OFFLINE LEARNING TO DIAGNOSE AND INTERVENE IN TOOL-USING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool-using agents act on stateful environments, where erroneous tool actions, such as API calls or file operations, can leave persistent effects that are not directly visible in later interactions. This creates a sequential safety problem: an external harness must track possible hidden damage, decide when costly diagnosis is worthwhile, and choose how to intervene while accounting for intervention cost. We study how to learn such a harness from fixed interaction logs in which damage labels are revealed only selectively. We propose Revealing-Action Pessimistic Value Iteration (RA-PEVI), which learns intervention dynamics shared across diagnostic branches from selectively revealed latent states and performs pessimistic planning in the belief space. We establish a finite-sample performance guarantee, derive the corresponding $\widetilde{O}(N^{-1/2})$ rate, and prove matching lower bounds and identifiability limits. Experiments on synthetic data and AppWorld support our theoretical results and show that RA-PEVI achieves better performance than standard baselines.

1 INTRODUCTION

Tool-using agents act on stateful environments through tools such as API calls that update databases, modify files, or trigger external workflows. Their mistakes can leave damage that is persistent yet only partially observable. For example, an agent may update the wrong database record without any explicit signal that the environment has become inconsistent; later tool calls may then be locally correct while still operating on the damaged state. Related risks from tool use and environment interaction have been documented in ToolEmu and AgentDojo (Ruan et al., 2024; Debenedetti et al., 2024). As a result, safety is inherently sequential: the current tool action alone may not reveal whether earlier interactions have already left the environment in a harmful state.

Handling hidden damage requires both gathering information and acting on it. Diagnostics can reduce uncertainty, but acquiring additional measurements, running audits, or involving human oversight may incur nontrivial cost (Yin et al., 2020; Krale et al., 2023). Interventions such as blocking, repair, or rollback can likewise consume resources, add execution overhead, or disrupt task progress. The two decisions are therefore inherently coupled: whether a diagnostic is worthwhile depends on how its outcome changes the subsequent intervention.

Meanwhile, tool-using agents naturally generate interaction trajectories as they operate, providing a source of data from which a harness can be learned. These logs, however, typically record observable actions and outcomes rather than the full state. Persistent damage may remain hidden for long stretches and be revealed only at selected steps, so only a subset of transitions is labeled with the underlying damage state. Because damage carries over across steps, this missing information affects not only the current decision but also how subsequent observations should be interpreted. The central challenge is therefore to learn the latent damage dynamics from selectively revealed trajectories and determine which diagnose-then-intervene decisions are sufficiently supported by the offline data. This leads to our central question: *Can we learn from fixed interaction logs a cost-sensitive harness that decides when to diagnose and how to intervene under persistent hidden damage?*

We formulate this problem as offline reinforcement learning in a finite-horizon partially observable Markov decision process (POMDP). We treat the base agent as fixed and learn only the external harness from a dataset of logged tool-use trajectories, without collecting additional interaction data.

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.

During deployment, the harness maintains a belief over the hidden damage state. At each step, it first chooses a diagnostic and then selects an intervention conditioned on the diagnostic outcome, minimizing terminal task loss together with diagnostic and intervention costs. This setting differs from standard offline RL because the control-relevant state is not fully observed in the logged trajectories (Jin et al., 2021; Uehara & Sun, 2021). The learner must therefore estimate latent intervention dynamics from selectively revealed damage states and plan over the resulting belief process.

To address this challenge, we propose *Revealing-Action Pessimistic Value Iteration* (RA-PEVI), a structure-aware offline RL method for selectively revealed latent damage. RA-PEVI exploits the shared-transition structure of our model: conditional on the latent damage state, the dynamics of an intervention are shared across diagnostic choices and outcomes. It therefore pools revealed evidence across diagnostic branches to estimate these latent intervention dynamics, constructs confidence sets for the shared transition components, and converts their uncertainty into belief-dependent pessimism. RA-PEVI then plans over the two-stage diagnose-then-intervene decision in belief space. This allows evidence collected under one diagnostic to support the same intervention under another, while concentrating pessimism on latent states that are plausible under the current belief.

We provide a finite-sample analysis of RA-PEVI. We show that transition confidence sets can be constructed under selective damage reveals, yielding a comparator-dependent suboptimality guarantee and an $\widetilde{O}(\sqrt{1/N})$ rate governed by coverage of value-relevant latent dynamics, which also matches the lower bound. We also conduct extensive experiments on both synthetic data and an offline dataset collected from AppWorld. The results show that RA-PEVI can outperform standard baselines and support our theoretical findings.

Overall, our main contributions can be summarized as:

- We formulate offline harness learning for tool-using agents with persistent latent damage, selective state reveals, and a diagnose-then-intervene decision structure.
- We develop RA-PEVI, which learns shared latent intervention dynamics from selectively labeled trajectories and performs belief-dependent pessimistic planning.
- We establish finite-sample performance guarantees, matching lower and identification bounds, and empirically validate the method on synthetic workflows and AppWorld data.

2 Related Work

Safety and control for tool-using agents. As language-model agents increasingly interact with stateful environments through external tools, their actions can directly affect persistent system states. This creates safety risks that extend beyond isolated model outputs and have been documented in ToolEmu and AgentDojo (Ruan et al., 2024; Debenedetti et al., 2024). Existing work therefore develops external control mechanisms: shielding constrains unsafe actions, including under partial observability (Alshiekh et al., 2018; Carr et al., 2023), while recent systems use trusted monitors, guard agents, or runtime specifications to mediate execution (Greenblatt et al., 2023; Xiang et al., 2025; Wang et al., 2026). We share this external-control perspective, but focus on learning the controller from fixed interaction logs when safety-relevant damage is latent and can persist across multiple tool calls.

Offline learning under partial observability. Offline RL uses pessimism to address distribution shift and limited coverage (Jin et al., 2021; Xie et al., 2021; Rashidinejad et al., 2021; Uehara & Sun, 2021), while partially observed settings require history-dependent reasoning (Ghosh et al., 2022; Guo et al., 2022; Simao et al., 2023). Related work also uses privileged or hindsight state information during training (Warrington et al., 2021; Lee et al., 2023; Cai et al., 2024). Our setting reveals latent damage only selectively, requiring uncertainty-aware learning of latent intervention dynamics from partially labeled offline data.

Selective and costly information. Missing-data and selective-label models study selective outcome observation (Rubin, 1976; Wei, 2021), while active feature acquisition and measurement study costly information gathering (Yin et al., 2020; Li & Oliva, 2021; Krale et al., 2023). Related work also considers help seeking, deferral, and resource constraints (Ren et al., 2023; Ghoummaid & Shalit, 2024; Lin et al., 2023). Our diagnose-then-intervene structure is related, but the controller is learned entirely from fixed logs with selectively revealed latent states.

3 PROBLEM FORMULATION

3.1 TOOL-USING AGENTS WITH PERSISTENT HIDDEN DAMAGE

We model the interaction among a frozen tool-using agent, an external harness, and a stateful environment as a finite-horizon partially observable Markov decision process (POMDP) with at most H decision steps. At step h, the agent proposes a concrete tool action for the environment, such as an API call that updates a database or a shell command that modifies a file.

The underlying state is $\boldsymbol{s}_h = (d_h, \boldsymbol{x}_h) \in \mathcal{S} := \mathcal{D} \times \mathcal{X}$, where $\boldsymbol{x}_h \in \mathcal{X}$ is observed by the harness and $d_h \in \mathcal{D}$ is hidden. The observed component $\boldsymbol{x}_h$ collects the current tool proposal together with other transition-relevant information available to the harness, such as the workflow stage, tool arguments, and observable environment state. The latent state d_h summarizes unresolved damage carried over from previous interactions, for example when an earlier tool action leaves the environment in an incorrect or inconsistent state. We take $\mathcal{X}$ and $\mathcal{D}$ to be finite.

Damage can persist across decision steps: even when the current proposal is itself safe, the environment may still be affected by an earlier interaction. Persistence does not imply irreversibility; subsequent interventions may repair or remove the damage. We assume $d_1 = d_{\text{clean}}$ almost surely. Because d_h is hidden, the harness maintains a belief $\boldsymbol{b}_h \in \Delta(\mathcal{D})$, where $\boldsymbol{b}_h(d)$ is the posterior probability of $d_h = d$ given the observable history up to step h.

3.2 DIAGNOSTIC AND INTERVENTION DECISIONS

At each step h, the harness first chooses a diagnostic $m_h \in \mathcal{M}$, which is read-only and produces $y_h \in \mathcal{Y}_{m_h}$ according to $y_h \sim O_{h,m_h}(\cdot \mid d_h, \boldsymbol{x}_h)$. Diagnostics range from noisy lightweight checks to a full audit m^{rev} that reveals the current damage state; in practice, such an audit may correspond to a trusted procedure with privileged access to the relevant environment state. A null diagnostic corresponds to skipping diagnosis.

Let $\mathcal{H}_h^-$ denote the observed history before diagnosis and $\boldsymbol{b}_h$ the corresponding belief over d_h. After observing $(m_h, y_h) = (m, y)$, the harness updates its belief to $\bar{\boldsymbol{b}}_h(d) \propto O_{h,m}(y \mid d, \boldsymbol{x}_h)\boldsymbol{b}_h(d)$ and chooses an intervention i_h from a finite set $\mathcal{I}$. Interventions may execute the proposal as given, block it, repair its effects, or roll the environment back. Thus, a null diagnostic followed by executing the proposal corresponds to taking no additional harness action.

Diagnostics and interventions incur known costs $c_{\text{M}}(\boldsymbol{x}_h, m_h)$ and $c_{\text{I}}(\boldsymbol{x}_h, m_h, y_h, i_h)$, respectively; either may be zero. A policy π consists of a diagnostic policy $\pi_h^{\text{M}}(\cdot \mid \mathcal{H}_h^-)$ and an intervention policy $\pi_h^{\text{I}}(\cdot \mid \mathcal{H}_h^-, m_h, y_h)$. We assume that diagnostics are read-only: conditional on the current state and the fully specified intervention, the subsequent environment transition does not depend on which diagnostic produced the information used to select that intervention. Hence,

$$(d_{h+1}, \boldsymbol{x}_{h+1}, e_h, z_h) \sim T^\star_{h,i_h}\big(\cdot \mid d_h, \boldsymbol{x}_h\big). \tag{1}$$

Here $e_h \in \{0, 1\}$ indicates termination and $z_h \in \mathcal{Z}$ is the terminal outcome when $e_h = 1$. For $e_h = 0$, we set $z_h = \perp$; after termination, the process remains in a fixed absorbing state $(d_\dagger, \boldsymbol{x}_\dagger)$. The episode ends at $\tau = \min\{h : e_h = 1\} \le H$. Appendix H relaxes this exact transition sharing to branch-dependent dynamics.

3.3 DEPLOYMENT OBJECTIVE

We assume that task-level loss is assessed only when the episode terminates; there is no intermediate task reward. Let $\ell : \mathcal{Z} \to [0, L_{\max}]$ denote the terminal loss. Together with the diagnostic and intervention costs, the expected deployment cost of a policy π is

$$J_{T^\star}(\pi) = \mathbb{E}_{T^\star,\pi}\Big[\ell(z_\tau) + \sum_{h=1}^{\tau}\Big(c_{\text{M}}(\boldsymbol{x}_h, m_h) + c_{\text{I}}(\boldsymbol{x}_h, m_h, y_h, i_h)\Big)\Big]. \tag{2}$$

Thus, the harness trades off task performance against the cost of obtaining information and intervening. We use additive costs in the main text; Appendix I extends the formulation to pathwise finite budgets. Let Π denote the class of admissible history-dependent harness policies, and define

$J^\star_{\mathrm{opt}} := \min_{\pi\in\Pi} J_{T^\star}(\pi)$ and $\Pi^\star := \arg\min_{\pi\in\Pi} J_{T^\star}(\pi)$. We evaluate a learned policy $\widehat{\pi}$ by its suboptimality,

$$\mathrm{SubOpt}(\widehat{\pi}) := J_{T^\star}(\widehat{\pi}) - J^\star_{\mathrm{opt}}.$$

3.4 Offline Data with Selective Damage Reveals

We now specify the offline data available for learning a policy that minimizes the deployment objective above. The learner receives N logged trajectories generated by behavior harness policies with the same diagnose-then-intervene information structure as the deployment policy. The behavior policies need not be known and may differ across trajectories.

The latent damage state is revealed only at selected steps. Let $\delta_h^{\mathrm{obs}} := \mathbf{1}\{d_h \text{ is revealed in the offline log}\}$, and define $\widetilde{d}_h = d_h$ when $\delta_h^{\mathrm{obs}} = 1$ and $\widetilde{d}_h = \text{missing}$ otherwise. Since the initial state $d_1 = d_{\mathrm{clean}}$ is known, the source damage state at the first step is always available. A full audit m^{rev} also reveals the current d_h, while additional labels may be provided by the data-collection process. The logged dataset is

$$\mathcal{D}_N = \left\{(\boldsymbol{x}_h, m_h, y_h, i_h, \boldsymbol{x}_{h+1}, e_h, z_h, \delta_h^{\mathrm{obs}}, \widetilde{d}_h)_{h=1}^{\tau_n}\right\}_{n=1}^{N}.$$

The learner uses these partially labeled trajectories to estimate the latent intervention dynamics in equation (1) and learn a harness with small $\mathrm{SubOpt}(\widehat{\pi})$.

4 Algorithm: Revealing-Action Pessimistic Value Iteration

Directly applying pessimistic offline RL at the level of observable diagnostic branches fails to exploit the structure of our problem. Such a method would estimate separate dynamics for each (m, y, i) branch and treat an unseen branch as unsupported, even when the same intervention is well covered under other diagnostics. In contrast, conditional on the latent damage state, a fully specified intervention shares the same dynamics across diagnostic branches.

We exploit this structure with *Revealing-Action Pessimistic Value Iteration* (RA-PEVI). Its key idea is to move both estimation and uncertainty from observable branches to shared latent intervention rows $r = (h, d, \boldsymbol{x}, i)$. RA-PEVI pools selectively revealed evidence across diagnostics to estimate these rows, constructs row-wise confidence radii, and aggregates their uncertainty according to the current posterior belief before performing pessimistic value iteration. As a result, evidence collected under one diagnostic can support decisions under another, while pessimism is focused on latent dynamics that are plausible under the current belief.

4.1 Shared Latent Transition Estimation

Let $u_h = (\boldsymbol{x}_{h+1}, e_h, z_h) \in \mathcal{U}$ denote the observable post-intervention outcome, where $\mathcal{U}$ is finite, and let $\mathcal{U}_0 \subseteq \mathcal{U}$ be its nonterminal subset. We factor the transition into an observable outcome model and a latent-state continuation model. For $u \in \mathcal{U}_0$,

$$T^\star_{h,i}(d', u \mid d, \boldsymbol{x}) = Q^\star_{h,i}(u \mid d, \boldsymbol{x}) K^\star_{h,i}(d' \mid d, \boldsymbol{x}, u). \tag{3}$$

This factorization matches the selective-reveal structure of the data: estimating $Q^\star$ requires only the source damage state d_h to be revealed, since u_h is always observed, whereas estimating $K^\star$ additionally requires the next damage state d_{h+1}.

For a latent row $r = (h, d, \boldsymbol{x}, i)$, let $n_Q(r)$ be the number of logged transitions with a revealed source state and $N_Q(r, u)$ the number among them with outcome u. We estimate $\widehat{Q}_{h,i}(u \mid d, \boldsymbol{x}) = N_Q(r, u)/n_Q(r)$. For $u \in \mathcal{U}_0$, let $n_K(r, u)$ count transitions with both source and next-state labels and $N_K(r, u, d')$ those additionally satisfying $d_{h+1} = d'$. We estimate $\widehat{K}_{h,i}(d' \mid d, \boldsymbol{x}, u) = N_K(r, u, d')/n_K(r, u)$. If a denominator is zero, the corresponding empirical distribution is set to an arbitrary fixed distribution and assigned confidence radius one.

