ACTING IS EASIER THAN ADAPTING: THE GEOMETRY AND DESIGN OF ADAPTATION-SUFFICIENT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A representation can be sufficient for acting in the training environment yet insufficient for adapting after the environment changes. The missing object is not the nominal action but its *prospective response*: how the optimal action moves with the environment. For smooth continuous actions, the curvature-weighted projection of this response onto a frozen code yields a local conservation law, $\Omega = \Gamma_\phi + \Lambda_\phi$, decomposing prospective sensitivity into irrecoverable representation debt and executable motion. That motion is useful only to the extent that deployment evidence can identify which response to execute. An exact three-policy Gaussian frontier prices this evidence and shows that representation refinement is discounted by the resolvable fraction of newly executable motion. Representation rankings are therefore evidence-dependent and can reverse even between equal-capacity codes with identical nominal control. AP-GFAR operationalizes this principle by learning executable prospective responses and selecting representation–evidence programs from calibrated branch values, with freezing as an explicit option. Controlled reward systems verify the refinement law and ranking reversal, and equal-capacity learned encoders reproduce it. In native HalfCheetah, pre-evaluation complete-program pair-value estimates predict heterogeneous rankings under fixed interfaces; a prospectively locked, cross-fitted interface design then yields the predicted reversal across all ten declared systems. Fresh learned-POMDP and fixed-checkpoint VariBAD studies extend the evidence beyond the analytic model, showing gains from complete-program valuation over frozen/no-probe and native online deployment, respectively. Representation value is therefore a property of the representation–evidence pair, not of the representation alone.

1 INTRODUCTION

A standard question in representation learning for control is which parts of an observation can be discarded without changing the action or value needed at the training environment (Rakelly et al., 2021; Huang et al., 2022; Gu et al., 2025). That criterion answers whether a representation can *act*. It does not answer whether the same frozen representation can *adapt*. A variable can be irrelevant to the nominal action and still determine how that action should change after a nearby shift. Once compressed away, merely learning which environment occurred cannot reconstruct the missing state dependence from the downstream code.

Consider $X, U \overset{\text{iid}}{\sim} \text{Unif}[-1, 1]$, actions $a \in \{-1, +1\}$, and reward $r_\delta((X, U), a) = a(X + \delta U)$. The representation $\phi(X, U) = X$ is exactly $Q_0^\star$-sufficient: every nominal action value is a function of (X, a) alone. After a revealed shift with $|\delta| \leq 1$, however, its best policy loses exactly $\delta^2/6$ (Theorem 17). The loss is not caused by uncertainty about the shift; it remains after the shift is revealed because the code has merged histories that require different shifted actions. A transition-only construction gives the same separation (Proposition 18). Nominal sufficiency can therefore fail arbitrarily close to the training environment. This binary example is deliberately nonsmooth: it isolates the information separation, not the derivative theorem below. The conservation law characterizes the complementary regime of smooth continuous actions with stable interior optima.

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.

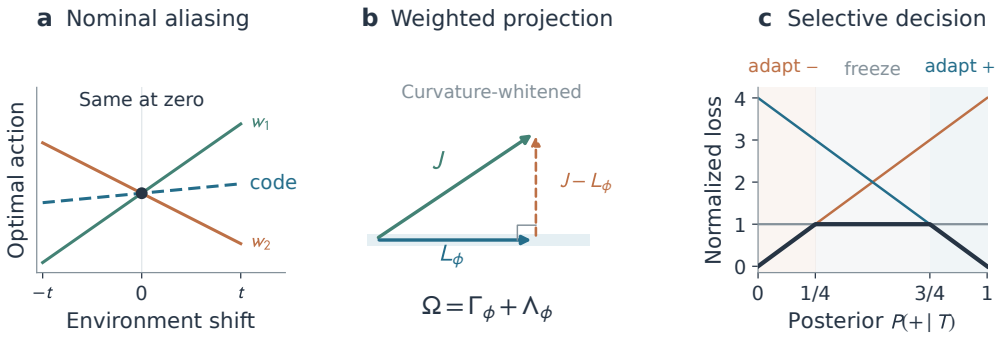

Figure 1: **The missing object is the prospective response.** (a) Histories with the same nominal action can have different action responses to the environment. (b) In curvature-whitened coordinates, projection onto functions of the frozen code separates executable motion from representation debt. (c) Deployment evidence moves posterior risk across two adaptation regions and a freeze region; acquiring that evidence is itself a costed decision.

This exposes a second bottleneck that nominal representation criteria do not price. A capacity-limited encoder must decide which prospective dependencies to retain, but a retained response is useful only when deployment evidence can identify when to execute it. One code may have lower loss after a fully revealed shift yet be worse under the evidence that is actually purchasable. Conversely, a code may preserve an excellent response direction that deployment cannot reliably resolve. Representation design and evidence acquisition are therefore coupled rather than separable decisions.

The primitive missing from nominal sufficiency is the *prospective response*, the local derivative of the optimal action with respect to environment change. We develop a theory that splits its value into two pieces: response motion executable from the frozen code and representation debt that task-only downstream evidence cannot repair. The statistical problem then asks how much of the executable motion can be resolved before its evidence cost exceeds its control value. Figure 1 shows these two bottlenecks in the order in which deployment encounters them.

The paper has one central claim: *adaptation value belongs to the representation–evidence pair*. A curvature-weighted projection yields the conservation law $\Omega = \Gamma_\phi + \Lambda_\phi$; refining a representation moves the same positive-semidefinite quantity from irrecoverable debt to executable motion. Under task-only evidence, an exact quadratic Bayes identity then couples that executable geometry to the task variation resolved by the evidence. A three-policy Gaussian specialization makes the acquisition tradeoff closed form. Together these results show that preserving more prospective motion has deployment value only to the extent that available evidence can resolve it. At fixed capacity this yields a sharp consequence—exchanging which prospective direction has the better evidence can reverse the preferred representation while nominal control is unchanged.

We call the operational consequence of this theory AP-GFAR, a prospective-response-and-evidence deployment rule. It learns an executable response, calibrates the losses of the branches that use it, and selects the representation and evidence allocation before test returns, with freezing as an explicit action. The experiments follow the same hierarchy: controlled systems test the value law and ranking reversal; equal-capacity neural encoders test whether the reversal survives representation learning; native HalfCheetah first tests whether pair values predict ranking heterogeneity under fixed measurement interfaces and then whether a locked interface-design rule can deliberately produce the predicted crossing on held-out banks; a prospectively locked cohort of 32 learned-POMDP systems tests complete-program deployment value; and an established VariBAD checkpoint tests the same valuation principle without retraining the host representation. The conservation law is a local value statement, with a finite-horizon extension under coefficient-stable occupancy in the appendix; the Gaussian frontier is exact for its declared three-policy library. In learned systems the fitted operator K_ϕ is valued as executed rather than identified with the oracle projection L_ϕ. A separate PEARL study isolates finite-sample freeze-or-acquire reliability rather than representation learning.

2 Representation and Deployment Evidence

Let $\{\mathcal{M}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a finite-horizon partially observable Markov decision process (POMDP) family, with nominal parameter $\theta = 0$. A frozen encoder maps history $\mathcal{H}_h$ to $C_h = \phi(\mathcal{H}_h)$. Write V_θ^π for expected cumulative reward. For representation debt, Π_ϕ contains policies whose stage-h action is measurable with respect to C_h and private randomization; the policy itself may be chosen with knowledge of θ. Define

$$V_\theta^\star = \sup_\pi V_\theta^\pi, \qquad V_{\theta,\phi}^\star = \sup_{\pi\in\Pi_\phi} V_\theta^\pi, \qquad \alpha_\phi(\theta) = V_\theta^\star - V_{\theta,\phi}^\star.$$

Thus the benchmark gives the compressed controller the shift for free: α_ϕ isolates what compression has made impossible even after the environment parameter is known.

A declared deployment experiment may additionally generate a protected transcript T_h and use policies measurable with respect to (C_h, T_h). In the common-floor setting of the selective frontier, T_h carries task evidence but no residual control-state information beyond C_h; it can identify which encoded response to execute without changing α_ϕ. For any deployed policy in this setting,

$$V_\theta^\star - V_\theta^\pi = \underbrace{\alpha_\phi(\theta)}_{\text{representation debt}} + \underbrace{V_{\theta,\phi}^\star - V_\theta^\pi}_{\text{encoded decision loss}}. \tag{1}$$

If an interactive transcript also refines the downstream control state, the augmented code (C_h, T_h) has its own debt. The common-floor frontier is then replaced by the augmented geometry or, in the learned sequential experiment, by direct complete-program value calibration.

For a declared experiment library $\mathcal{E}$, let P_θ^e denote the reward-inclusive transcript law under experiment e. For this definition assume the encoded optima are attained, and write $\theta \sim_\mathcal{E} \theta'$ when $P_\theta^e = P_{\theta'}^e$ for every $e \in \mathcal{E}$. Let $\Pi_{\theta,\phi}^\star = \arg\max_{\pi\in\Pi_\phi} V_\theta^\pi$.

Definition 1 (Adaptation sufficiency)**.** *A frozen representation is* prospectively realizable *if* $\alpha_\phi(\theta) = 0$ *throughout the declared family. It is* decision-compatible *with* $\mathcal{E}$ *if every observational indistinguishability class* B *satisfies*

$$\bigcap_{\theta\in B} \Pi_{\theta,\phi}^\star \neq \varnothing.$$

It is adaptation-sufficient when both conditions hold.

Prospective realizability asks whether the required response survives compression. Decision compatibility asks whether environments that no declared evidence can distinguish nevertheless admit a common encoded-optimal policy. The classwise intersection is essential for set-valued optima; pairwise overlap alone need not produce one policy that is optimal for an entire indistinguishable class. With unique encoded optima it reduces to the familiar pairwise separation condition. This is a qualitative structural notion: it rules out a permanent decision conflict inside an observational equivalence class, but it does not assert that one finite or affordable protocol uniformly identifies the classes. Sample complexity and acquisition cost are priced separately below. Family-wise raw-information sufficiency, neighborhood prospective realizability, first-order prospective sufficiency, and nominal sufficiency form a strict hierarchy (Theorem 19).

The geometry below is first stated for a single decision under a fixed history law. The dynamic extension in Appendix A requires stability only of the quadratic value coefficients under policy-dependent occupancy, not total-variation continuity of the full raw-history law. The protected-task-evidence condition above is characterized exactly in Appendix C; interactive trajectory probes need not satisfy it and are valued as complete programs instead.

3 Debt–Motion Geometry

Let W have the fixed nominal history law, $C = \phi(W)$, and $Q_\delta(W, a)$ be the single-decision action value for $a \in \mathbb{R}^m$. All expectations in this section use that law. Suppose the unique interior nominal maximizer $a_0(W) = a_0(C)$ is encoded. For some $\mu > 0$, write

$$H(W) = -\nabla_{aa}^2 Q_0(W, a_0(W)) \succeq \mu I,$$
$$G(W) = \nabla_a \nabla_\theta Q_\theta(W, a)|_{(0,a_0(W))}, \qquad J(W) = H(W)^{-1} G(W).$$

Here J is the raw optimal-action derivative. The encoded derivative is its curvature-weighted projection:

$$L_\phi(C) = \mathbb{E}[H \mid C]^{-1}\mathbb{E}[G \mid C]. \tag{2}$$

The weighting matters when histories merged by the code have different action curvatures. An ordinary conditional average $\mathbb{E}[J \mid C]$ then need not maximize encoded value. Define

$$\begin{aligned} \Omega &= \tfrac{1}{2}\mathbb{E}[J^\top H J], & \Lambda_\phi &= \tfrac{1}{2}\mathbb{E}[L_\phi^\top H L_\phi], \\ \Gamma_\phi &= \tfrac{1}{2}\mathbb{E}[(J - L_\phi)^\top H (J - L_\phi)]. \end{aligned} \tag{3}$$

Theorem 2 (Debt–motion conservation)**.** *Under Assumption 6—smooth action values, globally stable interior optimizers, strong concavity, and integrable derivative envelopes—uniformly over local directions δ,*

$$\begin{aligned} \alpha_\phi(\delta) &= \delta^\top \Gamma_\phi \delta + O(\|\delta\|^3), \\ V^\star_{\delta,\phi} - V^{\pi_0}_\delta &= \delta^\top \Lambda_\phi \delta + O(\|\delta\|^3), \end{aligned} \qquad \Omega = \Gamma_\phi + \Lambda_\phi.$$

For nested codes $\mathcal{C}_1 \subseteq \mathcal{C}_2$,

$$\Gamma_1 - \Gamma_2 = \Lambda_2 - \Lambda_1 = \tfrac{1}{2}\mathbb{E}[(L_2 - L_1)^\top H (L_2 - L_1)] \succeq 0. \tag{4}$$

The value laws are exact in the jointly quadratic subclass of Proposition 8.

The proof expands optimized values and uses $\mathbb{E}[G - HL_\phi \mid C] = 0$. This makes the debt a residual in the value metric: histories may be merged at no quadratic adaptation cost precisely when their optimal-action derivatives agree in that metric. The full proof works with integrable value remainders, without requiring an integrated second-order optimizer expansion. Appendix A supplies the regularity details and a finite-horizon extension under explicit coefficient-stability conditions.

4 The Cost-Aware Adaptation Frontier

The debt–motion geometry is d-dimensional. The closed-form frontier below solves a narrower identification problem: a direction u is declared or revealed before acquisition, and only its sign remains unknown. If direction is also latent, the scalar frontier is replaced by a multivariate experimental design problem.

Consider a declared alternative pair $\theta = \eta t u$, where $t > 0$ and the unit direction u are fixed and the unknown sign $\eta \in \{-1, +1\}$ has equal prior probability. Set $g_u = u^\top \Gamma_\phi u$ and $\ell_u = u^\top \Lambda_\phi u$. The three available policies execute the negative tangent, the nominal action, or the positive tangent. In the symmetric quadratic problem, their encoded losses are 0, $t^2\ell_u$, and $4t^2\ell_u$ for correct adaptation, freezing, and opposite adaptation.

An allocation n from a finite library $\mathcal{N} \subset \mathbb{Z}_+^{N_\mathrm{p}}$ for N_p probe types, containing 0, is chosen before seeing shift evidence. Independent Gaussian location blocks provide information $I_u(n) = I_{u,0} + \sum_j n_j i_{u,j}$, or equivalently a sufficient statistic $Z_n \mid \eta \sim N(\eta t\sqrt{I_u(n)}, 1)$. Let $c(n)$ include all acquisition penalties and foregone control opportunities. The canonical experiment has a fixed post-acquisition horizon H and downstream histories independent of the transcript conditional on the task. Its policy-loss matrix is therefore common to all transcript realizations.

Theorem 3 (Exact selective Gaussian frontier)**.** *Put $s = t\sqrt{I_u(n)}$, $\tau = \log 3$, and let Φ be the standard normal cumulative distribution function. For $Z_n \mid \eta \sim N(\eta s, 1)$ the log-likelihood ratio is $\mathcal{L} = 2sZ_n$. The Bayes rule over $\{\pi_-, \pi_0, \pi_+\}$ chooses π_+ for $\mathcal{L} > \tau$, π_- for $\mathcal{L} < -\tau$, and π_0 otherwise. Its normalized encoded risk is*

$$\psi(s) = \Phi\left(\frac{\tau - 2s^2}{2s}\right) + 3\Phi\left(\frac{-\tau - 2s^2}{2s}\right), \qquad \psi(0) = 1. \tag{5}$$

The exact pre-evidence cost is

$$\boxed{\mathfrak{C}^{\mathrm{G}}_\phi(t, u, H) = Ht^2 g_u + \min_{n \in \mathcal{N}} \left\{ c(n) + Ht^2 \ell_u \psi\left(t\sqrt{I_u(n)}\right) \right\}.} \tag{6}$$

If $n = 0$ is feasible and $c(0) = 0$, this cost is no greater than the no-probe Bayes cost. Mandatory sign commitment has encoded risk $4Ht^2\ell_u\Phi(-s)$ and is worse than freezing when $s < \Phi^{-1}(3/4)$ and $\ell_u > 0$.

The thresholds follow by comparing posterior risks $4p$, 1, and $4(1-p)$, where $p = \mathbb{P}(\eta = +1 \mid Z_n)$. This is a reject-option Bayes decision (Chow, 1970) with losses induced by the control geometry. The thresholds apply to the declared policy library. The nontrivial freeze interval is therefore a property of this discrete set of prevalidated updates, not an unavoidable feature of unconstrained continuous-action Bayesian control: a policy class that permits continuously interpolated actions has a different Bayes rule. The three-policy library models deployment settings with an explicit nominal fallback.

The representation–evidence coupling itself does not depend on this discrete library. The following exact quadratic result isolates the general object that the Gaussian frontier specializes.

Proposition 4 (Evidence-resolved quadratic value)**.** *Suppose*

$$Q_\vartheta(W, a) = q_\vartheta(W) - \tfrac{1}{2}\|a - a_0(C) - J(W)\vartheta\|^2_{H(W)}, \qquad \|x\|^2_H = x^\top H x,$$

where $C = \phi(W)$, $H(W) \succeq \mu I$ almost surely, $W \perp (\vartheta, Y)$, $\mathbb{E}\vartheta = 0$, $\mathrm{Cov}(\vartheta) = \Sigma$, and Y is task-only evidence. Assume the displayed quadratic moments are finite and the relevant actions lie in the admissible action domain. Set $m(Y) = \mathbb{E}[\vartheta \mid Y]$ and $R_Y = \mathbb{E}[m(Y)m(Y)^\top]$, so $0 \preceq R_Y \preceq \Sigma$. Then the Bayes-optimal encoded action is $a_0(C) + L_\phi(C)m(Y)$ and its expected regret relative to the raw, task-revealed optimum is

$$\mathcal{B}_\phi(Y) = \mathrm{tr}(\Gamma_\phi \Sigma) + \mathrm{tr}[\Lambda_\phi(\Sigma - R_Y)] = \mathrm{tr}(\Omega\Sigma) - \mathrm{tr}(\Lambda_\phi R_Y). \tag{7}$$

For any C-measurable deployed operator $K(C)$ with finite quadratic moment, define

$$M_K = \tfrac{1}{2}\mathbb{E}[(K - L_\phi)^\top H (K - L_\phi)].$$

The action $a_0(C) + K(C)m(Y)$ has expected regret

$$\mathcal{B}_{\phi,K}(Y) = \mathrm{tr}(\Omega\Sigma) - \mathrm{tr}[(\Lambda_\phi - M_K)R_Y]. \tag{8}$$

Thus R_Y is the task variation resolved by evidence, and $\mathrm{tr}(\Lambda_\phi R_Y)$ is recoverable prospective control value. For two representations of the same raw problem under a common task-only experiment, their recovered-value difference is $\mathrm{tr}[(\Lambda_1 - \Lambda_2)R_Y]$. Hence $\Lambda_1 - \Lambda_2 \succeq 0$ rules out an evidence-induced reversal under every such common experiment. Conversely, if $\Sigma \succ 0$, $\vartheta \sim N(0, \Sigma)$, and equally costed experiments can reveal arbitrary one-dimensional projections of the whitened task, an indefinite $\Sigma^{1/2}(\Lambda_1 - \Lambda_2)\Sigma^{1/2}$ yields opposite recovered-value preferences along positive and negative eigendirections. Appendix C proves these statements. Reversal therefore requires crossed recoverable response geometry together with evidence that resolves different directions; it is a sharp possibility, not a prevalence premise.

An allocation improves on no probing exactly when

$$c(n) - c(0) < Ht^2 \ell_u \left[\psi\Big(t\sqrt{I_u(0)}\Big) - \psi\Big(t\sqrt{I_u(n)}\Big)\right]. \tag{9}$$

This comparison credits natural evidence already available without a probe. For an allocation-dependent remaining horizon, the full objective must also use $H(n)$; a time penalty alone cannot replace the change in remaining decision value.

4.1 Evidence-priced representation design

Conservation rewrites the frontier as total prospective sensitivity minus the motion value that can be recovered after paying for evidence:

$$\mathfrak{C}^{\mathrm{G}}_\phi = Ht^2 u^\top \Omega u - \max_{n \in \mathcal{N}} \{Ht^2 \ell_u [1 - \psi(t\sqrt{I_u(n)})] - c(n)\}.$$

The first term is common to representations of the same nominal problem. Representation choice therefore depends on the recoverable motion value, after acquisition costs, rather than on revealed-shift debt alone.

