thereafter. The previous prediction reward, after clipping and scaling, forms the final observation channel. This channel always uses the base prediction reward, including during shaped-reward pretraining, so its meaning does not change between training stages.

Actions and reward. At each timestep, the policy produces a two-dimensional joint-torque command a_t^{ctrl} and a four-dimensional prediction a_t^{pred} of the current physical state. The formal task reward is $r_t^{\mathrm{pred}} = -\|a_t^{\mathrm{pred}} - s_t\|_2^2$. No upright bonus, survival reward, or fall penalty is present in the formal PPO–BRPO comparison.

Pretraining and noisy adaptation. The experiment uses three stages. First, PPO is trained from scratch under partial but noise-free observations with an upright-shaped reward based on the

Gymnasium `InvertedDoublePendulum` task. This reward encourages the tip to remain high and near the vertical axis, penalises control effort, and penalises falls. Its sole purpose is to acquire a viable stabilising controller. Second, the controller is adapted with the same shaped reward and observation noise $\sigma_{\mathrm{obs}} = 0.05$. We select the best checkpoint from this stage as the common initialization for both methods.

Formal comparison. In the third stage, we increase the observation noise to $\sigma_{\mathrm{obs}} = 0.07$ and remove every shaped-reward term, leaving only r_t^{pred}. PPO uses $\lambda_1 = \lambda_2 = 0$, while BRPO uses $\lambda_1 = \lambda_2 = 1$. The first coefficient regularises the learned representation directly; the second enters the augmented reward and can therefore change which states the policy visits. All other aspects of the initial policy and formal training protocol are shared. The experiment thus tests whether bounded-rational regularisation preserves an acquired low-complexity control regime after removing its explicit stabilisation reward. It does not test discovery of upright control from random initialization.

L Stochastic Drift Arena

Stochastic Drift Arena is a one-dimensional POMDP in which an unobserved binary drift direction follows a slowly switching Markov chain. The agent receives noisy position and drift cues together with task-irrelevant distractors. We use two variants sharing this construction, because the coefficient and encoder ablations test different mechanisms.

Coefficient ablation. For the 2×2 λ_1–λ_2 ablation, the observation contains the noisy position, a noisy drift cue, two phase features, and 16 Gaussian nuisance features. Region B is the right half of the arena; its nuisance standard deviation is 0.03, compared with 1.0 in region A. The four actions are the Cartesian product of two counter-drift choices and two navigation choices. A correct counter-drift choice gives reward 0.1 and an incorrect choice gives zero; occupying region B has no direct reward. Episodes contain 100 transitions, and the first 20 are treated as burn-in for diagnostics.

The four cells independently enable the direct encoder penalty λ_1 and the reward-side penalty λ_2, with each coefficient in $\{0, 0.5\}$. All cells use a 256-dimensional GRU, batches of 256 transitions, one PPO epoch, minibatches of 128, discount 0.99, GAE coefficient 0.95, and 150 updates. The score model is warmed up for 50 updates and uses four Monte Carlo samples. Evaluation is performed every 25 updates using 20 episodes, followed by 50 final episodes.

Encoder and horizon ablation. For the GRU–Transformer comparison, the observation instead contains noisy position and drift cues plus four uniform distractors. Region B is the central interval $|x| \leq 1$ in an arena of radius 4. The three actions apply left, neutral, or right control. Correctly opposing the drift gives $+1$ and an incorrect choice gives -1; being inside region B adds 0.2, being outside adds -0.2, and nonzero control costs 0.01. We compare a spectrally normalised GRU with a Transformer that receives the full history at horizons $H \in \{50, 100, 200, 500\}$; the first 20% of each episode is burn-in.

Spectral normalization controls layer operator norms (Miyato et al., 2018); its use here constrains the recurrent implementation without by itself establishing the theoretical update conditions. Both encoders use a 256-dimensional representation. The Transformer has two layers, four attention heads, and hidden dimension 256. Each condition uses rollout batches of 1024 transitions, one PPO epoch, minibatches of 256, discount 0.99, GAE coefficient 0.95, and shared coefficient $\lambda_1 = \lambda_2 = 0.01$. Training lasts 1000 updates. We evaluate 50 episodes every 50 updates and 100 episodes at the end; the reported comparison uses the post-burn-in fraction of steps in region B, p_B. Each launcher invocation runs one seed.

M Evaluation Protocol and Additional Ablations

This appendix gives the detailed evaluation rules and auxiliary results summarized in Section 8. Environment definitions and training configurations appear in their respective appendices.