When $Q^\star_{h,i}(u \mid d, \boldsymbol{x}) = 0$, the conditional kernel $K^\star_{h,i}(\cdot \mid d, \boldsymbol{x}, u)$ is immaterial to the joint transition law; we fix it arbitrarily on $\mathcal{D}$. For nonterminal outcomes, $\widehat{T}_{h,i}(d', u \mid d, \boldsymbol{x}) = \widehat{Q}_{h,i}(u \mid$

$d, \boldsymbol{x})\widehat{K}_{h,i}(d' \mid d, \boldsymbol{x}, u)$. For terminal u, we set $\widehat{T}_{h,i}(d_\dagger, u \mid d, \boldsymbol{x}) = \widehat{Q}_{h,i}(u \mid d, \boldsymbol{x})$ and assign zero mass to other next damage states.

Crucially, all counts are pooled across diagnostic choices and outputs. RA-PEVI therefore estimates the shared latent intervention dynamics rather than a separate model for each (m, y, i). This allows an intervention to be learned from data collected under different diagnostics. Appendix E formalizes this: an observable branch can remain identifiable without direct support when its latent rows are covered, whereas such transfer is impossible in general without the shared transition structure.

4.2 Pessimism over Latent Transition Rows

We next quantify uncertainty in the shared latent transition estimates. Let $|\mathcal{R}| = H|\mathcal{D}||\mathcal{X}||\mathcal{I}|$. For confidence level $\delta \in (0, 1)$, define

$$\Lambda_Q = |\mathcal{U}|\log(2) + \log(4N|\mathcal{R}|/\delta), \qquad \Lambda_K = |\mathcal{D}|\log(2) + \log(4N|\mathcal{R}||\mathcal{U}_0|/\delta),$$
$$\beta_Q(r) = \min\{1, \sqrt{\Lambda_Q/2n_Q(r)}\}, \qquad \beta_K(r, u) = \min\{1, \sqrt{\Lambda_K/2n_K(r, u)}\}.$$

Division by zero is interpreted as $+\infty$, so an unsupported component receives radius one. Combining uncertainty in the observable outcome model and latent continuation model gives the transition-row radius

$$\beta_h(d, \boldsymbol{x}, i) = \min\Big\{1,\ \beta_Q(r) + \sum_{u \in \mathcal{U}_0} \widehat{Q}_{h,i}(u \mid d, \boldsymbol{x})\beta_K(r, u)\Big\}. \tag{4}$$

The radii above are defined for latent rows, whereas the harness acts from a belief over the latent state. For a post-diagnostic belief $\bar{\boldsymbol{b}}$, RA-PEVI therefore uses the belief-weighted penalty

$$\Gamma_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, i) = W_{h+1} \sum_{d \in \mathcal{D}} \bar{\boldsymbol{b}}(d)\beta_h(d, \boldsymbol{x}, i), \tag{5}$$

where W_{h+1} upper bounds the remaining cost after the current intervention. Thus, uncertainty is weighted by how plausible each latent damage state is under the current posterior. An observable branch can therefore receive a small penalty even if it was never observed directly, provided that the latent rows it uses are well covered elsewhere in the data.

4.3 Pessimistic Diagnose-Then-Intervene Planning

Given the shared transition estimates and belief-dependent penalties, RA-PEVI performs pessimistic dynamic programming in belief space. For a post-diagnostic belief $\bar{\boldsymbol{b}}$, define $\widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}}, \boldsymbol{x}) = \sum_d \bar{\boldsymbol{b}}(d)\widehat{Q}_{h,i}(u \mid d, \boldsymbol{x})$. For $u = (\boldsymbol{x}', 0, \perp) \in \mathcal{U}_0$, the next belief is

$$\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}}, \boldsymbol{x}, u)(d') = \frac{\sum_d \bar{\boldsymbol{b}}(d)\widehat{Q}_{h,i}(u \mid d, \boldsymbol{x})\widehat{K}_{h,i}(d' \mid d, \boldsymbol{x}, u)}{\widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}}, \boldsymbol{x})}.$$

Fix an arbitrary fallback belief $\boldsymbol{b}_{\mathrm{fb}} \in \Delta(\mathcal{D})$ and set $\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}}, \boldsymbol{x}, u) = \boldsymbol{b}_{\mathrm{fb}}$ whenever $\widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}}, \boldsymbol{x}) = 0$. The choice is immaterial to the nominal Bellman backup because such an outcome has zero probability.

Starting from $\widehat{V}_{H+1} = 0$, let $\widehat{G}_{h+1}(u) = \ell(z)$ for $e = 1$, and $\widehat{G}_{h+1}(u) = \widehat{V}_{h+1}\big(\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}}, \boldsymbol{x}, u), \boldsymbol{x}'\big)$ for $e = 0$, with the current $(\bar{\boldsymbol{b}}, \boldsymbol{x}, i)$ understood. The intervention-stage backup is

$$\widehat{Q}^{\mathrm{I}}_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, m, y, i) = c_{\mathrm{I}}(\boldsymbol{x}, m, y, i) + \Gamma_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, i) + \sum_u \widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}}, \boldsymbol{x})\widehat{G}_{h+1}(u). \tag{6}$$

Before observing the diagnostic output, $p_{h,m}(y \mid \boldsymbol{b}, \boldsymbol{x}) = \sum_d \boldsymbol{b}(d)O_{h,m}(y \mid d, \boldsymbol{x})$. Using the corresponding posterior $\bar{\boldsymbol{b}}^{m,y}$, the diagnostic-stage backup is

$$\widehat{Q}^{\mathrm{M}}_h(\boldsymbol{b}, \boldsymbol{x}, m) = c_{\mathrm{M}}(\boldsymbol{x}, m) + \sum_{y \in \mathcal{Y}_m} p_{h,m}(y \mid \boldsymbol{b}, \boldsymbol{x}) \min_{i \in \mathcal{I}} \widehat{Q}^{\mathrm{I}}_h\big(\bar{\boldsymbol{b}}^{m,y}, \boldsymbol{x}, m, y, i\big). \tag{7}$$

Finally, $\widehat{V}_h(\boldsymbol{b}, \boldsymbol{x}) = \min_{m \in \mathcal{M}} \widehat{Q}^{\mathrm{M}}_h(\boldsymbol{b}, \boldsymbol{x}, m)$, and the minimizing actions define $\widehat{\pi}^{\mathrm{M}}_h$ and $\widehat{\pi}^{\mathrm{I}}_h$.

Algorithm 1 summarizes the full procedure. Although $\Delta(\mathcal{D})$ is continuous, the penalized finite-horizon POMDP retains the standard alpha-vector representation (Smallwood & Sondik, 1973; Kaelbling et al., 1998). RA-PEVI can therefore use exact alpha-vector dynamic programming in finite problems, while approximate belief-space planning can be used at larger scale.

Algorithm 1 Revealing-Action Pessimistic Value Iteration (RA-PEVI)

Require: Offline data $\mathcal{D}_N$, diagnostic kernels O, costs $c_{\mathrm{M}}, c_{\mathrm{I}}$, confidence level δ

for each latent row $r = (h, d, \boldsymbol{x}, i)$ **do**
  Estimate $\widehat{Q}_{h,i}$ from source-revealed transitions and $\widehat{K}_{h,i}$ from source–destination-revealed transitions.
  Form $\widehat{T}_{h,i}$ and construct $\beta_h(d, \boldsymbol{x}, i)$ via equation (4).
end for
Construct the nominal belief filter $\widehat{\Phi}$ from $\widehat{T}$ and O.
Initialize $\widehat{V}_{H+1} = 0$.
for $h = H, \ldots, 1$ **do**
  Define the belief-dependent penalty Γ_h via equation (5).
  Compute intervention values $\widehat{Q}^{\mathrm{I}}_h$ via equation (6).
  Compute diagnostic values $\widehat{Q}^{\mathrm{M}}_h$ via equation (7) and set $\widehat{V}_h(\boldsymbol{b}, \boldsymbol{x}) = \min_m \widehat{Q}^{\mathrm{M}}_h(\boldsymbol{b}, \boldsymbol{x}, m)$.
end for
return $\widehat{\pi}^{\mathrm{M}}$, $\widehat{\pi}^{\mathrm{I}}$, and the nominal belief filter $\widehat{\Phi}$.

5 Theoretical Analysis

The analysis of RA-PEVI faces two challenges beyond standard tabular offline RL. First, latent damage is revealed only at selected, trajectory-dependent steps, leading to random effective sample sizes. Second, the true and estimated models generally induce different beliefs after the same observable history. We handle these issues by deriving confidence directly from selectively revealed trajectories and comparing the two controlled processes along common observable histories.

Assumption 1 (Known diagnostic model)**.** The kernel $O_{h,m}(\cdot \mid d, \boldsymbol{x})$ of each diagnostic is known. The diagnostic set contains a full audit m^{rev} with $\mathcal{Y}_{m^{\mathrm{rev}}} = \mathcal{D}$ and $O_{h,m^{\mathrm{rev}}}(y \mid d, \boldsymbol{x}) = \mathbf{1}\{y = d\}$.

We treat the diagnostic kernels as known to isolate learning of the intervention dynamics; Appendix G extends the analysis to independently estimated diagnostics.

Assumption 2 (Conditionally representative damage reveals)**.** The logged trajectories use fresh environment randomness across episodes. Let $\mathcal{G}_h$ denote the observable history immediately before the environment transition, including $(\boldsymbol{x}_h, m_h, y_h, i_h)$. Define $R^Q_h := \delta^{\mathrm{obs}}_h$ and, for $e_h = 0$, $R^K_h := \delta^{\mathrm{obs}}_h \delta^{\mathrm{obs}}_{h+1}$. Then $u_h \perp R^Q_h \mid \mathcal{G}_h, d_h$ and $d_{h+1} \perp R^K_h \mid \mathcal{G}_h, d_h, u_h$ for $e_h = 0$.

Assumption 2 allows reveal decisions to depend on the observed trajectory, but not on the unobserved transition outcome being estimated. The fresh-randomness condition rules out cross-episode coordination that could bias the pooled revealed samples; Appendix A.2 formalizes this condition.

We further assume $0 \le c_{\mathrm{M}}(\boldsymbol{x}, m) + c_{\mathrm{I}}(\boldsymbol{x}, m, y, i) \le c_{\max}$ and set $W_{h+1} = L_{\max} + (H - h)c_{\max}$.

Lemma 1. *Under Assumption 2, with probability at least $1 - \delta$, simultaneously for every $r = (h, d, \boldsymbol{x}, i)$,* $\mathrm{TV}\left(T^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\right) \le \beta_h(d, \boldsymbol{x}, i)$. *Moreover, for every $\boldsymbol{b} \in \Delta(\mathcal{D})$,*

$$\mathrm{TV}\left(\sum_d \boldsymbol{b}(d) T^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \sum_d \boldsymbol{b}(d) \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\right) \le \sum_d \boldsymbol{b}(d) \beta_h(d, \boldsymbol{x}, i).$$

Thus, latent-row confidence extends uniformly to the continuous belief space.

For a policy π, define $\mathcal{B}_{\widehat{T}}(\pi) := \mathbb{E}_{\widehat{T},\pi}[\sum_{h=1}^{\tau} \Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)]$. Let ϵ_{plan} denote the planning error.

Theorem 1. *Under Assumptions 1–2, with probability at least $1-\delta$, every $\pi^\circ \in \Pi$ satisfies $J_{T^\star}(\widehat{\pi}) - J_{T^\star}(\pi^\circ) \le 2\mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}}$. Hence, for any $\pi^\star \in \Pi^\star$, $\mathrm{SubOpt}(\widehat{\pi}) \le 2\mathcal{B}_{\widehat{T}}(\pi^\star) + \epsilon_{\mathrm{plan}}$. For exact alpha-vector planning, $\epsilon_{\mathrm{plan}} = 0$.*

The bound depends only on the latent transition uncertainty encountered by the comparator. Thus, RA-PEVI can achieve a small suboptimality gap even when much of the offline data is sparse, provided that the latent dynamics relevant to a good comparator are well covered. In particular, it does not require direct support for every observable (m, y, i) branch.

To obtain an explicit rate, let $\rho^\pi(r)$ be the occupancy of latent row r under $(T^\star, \pi)$, and let $q_Q(r)$ and $q_K(r,u)$ be the offline sampling probabilities of the corresponding source-labeled and source–destination-labeled samples. Writing $Q_r^\star(u) := Q_{h,i}^\star(u \mid d, \boldsymbol{x})$, define

$$C_Q(\pi) := \max_{r:\rho^\pi(r)>0} \frac{\rho^\pi(r)}{q_Q(r)}, \qquad C_K(\pi) := \max_{\rho^\pi(r)Q_r^\star(u)>0} \frac{\rho^\pi(r)Q_r^\star(u)}{q_K(r,u)},$$

and $C_{\mathrm{lat}}(\pi) := \max\{C_Q(\pi), C_K(\pi)\}$.

Corollary 1 (Finite-sample rate)**.** *Suppose the N logged trajectories are independently drawn from a fixed offline data distribution. If some $\pi_{\mathrm{cov}}^\star \in \Pi^\star$ satisfies $C_{\mathrm{lat}}(\pi_{\mathrm{cov}}^\star) < \infty$, then with probability at least $1-\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi}) = \widetilde{O}\left(\sqrt{\frac{C_{\mathrm{lat}}(\pi_{\mathrm{cov}}^\star)}{N}}\right) + \epsilon_{\mathrm{plan}}.$$

Thus, RA-PEVI achieves the $N^{-1/2}$ rate when the value-relevant latent dynamics are covered, without requiring direct support for every observable (m, y, i) branch. Appendix D.4 gives the matching $\Omega(\sqrt{C_{\mathrm{lat}}/N})$ lower bound, and Appendix E gives the zero-coverage impossibility result.

6 Synthetic Experiments

We use synthetic experiments to evaluate RA-PEVI and study the underlying mechanisms.

Setup. Our main environment is a six-step tool-use workflow in which the observable state records task progress and whether the current proposal is safe or risky, while a latent state $d_h \in \{C, R, K\}$ represents persistent damage. At each step, the harness chooses one of three diagnostics followed by Execute, Block, Repair, or Rollback. These choices trade task progress and intervention cost against the risk of persistent damage. Episodes terminate by completion, catastrophic failure, or timeout. More details are introduced in Appendix J.

Baselines. We compare RA-PEVI with five baselines. The *Plug-in* planner uses the same shared latent transition model as RA-PEVI but removes the pessimistic uncertainty penalty. The *Myopic router* reacts only to the current proposal and does not track historical damage. The *observable-branch learner* estimates dynamics directly at the observable diagnostic branch level and therefore cannot transfer transition evidence across branches. *Recurrent-IQL* and *Decision Transformer* receive the full public interaction history but do not use the shared latent transition structure.

Table 1: Suboptimality comparison at $N = 5000$. Lower is better.

Method	Suboptimality ↓
RA-PEVI	0.243 ± 0.073
Plug-in planner	0.254 ± 0.087
Myopic router	5.792 ± 0.000
Observable-branch learner	6.827 ± 0.451
Recurrent-IQL	5.753 ± 0.248
Decision Transformer	9.616 ± 0.587

Suboptimality comparison. On a workflow with $N = 5000$ logged episodes, RA-PEVI achieves the lowest suboptimality among all learned methods. It outperforms the matched plug-in planner, supporting the benefit of pessimistic planning under transition uncertainty. The myopic router performs poorly because it ignores persistent historical damage, while the observable-branch learner suffers from estimating separate dynamics for each diagnostic branch. Recurrent-IQL and Decision Transformer also incur substantially larger gaps, indicating that generic history-based offline learners do not exploit the shared latent structure as effectively.

Finite-sample and persistent-damage behavior. Figure 1(a) varies the number of logged episodes from $N = 250$ to 10,000 in the same workflow. RA-PEVI has its largest advantage over the plug-in planner in the low-data regime, and the gap shrinks as more data become available. Panels (b)–(c) fix $N = 5000$ to study persistent-damage. Panel (b) conditions on states in which the current proposal

is safe but damage from earlier interactions remains unresolved. In these states, RA-PEVI still takes protective actions, whereas the myopic router never does so. Panel (c) shows the corresponding downstream consequence: the myopic router completes substantially more episodes with unresolved damage than RA-PEVI. Together, these results empirically support our analysis that the effect of pessimism decreases as uncertainty shrinks, while persistent historical damage remains important for sequential decision making.