Corollary 5 (Evidence-priced refinement)**.** *Under Theorems 2 and 3, fix common experiments, costs, allocation library, nominal problem, and an allocation-invariant post-acquisition horizon H for nested codes. Let $d_u = u^\top(\Gamma_1 - \Gamma_2)u \geq 0$ and $r_n = 1 - \psi(t\sqrt{I_u(n)})$. For any optimizing allocations n_1, n_2,*

$$Ht^2 d_u r_{n_1} \;\leq\; \mathfrak{C}^{\mathrm{G}}_1 - \mathfrak{C}^{\mathrm{G}}_2 \;\leq\; Ht^2 d_u r_{n_2}. \tag{10}$$

For a fixed allocation n, the gain is exactly $Ht^2 d_u r_n$. If $d_u > 0$, then $r_{n_2} \geq r_{n_1}$.

If both codes instead share an allocation-dependent remaining horizon $H(n)$, the same comparison gives

$$t^2 d_u H(n_1) r_{n_1} \le \mathfrak{C}_1 - \mathfrak{C}_2 \le t^2 d_u H(n_2) r_{n_2}.$$

Thus the fixed-H corollary is not used unchanged when probing consumes control time: the ordered quantity is effective recovered exposure $H(n)r_n$, not evidence resolution r_n alone. Appendix C gives the derivation.

For a fixed allocation, the fraction $1 - \psi(t\sqrt{I_u(n)})$ is the deployment discount on reducing debt. It is zero for uninformative evidence and approaches one as the sign is resolved. With optional acquisition, newly executable motion can justify a previously uneconomic probe. This comparison holds for nested codes under common experiments. At fixed capacity, non-nested codes can instead retain different directions: an ordering by average debt is not a positive-semidefinite ordering, and their deployment ranking can reverse when evidence channels are exchanged (Section 6.1). Appendix C gives the proof and the concavity of the frontier in executable motion.

When an evidence regime E is known before choosing between two representations with values $v_1(E), v_2(E)$, let $D(E) = v_1(E) - v_2(E)$. The ideal value of regime-dependent switching over committing ex ante to one representation is

$$G_{\text{switch}} = \frac{1}{2}\left\{\mathbb{E}|D(E)| - |\mathbb{E}D(E)|\right\}. \tag{11}$$

For two equally likely regimes this is $\{|D_A| + |D_B| - |D_A + D_B|\}/4$: it is zero when the rankings keep one sign, and when they cross it equals half the smaller absolute margin. Evidence can therefore change relative representation value substantially without creating a material switching payoff unless the change crosses zero with margin.

Calibration to a deployed update. For an encoded correction $K(C)\delta$, the correct-direction loss is governed by $M_K = \frac{1}{2}\mathbb{E}[(K - L_\phi)^\top H (K - L_\phi)]$. Its gain over freezing is $\delta^\top(\Lambda_\phi - M_K)\delta$ to second order. AP-GFAR therefore evaluates the loss matrix of the update actually executed (Proposition 12). The canonical loss matrix is recovered at $K = L_\phi$; otherwise the frontier must use the deployed operator's own branch losses.

5 AP-GFAR

AP-GFAR operationalizes the prospective response for a finite library of frozen representations and evidence allocations. It learns a response that the frozen code can execute and evaluates whether available evidence makes that response preferable to freezing.

Write $J = \partial_\theta a^\star_\theta|_0$ for the optimal-action derivative and L_ϕ for its curvature-weighted projection. Offline paired interventions can produce a response target $\widehat{T}(W)$ from reward-trained policies or an expert reference. The learned head $K_\phi(C)$ supplies the executable correction; its target need not recover J, and K_ϕ need not equal L_ϕ. A nominal action head and this response head share a variational bottleneck (Alemi et al., 2017). The learned-POMDP implementation minimizes

$$\widehat{\mathcal{L}}_{\text{nominal}} + \lambda_{\tan}\widehat{\mathbb{E}}\|\widehat{T}(W) - K_\phi(\phi(W))\|_F^2 + \beta\widehat{\mathcal{L}}_{\text{compression}}. \tag{12}$$

The Euclidean response term is an estimation surrogate, not the curvature-weighted projection itself. With heterogeneous action curvature its population target generally differs from L_ϕ. If a positive-semidefinite curvature estimate $\widehat{H}(W)$ is available, the replacement

$$\widehat{\mathbb{E}}\ \mathrm{tr}\Big[(\widehat{T} - K_\phi)^\top \widehat{H}(\widehat{T} - K_\phi)\Big]$$

targets L_ϕ when $\widehat{T} = J$ and $\widehat{H} = H$, by the same normal equation as Theorem 2. The learned-POMDP implementation does not assume such curvature recovery: it fits the Euclidean surrogate, treats the resulting K_ϕ as the deployed operator, and calibrates that operator's realized branch losses as required by Proposition 12. Evidence features use complete transition blocks, including rewards. For allocation n, let X_n contain the available transcript features and pre-probe descriptors, and let $\ell_{n,a}$ be the total loss of branch $a \in \mathcal{A}_n$, including acquisition and control. The sign experiments use

$\mathcal{A}_0 = \{0\}$ and $\mathcal{A}_n = \{0, +, -\}$ for $n > 0$, where branch 0 freezes at the nominal policy. AP-GFAR estimates conditional gain

$$g_{n,a}(x) = \mathbb{E}[\ell_{n,0} - \ell_{n,a} \mid X_n = x], \qquad \widehat{d}_n(x) \in \arg\max_{a \in \mathcal{A}_n} \widehat{g}_{n,a}(x). \tag{13}$$

An acquisition head estimates the total cost $\widehat{C}_n$ of this rule from pre-probe quantities only. Known deterministic charges are accounted for analytically rather than relearned: the implementation predicts the charge-neutral complete-program value and then restores the known charge when forming $\widehat{C}_n$. Development-seed cross-fitting supplies out-of-fold terminal outcomes for acquisition training (Chernozhukov et al., 2018); representations, evidence models, and deployment heads are fixed before evaluation. Appendix G gives the finite-library excess-cost bound. Appendix G.4 separately treats finite-sample confidence guarantees for bounded calibration returns.

For a finite library of encoders, the same calibrated cost comparison selects the representation as well as the allocation. All candidates are evaluated against a common nominal reference. If task direction is revealed only after code selection, the pre-deployment comparison averages over its declared prior, while the allocation may depend on the subsequently revealed direction. Corollary 14 controls excess cost within this library under the stated estimation errors.

Algorithm 1 (AP-GFAR deployment)**.** Given frozen encoder(s), fitted corrections K_ϕ, candidate branches, the declared task alternatives, and remaining horizon:

1. Value every feasible representation–allocation program, including no probing. Use either an analytic evidence model with operator-specific branch losses or a calibrated complete-program estimate; add known deterministic charges exactly. If direction is revealed only after code selection, compare codes by prior-averaged contingent allocation costs.
2. Select before observing test returns or the evidence to be purchased, execute the chosen allocation, retain the complete transcript, and form X_n.
3. Execute the branch with the largest value over freezing. With a specified evidence law and loss matrix, compute branch values from posterior risk; the canonical three-policy specialization has thresholds $\pm \log 3$.

The value-calibrated and Gaussian-rule implementations share the same encoder, response policies, evidence model, and allocation library; their comparison isolates deployment-value estimation. The sign is withheld from deployment inputs, while magnitude and direction are supplied to all compared methods.

6 Experiments

The experiments move outward from the analytic model while keeping one question fixed: does deployment value depend jointly on the response retained by a code and on the evidence available to identify that response? Controlled rewards test the value law, learned equal-capacity encoders test representation choice, native HalfCheetah systems test both pair-value prediction and evidence-interface design, and learned-POMDP plus established-method tests probe complete-program valuation. Protocol details and full comparator tables are in Appendices D, E, F, and G.

6.1 Mechanism: evidence prices prospective motion

The refinement law. A jointly quadratic reward family has three action and three shift coordinates, history-dependent non-diagonal curvature, and four nested codes. Across 20 independently rotated families, eight directions, and four magnitudes, direct reward evaluation matches the debt and motion laws to numerical precision. For each family, 50,000 sessions at each of six evidence strengths test refinement. Mean risk differs from the analytic frontier by at most 0.00128, and normalized refinement gain follows $1 - \psi(s)$ (Figure 2a). This isolates the evidence-priced value law from response-estimation error.

A ranking reversal at fixed capacity. Two non-nested codes have equal capacity and identical nominal control but retain different prospective dependencies. The first has lower average debt

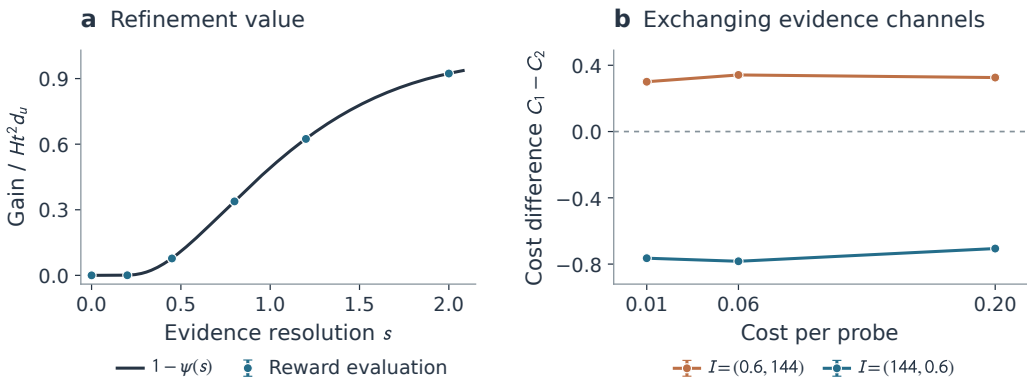

Figure 2: **The theory makes a falsifiable representation-design prediction.** (a) Normalized gain from nested refinement follows $1 - \psi(s)$; bars are paired 95% intervals over 20 independently rotated reward families. (b) At equal capacity and nominal control, exchanging the evidence channels reverses the preferred non-nested code. Each point averages all declared horizons and magnitudes at that price over 20 independent fits.

when the shift is fully revealed. Noisy antithetic reward queries estimate their responses and a separate sample calibrates code-specific deployment frontiers. When the second dependency has the stronger evidence channel, the first code costs 0.323 $[0.315, 0.331]$ more. Exchanging only the evidence channels changes the contrast to -0.751 $[-0.752, -0.750]$ across the complete declared grid (20 independent fits; Figure 2b). Capacity, nominal policy, and response models are unchanged.

6.2 Learned representations: choose the code before test returns

Equal-capacity neural encoders replace the prescribed codes. Each adds one learned bit to the same nominal code, while matched objectives emphasize different prospective directions. Independent reward measurements calibrate each frozen response. The code is selected before task sign or evaluation return, and evidence acquisition is chosen only afterward.

Across 20 fresh outer-training seeds, exchanging the evidence channels again reverses the learned-code ranking. Evidence-priced selection lowers cost relative to the held-out tangent-proxy rule by 0.1554 $[0.1534, 0.1574]$ while preserving nominal actions. A condition-indexed development lookup nearly saturates seen-grid cost—it is lower by only 2.8×10^{-5}—yet the channel exchange still reverses the learned representation ranking before test returns. Table 1 gives the complete comparison.

6.3 Native control: predict the ranking, then design the evidence

We next test pair-valued deployment in HalfCheetah-v5 (Todorov et al., 2012; Towers et al., 2025) under physical mass/gravity shifts, using native returns rather than the local differential expansion. A fixed-interface protocol with 16 independently trained systems does not show a population-wide crossing, but sealed pre-evaluation values predict its ranking heterogeneity. These forecasts are within-system plug-in complete-program values—not regressions fit to the 16 outcomes—constructed from disjoint pre-evaluation context and branch-return banks. The task-level fixed-effects slope is 0.994 $[0.774, 1.215]$ and the system-level evidence-exchange slope is 1.162 $[0.680, 1.643]$ (Figure 3; construction and full inference in Appendix E).

A prospectively locked follow-up uses two disjoint banks in each of 10 declared systems to choose feasible native measurement interfaces predicted to maximize or minimize the same program-value contrast, then evaluates them on the held-out bank before swapping roles. All ten cross: $D_A = 43.32$ $[19.37, 67.28]$ and $D_B = -24.50$ $[-44.75, -4.24]$ (IUT $p = 0.0115$). Thus reversal is a designable consequence of pair value, not a prevalence assumption about arbitrary native interfaces.

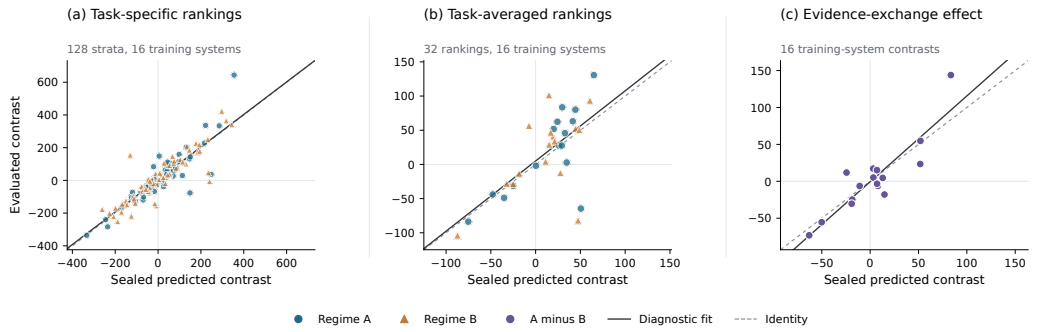

Figure 3: **Native pair value predicts heterogeneous rankings.** Sealed pre-evaluation contrasts track task-specific, task-averaged, and evidence-exchange effects across 16 independently trained systems. Solid lines are diagnostic regressions; dashed lines are identity.

6.4 Deployment: value the program actually executed

A prospectively locked cohort of 32 fresh learned-POMDP systems tests complete programs with interactive prefixes and allocation-dependent horizons. Calibration gains 0.4163 $[0.3865, 0.4461]$ over frozen/no-probe; its 0.0111 $[-0.0004, 0.0226]$ contrast with the operator-aware Gaussian rule is unresolved, and a matched generic acquisition head makes the same 256 interpolation decisions (Appendix G).

An external test freezes an established VariBAD PPO/GRU/VAE checkpoint (Zintgraf et al., 2020) and uses AP-GFAR only to select among its declared native online and frozen-context programs on HalfCheetah gravity shifts. Over 32 paired fresh reset blocks, the selector improves net return over native online no-prefix deployment by 73.74 $[20.42, 127.06]$ (one-sided $p = 0.0041$). This fixed-checkpoint test isolates deployment-program valuation rather than training superiority (Appendix H).

7 Related Work and Discussion

Statistical sufficiency and comparisons of experiments formalize the decision value of observations (Blackwell, 1953; Le Cam, 1964). RL abstractions, value equivalence, reward-predictive objectives, and action- or control-sufficient representations characterize information needed to predict, value, or act (Abel et al., 2016; Grimm et al., 2020; Lehnert et al., 2020; Huang et al., 2022; Rakelly et al., 2021; Gu et al., 2025). Contextual representation, meta-RL, and deployment-time adaptation already address task or dynamics changes (Rakelly et al., 2019; Zintgraf et al., 2020; Hansen et al., 2021; Zhang et al., 2024). We instead assume nominal control is representable and the code is frozen, then ask whether the *response of the optimum* to a nearby change survives compression.

Policy and value sensitivity to model perturbations has classical antecedents (Cao, 2003; Epshteyn et al., 2008; Kern et al., 2020), while prospective learning studies predictors for evolving data distributions (De Silva et al., 2024). Our contribution is the representation-level geometry of policy sensitivity and its coupling to deployment evidence. Work on latent-dynamics prediction, Bellman structure, and online representation–control coupling is complementary (Bagatella et al., 2026; Nabati et al., 2026; Martinez et al., 2026); sequential detection and experimental design price information for decisions (Page, 1954; Kiefer, 1961; Graves and Lai, 1997; Russo and Van Roy, 2018; Foster et al., 2021; Ziemann and Sandberg, 2021).

Interaction, rank crossing, and switching utility are distinct: evidence may change relative value without crossing zero, and Equation (11) prices the gain from crossing. The fixed-interface native study shows that a ranking reversal is not a prevalence premise; the locked follow-up instead uses pair values to design evidence interfaces whose held-out rankings cross. AP-GFAR operationalizes this pair-valued geometry through complete-program valuation.

Reproducibility Statement

Appendices specify the locked protocols, cohorts, statistical units, training–calibration–evaluation separation, and analysis contracts used in the experiments; complete proofs follow the references. The fresh learned-POMDP confirmation fixes 32 outer-system seeds and its analysis before any fresh outcome is generated; each system's allocation plan is written before its evaluation rollouts. In HalfCheetah, each 16-system fixed-interface forecast is a within-system complete-program plug-in value: branch frequencies come from 512 pre-evaluation context blocks and branch returns from a disjoint 128-block return bank; same-size independent banks are reserved for evaluation, and no cross-system outcome model is fitted. A distinct 10-system cross-fitted interface-design rule is locked before its validation-bank outcomes. The external VariBAD test uses a fixed checkpoint, fixed calibration selector, and 32 fresh paired reset blocks; it is reported conditionally on that trained system.

AI Use Statement

In this work, large language models (LLMs) are used exclusively for polishing the writing and checking grammar. They are not involved in research ideation, experimental design, data analysis, or the formulation of conclusions. The authors make all substantive intellectual contributions.

References

David Abel, David Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *ICML*, PMLR 48:2915–2923, 2016.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. `https://arxiv.org/abs/1612.00410`.

Marco Bagatella, Matteo Pirotta, Ahmed Touati, Alessandro Lazaric, and Andrea Tirinzoni. TD-JEPA: Latent-predictive representations for zero-shot reinforcement learning. In *International Conference on Learning Representations*, 2026.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Xi-Ren Cao. From perturbation analysis to Markov decision processes and reinforcement learning. *Discrete Event Dynamic Systems*, 13(1–2):9–39, 2003. doi:10.1023/A:1022188803039.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi:10.1109/TIT.1970.1054406.

Ashwin De Silva, Rahul Ramesh, Rubing Yang, Siyu Yu, Joshua T. Vogelstein, and Pratik Chaudhari. Prospective learning: Learning for a dynamic future. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Arkady Epshteyn, Adam Vogel, and Gerald DeJong. Active reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 296–303. ACM, 2008. doi:10.1145/1390156.1390194.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv:2112.13487*, 2021.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi:10.1214/aos/1013203451.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *NeurIPS*, 2020.

Yuliang Gu, Hongpeng Cao, Marco Caccamo, and Naira Hovakimyan. Observations meet actions: Learning control-sufficient representations for robust policy generalization. *arXiv:2507.19437*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, PMLR 80:1861–1870, 2018.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations*, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Biwei Huang, Chaochao Lu, Liu Leqi, Jose Miguel Hernandez-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *ICML*, PMLR 162:9260–9279, 2022.

Patrick Kern, Axel Simroth, and Henryk Zähle. First-order sensitivity of the optimal value in a Markov decision model with respect to deviations in the transition probability function. *Mathematical Methods of Operations Research*, 92:165–197, 2020.

Jack Kiefer. Optimum experimental designs V, with applications to systematic and rotatable designs. In Jerzy Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 381–405. University of California Press, 1961.

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

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964.

Lucas Lehnert, Michael L. Littman, and Michael J. Frank. Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Computational Biology*, 16(10):e1008317, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1711.05101.

Fernando Martinez, Tao Li, Yingdong Lu, and Juntao Chen. Stackelberg coupling of online representation learning and reinforcement learning. In *International Conference on Learning Representations*, 2026.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory*, 2009. https://arxiv.org/abs/0907.3740.

Ofir Nabati, Bo Dai, Shie Mannor, and Guy Tennenholtz. Spectral Bellman method: Unifying representation and exploration in RL. In *International Conference on Learning Representations*, 2026.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi:10.1214/07-AOAS122.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *ICML*, PMLR 97:5331–5340, 2019.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which mutual-information representation learning objectives are sufficient for control? In *NeurIPS*, 2021.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society, Series B*, 86(1):1–27, 2024. doi:10.1093/jrsssb/qkad009.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite. *arXiv:1801.00690*, 2018.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi:10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J. K. Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025. doi:10.52202/085713-4916.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi:10.1007/BF00992696.