M.1 BENCHMARK EVALUATION

Table 1 reports greedy evaluation return on the four benchmark environments. For each learning-based method (PPO-RNN, AIS, and BRPO), we first average the final five evaluations completed before the environment-specific training budget; evaluations are run every 100 training iterations. The budgets are 10,000 iterations for RockSample 15×15, 7,000 for LaserTag, 5,000 for NoisyTV and 1,000 for LightDark 2D. Planning methods have no training curve, so their per-seed score is the mean over all evaluation episodes. We report the mean of the three per-seed scores and their standard deviation.

M.2 SCALING EVALUATION

RockSample map size. Figure 2(a) plots the performance gap $\Delta r(k) = \bar{r}_{\mathrm{BRPO}}(k) - \bar{r}_{\mathrm{PPO}}(k)$ as a function of map size k. We use a common budget of 10,000 training iterations. A run's score is the mean of its final five greedy evaluations at or before the budget, and each method is then averaged over three seeds. The mean performance gap increases with map size, with a pronounced increase on the larger maps. This records how the observed benefit changes with map size at a fixed budget.

LightDark dimension. We independently stack $N \in \{2, 4, 8\}$ LightDark coordinates and scale the episode horizon as $12N$. Figure 2(b) reports $\Delta r(N) = \bar{r}_{\mathrm{BRPO}}(N) - \bar{r}_{\mathrm{PPO}}(N)$ after 1,000 training iterations. Scores are final greedy undiscounted returns per coordinate, averaged over three matched seeds. The gap is small at $N = 2$ and increases with dimension, becoming substantial at $N = 8$. This provides a second setting in which BRPO's advantage grows as the partially observed control problem becomes more complex.

M.3 COMPLETE COEFFICIENT ABLATIONS

We isolate the direct encoder penalty (λ_1) and the reward-side penalty (λ_2) through 2×2 ablations on two tasks. On Stochastic Drift Arena, each coefficient is chosen from $\{0, 0.5\}$. We average the final five evaluations for each seed and report the mean and standard deviation over five matched seeds. Here p_B denotes the fraction of evaluation steps after burn-in spent in the low-noise region B. See Appendix L for full details.

Table 6 shows that λ_1 alone reduces representation entropy but has little effect on task behaviour. In contrast, either setting with λ_2 substantially increases low-noise-region occupancy and task return. The reward-side penalty is therefore the primary mechanism driving behaviour on this task.

Table 6: 2×2 **coefficient ablation on Stochastic Drift Arena.** Greedy results are mean $\pm$ standard deviation over five seeds. Here $h(Z_\sigma)$ denotes the reported differential entropy of noisy representations and may be negative; p_B excludes burn-in.

Configuration	$\boldsymbol{\lambda_1}$	$\boldsymbol{\lambda_2}$	**Task return** $\uparrow$	$\boldsymbol{p_B} \uparrow$	$\boldsymbol{h(Z_\sigma)} \downarrow$
All off (PPO–RNN)	0	0	5.63 ± 0.24	0.203 ± 0.140	1.03 ± 0.79
λ_1 only	0.5	0	5.61 ± 0.24	0.199 ± 0.140	0.87 ± 0.79
λ_2 only	0	0.5	5.94 ± 0.16	0.657 ± 0.192	-4.27 ± 2.67
Both (BRPO)	0.5	0.5	5.94 ± 0.16	0.659 ± 0.192	-4.41 ± 2.67

NoisyTVHard provides the complementary test: control depends on retaining a one-step cue despite a large stochastic visual distractor. Each active coefficient is set to 3×10^{-5}, and we report the final discounted greedy return. See Appendix H.1 for full details.

As shown in Table 7, both settings without λ_1 collapse to cue-independent, single-direction policies, whereas the settings with λ_1 use the cue to select the appropriate goal. In these runs, the encoder-side term is associated with cue retention; the reward-side term alone does not yield this behavior.

Together, the two tasks distinguish the roles of the two penalty terms: λ_1 directly penalizes the encoder on recorded histories; λ_2 subtracts representation costs from rewards and updates action probabilities through the advantages. Their effects are task-dependent and complementary.

Table 7: 2×2 **coefficient ablation on NoisyTVHard.** Discounted greedy returns are mean $\pm$ standard deviation over three seeds.

Configuration	$\boldsymbol{\lambda_1}$	$\boldsymbol{\lambda_2}$	**Discounted return** $\uparrow$
All off (PPO–RNN)	0	0	-0.1407 ± 0.0393
λ_1 only	3×10^{-5}	0	0.3340 ± 0.0000
λ_2 only	0	3×10^{-5}	-0.1732 ± 0.0365
Both (BRPO)	3×10^{-5}	3×10^{-5}	0.3340 ± 0.0000

M.4 Representation Drift During Optimisation

Proposition F.2 bounds noisy-occupancy KL from a bound on the clean representation laws. PPO clipping alone does not supply that condition. We examine empirical representation stability during clipped updates; these diagnostics do not directly estimate the occupancy KL in the proposition.