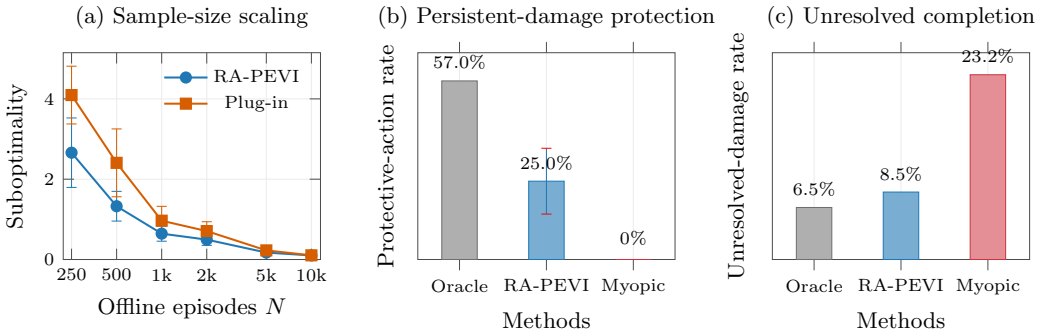

Figure 1: **Finite-sample performance and persistent damage.** (a) Suboptimality of RA-PEVI and the matched plug-in planner as the number of offline episodes increases. (b) Protective-action rate when the current proposal is safe but unresolved damage from earlier interactions remains. (c) Fraction of completed episodes that retain unresolved damage.

Cross-diagnostic transition sharing. A key design choice in RA-PEVI is to share latent intervention dynamics across diagnostic branches, allowing transition evidence collected under one branch to support another. We isolate this effect by comparing RA-PEVI with a Branch-complete variant. Shared PEVI (RA-PEVI in this ablation) and Branch-complete use the same labeled source–destination transitions, prior, and pessimistic planner; Branch-complete instead estimates a separate transition row for each diagnostic-output branch. We fix $N = 2000$ while varying the fraction of available damage labels retained for estimation. As shown in Table 2, sharing consistently reduces policy suboptimality across all reveal rates, demonstrating the value of cross-diagnostic support transfer.

Table 2: **Cross-diagnostic transition sharing.** Suboptimality at $N = 2000$ as the fraction of revealed damage labels varies. Lower is better.

Retained label probability	Shared PEVI	Branch-complete PEVI
1.00	0.222 ± 0.134	4.015 ± 0.062
0.50	1.300 ± 0.419	4.320 ± 0.203
0.20	5.436 ± 1.277	8.315 ± 1.527
0.10	5.966 ± 2.901	10.712 ± 2.299
0.05	6.807 ± 0.884	10.963 ± 1.425

7 Application to Stateful Tool Use in AppWorld

We next evaluate RA-PEVI in AppWorld (Trivedi et al., 2024), which provides stateful applications that agents manipulate through executable tool calls. We use AppWorld as the underlying execution environment and build a protocol on top of these interactions. Full details are given in Appendix K.

Task definition. We evaluate RA-PEVI in a stateful protocol implemented in AppWorld. Each episode contains eight candidate file-system operations executed in a live AppWorld instance. At each step, the harness first chooses whether to request a read-only diagnosis and then commits an intervention from Execute, Block, and Repair.

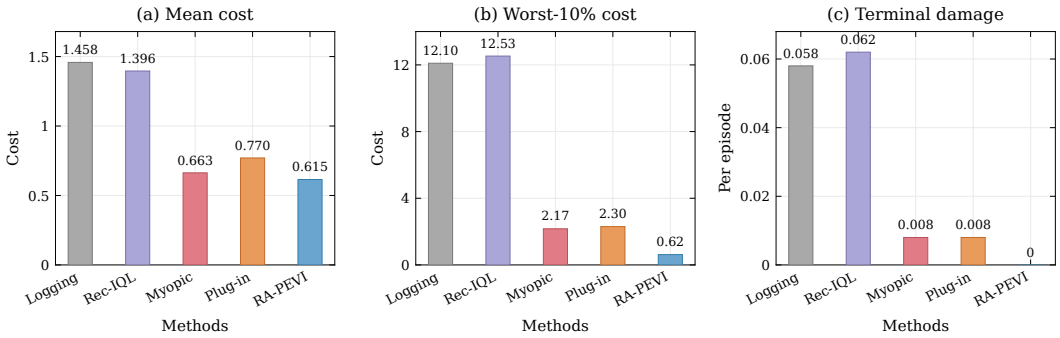

Figure 2: **Performance and safety in AppWorld.** (a) Mean episode cost. (b) Mean cost over the worst 10% of evaluation scenarios. (c) Terminal damage per episode. All methods are evaluated on the held-out evaluation scenarios.

The hidden condition for the current operation is whether its target already exists. Executing when the target is absent creates one erroneous operation and increases the unresolved-damage count by one; otherwise Execute is harmless. Diagnose queries target existence through the native AppWorld file-system interface, with its binary outcome flipped with probability 0.2. Repair removes one previously unresolved erroneous effect. Thus, local uncertainty about the current operation is coupled with damage accumulated from earlier steps. Diagnosis costs 0.02, Block and Repair cost 0.1 and 0.5, and unresolved damage remaining at the end of the episode incurs a terminal penalty.

Offline data, baselines and metrics. We collect an offline dataset using a fixed mixture of scripted logging policies on nine AppWorld training tasks, with 80 episodes per task and 720 trajectories in total. The mixture contains an Execute-heavy policy, a diagnosis-heavy policy, and a mixed policy, providing heterogeneous coverage of diagnostic and intervention choices. We compare RA-PEVI with the plug-in planner, a myopic router, and Recurrent-IQL as a generic history-based offline learner; we additionally report the logging policy as a data-collection reference.

We report three complementary measures. Mean cost is the average total episode cost, including diagnostic and intervention costs and terminal damage penalties. *Worst-10% cost* averages the highest-cost 10% of validation episodes and measures tail performance. *Terminal damage* is the average amount of unresolved damage remaining at the end of an episode.

Results. Figure 2 shows that RA-PEVI achieves the lowest mean and worst-10% cost and is the only method with no terminal damage. Relative to the plug-in planner, RA-PEVI reduces mean cost by 20.1% and worst-10% cost by 73.0%. These results show that explicitly reasoning about hidden persistent state and the value of diagnosis improves both average and tail outcomes in the stateful AppWorld setting. More results are reported in Appendix K.

8 Conclusion

We study offline learning of external harnesses for tool-using agents when safety-relevant damage is persistent, partially observed, and selectively revealed in logged trajectories. We propose RA-PEVI, which estimates shared latent intervention dynamics and performs belief-dependent pessimistic planning over diagnose–then–intervene decisions. Our analysis provides finite-sample guarantees governed by latent-state coverage, a lower bound matching the dependence on sample size and coverage, and complementary identification limits. Synthetic experiments illustrate the roles of transition sharing, limited coverage, and persistent damage, while the AppWorld study evaluates the approach in a stateful tool-use setting. Together, these results provide a principled framework for learning diagnostic and intervention policies from partially labeled offline interaction data.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, design and provide feedback on research methodology and experiments, and assist with method implementation. We also used generative AI tools to draft parts of the paper, summarize and analyze existing literature, and edit the manuscript to improve readability. We did not use generative AI tools for other tasks requiring disclosure. We reviewed all AI-assisted work. Implementation suggestions were tested through our experimental pipeline, and literature-based statements were verified against the cited sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the assumptions, auxiliary results, and complete proofs for the theoretical claims in Appendices A–I. Appendix J specifies the synthetic workflow, transition and observation models, costs, offline logging policy, selective-label mechanism, comparison methods, and evaluation procedures. Appendix K gives the corresponding details for the AppWorld study, including the diagnose–then–intervene protocol, offline data collection, baseline training, evaluation scenarios, and metric definitions. Together, these sections specify the models, data-generation procedures, learning methods, and evaluation protocols used for the reported theoretical and empirical results.

REFERENCES

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. *Advances in Neural Information Processing Systems*, 37:63790–63857, 2024.

Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. Safe reinforcement learning via shielding under partial observability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14748–14756, 2023.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in neural information processing systems*, 37:82895–82920, 2024.

Dibya Ghosh, Anurag Ajay, Pulkit Agrawal, and Sergey Levine. Offline rl policies should be trained to be adaptive. In *International Conference on Machine Learning*, pp. 7513–7530. PMLR, 2022.

Marah Ghoummaid and Uri Shalit. When to act and when to ask: policy learning with deferral under hidden confounding. *Advances in Neural Information Processing Systems*, 37:56108–56135, 2024.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023.

Hongyi Guo, Qi Cai, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Provably efficient offline reinforcement learning for partially observable markov decision processes. In *International Conference on Machine Learning*, pp. 8016–8038. PMLR, 2022.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Merlijn Krale, Thiago D Simao, and Nils Jansen. Act-then-measure: Reinforcement learning for partially observable environments with active measuring. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 33, pp. 212–220, 2023.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pp. 18733–18773. PMLR, 2023.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International conference on machine learning*, pp. 6450–6459. PMLR, 2021.

Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Safe offline reinforcement learning with real-time budget constraints. In *International Conference on Machine Learning*, pp. 21127–21152. PMLR, 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, volume 2024, pp. 27031–27098, 2024.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Thiago D Simao, Marnix Suilen, and Nils Jansen. Safe policy improvement for pomdps via finite-state controllers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15109–15117, 2023.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Haoyu Wang, Christopher M Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering*, pp. 2938–2950, 2026.

Andrew Warrington, Jonathan W Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in pomdps. In *International Conference on Machine Learning*, pp. 11013–11023. PMLR, 2021.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International conference on machine learning*, pp. 11035–11046. PMLR, 2021.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Nathaniel D Bastian, et al. Guardagent: safeguard llm agents via knowledge-enabled reasoning. In *ICML 2025 workshop on computer use agents*, 2025.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Haiyan Yin, Yingzhen Li, Sinno Jialin Pan, Cheng Zhang, and Sebastian Tschiatschek. Reinforcement learning with efficient active feature acquisition. *arXiv preprint arXiv:2011.00825*, 2020.

A PROOF PRELIMINARIES

This appendix fixes the policy, sampling, and boundary conventions used throughout the proofs and collects the total-variation tools needed to propagate latent-row estimation error to belief-space control. The proofs follow the main text: Appendix B establishes transition confidence, Appendix C turns it into a policy guarantee, and Appendix D derives the sample-size rate and its lower bound. Appendix E explains what transition sharing enables and where it stops helping. The remaining theory sections cover exact planning, calibrated diagnostics, approximate sharing, and hard budgets.

A.1 COMPARING THE SAME POLICY ACROSS MODELS

Let $\mathcal{H}_h^-$ denote the observable history immediately before the step-h diagnostic decision, and let $\mathcal{H}_h^+ = (\mathcal{H}_h^-, m_h, y_h)$ denote the observable history immediately before the intervention decision. A *public-history policy* π consists of stochastic kernels $\pi_h^{\mathrm{M}}(\cdot \mid \mathcal{H}_h^-)$ and $\pi_h^{\mathrm{I}}(\cdot \mid \mathcal{H}_h^+)$ for $h = 1, \dots, H$. The policy never receives an unrevealed damage state as input.

A belief-based policy becomes a public-history policy once its filter and action tie-breaking rule are fixed. RA-PEVI uses the nominal model $(O, \widehat{T})$. We use the fixed fallback belief $\boldsymbol{b}_{\mathrm{fb}}$ whenever either a diagnostic posterior or a transition update has zero normalizing probability. This specifies actions even on histories that are impossible under the nominal model. When evaluating a policy in another model, we keep its decision rule and internal filter unchanged; we do not replace them by that model's filter.

In particular, an optimal comparator retains its own history-to-action rule when run under $\widehat{T}$. The belief in $\mathcal{B}_{\widehat{T}}(\pi)$ is nevertheless the conditional damage distribution under $(O, \widehat{T}, \pi)$, not necessarily the belief used internally by that comparator. All deterministic comparisons below hold after fixing the dataset, and hence apply to data-dependent learned policies as well.

A.2 SAMPLING SCOPE AND BOUNDARY CONVENTIONS

Sampling scope. The fresh-randomness condition in Assumption 2 ensures that revealed transitions can be pooled across episodes without introducing selection bias. Let $\mathcal{F}_{t-1}$ contain all information from the first $t-1$ logged episodes. Conditional on $\mathcal{F}_{t-1}$, episode t follows the controlled Markov law with fresh environment randomness, and the reveal mechanism remains representative:

$$u_{t,h} \perp R_{t,h}^{Q} \mid \mathcal{F}_{t-1}, \mathcal{G}_{t,h}, d_{t,h}, \qquad d_{t,h+1} \perp R_{t,h}^{K} \mid \mathcal{F}_{t-1}, \mathcal{G}_{t,h}, d_{t,h}, u_{t,h},$$

for $e_{t,h} = 0$.

This condition holds, for example, when each logged episode is generated independently given its behavior harness, even if the behavior harnesses differ across episodes. It also allows behavior and reveal rules to depend on previously collected data. Within an episode, a reveal may depend arbitrarily on information already observed when the reveal decision is made. Thus, a source-state reveal used to estimate $Q^\star$ may depend on the observed history before $u_{t,h}$ is generated, while a next-state reveal used to estimate $K^\star$ may additionally depend on the observed outcome $u_{t,h}$. What is excluded is selecting a sample based on the unrevealed outcome being estimated, or coordinating reveal decisions across episodes through shared hidden randomness correlated with those outcomes. These conditions are needed only to establish the confidence event below; the subsequent policy-comparison results are conditional on that event.

Initial state, termination, and empty cells. Model comparisons use the same initial observed context, or average over the same initial-context distribution, with $d_1 = d_{\mathrm{clean}}$. The loss function and control costs are the given functions used by the planner. All kernels, including zero-count defaults, respect the known terminal and absorbing structure; in particular, a live episode terminates by step H. After termination, padded contributions to costs, penalties, and counts are zero. Row occupancies below count only live steps $h \le \tau$.

We use $\max \varnothing = 0$, and all continuation-cell sums range over $u \in \mathcal{U}_0$. If $\mathcal{U}_0 = \varnothing$, the K model and its radii are omitted, $C_K = 0$, and no logarithm involving $|\mathcal{U}_0|$ is evaluated. For formulas containing the otherwise unused Λ_K, set $\Lambda_K = 0$ in this case. Zero-probability conditional kernels can be fixed arbitrarily because they do not alter the joint law.

A.3 Basic total-variation inequalities

For probability distributions P and Q on a finite set $\mathcal{A}$, we use

$$\mathrm{TV}(P,Q) := \frac{1}{2}\sum_{a\in\mathcal{A}} |P(a) - Q(a)|.$$

The first inequality converts transition error into cost error. The second averages latent-row errors using a belief. The decomposition of the joint transition into its Q and K components is proved directly in the proof of Lemma 1.