Xinyu Zhang, Wenjie Qiu, Yi-Chen Li, Lei Yuan, Chengxing Jia, Zongzhang Zhang, and Yang Yu. Debiased offline representation learning for fast online adaptation in non-stationary dynamics. In *ICML*, PMLR 235:59741–59758, 2024.

Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive LQR with unknown B-matrix. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, PMLR 144:213–226, 2021.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep reinforcement learning via meta-learning. In *International Conference on Learning Representations*, 2020.

APPENDIX GUIDE

The submission appendix follows the dependency chain of the main claim: proof of the prospective-response geometry, proof of the selective frontier, the representation-design experiments, and the learned deployment protocols. Results that broaden the framework without being needed to verify the main argument are kept in the extended technical record.

Core support.

Topic	Page
Smooth value geometry and coefficient-stable dynamic occupancy	13
Selective frontier and operator-specific losses	18
Evidence-priced refinement and direct reward tests	20
Learned-code selection and frozen protocol	23
Native HalfCheetah prediction and locked interface design	26
Learned POMDP model and matched deployment protocol	30
Fresh locked program-value confirmation and stability	31
Finite-sample acquisition in the official context system	35
Established-method VariBAD deployment test	39
Exact small-shift counterexamples and sufficiency hierarchy	39

A SMOOTH CONTINUOUS-ACTION DEBT–MOTION GEOMETRY

This appendix proves Theorem 2, records the exact quadratic case, and states a finite-horizon extension for policy-dependent occupancies. All conditional expectations below are taken under the nominal law. Matrices are identified up to almost-sure equality under that law.

A.1 ONE-DECISION REGULARITY AND OPTIMIZED-VALUE EXPANSIONS

Let W be the raw decision history, $C = \phi(W)$, and let the action set $\mathcal{A} \subset \mathbb{R}^m$ be compact and convex with nonempty interior. For $\delta \in \mathbb{R}^d$, write $Q_\delta(W, a)$ for the action value when the post-decision continuation is optimized in the relevant policy class, and define the conditional objective

$$\bar{Q}_\delta(C, a) := \mathbb{E}[Q_\delta(W, a) \mid C].$$

The global-value theorem requires global optimizer stability; local regularity around a nominal stationary point is not enough. Its proof is stated directly at the optimized-value level and does not require a second-order L_2 expansion of the conditional optimizer.

Assumption 6 (Globally stable smooth control)**.** *There are a neighborhood $\mathcal{N}$ of zero, constants $0 < \mu \le L < \infty$, and random envelopes $K_2(W), K_3(W) \ge 1$ such that:*

(i) for almost every W, $(\delta, a) \mapsto Q_\delta(W, a)$ is three times continuously differentiable on an open set containing $\mathcal{N} \times \mathcal{A}$;

(ii) for every $\delta \in \mathcal{N}$, $a \mapsto Q_\delta(W, a)$ is μ-strongly concave on $\mathcal{A}$ and its unique global maximizer $a^\star_\delta(W)$ lies in $\mathrm{int}(\mathcal{A})$; the conditional objective $\bar{Q}_\delta(C, \cdot)$ has a unique global maximizer $a^\star_{\delta,\phi}(C)$ in $\mathrm{int}(\mathcal{A})$, and the nominal maximizer $a_0(W) = a_0(C)$ is C-measurable;

(iii) $H(W) := -\nabla^2_{aa} Q_0(W, a_0(W))$ satisfies $\mu I_m \preceq H(W) \preceq L I_m$;

(iv) throughout $\mathcal{N} \times \mathcal{A}$, every second derivative carrying at least one δ index is bounded in norm by $K_2(W)$ and every third derivative is bounded by $K_3(W)$, with

$$\mathbb{E}\big[|Q_0(W, a_0)| + \|\nabla_\delta Q_0(W, a_0)\| + K_2(W)^6 + K_3(W)^2\big] < \infty;$$

(v) for every derivative of total order at most three used below, a conditional version satisfies $\partial^\alpha \bar{Q}_\delta(C, a) = \mathbb{E}[\partial^\alpha Q_\delta(W, a) \mid C]$.

Global strong concavity is a transparent sufficient condition rather than a necessity. It may be replaced by an explicit localization assumption stating that, for all sufficiently small shifts, the raw and conditional *global* maximizers are unique and remain in common neighborhoods on which the same smoothness, curvature, and integrable optimized-value envelopes hold. Merely following an implicit-function local branch does not justify a theorem about the global value.

Set

$$b(W) = \nabla_\delta Q_0(W, a_0), \quad D(W) = \nabla^2_{\delta\delta} Q_0(W, a_0),$$
$$\bar{b}(C) = \mathbb{E}[b(W) \mid C], \quad \bar{D}(C) = \mathbb{E}[D(W) \mid C],$$
$$\bar{K}_j(C) = \mathbb{E}[K_j(W) \mid C], \quad j \in \{2, 3\}.$$

Lemma 7 (Global-action derivatives and optimized-value Taylor laws)**.** *Under Assumption 6, the raw and conditional global optimizers are differentiable at zero, with*

$$\nabla_\delta a^\star_\delta\big|_0 = J = H^{-1}G, \qquad \nabla_\delta a^\star_{\delta,\phi}\big|_0 = L_\phi = \bar{H}^{-1}\bar{G}, \tag{14}$$

where $G = \nabla_a \nabla_\delta Q_\delta|_{(0,a_0)}$, $\bar{H} = \mathbb{E}[H \mid C]$, and $\bar{G} = \mathbb{E}[G \mid C]$. Let $v_W(\delta) = \max_a Q_\delta(W, a)$ and $\bar{v}_C(\delta) = \max_a \bar{Q}_\delta(C, a)$. There is a constant $c_\mu < \infty$, depending only on μ and the finite dimensions, such that, uniformly over directions,

$$v_W(\delta) = v_W(0) + b(W)^\top \delta + \frac{1}{2}\delta^\top \big(D(W) + G(W)^\top H(W)^{-1} G(W)\big)\,\delta + \rho_W(\delta), \tag{15}$$

$$\bar{v}_C(\delta) = \bar{v}_C(0) + \bar{b}(C)^\top \delta + \frac{1}{2}\delta^\top \big(\bar{D}(C) + \bar{G}(C)^\top \bar{H}(C)^{-1} \bar{G}(C)\big)\,\delta + \bar{\rho}_C(\delta), \tag{16}$$

with

$$|\rho_W(\delta)| \le c_\mu K_3(W)(1 + K_2(W))^3 \|\delta\|^3, \qquad |\bar{\rho}_C(\delta)| \le c_\mu \bar{K}_3(C)(1 + \bar{K}_2(C))^3 \|\delta\|^3. \tag{17}$$

Both envelopes are integrable. No L_2 second-order remainder for $a^\star_{\delta,\phi}$ is asserted or needed.

Proof. Fix a unit direction u, write $q(t, a) = Q_{tu}(W, a)$, and let $a(t)$ be its unique interior global maximizer. The implicit-function theorem and global uniqueness give a C^2 optimizer path. Differentiating $q_a(t, a(t)) = 0$ once and twice yields

$$\dot{a} = -q_{aa}^{-1} q_{at}, \qquad q_{aa}\ddot{a} + q_{att} + 2q_{aat}\dot{a} + q_{aaa}[\dot{a}, \dot{a}] = 0.$$

Hence $\|\dot{a}\| \le K_2/\mu$ and $\|\ddot{a}\| \le c_\mu K_3(1 + K_2)^2$. For the optimized value $v(t) = q(t, a(t))$, the envelope identity gives

$$\dot{v} = q_t, \qquad \ddot{v} = q_{tt} + q_{ta}\dot{a},$$

and therefore

$$v^{(3)} = q_{ttt} + 2q_{tta}\dot{a} + q_{taa}[\dot{a}, \dot{a}] + q_{ta}\ddot{a}.$$

Thus $|v^{(3)}(t)| \le c_\mu K_3(1 + K_2)^3$. At $t = 0$, $\dot{a}(0) = H^{-1}Gu$ and

$$\ddot{v}(0) = u^\top \big(D + G^\top H^{-1} G\big)\, u.$$

Taylor's theorem with integral remainder proves (15) and its first bound in (17).

Taking $t = \|\delta\|$ and $u = \delta/\|\delta\|$ shows that the bound is uniform over directions. Conditional expectation preserves strong concavity and gives derivative envelopes $\bar{K}_2, \bar{K}_3$, so the identical argument applied to $\bar{Q}_\delta$ proves (14) and (16). Finally,

$$\mathbb{E}[K_3(1 + K_2)^3] \le \{\mathbb{E}K_3^2\}^{1/2} \{\mathbb{E}(1 + K_2)^6\}^{1/2} < \infty,$$

and conditional Jensen followed by Cauchy–Schwarz gives the same conclusion for $\bar{K}_3(1 + \bar{K}_2)^3$. This justifies integration of both direction-uniform remainders. □

Proposition 8 (Exact jointly quadratic subclass)**.** *Suppose, for almost every W and every sufficiently small δ,*

$$Q_\delta(W, a) = q(W) + r(W)^\top \delta + \tfrac{1}{2}\delta^\top D(W)\delta \tag{18}$$
$$+ \{h(W) + G(W)\delta\}^\top a - \tfrac{1}{2} a^\top H(W) a, \tag{19}$$

where $H(W) \succeq \mu I$ is independent of δ, $a_0(W) = H(W)^{-1}h(W)$ is C-measurable, and the raw and conditional maximizers remain in the common action domain. Then

$$a^\star_\delta = a_0 + H^{-1}G\delta, \qquad a^\star_{\delta,\phi} = a_0 + \bar{H}^{-1}\bar{G}\delta$$

exactly, and both value identities in Theorem 2 hold with zero remainder. A sufficient extension is obtained when, on every segment joining a_0 to the raw or conditional optimizer, the action objective has a δ-independent constant Hessian and the optimizer is affine in δ. Affinity of the optimizer alone, or quadraticity in a with a δ-dependent Hessian, does not imply exact value laws; boundary contact also destroys the conclusion.

Proof. For a positive-definite H, $\max_a\{b^\top a - a^\top Ha/2\} = b^\top H^{-1}b/2$ and the maximizer is $H^{-1}b$. Apply this identity to (19). Since $h = Ha_0$ and a_0 is C-measurable, $\mathbb{E}[h \mid C] = \mathbb{E}[H \mid C]a_0$, which gives the two affine optimizer laws. After maximizing, the constant, linear, and direct δ-quadratic terms cancel between raw and conditional values by the tower property. The remaining quadratic difference is

$$\tfrac{1}{2}\delta^\top \mathbb{E}\big[G^\top H^{-1}G - \bar{G}^\top \bar{H}^{-1}\bar{G}\big]\,\delta = \delta^\top \Gamma_\phi \delta.$$

Evaluating the conditional quadratic objective at its optimizer and at a_0 gives $\delta^\top \Lambda_\phi \delta$ exactly. Weighted orthogonality then gives $\Omega = \Gamma_\phi + \Lambda_\phi$ with no remainder. □

A.2 Proof of Theorem 2

For the one-decision problem, $V^\star_\delta = \mathbb{E}[v_W(\delta)]$ and $V^\star_{\delta,\phi} = \mathbb{E}[\bar{v}_C(\delta)]$. Apply Lemma 7 and take expectations in (15)–(16); the constant, linear, and direct parameter-curvature terms cancel by the tower property. Hence

$$V^\star_\delta - V^\star_{\delta,\phi} = \frac{1}{2}\delta^\top \mathbb{E}\big[G^\top H^{-1}G - \bar{G}^\top \bar{H}^{-1}\bar{G}\big]\,\delta + O(\|\delta\|^3). \tag{20}$$

The nominal encoded policy uses the fixed action $a_0(C)$. Taylor's theorem gives

$$\mathbb{E}[Q_\delta(W, a_0)] = \mathbb{E}[v_W(0)] + \mathbb{E}[b]^\top \delta + \tfrac{1}{2}\delta^\top \mathbb{E}[D]\delta + O(\|\delta\|^3),$$

where the remainder is integrable by Assumption 6(iv). Thus this fixed-action value has the same constant, linear, and direct parameter-curvature terms but no optimizer-curvature term. Subtracting it from (16) gives

$$V^\star_{\delta,\phi} - V^{\pi_0}_\delta = \frac{1}{2}\delta^\top \mathbb{E}[\bar{G}^\top \bar{H}^{-1}\bar{G}]\delta + O(\|\delta\|^3). \tag{21}$$

The remainder bounds are direction-uniform by (17); no integrated second-order expansion of the optimizer is used.

It remains to identify the matrices. Since $HJ = G$ and $\bar{H}L_\phi = \bar{G}$,

$$\begin{aligned}\mathbb{E}[(J - L_\phi)^\top H L_\phi] &= \mathbb{E}[(G - HL_\phi)^\top L_\phi] \\ &= \mathbb{E}\big[\mathbb{E}[G - HL_\phi \mid C]^\top L_\phi\big] = 0.\end{aligned} \tag{22}$$

Thus

$$\mathbb{E}[G^\top H^{-1}G - \bar{G}^\top \bar{H}^{-1}\bar{G}] = \mathbb{E}[(J - L_\phi)^\top H (J - L_\phi)],$$

while $\mathbb{E}[\bar{G}^\top \bar{H}^{-1}\bar{G}] = \mathbb{E}[L_\phi^\top H L_\phi]$. Substitution into (20)–(21) proves the two value laws, and expanding $J = (J - L_\phi) + L_\phi$ proves conservation.

For refinement, the matrix claim is proved direction by direction. Fix $u \in \mathbb{R}^d$ and equip square-integrable m-vectors with $\langle x, y\rangle_H = \mathbb{E}[x^\top Hy]$. For a sigma-field $\mathcal{C}$, $L_\mathcal{C} u = \mathbb{E}[H \mid \mathcal{C}]^{-1}\mathbb{E}[HJu \mid \mathcal{C}]$ is the H-orthogonal projection of Ju onto the closed subspace of $\mathcal{C}$-measurable vectors. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, Pythagoras gives, for every u,

$$\begin{aligned}2u^\top(\Gamma_{\mathcal{C}_1} - \Gamma_{\mathcal{C}_2})u &= \mathbb{E}\big[((L_{\mathcal{C}_2} - L_{\mathcal{C}_1})u)^\top H((L_{\mathcal{C}_2} - L_{\mathcal{C}_1})u)\big], \\ 2u^\top(\Lambda_{\mathcal{C}_2} - \Lambda_{\mathcal{C}_1})u &= \mathbb{E}\big[((L_{\mathcal{C}_2} - L_{\mathcal{C}_1})u)^\top H((L_{\mathcal{C}_2} - L_{\mathcal{C}_1})u)\big].\end{aligned}$$

Because these quadratic forms agree for all u, the symmetric matrices are equal; the common matrix is positive semidefinite. This proves (4) without inferring a matrix identity from a trace norm. Finally, under Proposition 8—joint quadraticity in (a, δ), a δ-independent constant action Hessian, and interior affine optimizers—the optimized-value formulas and both debt–motion laws are exact.

A.3 FINITE-HORIZON EXTENSION WITH POLICY-DEPENDENT OCCUPANCY

First-order sensitivity of optimal values to transition perturbations has classical antecedents in Markov decision models (Kern et al., 2020); the result below instead decomposes frozen-representation debt and executable motion through Bellman action curvature.

Let $\mathcal{H}_h$ denote the raw history before action A_h, $C_h = \phi_h(\mathcal{H}_h)$, and $d_{0,h}$ the nominal occupancy induced by the common nominal optimal policy. Write $Q^\star_{\delta,h}(\mathcal{H}_h, a)$ for the raw optimal Bellman action value and set

$$H_h = -\nabla^2_{aa} Q^\star_{0,h}(\mathcal{H}_h, a_{0,h}), \qquad G_h = \nabla_a \nabla_\delta Q^\star_{\delta,h}(\mathcal{H}_h, a)|_{(0,a_{0,h})}.$$

The cross derivative G_h contains the immediate reward derivative, transition derivative, derivative of future value, and any smooth dual-control effect already valued by $Q^\star_{\delta,h}$.

Assumption 9 (Dynamic continuous-control regularity and coefficient-stable occupancy)**.** *The horizon is finite and the following hold uniformly over stages.*

(i) The Bellman action values $Q^\star_{\delta,h}$ satisfy Assumption 6, with common optimizer-stability, curvature, and optimized-value domination constants; the nominal optimal action is C_h-measurable.

(ii) Let $\mathcal{Q}^{\phi,\star}_{\delta,h}(C_h, a)$ be the true Bellman action objective of the encoded policy class. It is C^3 on a common neighborhood, uniformly strongly concave in a, and has a unique interior global maximizer. At the nominal optimum,

$$-\nabla^2_{aa} \mathcal{Q}^{\phi,\star}_{0,h}(C_h, a_{0,h}) = \mathbb{E}_{d_{0,h}}[H_h \mid C_h], \qquad \nabla_a \nabla_\delta \mathcal{Q}^{\phi,\star}_{\delta,h}(C_h, a)|_{(0,a_{0,h})} = \mathbb{E}_{d_{0,h}}[G_h \mid C_h]. \tag{23}$$

The identities in (23) are compatibility conditions, not automatic Bellman identities for an arbitrary lossy history encoder. At the common nominal optimum, an envelope-theorem argument removes first-order variation from locally optimal downstream policy changes; the remaining action curvature and parameter cross-derivative must still match the stated conditional raw quantities. The dynamic theorem is applied only when this compatibility holds.

(iii) Write $J_h = H_h^{-1} G_h$ and $L_h = \mathbb{E}[H_h \mid C_h]^{-1} \mathbb{E}[G_h \mid C_h]$. Let $a^r_{\delta,h}$ and $a^e_{\delta,h}$ be the raw and encoded global actions. Define the actual stagewise advantages

$$A^D_{\delta,h} = Q^\star_{\delta,h}(\mathcal{H}_h, a^r_{\delta,h}) - Q^\star_{\delta,h}(\mathcal{H}_h, a^e_{\delta,h}),$$
$$A^M_{\delta,h} = Q^\star_{\delta,h}(\mathcal{H}_h, a^e_{\delta,h}) - Q^\star_{\delta,h}(\mathcal{H}_h, a_{0,h}).$$

There is a measurable $W_h \geq 0$ with $\mathbb{E}_{d_{0,h}} W_h < \infty$ such that, uniformly over local directions,

$$A^D_{\delta,h} = \tfrac{1}{2} \delta^\top (J_h - L_h)^\top H_h (J_h - L_h)\delta + R^D_{\delta,h}, \tag{24}$$
$$A^M_{\delta,h} = \delta^\top G_h^\top L_h \delta - \tfrac{1}{2} \delta^\top L_h^\top H_h L_h \delta + R^M_{\delta,h}, \tag{25}$$

with $|R^D_{\delta,h}| + |R^M_{\delta,h}| \leq W_h \|\delta\|^3$. These expansions follow from the one-decision optimized-value theorem under a common deterministic third-derivative envelope and bounded local tangents; in the random-envelope case they are the explicit integrability obligation needed for the dynamic lift.

(iv) For a unit direction u, define the two quadratic coefficient functions

$$b^D_h(u, \mathcal{H}_h) = \tfrac{1}{2} u^\top (J_h - L_h)^\top H_h (J_h - L_h) u,$$
$$b^M_h(u, \mathcal{H}_h) = u^\top \left\{ \mathrm{sym}(G_h^\top L_h) - \tfrac{1}{2} L_h^\top H_h L_h \right\} u.$$

For every relevant shifted raw, encoded, or nominal policy π, there are finite constants K_{occ}, K_W and a common local radius such that, uniformly in h, u, and nonzero local δ,

$$\max_{X \in \{D,M\}} \left| \mathbb{E}_{d^\pi_{\delta,h}} b^X_h(u, \mathcal{H}_h) - \mathbb{E}_{d_{0,h}} b^X_h(u, \mathcal{H}_h) \right| \leq K_{\mathrm{occ}} \|\delta\|, \tag{26}$$

and

$$\sup_{\|\delta\| \leq \delta_0} \mathbb{E}_{d^\pi_{\delta,h}} W_h \leq K_W. \tag{27}$$

The compatibility identities, cubic advantage expansions, and coefficient stability are substantive rather than automatic. Condition (iv) is deliberately *not* a total-variation continuity assumption on the full raw-history law. For deterministic policies, raw histories that record past actions can be mutually singular even when the action changes by only $O(\|\delta\|)$; for example, the event $\{A_1 = 0\}$ separates $A_1 = 0$ from $A_1 = \delta$ for every $\delta \neq 0$. Such a singularity is irrelevant to the proof if the two quadratic coefficient expectations in (26) remain first-order stable. One sufficient route is a metric on histories for which the shifted occupancy is $O(\|\delta\|)$ in Wasserstein–1 and the coefficient functions b_h^D, b_h^M are uniformly Lipschitz, together with the shifted remainder moment (27). This route allows smooth deterministic control perturbations without requiring total-variation overlap of action-labelled histories.