Measuring representation drift. For each PPO update, we evaluate the encoder before and after optimisation on the same histories. Holding the score-network parameters fixed and using shared diffusion times and noise samples, we measure the root-mean-square change in the outputs of the diffusion denoising network,

$$d_{\text{score}} = \left[\frac{1}{N} \sum_{t=1}^{N} \| s_\psi(\alpha_{\tau_t} z'_t + \sigma_{\tau_t} \epsilon_t, \tau_t) - s_\psi(\alpha_{\tau_t} z_t + \sigma_{\tau_t} \epsilon_t, \tau_t) \|_2^2 \right]^{1/2},$$

as an empirical proxy for representation drift. Here N is the number of recorded samples, z_t and z'_t are the encoder outputs on the same history before and after the update, and τ_t, ϵ_t are shared diffusion times and Gaussian draws. Thus d_{score} measures score-output drift caused by changing the representation input. We additionally record the realised KL divergence between the policies before and after each update.

Drift under clipped policy updates. Table 4 summarises the results over three seeds on RockSample15×15. The realised policy KL remains small, with 96.5% of post-burn-in updates below 0.02. Meanwhile, both the encoder representation drift and the score-output drift remain consistently low across all seeds. Although their update-wise correlation is weak, their joint stability throughout optimisation documents small empirical displacements in these runs; it does not establish that clipping enforces an occupancy-divergence bound.

Main-text Table 4 reports these diagnostics.

M.5 Recurrent and Transformer encoders

We compare a deterministic recurrent encoder with a Transformer (Vaswani et al., 2017) that maps the full history on Stochastic Drift Arena. We vary the episode horizon over $T \in \{50, 100, 200, 500\}$ while holding the entropy-penalty coefficients and PPO settings fixed.

In this variant, p_B is the fraction of post-burn-in steps in the central region $|x| \leq 1$, whose occupancy also earns task reward. Figure 4 reports final p_B over 50 greedy episodes and the transitions required to stably reach $p_B \geq 0.8$. The recurrent encoder maintains $p_B \in [0.920, 0.958]$ and reaches the threshold within 1.55×10^5 transitions, whereas the Transformer's p_B falls from 0.414 to 0.064 and never stably reaches the threshold within the 1.024×10^6-transition budget. These results use one training seed, so variation across seeds remains unmeasured. The statistical theorems require distribution and update conditions beyond this architectural comparison.

M.6 Representation Entropy versus Standard Regularisation

We compare the representation-entropy penalty with KL-to-prior, weight decay, and unregularised PPO-RNN on RockSample 15×15. KL-to-prior is motivated by variational information-bottleneck regularization (Alemi et al., 2017). Our ablation penalizes the aggregate representation law and includes prior-matching effects; it does not reproduce a full conditional variational bottleneck

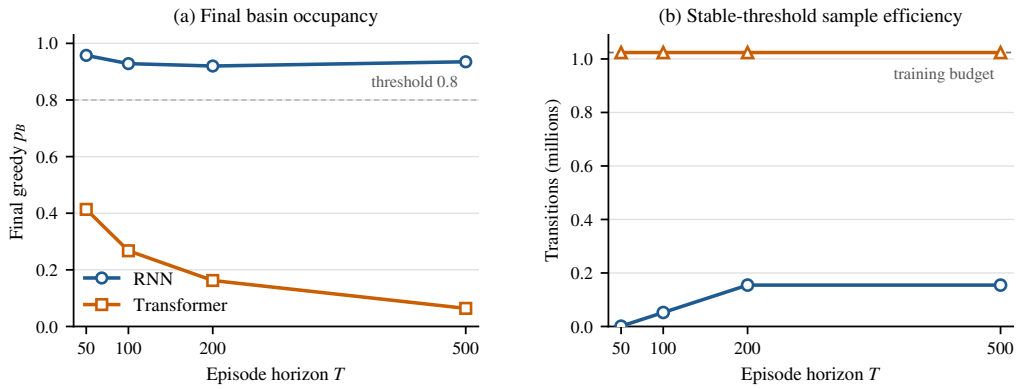

Figure 4: **Recurrent and Transformer encoders on Stochastic Drift Arena.** (a) final p_B over 50 greedy episodes. (b) transitions to stably reach $p_B \geq 0.8$; triangles mark runs that did not stably reach the threshold within the training budget.

objective. Weight decay controls parameter magnitude, whereas the entropy penalty measures visited representations. Under a common budget of 8,000 training iterations, Table 3 reports the mean episode return over the final five greedy evaluations. BRPO achieves 44.00, compared with 10.00 for both standard regularisers and 15.20 for PPO-RNN.

In Table 3, q_Z denotes the learned aggregate representation density and p_Z the KL reference prior.