Lemma 2 (Bounded test functions)**.** *If $f : \mathcal{A} \to [0, W]$, then*

$$\left|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\right| \le W\,\mathrm{TV}(P,Q).$$

Proof. Let $\mu = P - Q$, so that $\sum_a \mu(a) = 0$. Write $\mu_+(a) = \max\{\mu(a), 0\}$ and $\mu_-(a) = \max\{-\mu(a), 0\}$. Because the total mass of μ is zero,

$$\sum_a \mu_+(a) = \sum_a \mu_-(a) = \mathrm{TV}(P,Q).$$

Using $0 \le f \le W$,

$$\begin{aligned}\mathbb{E}_P[f] - \mathbb{E}_Q[f] &= \sum_a f(a)\mu_+(a) - \sum_a f(a)\mu_-(a)\\ &\le W\sum_a \mu_+(a) = W\,\mathrm{TV}(P,Q).\end{aligned}$$

Interchanging P and Q gives the reverse inequality and hence the absolute value bound. □

Lemma 3 (Convexity under a common mixture)**.** *Let $\{P_d\}_{d\in\mathcal{D}}$ and $\{Q_d\}_{d\in\mathcal{D}}$ be probability distributions on the same finite space, and let $\boldsymbol{b} \in \Delta(\mathcal{D})$. Then*

$$\mathrm{TV}\left(\sum_d \boldsymbol{b}(d)P_d, \sum_d \boldsymbol{b}(d)Q_d\right) \le \sum_d \boldsymbol{b}(d)\,\mathrm{TV}(P_d, Q_d).$$

Proof. Set $\bar{P} = \sum_d \boldsymbol{b}(d)P_d$ and $\bar{Q} = \sum_d \boldsymbol{b}(d)Q_d$. Then

$$\begin{aligned}2\,\mathrm{TV}(\bar{P},\bar{Q}) &= \sum_a \left|\sum_d \boldsymbol{b}(d)\{P_d(a) - Q_d(a)\}\right|\\ &\le \sum_d \boldsymbol{b}(d)\sum_a |P_d(a) - Q_d(a)|.\end{aligned}$$

Dividing by two gives the result. □

B Confidence under Selective Damage Reveals

This section proves the transition-confidence guarantee in Lemma 1 under selective damage reveals. The main complication is that the effective sample sizes for Q and K are random because only appropriately revealed transitions enter their estimators. Under Assumption 2, however, the selected observations remain conditionally representative of the transition component being estimated. A stopped-martingale argument therefore yields uniform confidence bounds for Q and K, which we combine into the latent-row and belief-mixture bounds.

B.1 Concentration for selectively observed categorical data

The next lemma treats a single distribution being estimated. Here $A_t = 1$ means that observation t is included in the estimate, and Y_t is the outcome being counted. Selection may depend on earlier observations, but, conditional on selection and the earlier data, Y_t must have the target law p. We first prove a bound that holds for every realized sample count; we then apply it to the two transition components.

Lemma 4 (Uniform categorical concentration under selective observation)**.** *Let $N \geq 1$ and let $(Y_t, A_t)_{t=1}^N$ be adapted to $(\mathcal{F}_t)_{t=0}^N$, with $A_t \in \{0,1\}$ and Y_t taking values in a finite set $\mathcal{A}$. Suppose $\Pr(Y_t = a \mid \mathcal{F}_{t-1}, A_t = 1) = p(a)$ whenever the conditioning event has positive probability. Let $n_t = \sum_{s=1}^t A_s$ and, when $n_t > 0$, let $\widehat{p}_t(a) = n_t^{-1} \sum_{s=1}^t A_s \mathbf{1}\{Y_s = a\}$. For any $\eta \in (0,1)$, with probability at least $1-\eta$, simultaneously for all $t \leq N$ with $n_t > 0$,*

$$\mathrm{TV}(\widehat{p}_t, p) \leq \min\left\{1, \sqrt{\frac{|\mathcal{A}| \log 2 + \log(N/\eta)}{2n_t}}\right\}.$$

For $n_t = 0$, an arbitrary empirical distribution and radius one give a valid bound.

Proof. For any $f : \mathcal{A} \to [-1,1]$, write $pf = \sum_a p(a) f(a)$. We first show that selection preserves the conditional mean and the exponential bound needed for concentration. Conditional on $(\mathcal{F}_{t-1}, A_t)$, the increment $A_t(f(Y_t) - pf)$ is zero when $A_t = 0$. When $A_t = 1$, the outcome has law p, so $f(Y_t) - pf$ has mean zero and range at most two. The bounded-variable exponential inequality therefore gives $\mathbb{E}[\exp\{\lambda(f(Y_t) - pf)\} \mid \mathcal{F}_{t-1}, A_t = 1] \leq \exp(\lambda^2/2)$. Averaging over A_t yields, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[A_t(f(Y_t) - pf) \mid \mathcal{F}_{t-1}] = 0,$$
$$\mathbb{E}\left[\exp\left\{\lambda A_t(f(Y_t) - pf) - \frac{\lambda^2}{2} A_t\right\} \middle| \mathcal{F}_{t-1}\right] \leq 1.$$

Fix a sign vector $s \in \{-1,1\}^{\mathcal{A}}$ and define $S_t^s = \sum_{j=1}^t A_j\{s(Y_j) - ps\}$, where $ps = \sum_a p(a)s(a)$. The conditional exponential bound just proved implies that

$$M_t^s(\lambda) = \exp\left\{\lambda S_t^s - \frac{\lambda^2}{2} n_t\right\}, \qquad M_0^s(\lambda) = 1,$$

is a nonnegative supermartingale for each fixed λ.

For $1 \leq n \leq N$, let $\tau_n = \inf\{t : n_t = n\}$, with $\tau_n = +\infty$ if the count is never reached. Bounded stopping at $\tau_n \wedge N$ gives $\mathbb{E}[M^s_{\tau_n \wedge N}(\lambda)] \leq 1$. Consequently, for $x > 0$ and $\lambda > 0$,

$$\Pr(\tau_n \leq N,\ S^s_{\tau_n} \geq x) \leq \exp\left(-\lambda x + \frac{\lambda^2 n}{2}\right).$$

Choosing $\lambda = x/n$ gives $\exp(-x^2/(2n))$. This bound includes the event that the nth selected observation exists; no conditioning on future sample availability is needed.

The identity $\|v\|_1 = \max_s \sum_a s(a)v(a)$ gives, on $\{\tau_n \leq N\}$, $2n\,\mathrm{TV}(\widehat{p}_{\tau_n}, p) = \max_s S^s_{\tau_n}$. A union bound over the $2^{|\mathcal{A}|}$ sign vectors therefore yields

$$\Pr(\tau_n \leq N, \mathrm{TV}(\widehat{p}_{\tau_n}, p) > \varepsilon) \leq 2^{|\mathcal{A}|} \exp(-2n\varepsilon^2).$$

Set $\varepsilon_n^2 = (|\mathcal{A}| \log 2 + \log(N/\eta))/(2n)$ and sum the failure probabilities over $n = 1, \ldots, N$. Between two selected observations neither the count nor the empirical distribution changes, so the result holds at every $t \leq N$. Truncation at one is valid because total variation is at most one. □

B.2 Proof of Lemma 1

Write

$$r = (h, d, \boldsymbol{x}, i) \in \mathcal{R} := [H] \times \mathcal{D} \times \mathcal{X} \times \mathcal{I}$$

for a latent intervention row. We use the shorthand $Q^\star_r(u) := Q^\star_{h,i}(u \mid d, \boldsymbol{x})$ and, for $u \in \mathcal{U}_0$, $K^\star_{r,u}(d') := K^\star_{h,i}(d' \mid d, \boldsymbol{x}, u)$. The same convention is used for the empirical kernels.

Define the componentwise confidence event

$$\mathcal{E}_{QK} := \left\{ \begin{array}{ll} \mathrm{TV}\left(Q^\star_r, \widehat{Q}_r\right) \leq \beta_Q(r) & \text{for every } r \in \mathcal{R}, \\ \mathrm{TV}\left(K^\star_{r,u}, \widehat{K}_{r,u}\right) \leq \beta_K(r,u) & \text{for every } r \in \mathcal{R},\ u \in \mathcal{U}_0 \end{array} \right\}. \tag{8}$$

Define the joint transition-row confidence event

$$\mathcal{E}_T := \left\{ \begin{array}{r} \mathrm{TV}\Big(T^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\Big) \le \beta_h(d, \boldsymbol{x}, i) \\ \text{for every } r = (h, d, \boldsymbol{x}, i) \in \mathcal{R} \end{array} \right\}. \tag{9}$$

Proof of Lemma 1. We use the episode-level sampling condition specified in Appendix A.2. The first two steps establish $\Pr(\mathcal{E}_{QK}) \ge 1-\delta/2$. The last two steps convert that event into the transition-row and belief-mixture statements.

Fix $r = (h, d, \boldsymbol{x}, i)$ and index episodes by t. For the Q estimate, select episode t when its live step-h transition has row r and a revealed source label; denote this indicator by $A^Q_t(r)$. The controlled Markov law and the conditional reveal condition give

$$\Pr\Big(u_{t,h} = u \mid \mathcal{F}_{t-1}, A^Q_t(r) = 1\Big) = Q^\star_r(u). \tag{10}$$

Indeed, conditional on the current public history and $d_{t,h} = d$, the row fixes the law of $u_{t,h}$. Revealing the source does not change this law. Averaging over the current histories that select the row gives equation (10). The conditioning on earlier episodes is justified by Appendix A.2.

Each episode contributes at most one observation to this row, so $\sum_t A^Q_t(r) = n_Q(r) \le N$. Apply Lemma 4 with $\mathcal{A} = \mathcal{U}$ and $\eta = \delta/(4|\mathcal{R}|)$. A union bound over rows gives

$$\mathrm{TV}(Q^\star_r, \widehat{Q}_r) \le \beta_Q(r) \qquad \text{for every } r, \tag{11}$$

with probability at least $1 - \delta/4$. This includes zero-count rows by the radius-one convention.

For K, fix a nonterminal cell (r, u) with $Q^\star_r(u) > 0$ and let $A^K_t(r, u)$ select live transitions from row r with outcome u and both source and destination labels. The second reveal condition gives

$$\Pr\big(d_{t,h+1} = d' \mid \mathcal{F}_{t-1}, A^K_t(r, u) = 1\big) = K^\star_{r,u}(d'). \tag{12}$$

Apply Lemma 4 with $\mathcal{A} = \mathcal{D}$ and $\eta = \delta/(4|\mathcal{R}||\mathcal{U}_0|)$. Union-bounding over cells establishes

$$\mathrm{TV}(K^\star_{r,u}, \widehat{K}_{r,u}) \le \beta_K(r, u) \qquad \text{for every } r,\ u \in \mathcal{U}_0, \tag{13}$$

with probability at least $1 - \delta/4$. If $Q^\star_r(u) = 0$, the cell has zero count almost surely and is covered by radius one regardless of the chosen conditional kernel. If $\mathcal{U}_0 = \varnothing$, the entire K statement is vacuous and no K union bound is needed.

Combining the failure probabilities in equation (11) and equation (13) gives

$$\Pr(\mathcal{E}_{QK}) \ge 1 - \frac{\delta}{2}. \tag{14}$$

When $\mathcal{U}_0 = \varnothing$, only the Q event is needed, so this probability bound remains valid.

Fix $r = (h, d, \boldsymbol{x}, i)$. For terminal u, extend $K^\star_{r,u}$ and $\widehat{K}_{r,u}$ to the common point mass at $d_\dagger$. Introduce the intermediate kernel $\widetilde{T}_r(d', u) = \widehat{Q}_r(u) K^\star_{r,u}(d')$. This kernel has the estimated outcome distribution but the true conditional next-state distribution. By the triangle inequality,

$$\mathrm{TV}(T^\star_r, \widehat{T}_r) \le \mathrm{TV}(T^\star_r, \widetilde{T}_r) + \mathrm{TV}(\widetilde{T}_r, \widehat{T}_r).$$

For the first term, the conditional next-state distributions are identical, so

$$\begin{aligned} 2\,\mathrm{TV}(T^\star_r, \widetilde{T}_r) &= \sum_{u, d'} \Big| Q^\star_r(u) K^\star_{r,u}(d') - \widehat{Q}_r(u) K^\star_{r,u}(d') \Big| \\ &= \sum_u |Q^\star_r(u) - \widehat{Q}_r(u)| \sum_{d'} K^\star_{r,u}(d') \\ &= 2\,\mathrm{TV}(Q^\star_r, \widehat{Q}_r). \end{aligned}$$

For the second term, terminal outcomes contribute zero because their next damage state is fixed. Hence

$$\begin{aligned} 2\,\mathrm{TV}(\widetilde{T}_r, \widehat{T}_r) &= \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u) \sum_{d'} |K^\star_{r,u}(d') - \widehat{K}_{r,u}(d')| \\ &= 2 \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u)\, \mathrm{TV}(K^\star_{r,u}, \widehat{K}_{r,u}). \end{aligned}$$

On $\mathcal{E}_{QK}$, these identities give

$$\begin{aligned}\mathrm{TV}(T_r^\star, \widehat{T}_r) &\le \mathrm{TV}(Q_r^\star, \widehat{Q}_r) + \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u)\, \mathrm{TV}(K_{r,u}^\star, \widehat{K}_{r,u}) \\ &\le \beta_Q(r) + \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u) \beta_K(r,u).\end{aligned}$$

Total variation is also at most one. Taking the minimum of one and the last expression therefore gives exactly $\beta_h(d, \boldsymbol{x}, i)$ from equation (4). Since the argument holds for every row,

$$\mathcal{E}_{QK} \subseteq \mathcal{E}_T. \tag{15}$$

Fix any $\boldsymbol{b} \in \Delta(\mathcal{D})$ and apply Lemma 3 with $P_d = T_{h,i}^\star(\cdot \mid d, \boldsymbol{x})$ and $Q_d = \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})$:

$$\begin{aligned}&\mathrm{TV}\Big(\sum_d \boldsymbol{b}(d) T_{h,i}^\star(\cdot \mid d, \boldsymbol{x}), \sum_d \boldsymbol{b}(d) \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\Big) \\ &\quad \le \sum_d \boldsymbol{b}(d)\, \mathrm{TV}\big(T_{h,i}^\star(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \\ &\quad \le \sum_d \boldsymbol{b}(d) \beta_h(d, \boldsymbol{x}, i).\end{aligned}$$

The bound holds for every belief on the same finite collection of row-wise confidence events, so no union bound over the continuous belief simplex is needed. By equation (14) and equation (15), both conclusions hold with probability at least $1 - \delta/2$, and hence at least $1 - \delta$ as stated in Lemma 1. □

C Proof of the RA-PEVI Performance Guarantee

This section converts the transition-confidence guarantee of Appendix B into the performance guarantee for RA-PEVI. Under $T^\star$ and $\widehat{T}$, the same observable history can induce different posterior beliefs, so directly comparing belief-state trajectories is inconvenient. We instead fix the policy as a public-history decision rule and use a one-transition-at-a-time hybrid argument to obtain a policy-wise simulation bound. We then identify the RA-PEVI recursion with a penalized nominal objective and combine the two results to prove Theorem 1.

Proof of Theorem 1. Fix the logged dataset and suppose that $\mathcal{E}_T$ holds. We first compare the true and estimated costs of any fixed policy, then identify the objective minimized by RA-PEVI, and finally compare the learned policy with an arbitrary admissible policy π°.

Recall that

$$\mathcal{B}_{\widehat{T}}(\pi) = \mathbb{E}_{\widehat{T}, \pi}\Big[\sum_{h=1}^{\tau} \Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)\Big],$$

where $\bar{\boldsymbol{b}}_h$ is the post-diagnostic posterior produced by the nominal filter $(O, \widehat{T})$ along the observed history. By padding with the absorbing state, we may equivalently sum from $h = 1$ to H.

We will prove that every fixed public-history policy π satisfies

$$\big|J_{T^\star}(\pi) - J_{\widehat{T}}(\pi)\big| \le \mathcal{B}_{\widehat{T}}(\pi). \tag{16}$$

For $k = 0, 1, \ldots, H$, define a hybrid controlled process $M^{(k)}$ that uses $\widehat{T}$ for transitions $1, \ldots, k$ and $T^\star$ for transitions $k+1, \ldots, H$. The diagnostic kernel O, the cost functions, and the fixed public-history policy π are the same in every hybrid. Let $J^{(k)}(\pi)$ be the expected deployment cost in $M^{(k)}$. By construction, $J^{(0)}(\pi) = J_{T^\star}(\pi)$ and $J^{(H)}(\pi) = J_{\widehat{T}}(\pi)$. Therefore

$$J_{T^\star}(\pi) - J_{\widehat{T}}(\pi) = \sum_{h=1}^{H} \big\{J^{(h-1)}(\pi) - J^{(h)}(\pi)\big\}. \tag{17}$$

Fix a step h. The two hybrids $M^{(h-1)}$ and $M^{(h)}$ use the same transition kernels during the first $h-1$ steps, namely $\widehat{T}$. Consequently, they induce exactly the same distribution over the public history $\mathcal{H}_h^-$, the diagnostic action m_h, the diagnostic output y_h, and the intervention action i_h. Conditional on this common public history after observing y_h, the latent state d_h has the nominal posterior $\bar{\boldsymbol{b}}_h$ under both hybrids. The only difference between the hybrids is the step-h environment transition: $M^{(h-1)}$ uses $T^\star$ and $M^{(h)}$ uses $\widehat{T}$.