The pointwise expansions in (iii) themselves admit a derivative-level sufficient condition that does not assume a value-loss expansion: uniform strong concavity, a common bound on the raw and encoded Bellman derivatives through order three, and a common interior neighborhood of the optimizers. Implicit differentiation then gives $a^r_{\delta,h} = a_{0,h} + J_h\delta + O(\|\delta\|^2)$ and $a^e_{\delta,h} = a_{0,h} + L_h\delta + O(\|\delta\|^2)$ with bounded tangents and remainder constants. Taylor expansion of the raw Bellman objective yields (24)–(25): the linear-in-action terms vanish at the nominal optimum, and every product of an $O(\|\delta\|^2)$ action remainder with a first-order displacement is cubic. A nominal tie whose value comes from leading-order information acquisition requires a separate active-design envelope, retained in the extended technical record.

Theorem 10 (Finite-horizon continuous-action tensors)**.** *Under Assumption 9, let*

$$J_h = H_h^{-1} G_h, \quad L_h(C_h) = \mathbb{E}_{d_{0,h}}[H_h \mid C_h]^{-1} \mathbb{E}_{d_{0,h}}[G_h \mid C_h].$$

Define

$$\Omega^{\mathrm{cc}}_{\mathrm{dyn}} = \frac{1}{2} \sum_{h=1}^{H} \mathbb{E}_{d_{0,h}}[J_h^\top H_h J_h],$$

$$\Gamma^{\mathrm{cc}}_{\phi,\mathrm{dyn}} = \frac{1}{2} \sum_{h=1}^{H} \mathbb{E}_{d_{0,h}}[(J_h - L_h)^\top H_h (J_h - L_h)],$$

$$\Lambda^{\mathrm{cc}}_{\phi,\mathrm{dyn}} = \frac{1}{2} \sum_{h=1}^{H} \mathbb{E}_{d_{0,h}}[L_h^\top H_h L_h].$$

Then

$$V^\star_\delta - V^\star_{\delta,\phi} = \delta^\top \Gamma^{\mathrm{cc}}_{\phi,\mathrm{dyn}} \delta + O(\|\delta\|^3),$$
$$V^\star_{\delta,\phi} - V^{\pi_0}_\delta = \delta^\top \Lambda^{\mathrm{cc}}_{\phi,\mathrm{dyn}} \delta + O(\|\delta\|^3),$$

with direction-uniform remainders, and $\Omega^{\mathrm{cc}}_{\mathrm{dyn}} = \Gamma^{\mathrm{cc}}_{\phi,\mathrm{dyn}} + \Lambda^{\mathrm{cc}}_{\phi,\mathrm{dyn}}$.

Proof. The raw-optimal performance-difference identity gives

$$V^\star_\delta - V^\star_{\delta,\phi} = \sum_h \mathbb{E}_{d^{\pi^e}_{\delta,h}} A^D_{\delta,h}, \qquad V^\star_\delta - V^{\pi_0}_\delta = \sum_h \mathbb{E}_{d^{\pi_0}_{\delta,h}} (A^D_{\delta,h} + A^M_{\delta,h}).$$

Subtracting these identities expresses encoded motion as the nominal-policy expectation of A^M plus an occupancy difference multiplying A^D. In particular, A^M need not be the encoded Bellman advantage. Assumption 9(iii) supplies the explicit pointwise expansions (24)–(25); no L_2 second-order optimizer remainder is used. For $\delta \neq 0$, write $t = \|\delta\|$ and $u = \delta/t$. Then, for $X \in \{D, M\}$,

$$A^X_{\delta,h} = t^2 b^X_h(u, \mathcal{H}_h) + R^X_{\delta,h}.$$

By (26) and (27), for every relevant occupancy,

$$\mathbb{E}_{d^\pi_{\delta,h}} A^X_{\delta,h} = t^2 \mathbb{E}_{d_{0,h}} b^X_h(u, \mathcal{H}_h) + O(t^3),$$

uniformly over local directions. Thus only the expectations of the two quadratic coefficient functions, not the entire raw-history law, need to be replaced by their nominal counterparts. Summing over the finite horizon gives the two value expansions. Curvature-weighted projection orthogonality holds separately at each stage under $d_{0,h}$, so the stagewise Pythagorean identities prove conservation. The comparison with π_0 uses the same coefficient-stability argument for the displacement $L_h\delta$. □

The local quadratic tensor applies within a smooth decision regime. With several nominally tied probes, the minimizing experiment can change with direction; the resulting active-design envelope is retained in the extended technical record.

B Adaptation Geometry and Identification

B.1 The exact cost-aware selective frontier

We formalize the canonical experiment used in Theorem 3. A sign $\eta \in \{-1, +1\}$ has equal prior probability. For a fixed direction u, shift magnitude t, and allocation $n = (n_1, \dots, n_J)$ selected before observing any shift evidence, the natural statistic and protected probe responses are conditionally independent Gaussian location observations with directional information

$$I_u(n) = I_{u,0} + \sum_{j=1}^{J} n_j i_{u,j}.$$

Equivalently, their joint sufficient statistic is

$$Z_n \mid \eta \sim \mathcal{N}\Big(\eta t\sqrt{I_u(n)}, 1\Big). \tag{28}$$

The regret-equivalent acquisition cost $c(n)$ is paid before the post-evidence decision and includes direct probe penalties, foregone control opportunities, switching costs, and additional interaction costs. In the canonical theorem, the post-acquisition horizon H and downstream geometry are common to all allocations. If probes consume a fixed overall control budget, one must instead minimize the complete objective $c(n) + H(n)t^2\{g_u + \ell_u\psi(t\sqrt{I_u(n)})\}$. Any regret incurred during acquisition belongs in $c(n)$ once; its effect on the remaining horizon belongs in $H(n)$.

The local encoded policy library is $\{\pi_-, \pi_0, \pi_+\}$. After removing the common representation debt $t^2 g_u$, its per-decision losses under sign η are

$$L_\eta(\pi_\eta) = 0, \qquad L_\eta(\pi_0) = t^2\ell_u, \qquad L_\eta(\pi_{-\eta}) = 4t^2\ell_u. \tag{29}$$

The factor four is not a convention: for a symmetric quadratic action value, the two shifted optima are twice as far apart as either is from the nominal action. Thus the nominal-to-shifted value gain is $t^2\ell_u$, whereas choosing the opposite shifted optimum loses $4t^2\ell_u$. The freeze interval derived below belongs to this three-policy deployment library. An unconstrained continuous-action Bayes rule can instead interpolate between signed corrections and generally has no identical deadband.

B.2 Proof of Theorem 3

If $\ell_u = 0$, every encoded loss is zero and the stated thresholds remain Bayes-optimal. In the following calculation assume $\ell_u > 0$. For a fixed allocation, let $s = t\sqrt{I_u(n)}$. The log-likelihood ratio of $\eta = +1$ against $\eta = -1$ is

$$\mathcal{L} = 2sZ_n.$$

Under $\eta = +1$, $\mathcal{L} \sim \mathcal{N}(2s^2, 4s^2)$; the law under $\eta = -1$ is its reflection. With equal priors, the posterior probability $p = \mathbb{P}(\eta = +1 \mid \mathcal{L})$ satisfies $p = (1 + e^{-\mathcal{L}})^{-1}$. Dividing the three posterior expected losses by $t^2\ell_u$ gives

$$4(1-p), \qquad 1, \qquad 4p$$

for π_+, π_0, π_-, respectively. Hence the Bayes rule chooses π_+ when $p \geq 3/4$, chooses π_- when $p \leq 1/4$, and freezes otherwise; equivalently its likelihood-ratio thresholds are $\pm\tau$ with $\tau = \log 3$.

By symmetry it suffices to evaluate the motion risk under $\eta = +1$. The normalized risk is

$$\begin{aligned}
&\mathbb{P}_+\{-\tau < \mathcal{L} < \tau\} + 4\mathbb{P}_+\{\mathcal{L} \leq -\tau\} \\
&\quad = \Phi\left(\frac{\tau - 2s^2}{2s}\right) - \Phi\left(\frac{-\tau - 2s^2}{2s}\right) + 4\Phi\left(\frac{-\tau - 2s^2}{2s}\right) \\
&\quad = \psi(s).
\end{aligned}$$

At $s = 0$ the transcript is uninformative and the Bayes action is π_0, so the continuous extension is $\psi(0) = 1$. Adding the common debt $Ht^2 g_u$ and the pre-evidence cost $c(n)$, then minimizing over the feasible allocations, proves (6).

If freezing is removed, the likelihood-ratio test commits to the sign of $\mathcal{L}$ and errs with probability $\Phi(-s)$. Its motion risk is therefore $4Ht^2\ell_u\Phi(-s)$, which is smaller than the frozen motion risk $Ht^2\ell_u$ exactly when $\Phi(-s) < 1/4$, or $s > \Phi^{-1}(3/4)$. Because the selective decision set contains both the frozen policy and the always-commit likelihood-ratio rule,

$$0 \le \psi(s) \le \min\{1, 4\Phi(-s)\}. \tag{30}$$

Finally, $n = 0$ is feasible and has zero acquisition cost, so the minimized total frontier cannot exceed the frozen raw regret in the canonical model.

Corollary 11 (Exact no-probe comparison)**.** *For any feasible candidate allocation n, probing is strictly preferred to the no-probe allocation if and only if*

$$c(n) - c(0) < Ht^2\ell_u\left\{\psi\left(t\sqrt{I_u(0)}\right) - \psi\left(t\sqrt{I_u(n)}\right)\right\}. \tag{31}$$

In the special case of no natural evidence, $c(0) = 0$, and one indivisible probe with information i and cost c_p, this reduces to

$$c_p < Ht^2\ell_u\{1 - \psi(t\sqrt{i})\}.$$

Proof. The debt term in (6) is common to both allocations. Subtract the exact no-probe objective $c(0) + Ht^2\ell_u\psi(t\sqrt{I_u(0)})$ from the objective at n. The special case uses $I_u(0) = 0$ and $\psi(0) = 1$. □

Proposition 12 (The selective frontier is adaptation-operator specific)**.** *Under Assumption 6, fix the feasible encoded deployment family $a_0 + K(C)\delta$ on a common local neighborhood. In addition to square-integrability of K, assume $\mathbb{E}[\bar{K}_3(C)(1 + \|K(C)\|)^3] < \infty$, so its conditional action-value Taylor remainder is integrably cubic. The same conclusion holds for an update with an $O(\|\delta\|^2)$ action remainder under an integrable cubic value-error envelope. Let*

$$M_K = \tfrac{1}{2}\mathbb{E}[(K-L_\phi)^\top H(K-L_\phi)], \qquad S_K = \tfrac{1}{2}\mathbb{E}[L_\phi^\top HK + K^\top HL_\phi], \qquad C_K = \tfrac{1}{2}\mathbb{E}[K^\top HK].$$

Then, uniformly over local δ,

$$V^\star_{\delta,\phi} - V^{\pi_{K,\delta}}_\delta = \delta^\top M_K\delta + O(\|\delta\|^3), \tag{32}$$

$$V^{\pi_{K,\delta}}_\delta - V^{\pi_0}_\delta = \delta^\top(\Lambda_\phi - M_K)\delta + O(\|\delta\|^3), \tag{33}$$

$$V^\star_{\delta,\phi} - V^{\pi_{K,-\delta}}_\delta = \delta^\top(\Lambda_\phi + S_K + C_K)\delta + O(\|\delta\|^3). \tag{34}$$

Consequently the canonical post-evidence loss coefficients $(0, 1, 4)\ell_u$ are valid for the first-order update family precisely when $K = L_\phi$ in the H-seminorm. For a general deployment operator, its three losses must be estimated or derived for that operator before solving a selective Bayes frontier; substituting Λ_ϕ from the encoded-optimal tangent into an unrelated update is not justified.

Proof. Condition on C and expand the conditional action value at a_0. Since $\bar{G} = \bar{H}L_\phi$,

$$\bar{Q}_\delta(C, a_0 + K\delta) - \bar{Q}_\delta(C, a_0) = \delta^\top\big(L_\phi^\top\bar{H}K - \tfrac{1}{2}K^\top\bar{H}K\big)\delta + O(\|\delta\|^3).$$

After expectation the symmetric quadratic coefficient is $S_K - C_K = \Lambda_\phi - M_K$, proving (33). Subtracting this value from the encoded optimum law gives (32). Replacing K by $-K$ gives (34). The correct-direction encoded regret vanishes to second order for every direction if and only if $M_K = 0$, equivalently $K = L_\phi$ in the H-seminorm. In that case $S_K = 2C_K = 2\Lambda_\phi$, and the opposite-direction coefficient is $4\Lambda_\phi$. □

B.3 Locked canonical frontier executor

The canonical experiment fixes a $3 \times 3 \times 3$ grid of shift magnitudes, remaining horizons, and per-probe costs, together with g_u, ℓ_u, probe information, a twelve-probe cap, and the Monte Carlo seed. For each of the 27 cells it enumerates every allocation, selects the exact minimizer of (6), draws the reward-inclusive Gaussian transcript, and applies the likelihood-ratio freeze/adapt rule

from Theorem 3. The selector chooses zero probes in 6 cells. Across 100,000 draws per cell, the largest absolute discrepancy between empirical total regret and the analytic frontier is 0.00208.

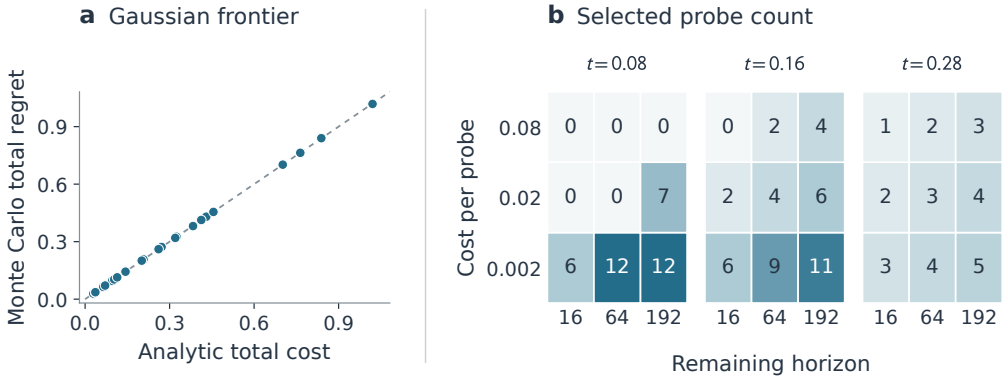

Figure 4: **Execution of the canonical acquisition rule.** Left: locked Monte Carlo risk against the exact Bayes frontier. Right: the selected allocation over the complete shift–horizon–cost grid. Each cell contains the selected probe count; zero is the no-probe action. These are canonical Gaussian experiments with a frozen selector.

C The Deployment Value of Representation Refinement

C.1 Common deployment experiments

Fix t, u, H, a finite allocation set $\mathcal{N}$, and common costs and Gaussian information $\{c(n), I_u(n)\}_{n\in\mathcal{N}}$ for two nested codes. The restriction to common experiments holds, for example, when a protected task-identification channel is unchanged by encoder refinement. Let $A = Ht^2$, $\omega_u = u^\top \Omega u$, and $r_n = 1 - \psi(t\sqrt{I_u(n)}) \in [0, 1]$.

C.2 Proof of Corollary 5

In the setting of Theorem 3, suppose $\mathcal{C}_1 \subseteq \mathcal{C}_2$ and the downstream nominal problem is unchanged. Put $d = u^\top(\Gamma_1 - \Gamma_2)u \geq 0$ and $\ell_i = u^\top \Lambda_i u$. Then $\ell_2 = \ell_1 + d$ and

$$\mathfrak{C}_i^{\mathrm{G}} = A\omega_u - \max_{n\in\mathcal{N}}\{A\ell_i r_n - c(n)\}. \tag{35}$$

For any minimizing allocations n_1, n_2,

$$Adr_{n_1} \leq \mathfrak{C}_1^{\mathrm{G}} - \mathfrak{C}_2^{\mathrm{G}} \leq Adr_{n_2}.$$

If both codes use the same fixed allocation n, their cost difference is exactly Adr_n. If $d > 0$, then $r_{n_2} \geq r_{n_1}$. As a function of executable motion $\ell \in [0, \omega_u]$, the frontier is nonincreasing and concave. On an interval with a unique optimizing allocation, its derivative is $-Ar_{n^\star}$.

Proof. Conservation gives $g_i = \omega_u - \ell_i$. Substitute into the frontier:

$$Ag_i + c(n) + A\ell_i\psi(t\sqrt{I_u(n)}) = A\omega_u + c(n) - A\ell_i r_n.$$

This proves (35). Evaluating the refined objective at n_1 gives $\mathfrak{C}_2^{\mathrm{G}} \leq \mathfrak{C}_1^{\mathrm{G}} - Adr_{n_1}$. Evaluating the coarse objective at n_2 gives $\mathfrak{C}_1^{\mathrm{G}} \leq \mathfrak{C}_2^{\mathrm{G}} + Adr_{n_2}$. Together these prove the bounds and, for $d > 0$, the ordering of resolutions. A pointwise minimum of affine functions of ℓ is concave. Each affine function has slope in $[-A, 0]$, proving monotonicity and the stated derivative. □

If the two codes share an allocation-dependent post-acquisition horizon $H(n)$, set $A_n = t^2H(n)$. Conservation gives the fixed-allocation difference $A_n dr_n$, and the same two plug-in comparisons yield

$$t^2 dH(n_1)r_{n_1} \leq \mathfrak{C}_1 - \mathfrak{C}_2 \leq t^2 dH(n_2)r_{n_2}.$$

Thus only $H(n_2)r_{n_2} \geq H(n_1)r_{n_1}$ follows when $d > 0$; the pure resolution ordering need not. This is the form relevant when probing consumes control steps. The fixed-H statement in Corollary 5 is its simpler specialization.

The ordering concerns achieved resolution, not the number of heterogeneous probes. A larger count need not be more informative when the probe types differ. If refinement changes the evidence channel, the common-experiment comparison no longer applies; both information functions enter their respective frontiers. In a smooth nonquadratic local model, the same substitution applies to the leading quadratic costs under the uniform remainder assumptions.

C.3 When a protected transcript leaves the debt unchanged

For the downstream geometry, let L_C denote the H-weighted projection of J onto functions of C. After revealing a transcript T, the projection onto the augmented code is unchanged if and only if

$$\mathbb{E}[H(J - L_C) \mid C, T] = 0 \quad \text{almost surely.} \tag{36}$$

Indeed, the conditional normal equation uniquely specifies the projection because $\mathbb{E}[H \mid C, T] \succeq \mu I$. A sufficient condition is that the transcript be independent of (H, J) conditional on C and the revealed task parameter. This models information about which task is present, without additional information about its downstream residual state. If the condition fails, Theorem 2 instead assigns the augmented code the smaller debt $\Gamma_{(C,T)}$. This is an acquisition-dependent representation problem, covered by recomputing the geometry rather than by a common-floor substitution. This condition characterizes preservation of the projection. The scalar Gaussian frontier also requires a transcript-independent downstream policy-loss matrix, as supplied by conditionally independent post-acquisition histories in the canonical experiment.