Condition on the common post-diagnostic public history and the chosen intervention i_h. For each possible pair (d', u) produced by the step-h transition, let $F_h(d', u)$ be the conditional expected cost from the transition outcome onward, under the common continuation dynamics and policy of the two hybrids. This continuation includes the terminal loss if u is terminal and all later diagnostic and intervention costs if u is nonterminal. The terminal loss and each total step cost are nonnegative, and after the intervention at step h there are at most $H-h$ further harness decisions. Hence

$$0 \le F_h(d', u) \le L_{\max} + (H-h)c_{\max} = W_{h+1}.$$

Let $P_h^\star$ and $\widehat{P}_h$ denote the true and nominal belief-mixed laws of (d_{h+1}, u_h) conditional on the common history. Lemma 2 and Lemma 1 give

$$\begin{aligned}\left|\mathbb{E}_{P_h^\star}[F_h] - \mathbb{E}_{\widehat{P}_h}[F_h]\right| &\le W_{h+1} \sum_d \bar{\boldsymbol{b}}_h(d)\beta_h(d, \boldsymbol{x}_h, i_h) \\ &= \Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h).\end{aligned}$$

The diagnostic and intervention costs paid before the environment transition are identical in the two hybrids and therefore cancel. Taking expectation over the common public-history distribution gives

$$\left|J^{(h-1)}(\pi) - J^{(h)}(\pi)\right| \le \mathbb{E}_{\widehat{T},\pi}\left[\Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)\right].$$

The expectation is under $\widehat{T}$ because the first $h-1$ transitions in both hybrids are nominal. Summing this inequality over h and applying the triangle inequality to equation (17) gives

$$\left|J_{T^\star}(\pi) - J_{\widehat{T}}(\pi)\right| \le \mathbb{E}_{\widehat{T},\pi}\left[\sum_{h=1}^{H} \Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)\right] = \mathcal{B}_{\widehat{T}}(\pi).$$

This proves equation (16).

The continuation function in this argument uses the original costs, not the penalized value $\widehat{V}$. Thus W_{h+1} need not upper bound $\widehat{V}_{h+1}$. Because the proof is deterministic on $\mathcal{E}_T$, it holds for all fixed public-history policies on that event without a union bound over policies.

Define the penalized nominal objective

$$\mathcal{U}_{\widehat{T}}(\pi) := J_{\widehat{T}}(\pi) + \mathcal{B}_{\widehat{T}}(\pi). \tag{18}$$

We next show that the recursion in equation (6)–equation (7) minimizes this objective. With planning error ϵ_{plan}, the returned policy must therefore satisfy

$$\mathcal{U}_{\widehat{T}}(\widehat{\pi}) \le \inf_\pi \mathcal{U}_{\widehat{T}}(\pi) + \epsilon_{\text{plan}}. \tag{19}$$

For a latent state d, observed state $\boldsymbol{x}$, and intervention i, define the additional latent stage cost

$$g_h(d, \boldsymbol{x}, i) := W_{h+1}\beta_h(d, \boldsymbol{x}, i).$$

Conditional on a post-diagnostic belief $\bar{\boldsymbol{b}}$, the expected value of this latent cost is

$$\begin{aligned}\sum_d \bar{\boldsymbol{b}}(d) g_h(d, \boldsymbol{x}, i) &= W_{h+1} \sum_d \bar{\boldsymbol{b}}(d)\beta_h(d, \boldsymbol{x}, i) \\ &= \Gamma_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, i).\end{aligned}$$

Define $\widetilde{c}_h := c_{\text{M}}(\boldsymbol{x}_h, m_h) + c_{\text{I}}(\boldsymbol{x}_h, m_h, y_h, i_h) + g_h(d_h, \boldsymbol{x}_h, i_h)$. Then the penalized objective is

$$\mathcal{U}_{\widehat{T}}(\pi) = \mathbb{E}_{\widehat{T},\pi}\left[\ell(z_\tau) + \sum_{h=1}^{\tau} \widetilde{c}_h\right].$$

This is an ordinary finite-horizon POMDP objective under $(O, \widehat{T})$ with an additional latent intervention cost g_h.

We verify the recursion by backward induction. At stage $H+1$, no future cost remains, so $\widehat{V}_{H+1} = 0$ is correct. Suppose $\widehat{V}_{h+1}$ equals the minimum penalized continuation cost from stage $h+1$ onward for every belief and observed state. Fix a post-diagnostic belief $\bar{\boldsymbol{b}}$, observed state $\boldsymbol{x}$, diagnostic branch (m, y), and intervention i. The immediate cost is $c_{\mathrm{I}}(\boldsymbol{x}, m, y, i)$, and the conditional expectation of the latent penalty is $\Gamma_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, i)$. If the transition outcome is terminal, the remaining cost is $\ell(z)$; if it is nonterminal, the induction hypothesis gives the optimal continuation cost $\widehat{V}_{h+1}(\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}}, \boldsymbol{x}, u), \boldsymbol{x}')$. Taking the nominal expectation over u gives exactly equation (6). Minimizing this quantity over i is valid because the intervention is chosen after (m, y) is observed.

Before the diagnostic output is observed, choosing diagnostic m incurs $c_{\mathrm{M}}(\boldsymbol{x}, m)$ and produces output y with probability $p_{h,m}(y \mid \boldsymbol{b}, \boldsymbol{x})$. For each realized y, the best subsequent intervention value is the minimum just described. Taking the expectation over y gives equation (7), and minimizing over m yields $\widehat{V}_h$. This completes the backward induction. Because all action spaces are finite and the objective is linear in randomized action probabilities, deterministic minimizers attain the optimum. The approximate statement equation (19) is precisely the definition of ϵ_{plan}.

Fix $\pi^\circ \in \Pi$. The bound in equation (16) applies to both $\widehat{\pi}$ and π°, with their history-to-action rules held fixed in the two models. First apply its upper side to the learned policy:

$$J_{T^\star}(\widehat{\pi}) \le J_{\widehat{T}}(\widehat{\pi}) + \mathcal{B}_{\widehat{T}}(\widehat{\pi}) = \mathcal{U}_{\widehat{T}}(\widehat{\pi}).$$

Next, use the planning guarantee equation (19) and compare against π°:

$$\mathcal{U}_{\widehat{T}}(\widehat{\pi}) \le \mathcal{U}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}} = J_{\widehat{T}}(\pi^\circ) + \mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}}.$$

Finally, apply the other side of equation (16) to the comparator:

$$J_{\widehat{T}}(\pi^\circ) \le J_{T^\star}(\pi^\circ) + \mathcal{B}_{\widehat{T}}(\pi^\circ).$$

Combining the three inequalities gives

$$J_{T^\star}(\widehat{\pi}) \le J_{T^\star}(\pi^\circ) + 2\mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}},$$

or, equivalently,

$$J_{T^\star}(\widehat{\pi}) - J_{T^\star}(\pi^\circ) \le 2\mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}}. \tag{20}$$

The argument is deterministic on $\mathcal{E}_T$ and applies to every admissible π° on that event. Lemma 1, with the sampling condition in Appendix A.2, gives $\Pr(\mathcal{E}_T) \ge 1 - \delta$. Thus the comparator inequality holds with the probability stated in Theorem 1. Taking $\pi^\circ \in \Pi^\star$ gives the suboptimality bound. Exact alpha-vector planning minimizes the penalized objective without optimization error, so $\epsilon_{\mathrm{plan}} = 0$ in that case. □

The comparison charges uncertainty only along the comparator's trajectories under the estimated model. The learned policy need not itself have small uncertainty: minimizing the penalized objective is what removes its uncertainty term from the final inequality. For a nonoptimal comparator, a bound relative to $J^\star_{\mathrm{opt}}$ also includes its excess true cost $J_{T^\star}(\pi^\circ) - J^\star_{\mathrm{opt}}$.

D Finite-Sample Rates and Statistical Lower Bounds

The performance in Theorem 1 is expressed through transition uncertainty rather than the number of logged episodes. We prove the finite-sample result in Corollary 1 in this section. We first write that uncertainty in terms of the labels actually observed. We then use independent-episode sampling to obtain the population $N^{-1/2}$ rate, including its dependence on latent coverage. The final subsection gives a lower bound matching the dependence on sample size and coverage, not on all horizon and dimension parameters.

D.1 Realized-count form of the performance bound

This consequence requires only $\mathcal{E}_T$, not independent or identically distributed episodes. It displays how the revealed source labels and paired source–destination labels enter the performance bound.

Write $r_h = (h, d_h, \boldsymbol{x}_h, i_h)$ for the live latent row and define

$$\widetilde{\beta}(r) := \beta_Q(r) + \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u)\beta_K(r,u).$$

On $\mathcal{E}_T$, the proof of Theorem 1 gives, for every $\pi^\star \in \Pi^\star$, $\mathrm{SubOpt}(\widehat{\pi}) \le 2\mathcal{B}_{\widehat{T}}(\pi^\star) + \epsilon_{\mathrm{plan}}$. By the definitions of $\mathcal{B}_{\widehat{T}}$ and Γ_h,

$$\begin{aligned}\mathcal{B}_{\widehat{T}}(\pi^\star) &= \mathbb{E}_{\widehat{T},\pi^\star}\left[\sum_{h=1}^{\tau} W_{h+1} \sum_d \bar{\boldsymbol{b}}_h(d)\beta_h(d, \boldsymbol{x}_h, i_h)\right] \\ &= \mathbb{E}_{\widehat{T},\pi^\star}\left[\sum_{h=1}^{\tau} W_{h+1}\beta_h(d_h, \boldsymbol{x}_h, i_h)\right].\end{aligned}$$

The second equality is the tower property: under the nominal model, $\bar{\boldsymbol{b}}_h$ is the conditional distribution of d_h after the diagnostic output. Substituting the definition equation (4) and dropping the outer truncation at one can only increase the right-hand side. Therefore

$$\mathrm{SubOpt}(\widehat{\pi}) \le 2\mathbb{E}_{\widehat{T},\pi^\star}\left[\sum_{h=1}^{\tau} W_{h+1}\widetilde{\beta}(r_h)\right] + \epsilon_{\mathrm{plan}}. \tag{21}$$

The two parts of $\widetilde{\beta}$ depend on the observed counts $n_Q(r)$ and $n_K(r,u)$ through the radii in the main text. A large total log does not by itself make this bound small: the needed damage labels must be available for the transition rows used by the comparator.

D.2 Relating nominal uncertainty to true occupancy

The certificate counts uncertainty along trajectories generated by $\widehat{T}$, but the coverage coefficients use trajectories generated by $T^\star$. We must connect these two distributions; otherwise good coverage under the true model would not yet control the certificate. For a public-history policy π, let

$$\rho^\pi(r) := \Pr_{T^\star,\pi}\big(h \le \tau,\ (h, d_h, \boldsymbol{x}_h, i_h) = r\big)$$

denote the true-model occupancy of row r. Define the cumulative weight

$$\overline{W}_h := \sum_{s=h}^{H} W_{s+1}.$$

The larger weight $\overline{W}_h$ appears because replacing an earlier transition can change the distribution of all later uncertainty costs.

Lemma 5 (Nominal uncertainty under true occupancy)**.** *On $\mathcal{E}_T$, every public-history policy π satisfies*

$$\mathcal{B}_{\widehat{T}}(\pi) \le \sum_{r \in \mathcal{R}} \overline{W}_{h(r)}\rho^\pi(r)\beta(r), \tag{22}$$

where $h(r)$ is the stage contained in r and $\beta(r) = \beta_h(d, \boldsymbol{x}, i)$.

Proof. Define the auxiliary cost $a_h(d, \boldsymbol{x}, i) = W_{h+1}\beta_h(d, \boldsymbol{x}, i)$. Then

$$\mathcal{B}_{\widehat{T}}(\pi) = \mathbb{E}_{\widehat{T},\pi}\left[\sum_{h=1}^{H} a_h(d_h, \boldsymbol{x}_h, i_h)\right].$$

We compare this expectation with the corresponding true-model expectation by another hybrid argument. For $k = 0, \dots, H$, let $N^{(k)}$ use $T^\star$ for the first k transitions and $\widehat{T}$ thereafter. Thus $N^{(0)}$ is the nominal model and $N^{(H)}$ is the true model.

The hybrids $N^{(k-1)}$ and $N^{(k)}$ have the same law up to the intervention at step k, because both use $T^\star$ for the first $k-1$ transitions. The step-k auxiliary cost a_k is incurred before the step-k transition and is therefore identical in the two hybrids. Only future auxiliary costs can change. Their total range is at most $\sum_{s=k+1}^{H} W_{s+1}$. Conditional on the common public history, the true post-diagnostic belief is the same in the two hybrids, and the belief-mixture confidence bound gives

$$\begin{aligned}&\left|\mathbb{E}_{N^{(k-1)},\pi}\Big[\sum_{s=k+1}^{H} a_s\Big] - \mathbb{E}_{N^{(k)},\pi}\Big[\sum_{s=k+1}^{H} a_s\Big]\right| \\ &\qquad \le \Big(\sum_{s=k+1}^{H} W_{s+1}\Big)\mathbb{E}_{T^\star,\pi}\big[\beta_k(d_k, \boldsymbol{x}_k, i_k)\big].\end{aligned}$$

The expectation is under the true model because the first $k-1$ transitions of these hybrids are true, and the row at stage k is determined before the transition being replaced.

Telescoping the hybrids and using the triangle inequality gives

$$\begin{aligned}\mathcal{B}_{\widehat{T}}(\pi) &\le \mathbb{E}_{T^\star,\pi}\Big[\sum_{h=1}^{H} W_{h+1}\beta_h(d_h, \boldsymbol{x}_h, i_h)\Big] \\ &\quad + \sum_{h=1}^{H-1}\Big(\sum_{s=h+1}^{H} W_{s+1}\Big)\mathbb{E}_{T^\star,\pi}\big[\beta_h(d_h, \boldsymbol{x}_h, i_h)\big] \\ &= \sum_{h=1}^{H} \overline{W}_h \mathbb{E}_{T^\star,\pi}\big[\beta_h(d_h, \boldsymbol{x}_h, i_h)\big].\end{aligned}$$

Expanding the final expectation over the true row occupancy gives equation (22). □

This lemma lets us evaluate coverage using the true trajectories of the comparator. The larger weight $\overline{W}_h$ accounts for the fact that an error at step h can change which uncertain transitions are encountered later. The definitions of C_Q and C_K can now be used to bound the right-hand side of equation (22).

D.3 Proof of Corollary 1

For the population rate, additionally suppose that the N logged episodes are independently drawn from a fixed data-collection distribution. This distribution may randomize over behavior harnesses. The additional condition lets us relate a row's observed count to its probability of appearing in one episode. For $r = (h, d, \boldsymbol{x}, i)$ and $u \in \mathcal{U}_0$, define

$$\begin{aligned}q_Q(r) &:= \Pr_{\text{off}}\big(h \le \tau,\ (d_h, \boldsymbol{x}_h, i_h) = (d, \boldsymbol{x}, i),\ \delta_h^{\text{obs}} = 1\big), \\ q_K(r, u) &:= \Pr_{\text{off}}\big(h \le \tau,\ (d_h, \boldsymbol{x}_h, i_h) = (d, \boldsymbol{x}, i),\ u_h = u, \\ &\qquad\qquad \delta_h^{\text{obs}} = \delta_{h+1}^{\text{obs}} = 1\big).\end{aligned}$$

Because the stage is included in r, each episode contributes a Bernoulli indicator to each fixed cell. Thus $n_Q(r) \sim \text{Binomial}(N, q_Q(r))$ and $n_K(r, u) \sim \text{Binomial}(N, q_K(r, u))$ marginally, even though different cells within the same episode are dependent.

Proof of Corollary 1. Fix the covered optimal comparator $\pi^\star := \pi^\star_{\text{cov}} \in \Pi^\star$ from the corollary. We combine the componentwise confidence event with a bound on the random label counts, and then sum the resulting radii using the true occupancy of $\pi^\star$.

Let $M_{\text{cnt}} = |\mathcal{R}|(1 + |\mathcal{U}_0|)$ and $L_{\text{cnt}} = \log(4M_{\text{cnt}}/\delta)$. Define $\mathcal{E}_{\text{cnt}}$ to be the event that, simultaneously for every cell with positive sampling probability,

$$\frac{1}{n_Q(r) \vee 1} \le \frac{8L_{\text{cnt}}}{N q_Q(r)}, \qquad \frac{1}{n_K(r, u) \vee 1} \le \frac{8L_{\text{cnt}}}{N q_K(r, u)}. \tag{23}$$

To bound its failure probability, consider a generic count $n = \sum_{j=1}^{N} B_j$ with independent Bernoulli variables of mean $q > 0$. There are two cases.

If $Nq < 8L_{\mathrm{cnt}}$, then $1/(n \vee 1) \le 1 < 8L_{\mathrm{cnt}}/(Nq)$ deterministically, so the desired inequality holds without a concentration argument.

If $Nq \ge 8L_{\mathrm{cnt}}$, the multiplicative Chernoff inequality gives

$$\Pr\big(n < Nq/2\big) \le \exp(-Nq/8) \le \exp(-L_{\mathrm{cnt}}) = \frac{\delta}{4M_{\mathrm{cnt}}}.$$

On the complementary event, $1/(n \vee 1) \le 2/(Nq) \le 8L_{\mathrm{cnt}}/(Nq)$. There are at most M_{cnt} source and source–destination cells. A union bound therefore gives total failure probability at most $\delta/4$; $\Pr(\mathcal{E}_{\mathrm{cnt}}) \ge 1 - \delta/2$ follows as well.

On $\mathcal{E}_{\mathrm{cnt}}$, the definitions of β_Q and β_K imply

$$\beta_Q(r) \le 2\sqrt{\frac{\Lambda_Q L_{\mathrm{cnt}}}{Nq_Q(r)}}, \qquad \beta_K(r,u) \le 2\sqrt{\frac{\Lambda_K L_{\mathrm{cnt}}}{Nq_K(r,u)}}.$$

For positive counts this follows directly from equation (23). At zero count, it follows as well because $\Lambda_Q \ge 2$ and, for nonempty continuation cells, $\Lambda_K \ge 2$. Thus the radius-one convention obeys $\beta_Q(r) \le \sqrt{\Lambda_Q/(2(n_Q(r) \vee 1))}$, and analogously for K. Substituting the inverse-count bounds gives the displayed inequalities. All sums of ratios below are restricted to positive numerators; covered comparators then have positive denominators on those terms.

The proof of Lemma 1 established $\Pr(\mathcal{E}_{QK}) \ge 1-\delta/2$; see equation (14). Consequently, $\mathcal{E}_{QK} \cap \mathcal{E}_{\mathrm{cnt}}$ has probability at least $1 - \delta$. We work on this intersection for the rest of the proof. In particular, $\mathcal{E}_T$ also holds.

The radius in equation (4) uses $\widehat{Q}$ as a weight. For the population bound it is convenient to replace this weight by $Q^\star$. The following comparison uses the stronger componentwise event $\mathcal{E}_{QK}$ rather than only the induced joint-row event $\mathcal{E}_T$. On $\mathcal{E}_{QK}$, because $0 \le \beta_K(r,u) \le 1$,

$$\begin{aligned}
\beta(r) &\le \beta_Q(r) + \sum_{u \in \mathcal{U}_0} \widehat{Q}_r(u)\beta_K(r,u) \\
&\le \beta_Q(r) + \sum_{u \in \mathcal{U}_0} Q^\star_r(u)\beta_K(r,u) + \sum_{u \in \mathcal{U}_0} \left|\widehat{Q}_r(u) - Q^\star_r(u)\right| \\
&\le \beta_Q(r) + \sum_{u \in \mathcal{U}_0} Q^\star_r(u)\beta_K(r,u) + \left\|\widehat{Q}_r - Q^\star_r\right\|_1 \\
&\le 3\beta_Q(r) + \sum_{u \in \mathcal{U}_0} Q^\star_r(u)\beta_K(r,u). \qquad (24)
\end{aligned}$$

The last inequality uses

$$\left\|\widehat{Q}_r - Q^\star_r\right\|_1 = 2\,\mathrm{TV}\left(\widehat{Q}_r, Q^\star_r\right) \le 2\beta_Q(r),$$

which follows from $\mathcal{E}_{QK}$.

The coefficients $C_Q(\pi^\star)$ and $C_K(\pi^\star)$ are those defined in the main text. Write

$$S_{Q,h}(\pi) := \big|\{r : h(r) = h,\ \rho^\pi(r) > 0\}\big|,$$

$$S_{K,h}(\pi) := \big|\{(r,u) : h(r) = h,\ u \in \mathcal{U}_0,\ \rho^\pi(r)Q^\star_r(u) > 0\}\big|.$$

These are the numbers of source rows and continuation cells that the policy uses with positive probability at stage h. For compactness, set

$$A_{Q,h} := \sqrt{\Lambda_Q C_Q(\pi^\star) S_{Q,h}(\pi^\star)}, \qquad A_{K,h} := \sqrt{\Lambda_K C_K(\pi^\star) S_{K,h}(\pi^\star)}.$$

Fix a stage h. By the definition of $C_Q(\pi^\star)$, $\rho^{\pi^\star}(r)/q_Q(r) \le C_Q(\pi^\star)$ for every relevant row. Therefore

$$\begin{aligned}\sum_{r:h(r)=h} \frac{\rho^{\pi^\star}(r)}{\sqrt{q_Q(r)}} &= \sum_{r:h(r)=h} \sqrt{\rho^{\pi^\star}(r)}\sqrt{\frac{\rho^{\pi^\star}(r)}{q_Q(r)}} \\ &\le \sqrt{C_Q(\pi^\star)} \sum_{r:h(r)=h} \sqrt{\rho^{\pi^\star}(r)} \\ &\le \sqrt{C_Q(\pi^\star) S_{Q,h}(\pi^\star)}.\end{aligned}$$

The last inequality is Cauchy–Schwarz together with $\sum_{r:h(r)=h} \rho^{\pi^\star}(r) \le 1$.

For the K term, let $w(r,u) = \rho^{\pi^\star}(r) Q^\star_r(u)$. Coverage gives $w(r,u)/q_K(r,u) \le C_K(\pi^\star)$, and these weights sum to at most one at each stage. The same calculation gives

$$\sum_{r,u:h(r)=h} \frac{\rho^{\pi^\star}(r) Q^\star_r(u)}{\sqrt{q_K(r,u)}} \le \sqrt{C_K(\pi^\star) S_{K,h}(\pi^\star)}.$$

Substituting equation (24) and the radius bounds from Step 1 into Lemma 5, then applying the sums bounded in Step 3, gives

$$\mathcal{B}_{\widehat{T}}(\pi^\star) \le \frac{2\sqrt{L_{\mathrm{cnt}}}}{\sqrt{N}} \sum_{h=1}^{H} \overline{W}_h (3A_{Q,h} + A_{K,h}).$$

The comparator inequality proved in Appendix C holds on $\mathcal{E}_T$. Applying it on our intersection event yields the dimension-explicit bound

$$\mathrm{SubOpt}(\widehat{\pi}) \le \frac{4\sqrt{L_{\mathrm{cnt}}}}{\sqrt{N}} \sum_{h=1}^{H} \overline{W}_h (3A_{Q,h} + A_{K,h}) + \epsilon_{\mathrm{plan}}. \tag{25}$$

Finally, $C_Q(\pi^\star), C_K(\pi^\star) \le C_{\mathrm{lat}}(\pi^\star)$, $S_{Q,h}(\pi^\star) \le |\mathcal{D}||\mathcal{X}||\mathcal{I}|$, and $S_{K,h}(\pi^\star) \le |\mathcal{D}||\mathcal{X}||\mathcal{I}||\mathcal{U}_0|$. Keeping the finite state, action, outcome, and horizon dimensions and the cost bounds fixed, equation (25) therefore implies

$$\mathrm{SubOpt}(\widehat{\pi}) \le \widetilde{O}\left(\sqrt{\frac{C_{\mathrm{lat}}(\pi^\star_{\mathrm{cov}})}{N}}\right) + \epsilon_{\mathrm{plan}}.$$

The $\widetilde{O}$ notation hides logarithmic factors and the listed problem parameters, but no further inverse-coverage factor. If $\mathcal{U}_0 = \varnothing$, every K contribution is omitted under the conventions of Appendix A.2. The intersection event has probability at least $1 - \delta$, completing the proof. □

D.4 A LOWER BOUND MATCHING SAMPLE SIZE AND COVERAGE

The main corollary gives an $N^{-1/2}$ rate. The next result shows that this dependence on the number of logged episodes and on the effective coverage of a relevant row cannot be improved in general. The lower bound uses a fully observed one-step subclass, so it also applies to the larger diagnostic POMDP class. It does not establish separate optimality in the horizon, state dimensions, or paired-label coverage C_K.

Theorem 2 (Minimax lower bound under positive coverage)**.** *There is a universal constant $c > 0$ with the following property. For every $q \in (0, 1/2]$ and every N with $Nq \ge 1$, there exist two one-step instances $T^{(0)}$ and $T^{(1)}$ and a common behavior policy such that:*

1. *the two environments differ on only one latent transition row, and the behavior policy samples that row with probability q;*
2. *in each environment $T^{(j)}$, an optimal policy $\pi^\star_j$ satisfies*

$$C_{\mathrm{lat}}(\pi^\star_j) = \frac{1}{q};$$

3. every offline learner $\mathfrak{A}$ satisfies

$$\sup_{j\in\{0,1\}} \mathbb{E}_{T^{(j)}}\left[\mathrm{SubOpt}_{T^{(j)}}\left(\mathfrak{A}(\mathcal{D}_N)\right)\right] \geq \frac{c}{\sqrt{Nq}}.$$

Consequently, the dependence $\Omega(\sqrt{C_{\mathrm{lat}}/N})$ is unavoidable in general.

Proof. Consider $H = 1$, one observed state, and a known latent state, so diagnostics are unnecessary and the source state is observed in every episode. There are three interventions, i_0, i_1, and i_2, and no action cost. Include both the null diagnostic and the exact audit at zero cost; with one known live latent state they provide the same information. The terminal loss lies in $\{0, 1\}$, and every terminal outcome leads to the fixed absorbing state.

Intervention i_0 has Bernoulli loss with mean $1/2$ in both environments. For a parameter $\Delta \in (0, 1/4]$, intervention i_1 has Bernoulli loss with mean

$$\frac{1}{2} - \Delta \quad \text{in } T^{(0)}, \qquad \frac{1}{2} + \Delta \quad \text{in } T^{(1)}.$$

Intervention i_2 has deterministic loss one in both environments. Thus i_1 is the unique optimal intervention in $T^{(0)}$, whereas i_0 is the unique optimal intervention in $T^{(1)}$. Choosing i_0 instead of i_1 in $T^{(0)}$, or choosing i_1 instead of i_0 in $T^{(1)}$, incurs expected excess loss Δ. Intervention i_2 incurs excess loss at least Δ in both environments.

The behavior policy selects i_0 with probability q, selects i_1 with probability q, and selects i_2 with probability $1-2q$, independently across episodes. Only the transition row corresponding to i_1 differs between the two environments, and this row is sampled with probability q.

Let P_0^N and P_1^N denote the resulting offline-data laws. By the chain rule for relative entropy,

$$\mathrm{KL}(P_0^N \| P_1^N) = Nq \ \mathrm{kl}\left(\mathrm{Bern}(1/2 - \Delta) \| \mathrm{Bern}(1/2 + \Delta)\right).$$

Let $p_- = 1/2 - \Delta$ and $p_+ = 1/2 + \Delta$. For $\Delta \leq 1/4$,

$$\mathrm{KL}(p_- \| p_+) \leq \frac{(p_- - p_+)^2}{p_+(1 - p_+)} \leq \frac{64}{3}\Delta^2.$$

Choose

$$\Delta = \frac{1}{16\sqrt{Nq}}.$$

Since $Nq \geq 1$, this choice satisfies $\Delta \leq 1/16$ and is therefore valid. It gives

$$\mathrm{KL}(P_0^N \| P_1^N) \leq \frac{1}{12}.$$

Pinsker's inequality then yields

$$\mathrm{TV}(P_0^N, P_1^N) \leq \sqrt{\frac{\mathrm{KL}(P_0^N \| P_1^N)}{2}} < \frac{1}{4}.$$

Let $\varphi(\mathcal{D}_N) \in [0, 1]$ denote the probability with which the policy output by the learner selects i_1, conditional on the logged dataset. This definition also covers randomized learners and randomized output policies. In environment $T^{(0)}$, every intervention other than i_1 has excess loss at least Δ, and hence

$$\mathbb{E}_{T^{(0)}}\left[\mathrm{SubOpt}_{T^{(0)}}\left(\mathfrak{A}(\mathcal{D}_N)\right)\right] \geq \Delta\, \mathbb{E}_{P_0^N}[1 - \varphi(\mathcal{D}_N)].$$

In environment $T^{(1)}$, selecting i_1 incurs excess loss Δ, so

$$\mathbb{E}_{T^{(1)}}\left[\mathrm{SubOpt}_{T^{(1)}}\left(\mathfrak{A}(\mathcal{D}_N)\right)\right] \geq \Delta\, \mathbb{E}_{P_1^N}[\varphi(\mathcal{D}_N)].$$

Therefore,

$$\begin{aligned}
&\max_{j\in\{0,1\}} \mathbb{E}_{T^{(j)}}\left[\mathrm{SubOpt}_{T^{(j)}}\left(\mathfrak{A}(\mathcal{D}_N)\right)\right] \\
&\quad \geq \frac{\Delta}{2}\left\{\mathbb{E}_{P_0^N}[1 - \varphi(\mathcal{D}_N)] + \mathbb{E}_{P_1^N}[\varphi(\mathcal{D}_N)]\right\} \\
&\quad \geq \frac{\Delta}{2}\left\{1 - \mathrm{TV}(P_0^N, P_1^N)\right\} \\
&\quad \geq \frac{3\Delta}{8} = \frac{3}{128\sqrt{Nq}}.
\end{aligned}$$

The second inequality follows from the bounded-test-function characterization of total variation, since $\varphi \in [0,1]$. Thus one may take $c = 3/128$.

It remains to verify the coverage claim. Let

$$r_0 = (1, d_{\text{clean}}, \boldsymbol{x}, i_0), \qquad r_1 = (1, d_{\text{clean}}, \boldsymbol{x}, i_1).$$

Because the initial latent state is known, every step-one transition is source-labeled. Under the behavior policy,

$$q_Q(r_0) = q_Q(r_1) = q.$$

In $T^{(0)}$, the optimal policy always selects i_1, so $\rho^{\pi_0^\star}(r_1) = 1$. In $T^{(1)}$, the optimal policy always selects i_0, so $\rho^{\pi_1^\star}(r_0) = 1$. Since the episode terminates after one step, there are no nonterminal continuation rows, and only C_Q is relevant. Consequently,

$$C_{\text{lat}}(\pi_0^\star) = C_Q(\pi_0^\star) = \frac{1}{q}, \qquad C_{\text{lat}}(\pi_1^\star) = C_Q(\pi_1^\star) = \frac{1}{q}.$$

Substituting $1/q = C_{\text{lat}}(\pi_j^\star)$ into the lower bound gives the claimed $\Omega(\sqrt{C_{\text{lat}}/N})$ dependence. □

Theorem 2 identifies the sample-size limitation behind the upper bound. Only about Nq logged episodes contain the intervention whose outcome distinguishes the two environments. When the outcome probabilities differ by order $1/\sqrt{Nq}$, the data do not reliably reveal which intervention is better, and choosing incorrectly incurs an error of that same order. Since $C_{\text{lat}} = 1/q$ in the construction, this matches the upper bound's dependence on N and latent coverage. Thus the upper and lower bounds are minimax-matched in their dependence on sample size and latent coverage.