C.4 Evidence-resolved quadratic value

We prove Proposition 4. Conditional on (C, Y), the only action-dependent term in expected regret is

$$\tfrac{1}{2}\mathbb{E}\big[\|a - a_0(C) - J\vartheta\|_H^2 \mid C, Y\big].$$

Because $W \perp (\vartheta, Y)$ and $HJ = G$,

$$\mathbb{E}[HJ\vartheta \mid C, Y] = \mathbb{E}[HJ \mid C]m(Y) = \mathbb{E}[H \mid C]L_\phi(C)m(Y),$$

so strong positivity of $\mathbb{E}[H \mid C]$ gives the unique conditional minimizer $a^\star(C, Y) = a_0(C) + L_\phi(C)m(Y)$. Write $m = m(Y)$. Then

$$J\vartheta - L_\phi m = (J - L_\phi)\vartheta + L_\phi(\vartheta - m).$$

The weighted projection satisfies the conditional normal equation

$$\mathbb{E}[H(J - L_\phi) \mid C] = 0,$$

so $(J - L_\phi)$ is H-orthogonal to every square-integrable C-measurable operator. Independence of W from (ϑ, Y) therefore eliminates the cross term above. The remaining terms factor into a representation part and a task part:

$$\begin{aligned}\mathcal{B}_\phi(Y) = {} & \tfrac{1}{2}\operatorname{tr}\big(\mathbb{E}[(J - L_\phi)^\top H(J - L_\phi)]\Sigma\big) \\ & + \tfrac{1}{2}\operatorname{tr}\big(\mathbb{E}[L_\phi^\top H L_\phi]\,\mathbb{E}[(\vartheta - m)(\vartheta - m)^\top]\big).\end{aligned}$$

The posterior-mean identity $\mathbb{E}[(\vartheta - m)m^\top] = 0$ and the law of total covariance give

$$\mathbb{E}[(\vartheta - m)(\vartheta - m)^\top] = \Sigma - R_Y, \qquad 0 \preceq R_Y \preceq \Sigma.$$

Using the definitions of Γ_ϕ and Λ_ϕ, followed by $\Omega = \Gamma_\phi + \Lambda_\phi$, proves Equation (7).

For a C-measurable deployed operator K, decompose

$$J\vartheta - Km = (J - L_\phi)\vartheta + L_\phi(\vartheta - m) + (L_\phi - K)m.$$

The conditional projection equation makes the first term orthogonal, in the H metric, to both L_ϕ and $L_\phi - K$. Posterior orthogonality makes the second and third terms orthogonal because $\mathbb{E}[(\vartheta - m)m^\top] = 0$. Hence the only additional quadratic term relative to the encoded Bayes action is

$$\tfrac{1}{2}\operatorname{tr}\big(\mathbb{E}[(K - L_\phi)^\top H(K - L_\phi)]R_Y\big) = \operatorname{tr}(M_K R_Y),$$

which proves Equation (8).

For two representations of the same raw problem under common task-only evidence, Ω is representation independent, so their recovered-value difference is $\mathrm{tr}[(\Lambda_1 - \Lambda_2)R_Y]$. If $\Lambda_1 - \Lambda_2 \succeq 0$, this quantity is nonnegative for every $R_Y \succeq 0$, ruling out reversal under a common, equally costed experiment. For the converse existence statement, let $\Sigma \succ 0$, $\vartheta \sim N(0, \Sigma)$, and suppose the experiment class can reveal $w^\top \Sigma^{-1/2}\vartheta$ for any unit vector w. Gaussian conditioning gives

$$R_w = \Sigma^{1/2} w w^\top \Sigma^{1/2}.$$

Thus the recovered-value difference equals

$$w^\top \Sigma^{1/2}(\Lambda_1 - \Lambda_2)\Sigma^{1/2} w.$$

If this symmetric matrix has positive and negative eigenvalues, unit eigenvectors from the corresponding eigenspaces produce opposite preferences. The signs persist for sufficiently high-signal noisy linear measurements by continuity.

C.5 Direct reward experiment

The numerical experiment has 16 equally likely histories $W = (X, Z_1, Z_2, Z_3) \in \{-1, +1\}^4$ and nested codes $C_k = (X, Z_1, \ldots, Z_k)$, $k = 0, 1, 2, 3$. For each of 20 seeds, independent orthogonal action and parameter rotations define a non-diagonal positive definite curvature $H(W)$ and a 3×3 tangent $J(W)$. The action value is

$$Q_\delta(W, a) = -\tfrac{1}{2}(a - a_0(X) - J(W)\delta)^\top H(W)(a - a_0(X) - J(W)\delta).$$

All tested optimizers lie in the interior of $[-4, 4]^3$. The code at $k = 0$ already retains a_0; added coordinates therefore affect prospective control without changing nominal optimality. The script evaluates rewards directly at the raw optimizer, conditional optimizer, nominal policy, and signed tangent policies. It reports every combination of eight directions, four magnitudes, four codes, and 20 seeds: 2,560 value comparisons.

For deployment, independent equal-prior signs and task-evidence Gaussian observations generate 50,000 sessions per seed at each of six signal strengths. The same downstream histories and random numbers are used for paired code comparisons. Empirical costs are computed from realized quadratic rewards; analytic predictions are computed separately from the matrices and Gaussian risk. The largest conservation error is 9.25×10^{-16}, and the largest direct optimized-value error is 3.13×10^{-17}. The largest absolute cohort-mean risk discrepancy is 0.00128. Quadrature verifies the selective risk formula to 1.12×10^{-16}. These are numerical mechanism checks in an exactly specified family, not tests of neural estimation or native optimality.

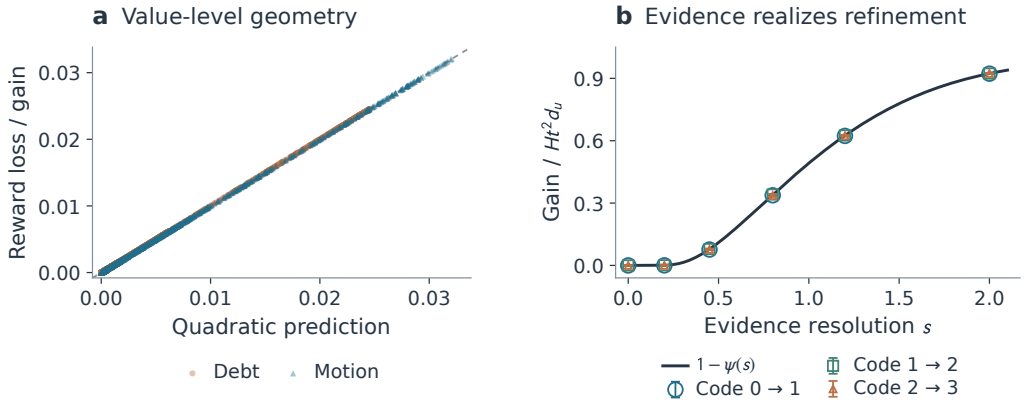

Figure 5: **Refinement converts debt into evidence-dependent value.** (a) Direct reward losses against their quadratic predictions, over all 2,560 tested code–direction–magnitude combinations. (b) Deployment gain from successive code refinements, divided by $Ht^2 d_u$. The solid curve is $1 - \psi(s)$; points and 95% paired outer-seed intervals come from reward evaluations over 20 independently rotated families. At zero evidence, every code uses the same nominal policy.

The fixed configuration is `config/release_v84_value_experiment.json`; the full value, risk, refinement, and allocation tables are under `results/raw/release_v84_*`. The experiment enumerates finite allocations with positive natural information and verifies both sides of (10). No cell or seed is excluded on the basis of its result.

C.6 Fixed-capacity representation choice under channel exchange

Let $W = (X, B_1, B_2)$ be three independent uniform signs and $a_0(X) = X(0.2, -0.1, 0.15)^\top$. Independent orthogonal action and parameter rotations U, V define

$$J(W) = U \begin{pmatrix} 1.4B_1 & 0 \\ 0 & B_2 \\ 0 & 0 \end{pmatrix} V^\top, \qquad Q_\delta(W, a) = -\tfrac{1}{2}\|a - a_0(X) - J(W)\delta\|^2.$$

The equal-capacity codes $C_1 = (X, B_1)$ and $C_2 = (X, B_2)$ each use two bits and preserve the same nominal action. For a uniform prior over the declared directions Ve_1, Ve_2, their average debts are 0.25 and 0.49, respectively. This trace ordering is not Loewner dominance; the codes are non-nested. Thus channel-dependent rank reversal is consistent with Corollary 5.

No expert actions supply the tangent estimate. At each of the eight histories, two parameter coordinates and three action coordinates are queried with parameter radius $\tau = 0.2$ and action radius $b = 0.3$. Each corner has 64 reward samples with noise standard deviation 0.025. Their contrast estimates

$$4\tau b\, \widehat{J}_{kj}(w) = \widehat{Q}_{+\tau e_j}(w, a_0 + be_k) - \widehat{Q}_{+\tau e_j}(w, a_0 - be_k) \\ - \widehat{Q}_{-\tau e_j}(w, a_0 + be_k) + \widehat{Q}_{-\tau e_j}(w, a_0 - be_k).$$

This estimator is unbiased in the specified reward model. Averaging within each code cell fits its response head; held-out execution uses that fitted head rather than the population projection.

A separate sample of 1,024 signed Gaussian observations per direction calibrates response slope, intercept, and noise variance. The per-probe information pair is $(0.6, 144)$ or its exchange $(144, 0.6)$; no downstream state is revealed by these task-only channels. The experiment crosses $t \in \{0.15, 0.30\}$, $H \in \{12, 48\}$, and costs $c \in \{0.01, 0.06, 0.20\}$. Counts are $0, 1, 2, 4, 8$. Acquisition uses a separate sensing budget with fixed post-acquisition horizon H, so its complete cost is nc. For each candidate code, the selector minimizes its fitted three-policy expected loss plus cost, then selects the code with lower direction-averaged frontier. The geometric reference chooses C_1 independently of evidence.

Five development fits and 20 independent locked fits use fixed configurations. Each code–direction–condition cell has 12,000 deployment sessions, with paired task signs, evidence noise, and downstream histories. Realized costs come from executed actions and quadratic rewards. Aggregation includes the high-cost cells where both codes freeze and tie. Across all declared horizons, magnitudes, and costs, the $C_1 - C_2$ cost contrast is 0.323 [0.315,0.331] when channel two is informative, and -0.751 [-0.752,-0.750] after exchange. This tests finite-library representation selection with noisy reward and evidence estimates, not global neural co-design or general physical control.

D Evidence-Priced Selection of Learned Codes

D.1 Task, observation, and representation capacity

This experiment uses the channel-exchange reward family of Appendix C.6, replacing its prescribed codes with trained encoders. At each downstream control step, $W = (X, B_1, B_2)$ is a fresh uniform draw from $\{-1, +1\}^3$. There are three action coordinates and two shift coordinates. Independent outer seeds draw orthogonal matrices $U \in \mathbb{R}^{3\times 3}$ and $V \in \mathbb{R}^{2\times 2}$. Define

$$a_0(W) = X(0.2, -0.1, 0.15)^\top, \qquad J(W) = U \begin{pmatrix} 1.4B_1 & 0 \\ 0 & B_2 \\ 0 & 0 \end{pmatrix} V^\top,$$

$$Q_\delta(W, a) = -\tfrac{1}{2}\|a - a_0(W) - J(W)\delta\|^2.$$

Direction $u_j = Ve_j$ has prior probability $1/2$. A deployment condition specifies magnitude t, downstream horizon H, sensing price c, and which evidence channel has high precision. The code

is selected before j is drawn; j is then revealed for allocation. The sign $\eta \in \{-1, +1\}$ of $\delta = \eta t u_j$ remains hidden. This is an iid controlled reward family, not a policy-dependent dynamics benchmark.

The shared nominal subcode is X. The adaptive encoder receives 64-dimensional observations

$$O = \tanh\big([X, B_1, B_2, N^\top]A + b\big) + \varepsilon_O,$$

where $N \sim N(0, 0.25^2 I_5)$ and $\varepsilon_O \sim N(0, 0.02^2 I_{64})$. Each outer seed draws $A \in \mathbb{R}^{8\times 64}$ with independent $N(0, 1/3)$ entries and $b \sim N(0, 0.15^2 I_{64})$. Fresh nuisance and measurement noise are used for training, calibration, and evaluation. Neither B_1, B_2 nor J is an encoder input or supervision label.

A 64–64–32–1 tanh network gives a Bernoulli probability $p_\phi(O)$. Its sampled bit Z and the common nominal bit form $C_\phi = (X, Z)$, a four-symbol representation. Only these bits, not the encoder logit, reach the response table $K_\phi(X, Z) \in \mathbb{R}^{3\times 2}$, stored in the intervention basis (u_1, u_2). Each model has 6,297 trainable parameters, including the 24 table entries. The nominal controller reads X alone, so both codes execute exactly the same nominal action for every observation and draw of Z. The predeclared nominal equivalence tolerance is 10^{-10} in mean loss; the measured difference is zero in every seed. This isolates the allocation of one adaptive bit at fixed nominal control.

D.2 Matched reward-derived learning

At each of the eight histories, offline interventions query $\delta = \pm\tau u_j$ and actions $a_0 \pm b_a e_k$, with $\tau = 0.2$ and $b_a = 0.3$. Each query returns Q_δ plus independent $N(0, 0.025^2)$ noise. The response target is the four-corner contrast

$$\widehat{T}_{k,j}(W) = \frac{1}{4\tau b_a} \sum_{s,r\in\{-1,+1\}} sr\, \overline{R}_{s\tau u_j}(W, a_0 + r b_a e_k), \tag{37}$$

where each mean uses 64 queries. Its expectation is $(J(W)u_j)_k$ in this reward family; the learner receives the noisy contrasts, not optimal-action labels. One common pool of 12,288 queries and 4,096 noisy observations trains both encoders. The observations are balanced over the eight histories.

The two candidate objectives use directional weights $w_0 = (0.95, 0.05)$ and $w_1 = (0.05, 0.95)$:

$$\mathcal{L}_\phi = \widehat{\mathbb{E}} \sum_{z=0}^{1} p_\phi(z \mid O) \sum_{j=1}^{2} w_{\phi,j} \|K_\phi(X, z)e_j - \widehat{T}(W)e_j\|^2 + 10^{-4}\widehat{\mathbb{E}}\, h(p_\phi(O)),$$

with $h(p) = -p\log p - (1-p)\log(1-p)$. The small positive entropy penalty encourages concentrated bits. Both objectives include both directions. The expectation over the two symbols is evaluated exactly, giving ordinary gradients without a sampled discrete-gradient approximation. Initial parameters, examples, minibatch order, capacity, and updates are matched; only w_ϕ differs. Adam (Kingma and Ba, 2015) uses learning rate 0.003, batch size 256, and 80 epochs (1,280 updates per model). Five development outer seeds fix the protocol and development comparators; 20 disjoint outer seeds regenerate the environment, observation map, reward queries, and training. All 40 locked encoders are retained.

D.3 Independent branch calibration and code selection

The encoders and response tables are frozen before calibration. A separate set of 4,096 noisy observations, balanced over histories, supplies direct reward measurements for all three executed branches

$$a_b(W) = a_0(W) + b t K_\phi(C_\phi)e_j, \qquad b \in \{0, -1, +1\},$$

under both known offline task signs and both evaluation magnitudes. Their negative sample mean rewards give the 2×3 loss matrix $\widehat{\ell}_{\phi,j,t}(\eta, b)$. All branches share the nominal action; paired reward noise is common to each branch comparison. Each code receives the same calibration budget. The loss matrix is measured without imposing canonical $0/1/4$ ratios. In particular, residual response loss is not substituted for representation debt.

A probe independently returns $Y \sim N(\eta t, I_j^{-1})$. The two channel conditions set (I_1, I_2) to $(0.6, 144)$ or $(144, 0.6)$. They provide task evidence, not the future control variables W. For each precision, 2,048 independent calibration probes under balanced assignments $\pm\tau$ fit an intercept, slope, and residual variance. These fitted parameters determine the sign likelihood ratio. The selector uses the measured branch losses and this fitted evidence law, rather than population J, Γ, Λ, or information matrices.

For $n > 0$, the terminal branch minimizes posterior expected measured loss. Gaussian integration of this rule gives the pre-probe cost

$$\widehat{C}_{\phi,j}(n) = nc + H\,\widehat{\mathbb{E}}\min_b \widehat{\mathbb{E}}[\widehat{\ell}_{\phi,j,t}(\eta, b) \mid Y_{1:n}].$$

For $n = 0$, the branch is fixed to the nominal policy. Acquisition uses $n \in \{0, 1, 2, 4, 8\}$. The outer choice is

$$\widehat{\phi} \in \arg\min_{\phi\in\{0,1\}} \tfrac{1}{2}\sum_{j=1}^{2} \min_n \widehat{C}_{\phi,j}(n). \tag{38}$$

It is recorded before drawing the direction, evidence, or evaluation returns. Allocation can subsequently use j; neither choice observes η. All sensing charges are paid once, including when the final decision is to freeze. Probes use a separate sensing budget and leave the declared downstream horizon unchanged.

Exact risk computation. For posterior $p = \mathbb{P}(\eta = +1 \mid Y_{1:n})$, each branch risk is affine: $r_b(p) = (1-p)\widehat{\ell}(-1, b) + p\widehat{\ell}(+1, b)$. Pairwise crossings partition $[0, 1]$ into intervals with a constant minimizing branch. Under the fitted law, the log odds conditional on η have distribution $N(2\eta s^2, 4s^2)$, where $s = t\sqrt{n\widehat{I}_j}$. Gaussian CDF differences therefore integrate the lower envelope exactly. This calculation extends the canonical reject-option rule to the actual measured loss matrix (Chow, 1970); it is exact for the fitted experiment, whose calibration error remains empirical.

D.4 Comparators, locked conditions, and outcomes

The condition grid crosses both channel assignments, $t \in \{0.15, 0.30\}$, $H \in \{12, 48\}$, and $c \in \{0.01, 0.06, 0.20\}$, giving 24 conditions. Every condition and direction uses 4,096 sessions with balanced task signs. Random-number streams separate training, response calibration, proxy calibration, evidence fitting, and deployment. Methods share evaluation histories, encoder randomization, and evidence prefixes. Code selection precedes the direction draw and averages the two directions equally. Deployment costs evaluate the stated reward at the executed actions; Gaussian reward perturbations are measurement noise in the offline queries.

Information-matched controls. Fixed-code comparators retain code 0 or code 1, using that code's calibrated allocation and terminal rules. Tangent-proxy selection chooses the code with smaller unweighted response error on independent noisy reward-derived targets and observations, without evidence prices; its deployment rule is otherwise identical. A global development rule chooses one code and one requested count per direction from mean returns over the five development seeds and all conditions. A condition-indexed development rule chooses the same joint plan separately for every observed condition. Both use the fitted terminal rule of the new frozen encoder. They have the same condition information and candidate policies as AP-GFAR, including the option not to probe; task sign is excluded from their inputs. Their lookup tables and the full analysis specification are fixed before locked training begins.

Costs are averaged over directions and conditions within each outer-training seed. Paired Student-t intervals use these 20 independent means. The six deployment contrasts, two channel-specific code contrasts, and their interaction form one nine-comparison Holm family (Holm, 1979). No seed or condition is excluded. Code 0 has smaller held-out unweighted response error in every locked seed, so the tangent-proxy and fixed-code-0 rows coincide; they are not separate replications of the effect.

When channel 2 is informative, the code-0-minus-code-1 cost contrast is 0.3108 $[0.3069, 0.3147]$. Exchanging the channels gives -0.7499 $[-0.7512, -0.7485]$, with interaction 1.0607 $[1.0569, 1.0645]$. The calibrated and realized mean preferences agree in all 480 seed–condition

Table 1: **Evidence-priced selection among learned codes.** Positive values are comparator cost minus AP-GFAR cost. Intervals use paired means over 20 outer-training seeds, averaging all 24 conditions and both directions. Each comparator uses the same frozen response and evidence models.

Comparator	Cost reduction	95% CI
Fixed learned code 0	0.1554	$[0.1534, 0.1574]$
Fixed learned code 1	0.3749	$[0.3743, 0.3756]$
Tangent-proxy selection	0.1554	$[0.1534, 0.1574]$
Global development rule	0.2114	$[0.2101, 0.2127]$
Condition-indexed development rule	-0.000028	$[-0.000054, -0.000002]$
Frozen nominal policy	0.5699	$[0.5697, 0.5701]$

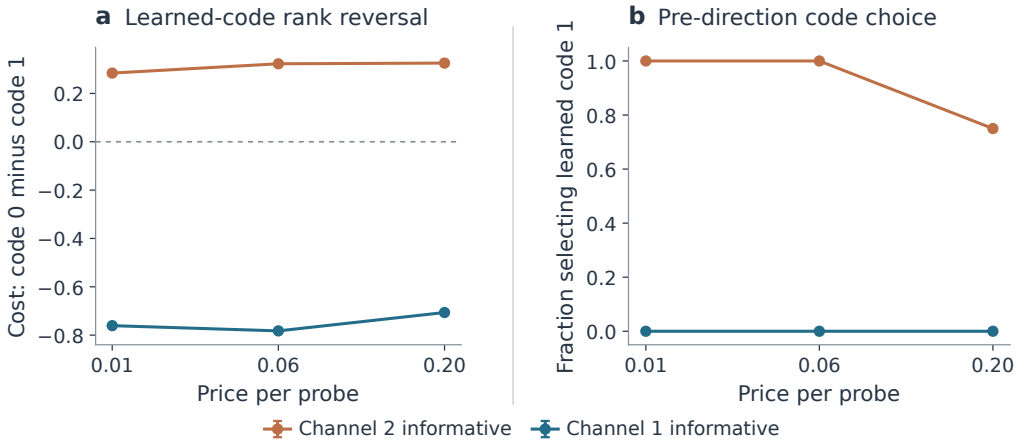

Figure 6: **Channel exchange changes which learned dependency is useful.** (a) Costs of the two non-nested learned codes change rank when only the evidence channels are exchanged. Points average magnitudes and horizons; bars are paired 95% intervals over 20 outer-training seeds. (b) The evidence-priced rule chooses a code before observing the task direction. The highest price includes conditions where both codes freeze; ties select the lower code index.

pairs, including 20 no-probe ties. The 480 pairs share the 20 training seeds and are summarized at the seed level. The mean within-seed RMSE of the predicted cost contrast is 0.0344. The positive selection gains and ranking interaction survive Holm adjustment. The condition-indexed development rule has a slightly lower mean cost: AP-GFAR's cost reduction against it is -0.000028 $[-0.000054, -0.000002]$.

Computational realization. The two-symbol training expectation and Gaussian lower-envelope integration are analytic. Deployment is streamed in batches of 128 sessions, with at most 4,096 observations per network call. Quadratic rewards permit exact reuse of branch energies across magnitudes and prices; paired horizon prefixes preserve the same rollout distribution. This reuse changes neither actions nor evaluation costs. The artifact retains encoder checkpoints, noisy reward queries, calibration outcomes, episode-level branch energies, and evidence draws, from which every selected return can be reconstructed. The experiment tests selection among learned codes in a controlled reward family. Sequential deployment and native reward-value measurements have the separate evidence roles described in the main text.

E Native evidence-conditioned representation value in HalfCheetah

This section supports Section 6.3. The native study has a different role from the controlled channel-exchange construction. The controlled experiment establishes that evidence can causally reverse a representation ranking. The native experiment asks whether the broader pair-valued criterion has

predictive content across independently trained continuous-control systems, without assuming that every system must cross zero.

E.1 SIXTEEN-SYSTEM FIXED-INTERFACE PROTOCOL

The environment is HalfCheetah-v5. Each of 16 independent training seeds produces two equal-capacity frozen representation–response programs and a shared nominal branch. Evaluation crosses four non-nominal physics tasks formed from two signs of a torso-mass perturbation and two signs of a gravity perturbation. Native episodic return is the outcome.

The evidence intervention uses one fixed physical probe (the locked `operating-front` probe) and the same underlying transition. Regime A exposes the root-state-plus-reward measurement interface; regime B exposes the joint-state-plus-reward interface. Evidence acquisition is a reset experiment; branch deployment then starts from an independent reset. The downstream controller observation, physical task, probe action, candidate branches, response programs, and charge are unchanged between regimes. The intervention therefore changes which part of the same native transition is available for identifying the task, rather than adding a synthetic side channel or changing the control problem.

For system s, physics task t, and evidence regime $e \in \{A, B\}$, let $D_{s,t,e}$ be realized return of program ϕ_1 minus ϕ_2 and let $\widehat{D}_{s,t,e}$ be the corresponding value contrast sealed before the evaluation bank is generated. Task-averaged contrasts are

$$D_{s,e} = \frac{1}{4}\sum_{t=1}^{4} D_{s,t,e}, \qquad \widehat{D}_{s,e} = \frac{1}{4}\sum_{t=1}^{4} \widehat{D}_{s,t,e}.$$

The prespecified reversal criterion tests whether the training-system mean satisfies $\mathbb{E}_s D_{s,A} > 0$ and $\mathbb{E}_s D_{s,B} < 0$ using an intersection–union test. Forecast and evaluation banks use disjoint sampling, and the cohort is stopped at the planned 16 systems with no performance-based extension.

Construction of the sealed value contrast. The forecast is a direct Monte Carlo plug-in estimate of the value of the *executed* finite branch program, not a regression trained across the 16 systems. Let $\mathcal{B}$ be the fixed nine-branch library. For representation $r \in \{1, 2\}$, task stratum t, and evidence interface e, define

$$\widehat{V}^{\mathrm{pre}}_{s,t,e,r} = \sum_{b\in\mathcal{B}} \widehat{p}^{\mathrm{pre}}_{s,t,e,r,b}\widehat{\mu}^{\mathrm{pre}}_{s,t,r,b} - c_e, \qquad \widehat{D}_{s,t,e} = \widehat{V}^{\mathrm{pre}}_{s,t,e,1} - \widehat{V}^{\mathrm{pre}}_{s,t,e,2}. \tag{39}$$

Here $\widehat{p}^{\mathrm{pre}}_{s,t,e,r,b}$ is the branch-selection frequency induced by the frozen evidence-to-branch rule, estimated from 512 native context blocks, while $\widehat{\mu}^{\mathrm{pre}}_{s,t,r,b}$ is the mean native episodic return of branch b, estimated from a separate bank of 128 return blocks; c_e is the fixed one-return-unit probe charge. Because deployment is reset independently after the probe, the branch-frequency and branch-return factors correspond to the declared executed program and can be estimated on separate banks. The physical probe, branch library, response programs, and branch selectors are fixed before these forecasts are formed. Evaluation repeats the same value functional with independent context and return banks of the same sizes. Thus $\widehat{D}$ is an empirical complete-program pair-value estimate, not a plug-in of the local tensor Λ_ϕ and not a regression fit to the realized 16-system outcomes. Task labels index the four evaluation strata and are not deployment inputs; evaluation returns enter only the realized contrast used by the audit.

The population-level reversal is not supported:

Contrast	Mean [95% CI]	Directional result
D_A	$18.79\ [-13.62, 51.21]$	positive side not established
D_B	$15.15\ [-15.35, 45.65]$	negative side not established
$D_A - D_B$	$3.64\ [-22.14, 29.42]$	interaction spans zero

The reversal IUT has $p = 0.847$. Two of 16 systems show the specified point reversal and eight of 16 have a positive interaction. Accordingly, the fixed-interface cohort characterizes heterogeneous pair values rather than establishing a population-wide reversal.

E.2 PREDICTIVE AUDIT OF SEALED NATIVE VALUES

After completion of the prespecified reversal analysis, we audit whether the already-sealed values predict the heterogeneous native outcomes. This is a secondary post-hoc analysis of prospectively sealed forecasts: no model is refit, no sensor or branch library is changed, and no rollout is added. The statistical unit for uncertainty remains the independent training system rather than the 128 task-level cells.

Ranking signs. The 128 cells contain 70 positive and 58 negative realized contrasts, so a majority-class rule achieves only 54.7% accuracy. The sealed values classify 116 of 128 signs correctly. After averaging over the four physics tasks, they classify 27 of 32 system–evidence ranking signs correctly. Cluster bootstrap intervals resample the 16 training systems.

Table 2: **Predictive validity of sealed native pair values.** Cell-level and task-averaged sign metrics use the same 16 independent training systems; constant-class performance is shown to rule out a class-imbalance explanation.

Metric	Cell	System × evidence
Accuracy (%)	90.62 [85.94, 95.31]	84.38 [68.75, 96.88]
Best constant class (%)	54.69 [50.00, 64.84]	62.50 [50.00, 81.25]
Accuracy advantage (pp)	35.94 [24.22, 42.97]	21.88 [0.00, 37.50]
Balanced accuracy (%)	90.69 [85.98, 95.34]	80.83 [60.45, 96.88]
MCC	0.812 [0.717, 0.906]	0.664 [0.289, 0.934]

Continuous value prediction. Sign accuracy does not establish quantitative calibration, so we also regress realized contrasts on the sealed contrasts. The main task-level regression includes training-system fixed effects. A separate evidence-difference regression asks whether the sealed values predict the change in relative representation value caused by exchanging the evidence interface. Wild cluster bootstrap inference treats the 16 systems as clusters; the seven reported regression tests form one Holm family.

Table 3: **Continuous native value prediction.** Slopes near one mean that realized representation contrasts change approximately one-for-one with the sealed value contrasts. The aggregate evidence-difference row is the training-system-level test of how the representation ranking moves when the evidence interface changes.

Regression	$\widehat{\beta}$ [95% CI]	Wild p	Holm p
Cell, pooled	0.994 [0.768, 1.219]	0.000031	0.000214
Cell, system FE	0.994 [0.774, 1.215]	0.000031	0.000214
Cell, system + stratum FE	0.996 [0.757, 1.234]	0.000031	0.000214
Aggregate, pooled	1.026 [0.475, 1.578]	0.036316	0.043579
Aggregate, system FE	1.160 [0.474, 1.847]	0.014526	0.043579
Evidence difference, cell	1.005 [0.648, 1.361]	0.000031	0.000214
Evidence difference, aggregate	1.162 [0.680, 1.643]	0.015320	0.043579

The system-fixed-effects task-level slope is 0.994 $[0.774, 1.215]$. After task averaging, the pooled slope is 1.026 $[0.475, 1.578]$. Most directly for the representation–evidence claim, the training-system-level evidence-exchange slope is 1.162 $[0.680, 1.643]$ (Holm-adjusted $p = 0.0436$). Thus heterogeneity is not merely that different systems happen to prefer different programs: value differences sealed before evaluation track both the realized ranking and its movement when the available native evidence changes.

E.3 PROSPECTIVELY LOCKED INTERFACE-DESIGN FOLLOW-UP

The fixed-interface cohort tests whether a particular pair of native measurements happens to reverse the population ranking. The pair-value theory makes a different, design-oriented prediction: if measurement interfaces can be chosen before evaluation, one can target interfaces that place the same two programs on opposite sides of the value ranking.

A separate 10-system follow-up uses the same HalfCheetah mass/gravity family, the same two frozen representation–execution programs, and one fixed physical probe. Each system contains two disjoint data banks. In one cross-fit fold, the training bank selects from the declared feasible stochastic measurement-interface family the interface with the largest sealed predicted contrast $V(\phi_1, E) - V(\phi_2, E)$ and the interface with the smallest predicted contrast; the other bank evaluates the realized contrasts. The bank roles are then swapped and the two held-out contrasts are averaged. Denote the resulting high- and low-value interfaces by E_A and E_B and define

$$D_A = V(\phi_1, E_A) - V(\phi_2, E_A), \qquad D_B = V(\phi_1, E_B) - V(\phi_2, E_B).$$

The interface-design rule, program pair, physical probe, and cross-fit contract were locked before any of these 10 validation-bank outcomes were inspected. The per-system cross-fitted values are:

Table 4: **Locked native interface design on held-out banks.** Positive D_A and negative D_B mean that changing only the native measurement interface changes which frozen representation–execution program has higher value. The table reports the cross-fitted held-out contrasts, not the training-bank predictions used to choose the interfaces.

Training seed	D_A	D_B
261003028	+76.99	-29.82
261005046	+20.58	-6.62
261006055	+36.76	-29.27
261008073	+50.34	-7.11
261009082	+0.01	-99.63
261010091	+103.43	-5.29
261012109	+24.87	-24.86
261013118	+45.06	-4.52
261016145	+72.64	-24.75
261017154	+2.57	-13.11

Across the 10 declared training systems, $D_A = 43.32\ [19.37, 67.28]$ and $D_B = -24.50\ [-44.75, -4.24]$; all 10 systems satisfy $D_A > 0$ and $D_B < 0$. The one-sided intersection–union test has $p = 0.0115$. This follow-up does not restore a claim that arbitrary native interfaces reverse rankings. Instead it validates the stronger design reading of the theory: pair value can be used prospectively to construct evidence interfaces for which the preferred representation changes on held-out data.

Prediction is distinct from switching utility. The same audit separates predictive validity from a stronger algorithmic claim. For two candidate program values, Equation (11) shows that evidence-conditioned code choice has no ideal advantage over the best fixed code unless the regime-specific contrasts cross zero with nonzero margin. In this cohort the plug-in hindsight room from that identity averages 1.47 return units; this is a descriptive sample quantity, not an unbiased population oracle-gain estimate or a confidence bound. Consistent with that small room, selecting from the sealed values improves mean return over frozen/no-probe and over fixed ϕ_2, while its contrast with fixed ϕ_1 spans zero. Allowing representation choice to depend on the evidence regime does not establish additional return over choosing one evidence-blind representation per system.

Table 5: **Native selection utility.** These are secondary seed-level contrasts. They show that accurate pair-value prediction need not imply a material benefit from switching representations across evidence regimes.

Comparator	Selected minus comparator [95% CI]
Fixed ϕ_1	10.15 [-3.28, 23.58]
Fixed ϕ_2	27.12 [5.61, 48.63]
Frozen, no probe	103.87 [29.79, 177.95]
Fixed-prior branch	23.70 [7.34, 40.05]
Evidence-blind code choice	-0.44 [-4.91, 4.02]

Scope. The native evidence separates prediction, prevalence, and design. Under two fixed measurement interfaces, a population-wide reversal is not supported, but sealed pair values track heterogeneous realized rankings and their movement under evidence exchange across the 16 systems in

this fixed task family. The prospectively locked 10-system follow-up addresses the design question: choosing the measurement interface from a disjoint bank to target the pair-value contrast produces the predicted crossing on every held-out system. The resulting native claim has three layers—pair-dependent value, prospective predictive content under fixed interfaces, and prospective interface design that changes the preferred representation. None requires reversal to be prevalent under arbitrary fixed interfaces.

F Experimental Models and Protocols

The learned representation comparisons and response-calibrated selector use the model and training protocol below. Value-calibrated deployment uses the same model family with a disjoint evaluation cohort and additional development-only branch-value fitting (Appendix G). The statistical unit is an independently trained outer seed. All representation, response, evidence, and deployment fits are sealed before the corresponding evaluation cohort is generated.

F.1 Policy-dependent learned POMDP

The environment has a controlled scalar state x_h, a latent distractor b_h, two eight-dimensional nuisance processes, and an eight-dimensional task parameter θ. An orthonormal matrix $V \in \mathbb{R}^{8\times 4}$ defines $z = V^\top \theta$. Its coordinates perturb reward, actuation, observation, and contextual drift. Each outer seed independently generates V, the observation map, training trajectories, and neural weights. Dynamics and observed reward are

$$\begin{aligned} x_{h+1} &= 0.86x_h + (0.38 + 0.20z_2)a_h + 0.18z_4 + \varepsilon_h^x, \\ r_h &= -\tfrac{1}{2}(x_h^2 + 0.12a_h^2) + z_1 x_h - c_h + \varepsilon_h^r, \qquad a_h \in [-4, 4], \\ b_{h+1} &= 0.82b_h + \varepsilon_h^b. \end{aligned}$$

Independent Gaussian standard deviations are 0.08 for process and reward noise and 0.52 for the distractor innovation. The sensor pair is

$$\begin{pmatrix} y_{1,h} \\ y_{2,h} \end{pmatrix} = \begin{pmatrix} \cos z_3 & \sin z_3 \\ -\sin z_3 & \cos z_3 \end{pmatrix} \begin{pmatrix} x_h \\ b_h \end{pmatrix} + \varepsilon_h^y, \qquad \varepsilon_h^y \sim N(0, 0.07^2 I).$$

A contextual cue has mean z_4 and noise standard deviation 0.26. The 26-dimensional natural feature vector combines the current and previous sensor pairs, cue, distractor, previous reward and action, time features, and nuisance coordinates. A fixed outer-seed map renders 64 observations as $o = W_1 v + 0.32\tanh(W_2 v + b) + \epsilon_o$, with output noise of standard deviation 0.015. State nuisances follow an AR(1) process with coefficient 0.75 and innovation standard deviation 0.65; task nuisances are fixed within an episode. The environment source specifies feature ordering and initialization. Actions affect later observations and rewards through x_{h+1}.

Training data and supervision. An episode has 56 steps. Each outer seed supplies 64 nominal training and 16 nominal validation episodes. Eight interventions $\theta = \pm 0.22 V e_j$ each supply 24 training, six evidence-calibration, and six held-out validation episodes. The nominal action reference is a finite-horizon Riccati controller evaluated at the nominal sensor estimate. Symmetric task perturbations give its expert tangent. These targets come from the reference controller; the primary deployment outcome is episodic return, with stagewise Riccati action regret used only as a reference diagnostic.

The reward-derived tangent starts from a nominal policy fitted to the same nominal action targets. A response network then receives 48 epochs of rollout-return training under antithetic task assignments: two rollouts per sign and relevant direction per epoch, Gaussian action exploration of standard deviation 0.35, discount 0.97, a learned value baseline, and clipped standardized advantages. A score-function policy-gradient update (Williams, 1992) includes a 10^{-4} tangent-norm penalty. If f_ψ denotes the trained reward-response network and $\widetilde{o}(W) = (o(W) - \mu_{\text{train}})/\sigma_{\text{train}}$ the standardized observation, the downstream regression target is exactly

$$\widehat{T}(W) = 2\, f_\psi(\widetilde{o}(W)).$$

The factor two is applied once after response-network training; it is distinct from the intervention radius and action bounds, and does not assert that $\widehat{T}$ is an unbiased estimate of the mathematical J.

No shifted expert actions or simulator derivatives enter this learner. Nominal imitation and known offline intervention assignments remain shared inputs; reward-derived adaptation does not imply expert-free nominal training.

Encoder and operator geometry. Two width-96 SiLU layers feed 12-dimensional Gaussian mean and log-variance heads. Width-96 action and tangent heads share the code. Nominal and prospective variants have identical architectures, initial weights, data order, and optimization budgets. AdamW (Loshchilov and Hutter, 2019) uses learning rate 10^{-3}, batch size 512, 24 epochs, and a 10^{-3} Gaussian KL penalty. Prospective weights are 0.35 for reward-derived targets, 0.05 for expert targets, and zero for nominal-VIB. After encoder freezing, tangent heads receive 24 fitting epochs from a common initialization. Deployment uses the latent mean.

For held-out response target $\widehat{T}$ and actual readout K_ϕ, the three estimated Euclidean response losses are

$$\widehat{E}_0 = \tfrac{1}{2}\widehat{\mathbb{E}}[\widehat{T}^\top \widehat{T}], \quad \widehat{E}_+ = \tfrac{1}{2}\widehat{\mathbb{E}}[(\widehat{T} - K_\phi)^\top (\widehat{T} - K_\phi)], \quad \widehat{E}_- = \tfrac{1}{2}\widehat{\mathbb{E}}[(\widehat{T} + K_\phi)^\top (\widehat{T} + K_\phi)].$$

These are operator-calibration surrogates. A finite neural residual includes target and approximation errors; it is not identified with irrecoverable value debt. Appendix C evaluates the exact value geometry.

Evidence and allocation. An evidence block contains eight completed transitions (o_h, a_h, r_h, o_{h+1}). Features are transition means and standard deviations, mean $a_h(o_{h+1} - o_h)$, and mean $r_h(o_{h+1} - o_h)$. Ridge regression with penalty 30 predicts offline task assignments. A separate fold fits $\widehat{\theta} = O\theta + b + \epsilon$; held-out covariance Σ includes ridge $10^{-3}I$. The information surrogate is $\widehat{F} = O^\top \Sigma^{-1} O$. Quality q is implemented by independent added noise with covariance $(q^{-1} - 1)\Sigma$. The Gaussian likelihood ratio therefore uses precision $q\Sigma^{-1}$. Task sign never enters deployment features.