E Structural Benefits and Identification Limits

This section characterizes both the benefit and the limits of the shared-transition structure. We first show that an observable diagnostic branch need not be directly supported in the offline data when the latent transition rows in its posterior mixture are identified elsewhere. We then establish two complementary identification limits: one when a value-relevant latent transition leaves no evidence in the log, and one when transition sharing across diagnostic branches is removed. Finally, we show that persistent damage can create a strict performance gap for policies that skip diagnosis solely because the current proposal appears safe.

E.1 Cross-diagnostic identification

The claim here concerns a fixed input belief: both models use the same mixture weights. It explains how identified latent rows determine an unobserved branch, without assuming that true and estimated filters agree along a trajectory. The latter issue is handled by Appendix C.

Fix a stage h, pre-diagnostic belief $\boldsymbol{b}$, observed state $\boldsymbol{x}$, diagnostic m, and output y with positive probability

$$p_{h,m}(y \mid \boldsymbol{b}, \boldsymbol{x}) = \sum_d \boldsymbol{b}(d) O_{h,m}(y \mid d, \boldsymbol{x}) > 0.$$

The corresponding posterior is

$$\omega_{m,y}^{\boldsymbol{b},\boldsymbol{x}}(d) := \frac{\boldsymbol{b}(d) O_{h,m}(y \mid d, \boldsymbol{x})}{\sum_{\tilde{d}} \boldsymbol{b}(\tilde{d}) O_{h,m}(y \mid \tilde{d}, \boldsymbol{x})}.$$

Proposition 1 (Identification of an unseen observable branch)**.** *Under the shared intervention model, the post-diagnostic transition law of intervention i is*

$$G_h^\star(\cdot \mid \boldsymbol{b}, \boldsymbol{x}, m, y, i) = \sum_d \omega_{m,y}^{\boldsymbol{b},\boldsymbol{x}}(d) T_{h,i}^\star(\cdot \mid d, \boldsymbol{x}). \tag{26}$$

On $\mathcal{E}_T$, its plug-in estimate satisfies

$$\mathrm{TV}(G_h^\star, \widehat{G}_h) \le \sum_d \omega_{m,y}^{\boldsymbol{b},\boldsymbol{x}}(d) \beta_h(d, \boldsymbol{x}, i). \tag{27}$$

Neither statement depends on the behavior probability of the complete branch (m, y, i).

Proof. Condition on the pre-diagnostic information $(\boldsymbol{b}, \boldsymbol{x})$, on choosing m, and on observing y. Bayes' rule gives the conditional distribution of the latent state as $\omega_{m,y}^{\boldsymbol{b},\boldsymbol{x}}$. Conditional on latent state d and intervention i, the shared model states that the next transition law is $T_{h,i}^{\star}(\cdot \mid d, \boldsymbol{x})$, independently of the identity of m and the realized output y. Applying the law of total probability over d gives equation (26).

The plug-in law replaces each latent row by $\widehat{T}_{h,i}$. Write $\omega(d) = \omega_{m,y}^{\boldsymbol{b},\boldsymbol{x}}(d)$. Then Lemma 3 and $\mathcal{E}_T$ give

$$\begin{aligned}\mathrm{TV}(G_h^{\star}, \widehat{G}_h) &\le \sum_d \omega(d)\, \mathrm{TV}\big(T_{h,i}^{\star}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \\ &\le \sum_d \omega(d)\beta_h(d, \boldsymbol{x}, i).\end{aligned}$$

The branch behavior probability never enters this calculation; only the latent rows used by the posterior mixture matter. □

Proposition 1 explains which evidence can support an unobserved diagnostic branch. For example, transitions recorded after a full audit can support the same intervention after a cheap check, provided they identify every damage-state row that the cheap check's posterior uses. The diagnostic choice and output change the weights on those rows, not the rows themselves. RA-PEVI can therefore reuse the transition evidence instead of requiring a separate dataset for each (m, y, i) branch. This is a statement at a fixed input belief with the known diagnostic model; it does not assume that the true and estimated models produce the same belief along a trajectory. Their resulting policy costs are compared in Appendix C.

The same argument applies to an entire contingent handling rule. Let $\kappa(i \mid y)$ be a stochastic intervention rule after diagnostic m. The joint law of (y, i, ξ), where ξ is the transition outcome, is

$$\begin{aligned}P_T^{m,\kappa}(y, i, \xi \mid \boldsymbol{b}, \boldsymbol{x}) = \sum_d \boldsymbol{b}(d) O_{h,m}(y \mid d, \boldsymbol{x}) \\ \times \kappa(i \mid y) T_{h,i}(\xi \mid d, \boldsymbol{x}).\end{aligned}$$

Convexity of total variation gives

$$\begin{aligned}\mathrm{TV}\big(P_{T^\star}^{m,\kappa}, P_{\widehat{T}}^{m,\kappa}\big) \le \sum_{d,y,i} \boldsymbol{b}(d) O_{h,m}(y \mid d, \boldsymbol{x}) \\ \times \kappa(i \mid y) \beta_h(d, \boldsymbol{x}, i).\end{aligned}$$

Thus, one latent-row confidence event controls every output-contingent intervention rule at the same input belief. Under independent episodes, positive $q_Q(r)$ on posterior-supported rows and positive $q_K(r, u)$ on their positive-probability continuation cells make these errors vanish as data increase. No direct sample of the complete branch is needed.

E.2 Non-Identifiability Under Missing Latent Support

Sharing cannot recover a transition that leaves no information in the observed data. We demonstrate this using two environments with identical log distributions but different optimal actions. The construction isolates a worst-case failure of identification when a value-relevant transition leaves no information in the complete offline-data distribution.

Theorem 3 (Latent-support impossibility)**.** *There exists a one-step instance satisfying the main model assumptions for which a value-relevant latent intervention row has zero offline support and, for every offline learner* $\mathfrak{A}$*,*

$$\sup_{j \in \{0,1\}} \mathbb{E}_{T^{(j)}}\big[\mathrm{SubOpt}_{T^{(j)}}(\mathfrak{A}(\mathcal{D}_N))\big] \ge \frac{L}{4},$$

where $L > 0$ *is the terminal-loss range. The two environments induce the same distribution over the offline data but have different optimal interventions.*

Proof. Consider $H = 1$, one observed state $\boldsymbol{x}$, and the known initial damage state $d_1 = d_{\text{clean}}$. Include both the null diagnostic and the required exact audit at zero cost. The audit is uninformative here because the initial state is already known. There are two interventions i_0 and i_1, and the episode terminates immediately after the intervention.

In both environments, i_1 produces deterministic terminal loss $L/2$. Intervention i_0 produces deterministic loss 0 in $T^{(0)}$ and deterministic loss L in $T^{(1)}$. The behavior policy always executes i_1. Therefore every logged trajectory consists of the same action and the same terminal loss $L/2$ in both environments. In particular, the row of i_0 has zero support and the full offline-data laws are identical. The optimal intervention is i_0 in $T^{(0)}$ and i_1 in $T^{(1)}$.

Because the data law is identical, an arbitrary learner deploys the same distribution over actions in both worlds. Let p be its probability of choosing i_0. In $T^{(0)}$, choosing i_1 incurs excess loss $L/2$, so the expected suboptimality is $(1-p)L/2$. In $T^{(1)}$, choosing i_0 incurs excess loss $L/2$, so the expected suboptimality is $pL/2$. Hence

$$\max\left\{\frac{(1-p)L}{2}, \frac{pL}{2}\right\} \geq \frac{L}{4}.$$

This lower bound holds for every p and therefore every offline learner. □

Theorem 3 gives a different situation from Proposition 1: the intervention i_0 is never tried under any diagnostic. The logs therefore cannot tell whether it is better or worse than i_1, and collecting more episodes with the same behavior policy does not resolve that uncertainty. The lower bound $L/4$ holds for every sample size. Since the damage state is known and there is only one decision, the difficulty comes from missing information in the data, not from belief estimation or planning.

The result rules out a vanishing worst-case error guarantee when an unobserved transition can change the optimal decision without changing the log distribution. The theorem establishes a worst-case necessity of information about value-relevant transitions, rather than necessity of the specific sufficient coverage condition used in Corollary 1. Here the impossibility follows because the two complete offline-data distributions are identical while their optimal interventions differ.

E.3 Necessity of Transition Sharing for Cross-Diagnostic Transfer

The preceding impossibility concerns a latent intervention row that is absent from all informative data. We next show why the shared transition model is essential for transferring information across diagnostics. If each complete observable branch is allowed to have unrelated dynamics, then zero support for that branch is itself a fundamental obstacle.

Theorem 4 (Zero-branch impossibility without transition sharing)**.** *Consider the larger model class in which the transition kernel may depend arbitrarily on (m, y, i). There exists a one-step instance with a value-relevant observable branch of zero behavior probability such that every offline learner has worst-case suboptimality at least $L/4$, even though the same intervention is well sampled under another diagnostic.*

Proof. There is one known latent state and two diagnostics, m_0 and m_1, each with a deterministic output y_0 and y_1, respectively. Take m_1 to be the exact audit, with $y_1 = d_{\text{clean}}$. There are two interventions, i_0 and i_1, and the episode terminates after the intervention. All diagnostic and intervention costs are zero. The behavior policy always selects m_1 and, conditional on m_1, selects each of i_0 and i_1 with probability $1/2$. Every branch under m_1 has deterministic loss L in both environments, so these samples are identical across worlds.

The behavior never visits (m_0, y_0, i_0). Under m_0, intervention i_1 has deterministic loss $L/2$ in both environments, whereas i_0 has loss 0 in $T^{(0)}$ and loss L in $T^{(1)}$. Hence the complete offline-data laws are identical. The unique optimal branch is (m_0, y_0, i_0) in $T^{(0)}$ and (m_0, y_0, i_1) in $T^{(1)}$.

Let p_0 and p_1 be the learner's probabilities of deploying these two m_0 branches, and let $p_{\text{else}} = 1 - p_0 - p_1$. In $T^{(0)}$, the expected suboptimality is at least $(p_1 + p_{\text{else}})L/2 = (1-p_0)L/2$. In $T^{(1)}$, it is at least $(p_0 + p_{\text{else}})L/2 = (1-p_1)L/2$. Since $p_0 + p_1 \leq 1$, at least one of $1-p_0$ and $1-p_1$ is at least $1/2$, yielding worst-case suboptimality at least $L/4$.

□

The distinction from Theorem 3 is that both interventions are now sampled under m_1. What is missing is evidence about the interventions' outcomes after m_0. If transitions can depend arbitrarily on the diagnostic branch, the m_1 samples impose no restriction on those missing transitions. Under the shared model, by contrast, fixing $(h, d, \boldsymbol{x}, i)$ fixes the same transition law under both diagnostics, so these samples also identify the law used after m_0.

Together with Proposition 1, this shows that transition sharing can do more than reduce estimation error: it can make a branch identifiable even when that branch has zero behavior probability. The theorem concerns the model class with unrestricted branch-dependent dynamics.

E.4 Limitations of Myopic Safe-Proposal Routing

The following construction formalizes a restricted class of routers that trigger diagnosis only from the immediate hazard of the current proposal. Let $\boldsymbol{x}_{\mathrm{safe}}$ denote an observed proposal type whose pass-through intervention cannot create any new damage. For the instance below, define

$$\Pi_{\mathrm{myop}} := \left\{\pi \in \Pi : \pi_2^{\mathrm{M}}\left(m_{\mathrm{null}} \mid \mathcal{H}_2^-\right) = 1 \text{ whenever } \boldsymbol{x}_2 = \boldsymbol{x}_{\mathrm{safe}}\right\}. \tag{28}$$

Thus, policies in Π_{myop} skip diagnosis whenever the current proposal has zero immediate hazard, regardless of the possibility of damage left by earlier steps.

Proposition 2 (Persistent-damage gap)**.** *For any $p \in (0,1)$, terminal-loss level $L > 0$, and audit cost $c \in (0, pL)$, there exists a two-step instance satisfying the main model such that*

$$\inf_{\pi \in \Pi_{\mathrm{myop}}} J_{T^\star}(\pi) = pL, \qquad \inf_{\pi \in \Pi} J_{T^\star}(\pi) = c.$$

Hence the additive gap is $pL - c$, and the approximation ratio pL/c can be made arbitrarily large.

Proof. Let

$$\mathcal{D} = \{d_{\mathrm{clean}}, d_{\mathrm{damaged}}\}.$$

The episode begins at $d_1 = d_{\mathrm{clean}}$ and an observed context distinct from $\boldsymbol{x}_{\mathrm{safe}}$, so its step-one costs can differ from the step-two costs defined below. At step one, all available diagnostic/intervention branches have zero cost and induce the same transition:

$$d_2 = \begin{cases} d_{\mathrm{damaged}}, & \text{with probability } p, \\ d_{\mathrm{clean}}, & \text{with probability } 1-p. \end{cases}$$

The transition produces no observable signal, and the episode does not terminate at step one. Thus no harness policy can affect or observe whether damage was created at the first step.

At step two, the observed proposal is $\boldsymbol{x}_{\mathrm{safe}}$. There are two diagnostics: a null diagnostic m_{null} with deterministic output $y_\varnothing$ and zero cost, and the full audit m^{rev}, which reveals d_2 and has cost

$$c_{\mathrm{M}}(\boldsymbol{x}_{\mathrm{safe}}, m^{\mathrm{rev}}) = c.$$

There are two interventions, i_{pass} and i_{rep}. The pass-through intervention preserves the latent state, while the repair intervention maps both latent states to d_{clean}. The intervention's damage outcome is recorded in z, and the next state is then set to $(d_\dagger, \boldsymbol{x}_\dagger)$. Define the terminal loss by

$$\ell(z) = \begin{cases} L, & \text{if the damage outcome recorded in } z \text{ is } d_{\mathrm{damaged}}, \\ 0, & \text{if the damage outcome recorded in } z \text{ is } d_{\mathrm{clean}}. \end{cases}$$

Thus i_{pass} cannot create new damage at step two; any loss under that intervention is caused entirely by persistent damage from step one.

To prevent free blind repair, define the intervention costs by

$$c_{\mathrm{I}}(\boldsymbol{x}_{\mathrm{safe}}, m_{\mathrm{null}}, y_\varnothing, i_{\mathrm{rep}}) = L,$$

and let all other step-two intervention costs be zero. These costs are known and remain compatible with the main model because c_{I} may depend on the diagnostic branch, while the transition induced by a fixed intervention remains shared.

Every policy in Π_{myop} selects m_{null} at step two. Let $a \in [0,1]$ be its probability of choosing the blind repair intervention. Since the step-one transition produced no observable signal, a cannot depend on d_2. Its expected cost is therefore

$$aL + (1-a)pL = pL + a(1-p)L \geq pL.$$

Choosing i_{pass} deterministically attains pL, so

$$\inf_{\pi \in \Pi_{\text{myop}}} J_{T^\star}(\pi) = pL.$$

A belief-based policy in the unrestricted class Π selects the full audit at step two, chooses i_{rep} if the audit reveals d_{damaged}, and chooses i_{pass} otherwise. It incurs audit cost c and zero terminal loss, so its total expected cost is c.

Conversely, any policy that audits incurs cost at least c. Conditional on not auditing, the policy has no information about d_2, and its minimum possible expected cost is $\min\{pL, L\} = pL$. A randomized mixture of auditing and not auditing therefore has expected cost at least $\min\{c, pL\} = c$. Since $c < pL$, the unrestricted optimum is exactly c. The additive and multiplicative gaps follow immediately. □

Proposition 2 shows that a proposal can be safe to execute in the immediate sense and still leave a costly problem unresolved. In the construction, passing through at step two creates no new damage, but it also does not remove the damage from step one. A rule that always skips diagnosis on such proposals pays pL, whereas the unrestricted harness can pay $c < pL$. Thus, the current proposal's immediate hazard alone does not justify excluding diagnosis; the accumulated damage and the cost of subsequent intervention also matter.

F Exact Planning over Continuous Beliefs

This section establishes that RA-PEVI admits exact finite-horizon planning despite operating on a continuous belief space. The key observation is that the belief-dependent pessimism penalty can be written as the expectation of an additional latent-state cost. The resulting penalized nominal problem therefore retains the standard finite-horizon POMDP structure, and its value function admits an exact alpha-vector representation when $\mathcal{M}$ and each $\mathcal{Y}_m$ are finite. This result establishes exact representability, but does not imply polynomial-time planning.