The grid crosses $t \in \{0.24, 0.48, 0.72\}$, $H \in \{14, 32, 48\}$, and the conditions below. The declared direction $u = V(1, 1, -1, -1)^\top/2$ is available to all direction-conditioned policies; sign is unknown and balanced. Each sign and condition has eight deployment replicates. Counts are $\{0, 1, 2, 4\}$. For directional response losses e_0, e_+, e_- and posterior p of the positive shift, the terminal rule minimizes

$$\{e_0,\ pe_+ + (1-p)e_-,\ pe_- + (1-p)e_+\}.$$

For $e_+ < e_0 < e_-$, the positive adaptation threshold is $\log\{(e_- - e_0)/(e_0 - e_+)\}$; it equals $\log 3$ only for the canonical loss ratios. Unprofitable operators freeze. Allocation integrates this same terminal rule under the fitted Gaussian experiment, using $nq\, u^\top \widehat{F} u$.

Condition	Quality	Step cost	Probe-action s.d.
High cost / low quality	0.1	0.12	0
Intermediate	0.5	0.015	0.04
Low cost / high quality	1	0.002	0.08

The selector accounts for $8n$ waiting steps and $(H - 8n)_+$ remaining control steps. It uses the estimated three-policy losses and charges the configured price plus predicted action-intervention penalty $0.06\sigma_a^2$. Actual return charges the price and measures the action's effect through reward. The primary implementation uses no additional information multiplier or condition-specific acquisition margin. These fitted response losses are calibration surrogates for control value; the population deployment comparison is measured directly from held-out returns.

G Value-Calibrated Deployment

G.1 Conditional branch values and acquisition costs

Fix the pre-probe descriptors, a finite allocation library $\mathcal{N}$, and finite permitted branch sets $\mathcal{A}_n$ containing the nominal branch 0. Allocation n produces transcript features X_n. Write $\ell_{n,a}$ for the loss of the complete deployment under branch a, including its acquisition prefix, remaining control,

and all charges. An additive charge already included in this loss is not added again. Expectations condition on the pre-probe descriptors and the fitted policy library. Let ℓ_{np} be the loss of no probing followed by the nominal policy. Centering at its value leaves allocation rankings unchanged:

$$C_n(d) = \mathbb{E}[\ell_{n,d(X_n)} - \ell_{\mathrm{np}}], \tag{40}$$

$$g_{n,a}(x) = \mathbb{E}[\ell_{n,0} - \ell_{n,a} \mid X_n = x], \qquad d_n^*(x) \in \arg\max_{a \in \mathcal{A}_n} g_{n,a}(x). \tag{41}$$

The nominal branch has $g_{n,0} = 0$. In the learned sign experiment, $\mathcal{A}_0 = \{0\}$ and $\mathcal{A}_n = \{0, +, -\}$ for $n > 0$. Matched rollouts share the acquisition prefix and random-number streams; different terminal actions can induce different subsequent states. Their return differences supply the conditional-gain targets.

The terminal rule maximizes $\widehat{g}_{n,a}(X_n)$. The acquisition rule minimizes a pre-probe estimate $\widehat{C}_n$ of $C_n(\widehat{d}_n)$. These costs include allocation-dependent occupancies and remaining horizons. When allocation n has a known deterministic charge q_n with $q_0 = 0$, write $\ell_{n,a} = \widetilde{\ell}_{n,a} + q_n$. Then

$$C_n(d) = \widetilde{C}_n(d) + q_n, \qquad \widetilde{C}_n(d) = \mathbb{E}[\widetilde{\ell}_{n,d(X_n)} - \ell_{\mathrm{np}}]. \tag{42}$$

Thus one may fit $\widetilde{C}_n$ and add q_n exactly at selection time; this changes neither the target deployment program nor the finite-library comparison. The canonical Gaussian calculation results when the branch losses and evidence experiment take the form of Theorem 3. For a fitted sequential system the non-deterministic part is estimated from the actual branches instead.

G.2 Finite-library estimation error

Proposition 13 (Stability of value-calibrated deployment)**.** *Suppose all branch and no-probe losses are integrable, learned heads are fixed before deployment, and $\widehat{g}_{n,0} = g_{n,0} = 0$. Define*

$$e_{T,n} = \mathbb{E} \max_{a \in \mathcal{A}_n} |\widehat{g}_{n,a}(X_n) - g_{n,a}(X_n)|, \qquad e_A = \max_{n \in \mathcal{N}} |\widehat{C}_n - C_n(\widehat{d}_n)|.$$

For $\widehat{n} \in \arg\min_n \widehat{C}_n$,

$$0 \le C_{\widehat{n}}(\widehat{d}_{\widehat{n}}) - \min_n C_n(d_n^*) \le 2e_A + 2 \max_n e_{T,n}. \tag{43}$$

The reference optimum uses the same allocations, candidate policies, and feature σ-fields as the deployed rule.

Proof. At each feature value, maximality of $\widehat{g}_{n,\widehat{d}_n}$ gives

$$g_{n,d_n^*} - g_{n,\widehat{d}_n} \le 2 \max_{a \in \mathcal{A}_n} |\widehat{g}_{n,a} - g_{n,a}|.$$

The tower property and integration bound the terminal excess by $2e_{T,n}$. For $n^* \in \arg\min_n C_n(d_n^*)$, finite-library minimization gives

$$\begin{aligned} C_{\widehat{n}}(\widehat{d}_{\widehat{n}}) &\le \widehat{C}_{\widehat{n}} + e_A \le \widehat{C}_{n^*} + e_A \le C_{n^*}(\widehat{d}_{n^*}) + 2e_A \\ &\le C_{n^*}(d_{n^*}^*) + 2e_A + 2e_{T,n^*}. \end{aligned}$$

The lower bound follows by minimizing over the same library. The common no-probe centering cancels from each difference. □

Corollary 14 (Joint representation and allocation selection)**.** *Fix a finite library Φ of encoders and their candidate response policies before calibration and evaluation. For each ϕ, let $\mathcal{N}_\phi$ be a finite set of feasible allocations, and put $\mathcal{I} = \{(\phi, n) : \phi \in \Phi,\ n \in \mathcal{N}_\phi\}$. All costs use the same physical units and one common no-probe reference. For $i \in \mathcal{I}$, use its permitted feature σ-field and branch set to define $C_i, g_{i,a}, d_i^*, \widehat{d}_i$ as in* (40)*. Suppose the integrability and zero-branch conditions of Proposition 13 hold, and set*

$$e_T = \max_{i \in \mathcal{I}} \mathbb{E} \max_{a \in \mathcal{A}_i} |\widehat{g}_{i,a}(X_i) - g_{i,a}(X_i)|, \qquad e_A = \max_{i \in \mathcal{I}} |\widehat{C}_i - C_i(\widehat{d}_i)|.$$

Then $\widehat{i} \in \arg\min_{i \in \mathcal{I}} \widehat{C}_i$ satisfies

$$0 \le C_{\widehat{i}}(\widehat{d}_{\widehat{i}}) - \min_{i \in \mathcal{I}} C_i(d_i^*) \le 2e_A + 2e_T. \tag{44}$$

Proof. For each fixed i, conditional maximization bounds terminal excess by $2e_T$. If i^* minimizes $C_i(d_i^*)$, comparison through the estimated costs gives

$$C_{\widehat{i}}(\widehat{d}_{\widehat{i}}) \le \widehat{C}_{\widehat{i}} + e_A \le \widehat{C}_{i^*} + e_A \le C_{i^*}(d_{i^*}^*) + 2e_A + 2e_T.$$

The lower bound follows from minimization over the same frozen library. □

If a finite-valued direction D is revealed after selecting the code but before acquisition, n can denote a contingent plan $D \mapsto n(D)$. The set of these plans is finite; include D in the permitted transcript and average costs over its pre-selection prior. The corollary then applies without letting the representation depend on D. All expectations condition on the fitted library and calibration data. The comparison is to the best member of that library with its own permitted terminal decisions, not to an unrestricted encoder or policy class. Applying the bound as a numerical certificate would require valid bounds on e_A and e_T; empirical training-seed intervals alone do not provide them.

The bound separates errors in terminal choice and in valuing acquisition. Representation debt is a different quantity: it compares the raw and encoded optimal policy classes under the assumptions of Theorem 2. Neither estimation term is substituted for that debt. If an independent calibration sample gives simultaneous intervals $C_n(\widehat{d}_n) \in [L_n, U_n]$, then $U_{\widehat{n}} - \min_n L_n$ bounds excess over the best frozen deployed rule on that event. Constructing distribution-free intervals requires additional boundedness or tail conditions. The paired empirical intervals below summarize training-seed contrasts, rather than simultaneous cost coverage.

G.3 Prospectively locked learned-system confirmation

The process, observation map, encoder, reward-trained response, and evidence model are those in Appendix F. The five development systems supply every fitted terminal and acquisition head. No model, feature set, hyperparameter, comparator, condition, or analysis rule is refitted after the fresh cohort is generated.

The original grid uses $t \in \{0.24, 0.48, 0.72\}$, $H \in \{14, 32, 48\}$, and $(q, c, \sigma_p) \in \{(0.1, 0.12, 0), (0.5, 0.015, 0.04), (1, 0.002, 0.08)\}$. The interpolation domain uses $t \in \{0.36, 0.60\}$, $H \in \{20, 40\}$, and $(q, c, \sigma_p) \in \{(0.2, 0.06, 0.02), (0.7, 0.008, 0.06)\}$. Positive allocations $n \in \{1, 2, 4\}$ consist of eight-step probe blocks and must satisfy $8n < H$; $n = 0$ freezes without acquisition. Each condition has 16 replicates for each balanced task sign. Matched branches share their probe prefix and exogenous random-number streams, while the post-prefix state evolves under the branch actually executed.

Complete-program target and matched heads. For a positive count n, the deterministic probe-step charge is $q_n = 8nc$. Let $R_{n,\widehat{d}_n^{(-s)}}$ denote the complete return obtained by the terminal rule trained without development system s, and let R_{np} be no-probe return. Acquisition fitting uses

$$Y_n = R_{n,\widehat{d}_n^{(-s)}} - R_{\mathrm{np}} + q_n, \qquad \widehat{G}_n = \widehat{\mathbb{E}}[Y_n \mid Z_n^{\mathrm{pre}}] - q_n, \qquad \widehat{G}_0 = 0. \tag{45}$$

Both returns contain the actual charge; the add-back and subtraction in (45) remove a known deterministic component from the regression target and restore it exactly at selection. Previous rewards remain part of the observation stream, so this accounting does not assume that probe price leaves trajectories unchanged.

Both acquisition heads use 180 histogram-boosted trees (Friedman, 2001), at most 12 leaves per tree, a 48-sample leaf minimum, learning rate 0.05, and ℓ_2 regularization 5. The generic head receives $(t, H, n, c, q, \sigma_p)$. The response-aware head adds held-out frozen-response, response-motion, cross-energy, and directional-information descriptors. They share terminal heads, response policies, evidence models, feasible counts, target construction, and tree budget. The operator-aware Gaussian comparator uses the same fitted response losses, the remaining horizon $H - 8n$, and the predeclared probe-action cost proxy; it is therefore distinct from substituting the canonical $0/1/4$ loss ratios for a learned operator. Global and condition-indexed rules are frozen from the same development data.

Prospective lock and statistical contract. Before any fresh system was trained, the analysis code, seed list, protocol, development artifacts, feature sets, tree settings, grids, comparators, and primary

Table 6: **Fresh learned-POMDP confirmation.** Entries are complete-program value calibration minus each comparator, with paired 95% intervals over 32 independently trained fresh outer systems. The two prespecified primary tests are the interpolation contrasts with the operator-aware Gaussian rule and frozen/no-probe; the generic-head row is a method diagnostic. Positive values favor complete-program calibration.

Comparator	Original grid	Interpolation
Operator-aware Gaussian	-0.0047 $[-0.0159, 0.0064]$	0.0111 $[-0.0004, 0.0226]$
Frozen / no probe	0.5826 $[0.5623, 0.6030]$	0.4163 $[0.3865, 0.4461]$
Capacity-matched generic head	-0.0131 $[-0.0226, -0.0035]$	0.0000 $[0, 0]$

tests were locally timestamped and SHA256-locked. The lock records that no fresh evaluation outcome existed at that time. The 32 fresh outer seeds have no overlap with 355 previously recorded configurations. Each outer system regenerates its task basis and nonlinear observation map and retrains the encoder, nominal policy, prospective response, and evidence model. Its predeployment allocation tables are written before evaluation outcomes. All 32 systems completed before the analysis was started; no seed was deleted, replaced, or added. Independent recomputation from 111,616 raw rows reproduces returns and gains to maximum absolute error 8.9×10^{-16} and verifies that the source lock is unchanged and every plan precedes its outcomes. The lock is local rather than externally registered.

The primary domain is the eight interpolation conditions. Two prespecified paired contrasts compare complete-program calibration with (i) the operator-aware Gaussian rule and (ii) frozen/no-probe. Conditions, signs, and replicates receive equal weight within an outer system. Two-sided paired Student-t tests use the 32 system-level differences and Holm correction over the two primary hypotheses. The joint success criterion requires both mean differences to be positive and both adjusted p values to be below 0.05. Original-grid and all remaining comparator contrasts are secondary.

Table 6 reports the full original-grid and interpolation contrasts.

On interpolation, complete-program calibration exceeds frozen/no-probe by 0.4163 $[0.3865, 0.4461]$ with Holm-adjusted $p = 1.61 \times 10^{-23}$. Its contrast with the operator-aware Gaussian rule is 0.0111 $[-0.0004, 0.0226]$ with adjusted $p = 0.0590$. The first primary comparison is therefore supported and the second does not cross the prespecified threshold; the joint primary success criterion is not met. The same Gaussian contrast on the original grid is -0.0047 $[-0.0159, 0.0064]$.

The response-aware and capacity-matched generic acquisition heads choose the same count in all $32 \times 8 = 256$ interpolation system–condition pairs and therefore have identical returns there. On the original grid, 831 of 864 pairs select the same count, and the response-aware-minus-generic contrast is -0.0131 $[-0.0226, -0.0035]$. Thus the extra response descriptors are not an empirical source of acquisition advantage in this cohort. The condition-indexed development rule has the same qualitative tradeoff: it is slightly better on the original grid, whereas complete-program selection is better on interpolation. These are secondary comparisons of estimator parameterization and interpolation, separate from the representation-level channel-exchange result.

Crossed acquisition and terminal controls. The two implementations differ in both pre-probe acquisition valuation and the terminal branch rule. Crossing these components on interpolation gives

Complete-program selector minus crossed comparator	Mean [95% CI]
Gaussian acquisition, calibrated terminal	0.0227 $[0.0102, 0.0351]$
Calibrated acquisition, Gaussian terminal	-0.0057 $[-0.0119, 0.0005]$

Hence, with calibrated terminal decisions fixed, changing acquisition valuation produces a positive secondary contrast. With calibrated acquisition fixed, the terminal contrast spans zero. These crossed controls localize the empirical difference within this finite branch library; they do not identify operator mismatch as its only possible cause.

Interpretation. Proposition 12 assigns local deployment loss to the operator actually executed when $K \neq L_\phi$. The fresh cohort establishes positive complete-program deployment value relative to freezing. The incremental contrast with the operator-aware Gaussian approximation remains unresolved under the prespecified primary test, while the matched generic head shows that the acquisition feature basis need not itself be response-aware. The structural object is the executable branch/program value; its finite-sample estimator may be generic or response-aware.

G.4 Finite-sample acquisition with a no-probe reference

Acquisition commits resources before the terminal decision. A positive point estimate of its value need not justify that commitment when the estimated margin is small relative to calibration error. We use paired returns to place the comparison with no probing on an explicit confidence scale. This controls acquisition for the fitted response library; it does not require that a response recover the optimal-action projection.

Condition on encoders, policies, evidence models, and terminal heads fitted without the acquisition-calibration sample. For a finite library of complete deployment programs i, including their pre-probe descriptors, let Y_i be return before a deterministic charge $q_i(p)$ at price p. Each program includes its acquisition prefix and the subsequent frozen terminal rule. Let Y_0 be the no-probe return under the same task and coupled rollout randomness. The random variable $Z_i = Y_i - Y_0$ has mean μ_i; its net deployment gain is $\Delta_i(p) = \mu_i - q_i(p)$. The reference program has $Z_0 = q_0 = r_0 = 0$. Price affects the charge, not the distribution of Z_i or the fixed terminal rule in this construction.

Corollary 15 (Confidence-calibrated acquisition)**.** *For each nonzero program i, suppose $Z_i \in [a_i, b_i]$ almost surely, with known width $w_i = b_i - a_i$. Calibration uses independent and identically distributed task–rollout units, independent of the fitted programs. Different programs may share each unit. Fix a finite set $\mathcal{S}$ of sample sizes at least two and let M count the nonzero program–sample-size pairs. At $N \in \mathcal{S}$, write $\bar{Z}_{i,N}$ and $s^2_{i,N}$ for the sample mean and unbiased sample variance, and define*

$$b_\alpha = \log(4M/\alpha), \qquad r_{i,N} = \sqrt{\frac{2s^2_{i,N} b_\alpha}{N}} + \frac{7 w_i b_\alpha}{3(N-1)}, \tag{46}$$

$$\widehat{i}_N(p) \in \operatorname*{arg\,max}_{i \text{ feasible}} \{\bar{Z}_{i,N} - q_i(p) - r_{i,N}\}, \tag{47}$$

with the no-probe program included. With probability at least $1-\alpha$ over calibration, simultaneously for the declared sample sizes and every price, the selected program satisfies

$$\Delta_{\widehat{i}_N(p)}(p) \geq 0, \qquad 0 \leq \max_i \Delta_i(p) - \Delta_{\widehat{i}_N(p)}(p) \leq 2 r_{i^*(p),N}, \tag{48}$$

where $i^*(p)$ is a best feasible frozen program. In particular, a program with $\Delta_i(p) > 2r_{i,N}$ makes the selected lower confidence bound strictly positive on this event.

Proof. The empirical Bernstein inequality of Maurer and Pontil (2009, Theorem 4), applied to each rescaled Z_i and its negative, and a union bound over the M pairs give $|\bar{Z}_{i,N} - \mu_i| \leq r_{i,N}$ simultaneously. No independence across programs or nested sample sizes is needed. Subtracting the known charge preserves these intervals for every price. Write $L_i = \bar{Z}_{i,N} - q_i(p) - r_{i,N}$. Then $\Delta_{\widehat{i}} \geq L_{\widehat{i}} \geq L_0 = 0$ and $L_{\widehat{i}} \geq L_{i^*} \geq \Delta_{i^*} - 2r_{i^*,N}$. The last assertion follows from $L_i \geq \Delta_i - 2r_{i,N} > 0$. □

The result instantiates the acquisition-error term in Proposition 13. Its reference uses the actual frozen terminal rules, so it makes no assertion that terminal estimation error vanishes. The nonnegative-gain statement is a conditional population expectation; individual trajectory outcomes remain random. The price-uniform statement relies on price entering only through a known charge; changing the evidence channel requires its own calibrated program. The confidence penalty is a standard concentration construction. Its role here is to resolve whether retained, executable motion is worth purchasing evidence for at the available calibration sample size.

Mixture-betting intervals. The additive range term in (46) can dominate a small but useful acquisition margin. We also instantiate the comparison with betting confidence sequences (Waudby-Smith and Ramdas, 2024). For K nonzero programs with known $|Z_i| \le B_i$, fix positive fractions $f_j < 1$ and weights v_j summing to one before calibration. For $m \in (-B_i, B_i)$ define

$$W_{i,N}(m) = \sum_j v_j \prod_{s=1}^{N} \left[1 + f_j \frac{Z_{i,s} - m}{B_i + m}\right]. \tag{49}$$

Let $L_{i,N}$ be the supremum of means rejected by $W_{i,N}(m) \ge 2K/\alpha$, with $\sup \varnothing = -B_i$, and obtain $U_{i,N}$ by negating the corresponding lower bound for $-Z_i$. Each factor is positive and nonincreasing in m, so inversion is one-dimensional. We use equal weights on $f_j \in \{1/16, 1/8, 1/4, 1/2\}$ and conservative bisection.

Lemma 16 (Simultaneous acquisition intervals)**.** *Under the independent-calibration conditions above, $\mathbb{P}\{L_{i,N} \le \mu_i \le U_{i,N}$ for all $i, N\} \ge 1 - \alpha$. Set $L_{0,N} = U_{0,N} = 0$ and select a feasible program maximizing $L_{i,N} - q_i(p)$. On this event, for all N and all prices, its expected net gain is nonnegative, and its excess below the best frozen program is at most*

$$\max_i \{U_{i,N} - q_i(p)\} - \{L_{\widehat{i},N} - q_{\widehat{i}}(p)\}. \tag{50}$$

Proof. At an interior true mean, each product in (49) is a nonnegative unit-mean martingale: its next multiplicative factor has conditional mean one. The fixed mixture has initial value one and the same property. Ville's inequality and a union bound over the $2K$ signed processes give simultaneous coverage for every calibration sample size. At a support endpoint the observations are constant almost surely and the corresponding endpoint bound holds directly. On the coverage event, the selected gain is at least its lower bound, which is at least the zero reference. Every program's gain is below its upper bound, giving (50). □

Here N counts offline calibration units; n denotes deployment probe episodes. The confidence sequence removes the need to pay a separate union-bound factor for the reported calibration sample sizes. Both interval constructions concern the same complete programs and the same known charges.

G.5 Independent calibration in the official context system

The experiment fixes six independently trained 32,000-update PEARL checkpoints and their already-fitted terminal regressors. There is no further policy, context-encoder, or terminal-head fitting. We draw 4,096 independent calibration goals and 2,048 independent evaluation goals uniformly from $[-1, 1]^2$, the original point-robot task-generating distribution. A goal and its random reset/latent stream form one statistical unit. The same units are shared across all models and compared programs; calibration and evaluation streams are disjoint. The six checkpoints are the independent training units; the added goals are new calibration and evaluation draws for those fixed runs.

Complete programs and permitted information. The original two-dimensional environment has position observation, additive bounded actions, and negative Euclidean distance to the hidden goal as reward. Each episode resets after 20 steps. Horizons are $h \in \{20, 60, 100\}$; counts $n \in \{0, 1, 2\}$ are feasible when $20n < h$. Each probe uses the current context-conditioned policy and updates the posterior with all accumulated reward-inclusive transitions. The terminal rule then chooses the prior or posterior policy from its frozen value head. Its inputs are horizon, count, context-reward mean and standard deviation, and posterior means and log variances. Neither acquisition nor terminal selection observes the goal, evaluation return, or task index. Evaluation of counterfactual branches is used for paired comparisons only. For the reported return per step, $q_n(p) = 20np/h$ at $p \in \{0, 0.25, 1\}$. Probe rewards, remaining control, and charges each enter once. No probing uses the prior throughout; buying evidence and then choosing the prior retains both the probe prefix and its charge.

A physical bound for paired returns. Two coupled episodes start at the same position and share the goal. With action coordinates bounded by a, their positions after step k differ by at most $2a\sqrt{2}\,k$.

The reverse triangle inequality bounds their reward difference by the same quantity. The first episode is identical for every program, giving

$$|Z_{h,n}| \le B_h = a\sqrt{2}(20+1)\frac{h-20}{h}, \qquad w_{h,n} = 2B_h. \tag{51}$$

The implementation uses $a = 0.10000001$, above the simulator's float32 action bound. The bound holds also when the terminal branch depends on the probe transcript. It does not use observed return extrema. Known charges shift the mean without enlarging the range or variance of the paired gain.

Frozen allocation and matched controls. The sample-size set is $\{64, 256, 1024, 4096\}$, with 4,096 designated as the primary budget. Nested prefixes measure sample-size dependence, not independent repetitions. At $\alpha = 0.05$, mixture betting covers $K = 6 \times 4 = 24$ model–horizon–positive-count programs at every calibration sample size, using rejection threshold $2K/\alpha = 960$. The empirical Bernstein comparator covers the 96 program–sample-size pairs at the same confidence level. No extra price factor is required. The reliable selector maximizes $L_{i,N} - q_i(p)$; ties prefer the smaller count. A paired plug-in selector maximizes the same independent-calibration mean without the confidence penalty, sharing every policy, terminal decision, and sample. A condition-indexed posterior rule chooses counts from the same calibration data but always executes the posterior after probing. We also retain the earlier AP-GFAR and condition-indexed plans and no probing. Comparisons with the paired plug-in rule isolate the confidence margin; comparisons with the earlier plans also include the change in calibration sample and budget. Every allocation table is sealed before evaluation trajectories are generated. The mixture-betting confirmation uses a fresh 4,096-goal calibration sample and fresh 2,048-goal evaluation sample. A separate empirical-Bernstein cohort has the same sample sizes and frozen models; its results are retained below. The two cohorts are not pooled, and both interval rules are compared on the fresh confirmation sample.

Evaluation and computation. We average conditions equally within each training run. Paired 95% product-bootstrap intervals resample the six training runs and the 2,048 shared evaluation tasks separately, using 20,000 replicates (Owen, 2007; Agarwal et al., 2021). These intervals describe empirical contrasts; the simultaneous finite-sample statement instead follows from the bounded-return construction above. Results include all three prices, the no-probe ties, and both active horizons. Each task requires 12 distinct episodes: identical prior and prefix rollouts are cached across allocations. An independent replay against the uncached executor agrees within 3.6×10^{-15}. Computation uses two single-threaded CPU workers, chunked transcript storage, and a 2 GiB resident-memory limit per worker.

Deployment value at finite calibration budgets. The mixture-betting rule gains 0.0253 $[0.0168, 0.0356]$ per step over no probing on the independent confirmation sample. Its contrast with the same-budget paired plug-in rule is -0.0001 $[-0.0008, 0.0004]$; the observed mean gain is 99.6% of the plug-in gain. This ratio is descriptive, not an equivalence test. The condition-indexed posterior comparison is 0.0053 $[0.0039, 0.0067]$. Table 7 also retains the original plans. Both original-versus-recalibrated condition rules select the same counts at the primary budget, while their lower-budget choices can differ.

The reliable rule purchases evidence in 16 of the 54 model–condition cells. Their observed mean net gains range from 0.0186 to 0.1884. All high-price and shortest-horizon cells select no probing. At the primary budget, the reliable and plug-in rules retain identical gains in the free evidence cells; the aggregate difference comes from marginal paid acquisitions. Thus the confidence rule retains useful response policies while testing the value of their evidence before incurring its cost.

At 1,024 calibration units, betting improves over the empirical-Bernstein rule by 0.0132 $[0.0106, 0.0160]$ per step, with the same programs and calibration sample. Its mean gain is 0.0243, compared with 0.0111 for Bernstein. At 64 units both confidence rules choose no probing; at 256 units betting begins to purchase evidence (Table 8). Figure 7 shows this sample-size dependence and the pre-test margins. These results assess the operational cost of finite-sample reliability, not recovery of an unrestricted optimal policy or an exact neural representation frontier.

Separate empirical-Bernstein cohort. A disjoint 4,096-goal calibration and 2,048-goal evaluation cohort uses (47) alone. Its gain over no probing is 0.0249 $[0.0177, 0.0347]$; the same-budget

Table 7: Independent acquisition confirmation. Gains are per control step, averaged over all nine horizon–price conditions. All rules share six frozen models and 2,048 evaluation goals. The two original plans use their 20-goal calibration; the other fitted plans use 4,096 independent calibration units. Intervals resample training runs and tasks.

Rule	Gain over no probing	Betting minus rule [95% CI]
No probing	0.0000	0.0253 [0.0168, 0.0356]
Original AP-GFAR plan	0.0247	0.0006 [−0.0001, 0.0016]
Original condition rule	0.0200	0.0053 [0.0039, 0.0067]
Recalibrated condition rule	0.0200	0.0053 [0.0039, 0.0067]
Paired plug-in	0.0254	−0.0001 [−0.0008, 0.0004]
Empirical Bernstein	0.0243	0.0010 [0.0000, 0.0024]
Mixture betting	0.0253	—

Table 8: Calibration-budget dependence on the same independent confirmation sample. The three selectors use identical calibration prefixes and frozen terminal heads. Means include no-probe ties; all purchased branches include their charges.

Calibration N	Plug-in gain	Bernstein gain	Betting gain	Betting probe rate
64	0.0248	0.0000	0.0000	0.0%
256	0.0250	0.0000	0.0035	1.9%
1,024	0.0255	0.0111	0.0243	25.9%
4,096	0.0254	0.0243	0.0253	29.6%

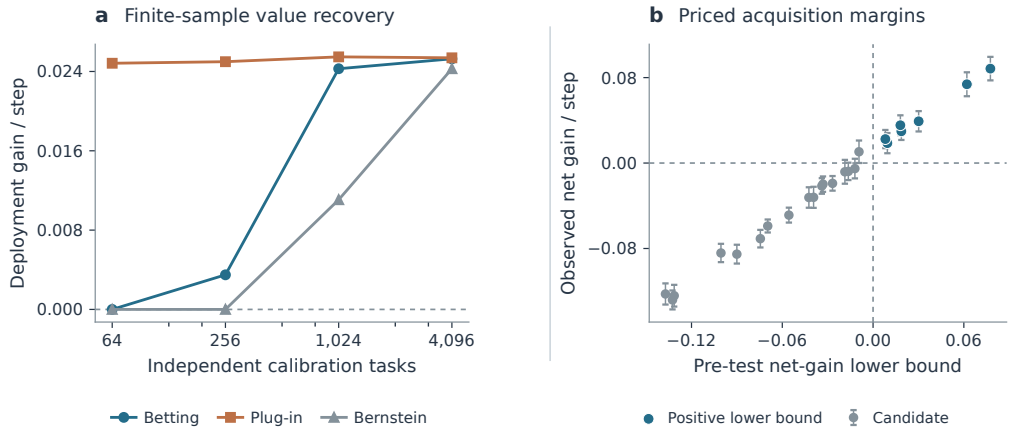

Figure 7: **Reliable acquisition retains deployment value.** (a) Mean net gain over no probing across all nine conditions and six frozen models; the three rules share each calibration prefix and terminal head. (b) At price 0.25, each point is one model–horizon–positive-count candidate. The horizontal coordinate is its 4,096-sample, pre-test net-gain lower bound; the vertical coordinate is its independent evaluation mean, with pointwise normal-approximation 95% task intervals. Blue candidates have positive lower bounds. Acquisition maximizes the feasible lower bound, including no probing at zero.

plug-in contrast is -0.0022 $[-0.0039, -0.0006]$. Free-evidence decisions coincide at the primary budget. The lower return arises from conservative paid acquisitions, demonstrating the opportunity cost of a valid but wider interval. The artifact retains both cohorts' complete branches, transcripts, frozen allocation tables, and all sample-size comparisons.

H ESTABLISHED-METHOD COMPLETE-PROGRAM DEPLOYMENT ON VARIBAD

The complete-program principle is also tested on an established adaptation architecture rather than an AP-GFAR-trained representation. We freeze one 500,000-transition PPO/GRU/VAE VariBAD checkpoint (Zintgraf et al., 2020) in HalfCheetah-v5 and leave its learned context system unchanged. AP-GFAR is used only as a pre-return deployment selector over a declared VariBAD program library: native online deployment with no purchased prefix, or frozen-context deployment after 0, 8, or 32 steps of the same shared nominal physical probe. No VariBAD parameter is retrained by this comparison.

The native parameter is gravity multiplied by $\exp(\theta)$, with magnitudes $0.15, 0.225, 0.30$ and balanced hidden signs. The episode horizon is 1000 and the per-probe-step prices are $0, 1, 4, 16, 24$. Calibration uses only magnitudes 0.15 and 0.30; the value at 0.225 is fixed by linear interpolation. For each magnitude and price, the selector maximizes the calibration mean of the complete native program minus the known price times its physical prefix length. It sees magnitude and price but not task sign, simulator state, true test gravity, or evaluation return. The selected program is written before the fresh evaluation resets.

Evaluation uses 32 paired reset blocks, balanced signs, and the full declared magnitude–price grid. All native rewards, including prefix rewards, are counted once inside the same 1000-step horizon; external cost is then subtracted as price times the declared prefix length. The primary contrast averages equally over the fixed grid within each reset block and compares the selector with VariBAD's native online no-prefix deployment. Independent recomputation from the raw evaluation JSONL gives

$$\Delta_{\text{selector}-\text{native}} = 73.74\ [20.42, 127.06], \qquad p_{\text{one-sided}} = 0.0041,$$

using a paired Student interval and test over the 32 reset-block vectors. The same aggregation gives $115.31\ [35.33, 195.30]$ relative to always taking the 32-step online prefix. The selector chooses the 32-step frozen-context program when low price makes its calibration value worthwhile and otherwise falls back to native online no-prefix deployment; no evaluation return enters this choice.

This experiment is an external deployment test conditional on one fixed trained VariBAD checkpoint. It is not a reproduction of the full official VariBAD benchmark and does not compare training algorithms. Its purpose is narrower: the pair-valued complete-program selector improves the deployment of an established adaptation system under a fixed native protocol without changing that system's representation learner.

I REPRESENTATION DEBT AND PROSPECTIVE-SUFFICIENCY COUNTEREXAMPLES

Theorem 17 (Exact arbitrarily small reward-shift separation)**.** *Let $X, U \overset{\text{iid}}{\sim} \text{Unif}[-1, 1]$, $a \in \{-1, +1\}$, $r_\delta((X, U), a) = a(X + \delta U)$, and $\phi(X, U) = X$. Then ϕ is exactly $Q_0^\star$-sufficient and, for every $|\delta| \le 1$,*

$$\alpha_\phi(\delta) = \delta^2/6. \tag{52}$$

The same per-round debt holds for every history-dependent encoded learner in the repeated fresh-context construction.

Proposition 18 (Exact transition-only separation)**.** *There is a two-stage MDP with parameter-independent initial law and terminal reward, and with only the transition kernel depending smoothly on δ, such that the same encoder is nominally exact and $\alpha_\phi(\delta) = \delta^2/6$ for $|\delta| \le 1$. The transition family is differentiable in total variation and satisfies a uniform $O(|\delta|)$ total-variation perturbation bound.*

I.1 A STRICT PROSPECTIVE-SUFFICIENCY HIERARCHY

For a declared neighborhood $\mathcal{N}$ of zero, call ϕ:

(i) *raw-information sufficient* if a parameter-independent measurable map reconstructs the raw history from the encoded experiment almost surely throughout the declared family;

(ii) *neighborhood prospectively realizable* if $\alpha_\phi(\theta) = 0$ for every $\theta \in \mathcal{N}$;

(iii) *first-order prospectively sufficient* if it is nominally exact and $\sup_{\|u\|=1} \alpha_\phi(tu) = o(t^2)$ as $t \to 0$;

(iv) *nominally control-sufficient* if $\alpha_\phi(0) = 0$.

In the smooth setting of Theorem 2, first-order prospective sufficiency is equivalent to $\Gamma_\phi = 0$.

Theorem 19 (Prospective-sufficiency hierarchy and strictness)**.** *On any declared family,*

$$\begin{aligned} \textit{raw-information sufficiency} &\Longrightarrow \textit{neighborhood prospective realizability} \\ &\Longrightarrow \textit{first-order prospective sufficiency} \\ &\Longrightarrow \textit{nominal control sufficiency}. \end{aligned}$$

Each implication is strict, even for one-step binary decisions with arbitrarily small smooth shifts.

Proof. A parameter-independent raw-history reconstruction lets every raw policy be executed from the encoded experiment, preserving its coupling to the actual decision state. Raw and encoded optimal values therefore agree throughout the family. Neighborhood realizability implies nominal equality and zero debt along every local path, hence $o(t^2)$. First-order prospective sufficiency includes nominal exactness by definition.

For strictness of the first implication, let $W = (X, Z)$ have independent uniform coordinates on $[-1, 1]$, set $r_\theta(W, a) = a(X + \theta)$, and retain $C = X$. The code determines the optimal action for every θ, but cannot reconstruct the independent coordinate Z.

Strictness of the last implication is Theorem 17: nominal exactness holds but $\alpha_\phi(t) = t^2/6$. For strictness of the preceding implication, retain the same X, U, a, ϕ but set $r_\delta((X, U), a) = a(X + \delta^2 U)$. Applying Theorem 17 with shift magnitude δ^2 gives

$$\alpha_\phi(\delta) = \delta^4/6.$$

Thus $\alpha_\phi(t) = o(t^2)$ and the first-order tangent is fully representable, yet every nonzero sufficiently small t has positive debt, so neighborhood realizability fails. □

I.2 Debt decomposition

For every episode or decision index t,

$$V^\star_{\theta_t} - V^{\pi_t}_{\theta_t} = \big(V^\star_{\theta_t} - V^\star_{\theta_t,\phi}\big) + \big(V^\star_{\theta_t,\phi} - V^{\pi_t}_{\theta_t}\big).$$

Summing proves (1). The first term depends only on the environment and the encoded policy class. Revealing θ_t, the complete model, or an exact optimizer cannot reduce it.

I.3 Proof of Theorem 17

At $\delta = 0$, $Q^\star_0((x, u), a) = ax$, so $\phi(x, u) = x$ is exactly $Q^\star$-sufficient. For fixed $c \in [-1, 1]$,

$$\frac{1}{2}\int_{-1}^{1} |x + c|\, dx = \frac{1}{4}\big((1 + c)^2 + (1 - c)^2\big) = \frac{1}{2} + \frac{c^2}{2}.$$

Since $|\delta U| \le 1$,

$$V^\star_\delta = \mathbb{E}|X + \delta U| = \frac{1}{2} + \frac{\delta^2}{2}\mathbb{E}U^2 = \frac{1}{2} + \frac{\delta^2}{6}.$$

For every randomized X-measurable action kernel, write $m(X) = \mathbb{E}[A \mid X] \in [-1, 1]$. Independence and centering of U give

$$\mathbb{E}[A(X + \delta U)] = \mathbb{E}[m(X)X] \le \mathbb{E}|X| = \frac{1}{2}.$$

Equality is attained by $A = \operatorname{sign}(X)$, hence $V^\star_{\delta,\phi} = 1/2$ and (52) follows.

For the repeated-decision statement, condition on the entire encoded past and the current X_t. The fresh U_t remains independent and centered. Thus every history-dependent encoded action has conditional expected reward at most $|X_t|$, while the raw action obtains $|X_t + \delta U_t|$. Expectations add over fresh rounds. The construction embeds as the terminal decision of any longer finite-horizon MDP by adding a parameter-independent zero-reward prefix.

I.4 Proof of Proposition 18

At stage one draw (X, U) as in Theorem 17, reveal it to the controller, and choose $a \in \{-1, +1\}$. The only parameter-dependent object is the transition

$$Y = a(X + \delta U) + \varepsilon,$$

where ε is independent, centered, and has a compactly supported C^1 density p_ε. There is no second decision and the fixed terminal reward is $r_2(Y) = Y$. The initial law and both reward functions are therefore parameter independent. Moreover,

$$Q^\star_{\delta,1}((x, u), a) = \mathbb{E}[Y \mid x, u, a] = a(x + \delta u).$$

Consequently the value calculation is identical to Theorem 17, proving exact debt $\delta^2/6$ while changing only the transition law.

For fixed (x, u, a), the transition density is $p_\delta(y \mid x, u, a) = p_\varepsilon(y - a(x+\delta u))$. Because a compactly supported C^1 density has $p'_\varepsilon \in L^1$, translations are differentiable in L^1:

$$\left\| \frac{p_\varepsilon(\cdot - a(x + \delta u)) - p_\varepsilon(\cdot - ax)}{\delta} + au\, p'_\varepsilon(\cdot - ax) \right\|_1 \longrightarrow 0.$$

Moreover, the fundamental theorem of calculus and translation invariance of the L^1 norm yield

$$d_{\mathrm{TV}}(P_\delta, P_0) \le \frac{|\delta u|}{2} \|p'_\varepsilon\|_1 \le \frac{|\delta|}{2} \|p'_\varepsilon\|_1.$$

This proves the claimed uniform $O(|\delta|)$ perturbation.