Proposition 3 (Alpha-vector representation)**.** *For every stage h and observed state $\boldsymbol{x}$, the penalized value $\widehat{V}_h(\cdot, \boldsymbol{x})$ is piecewise linear and concave on $\Delta(\mathcal{D})$. Specifically, there exists a finite set $\mathcal{A}_{h,\boldsymbol{x}} \subset \mathbb{R}^{|\mathcal{D}|}$ such that*

$$\widehat{V}_h(\boldsymbol{b}, \boldsymbol{x}) = \min_{\boldsymbol{\alpha} \in \mathcal{A}_{h,\boldsymbol{x}}} \boldsymbol{b}^\top \boldsymbol{\alpha}.$$

Each alpha vector can be stored together with the diagnostic, output-dependent interventions, and continuation vectors that generated it, thereby encoding an optimal contingent policy tree.

Proof. We give the recursive construction. At stage $H+1$, set $\mathcal{A}_{H+1,\boldsymbol{x}} = \{\mathbf{0}\}$ for every absorbing or nonabsorbing observed state. The representation then holds because $\widehat{V}_{H+1} \equiv 0$.

Assume the representation holds at stage $h+1$. Fix an observed state $\boldsymbol{x}$ at stage h. A deterministic one-step contingent plan consists of: (i) a diagnostic m; (ii) an intervention i_y for every possible diagnostic output $y \in \mathcal{Y}_m$; and (iii) for every pair (y, u) with nonterminal $u = (\boldsymbol{x}', 0, \perp)$, a continuation vector $\boldsymbol{\alpha}_{y,u} \in \mathcal{A}_{h+1,\boldsymbol{x}'}$. There are finitely many such choices because all spaces and the horizon are finite.

Associate with this contingent plan a vector $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{D}|}$. For each latent state d,

$$\alpha(d) = c_{\mathrm{M}}(\boldsymbol{x}, m) + \sum_{y \in \mathcal{Y}_m} O_{h,m}(y \mid d, \boldsymbol{x}) \Big[c_{\mathrm{I}}(\boldsymbol{x}, m, y, i_y) + W_{h+1} \beta_h(d, \boldsymbol{x}, i_y) + \sum_{u: e=1} \widehat{Q}_{h,i_y}(u \mid d, \boldsymbol{x}) \ell(z) + \sum_{u: e=0} \sum_{d'} \widehat{T}_{h,i_y}(d', u \mid d, \boldsymbol{x}) \alpha_{y,u}(d') \Big].$$

The four terms are the intervention cost, latent pessimism cost, terminal loss, and nonterminal continuation cost, respectively.

If the current belief is $\boldsymbol{b}$, the expected cost of this fixed contingent plan is

$$\sum_d \boldsymbol{b}(d) \alpha(d) = \boldsymbol{b}^\top \boldsymbol{\alpha}$$

by the law of total expectation. Conversely, every deterministic policy tree from stage h induces exactly one vector by this construction. Randomization cannot improve the minimum because it produces a convex combination of the costs of deterministic plans. Let $\mathcal{A}_{h,\boldsymbol{x}}$ be the finite set of all vectors generated by these plans. Minimizing over plans gives

$$\widehat{V}_h(\boldsymbol{b}, \boldsymbol{x}) = \min_{\boldsymbol{\alpha} \in \mathcal{A}_{h,\boldsymbol{x}}} \boldsymbol{b}^\top \boldsymbol{\alpha}.$$

The pointwise minimum of finitely many linear functions is piecewise linear and concave, completing the induction. □

Proposition 3 explains why the uncertainty penalty does not require a new representation of belief-space value functions. Each alpha vector records the penalized expected cost of one complete plan, separately for each possible current damage state. Averaging those entries under the current belief gives the cost of that plan, and minimizing over plans gives the value function. The pessimism term is included as an additional latent-state cost, so no belief discretization is needed to obtain an exact representation.

This is not a guarantee of efficient computation. Vectors that are never optimal can be removed by dominance pruning, but the number of remaining plans can still grow very quickly with the horizon. Approximate planning is covered by ϵ_{plan} only when that quantity bounds the returned policy's optimization error for the full penalized objective in equation (19). A fine belief grid or a short planning horizon does not by itself certify a small value of that error.

G Extension to Calibrated Diagnostic Kernels

Assumption 1 in the main text treats the diagnostic kernels as known. This section shows that the same pessimistic-planning framework extends when the non-audit diagnostic kernels must instead be estimated from a separate labeled calibration dataset. We retain the exact audit and attach an additional confidence penalty at the diagnostic decision node, before its output is observed. We first define the calibrated planner and then establish its performance guarantee without requiring a direct comparison of posterior ratios under the true and estimated diagnostic models. Thus, exact knowledge of all diagnostic kernels is not essential; what is needed is an independently justified confidence bound for each estimated diagnostic row.

G.1 Calibration model and confidence radii

Let

$$\mathcal{M}_{\mathrm{cal}} := \mathcal{M} \setminus \{m^{\mathrm{rev}}\}$$

denote the diagnostics whose kernels are estimated from the independent calibration data. For $m \in \mathcal{M}_{\mathrm{cal}}$, let $O^\star_{h,m}(\cdot \mid d, \boldsymbol{x})$ be the unknown true diagnostic kernel, and let $\widehat{O}_{h,m}(\cdot \mid d, \boldsymbol{x})$ be its empirical estimate.

The full audit remains known exactly:

$$\widehat{O}_{h,m^{\mathrm{rev}}}(y \mid d, \boldsymbol{x}) = O^{\star}_{h,m^{\mathrm{rev}}}(y \mid d, \boldsymbol{x}) = \mathbf{1}\{y = d\},$$

and we define

$$\alpha(h, d, \boldsymbol{x}, m^{\mathrm{rev}}) := 0.$$

Suppose an independent calibration dataset $\mathcal{C}_{N_O}$ contains labeled tuples $(h, d, \boldsymbol{x}, m, y)$ with $m \in \mathcal{M}_{\mathrm{cal}}$. The diagnostic action in the calibration data may be selected adaptively from previously observed calibration information. Conditional on that preceding information and the selected row $(h, d, \boldsymbol{x}, m)$, its output must have law $O^{\star}_{h,m}(\cdot \mid d, \boldsymbol{x})$. The transition log continues to satisfy the reveal and sampling conditions of Appendix B under the true diagnostic mechanism.

Define

$$\mathcal{R}_O := [H] \times \mathcal{D} \times \mathcal{X} \times \mathcal{M}_{\mathrm{cal}}.$$

For $r_O = (h, d, \boldsymbol{x}, m) \in \mathcal{R}_O$, let $n_O(r_O)$ be the calibration row count and define

$$\Lambda_O(m) := |\mathcal{Y}_m| \log 2 + \log \frac{2N_O|\mathcal{R}_O|}{\delta_O},$$

$$\alpha(r_O) := \min\left\{1, \sqrt{\frac{\Lambda_O(m)}{2n_O(r_O)}}\right\}.$$

These formulas apply when $N_O \geq 1$ and $\mathcal{R}_O \neq \varnothing$. A zero-count calibration row receives radius one. If $N_O = 0$, set every uncalibrated radius to one directly; if $\mathcal{R}_O = \varnothing$, there is no calibration error and the logarithm above is not evaluated.

Lemma 6 (Diagnostic-kernel confidence)**.** *With probability at least $1-\delta_O$, simultaneously for every $m \in \mathcal{M}_{\mathrm{cal}}$ and every $(h, d, \boldsymbol{x})$,*

$$\mathrm{TV}\Big(O^{\star}_{h,m}(\cdot \mid d, \boldsymbol{x}), \widehat{O}_{h,m}(\cdot \mid d, \boldsymbol{x})\Big) \leq \alpha(h, d, \boldsymbol{x}, m).$$

By the exact-audit convention, the same inequality also holds for $m = m^{\mathrm{rev}}$, with both sides equal to zero. Consequently, for every pre-diagnostic belief $\boldsymbol{b}$, if $P^{\star}_m = \sum_d \boldsymbol{b}(d) O^{\star}_{h,m}(\cdot \mid d, \boldsymbol{x})$ and $\widehat{P}_m = \sum_d \boldsymbol{b}(d) \widehat{O}_{h,m}(\cdot \mid d, \boldsymbol{x})$, then

$$\mathrm{TV}(P^{\star}_m, \widehat{P}_m) \leq \sum_d \boldsymbol{b}(d) \alpha(h, d, \boldsymbol{x}, m).$$

Proof. For a fixed row, let A_t indicate whether calibration draw t uses that row. The conditional calibration law gives the premise of Lemma 4 with $\mathcal{A} = \mathcal{Y}_m$, maximum count N_O, and $\eta = \delta_O/(2|\mathcal{R}_O|)$. A union bound over rows gives failure probability at most $\delta_O/2$, hence at most δ_O. Empty or zero-count cases are covered by the stated conventions. The exact audit has zero error, and Lemma 3 gives the mixture bound. □

G.2 Pessimistic planning with calibrated diagnostics

An inaccurate diagnostic model can change both the output and the subsequent intervention. We therefore charge its uncertainty before observing the output, using a bound that includes the current intervention and all later costs. The transition penalty remains at the intervention node.

Let

$$W_h := L_{\max} + (H - h + 1)c_{\max}$$

be an upper bound on the remaining cost before the diagnostic at step h. For a pre-diagnostic belief $\boldsymbol{b}$, define

$$\Gamma^{O}_h(\boldsymbol{b}, \boldsymbol{x}, m) := W_h \sum_d \boldsymbol{b}(d) \alpha(h, d, \boldsymbol{x}, m). \tag{29}$$

The extended planner uses $\widehat{O}$ for diagnostic-output probabilities and posterior updates, uses $\widehat{T}$ for environment transitions, and retains the transition penalty Γ_h.

Starting from

$$\widehat{V}^{O}_{H+1} \equiv 0,$$

define, for a post-diagnostic belief $\bar{\boldsymbol{b}}$,

$$\widehat{G}^{O}_{h+1}(u;\bar{\boldsymbol{b}},\boldsymbol{x},i) := \begin{cases} \ell(z), & u=(\boldsymbol{x}',1,z), \\ \widehat{V}^{O}_{h+1}\left(\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}},\boldsymbol{x},u),\boldsymbol{x}'\right), & u=(\boldsymbol{x}',0,\perp). \end{cases} \tag{30}$$

The intervention-stage backup is

$$\widehat{Q}^{I,O}_{h}(\bar{\boldsymbol{b}},\boldsymbol{x},m,y,i) := c_{\mathrm{I}}(\boldsymbol{x},m,y,i) + \Gamma_h(\bar{\boldsymbol{b}},\boldsymbol{x},i) + \sum_{u\in\mathcal{U}} \widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}},\boldsymbol{x})\widehat{G}^{O}_{h+1}(u;\bar{\boldsymbol{b}},\boldsymbol{x},i). \tag{31}$$

Define

$$\widehat{V}^{I,O}_{h}(\bar{\boldsymbol{b}},\boldsymbol{x},m,y) := \min_{i\in\mathcal{I}} \widehat{Q}^{I,O}_{h}(\bar{\boldsymbol{b}},\boldsymbol{x},m,y,i). \tag{32}$$

Before the diagnostic output is observed, define

$$\widehat{p}^{O}_{h,m}(y \mid \boldsymbol{b},\boldsymbol{x}) := \sum_{d\in\mathcal{D}} \boldsymbol{b}(d)\widehat{O}_{h,m}(y \mid d,\boldsymbol{x}).$$

When this probability is positive, define the nominal diagnostic posterior by

$$\widehat{\boldsymbol{b}}^{m,y}(d) := \frac{\boldsymbol{b}(d)\widehat{O}_{h,m}(y \mid d,\boldsymbol{x})}{\widehat{p}^{O}_{h,m}(y \mid \boldsymbol{b},\boldsymbol{x})}.$$

When $\widehat{p}^{O}_{h,m}(y \mid \boldsymbol{b},\boldsymbol{x}) = 0$, set $\widehat{\boldsymbol{b}}^{m,y} := \boldsymbol{b}_{\mathrm{fb}}$.

The diagnostic-stage backup is

$$\widehat{Q}^{M,O}_{h}(\boldsymbol{b},\boldsymbol{x},m) := c_{\mathrm{M}}(\boldsymbol{x},m) + \Gamma^{O}_{h}(\boldsymbol{b},\boldsymbol{x},m) + \sum_{y\in\mathcal{Y}_m} \widehat{p}^{O}_{h,m}(y \mid \boldsymbol{b},\boldsymbol{x})\widehat{V}^{I,O}_{h}\left(\widehat{\boldsymbol{b}}^{m,y},\boldsymbol{x},m,y\right). \tag{33}$$

Finally,

$$\widehat{V}^{O}_{h}(\boldsymbol{b},\boldsymbol{x}) = \min_{m\in\mathcal{M}} \widehat{Q}^{M,O}_{h}(\boldsymbol{b},\boldsymbol{x},m). \tag{34}$$

G.3 Performance guarantee with estimated diagnostics

To handle belief mismatch, we compare the joint law of damage and diagnostic output (d_h, y_h). Once this pair is fixed, both models have the same continuation rule. Replacing the diagnostic and transition kernels in turn then yields two additive error terms.

Let $\widehat{\pi}^{O}$ be an ϵ_{plan}-optimal policy for equation (30)– equation (34). Define the cumulative observation and transition uncertainty under the joint nominal model $(\widehat{O},\widehat{T})$ by

$$\mathcal{B}_O(\pi) := \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\sum_{h=1}^{\tau} \Gamma^{O}_{h}(\boldsymbol{b}_h,\boldsymbol{x}_h,m_h)\right],$$

and

$$\mathcal{B}_T(\pi) := \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\sum_{h=1}^{\tau} \Gamma_h(\bar{\boldsymbol{b}}_h,\boldsymbol{x}_h,i_h)\right].$$

Here $\boldsymbol{b}_h$ and $\bar{\boldsymbol{b}}_h$ are respectively the pre- and post-diagnostic beliefs generated by the nominal filter $(\widehat{O},\widehat{T})$.

Theorem 5 (Independently calibrated diagnostic kernels)**.** *On the joint confidence event of Lemmas 1 and 6, every admissible public-history comparator π° satisfies*

$$J_{O^\star,T^\star}(\widehat{\pi}^O) - J_{O^\star,T^\star}(\pi^\circ) \le 2\left\{\mathcal{B}_O(\pi^\circ) + \mathcal{B}_T(\pi^\circ)\right\} + \epsilon_{\mathrm{plan}}.$$

The joint event has probability at least $1 - \delta - \delta_O$. In particular, taking

$$\pi^\star_{OT} \in \arg\min_{\pi\in\Pi} J_{O^\star,T^\star}(\pi)$$

gives the corresponding suboptimality bound.

Proof. Fix both datasets and work on the joint confidence event in the theorem. All policy comparisons keep the same history-to-action rule in the true model $(O^\star, T^\star)$ and the estimated model $(\widehat{O}, \widehat{T})$.

We first prove, for every fixed public-history policy π,

$$\left|J_{O^\star,T^\star}(\pi) - J_{\widehat{O},\widehat{T}}(\pi)\right| \le \mathcal{B}_O(\pi) + \mathcal{B}_T(\pi). \tag{35}$$

We interpolate between the fully true model and the fully nominal model one stage at a time. Let A_{h-1} be the hybrid that uses $(\widehat{O}, \widehat{T})$ for stages $1, \dots, h-1$ and $(O^\star, T^\star)$ from stage h onward. Define an intermediate hybrid B_h that agrees with A_{h-1} except that it uses $\widehat{O}$ for the step-h diagnostic while still using $T^\star$ for the step-h transition. Finally, let A_h additionally replace the step-h transition by $\widehat{T}$. Thus, A_0 is the fully true model, A_H is the fully nominal model, and

$$J_{O^\star,T^\star}(\pi) - J_{\widehat{O},\widehat{T}}(\pi) = \sum_{h=1}^{H}\Big[J_{A_{h-1}}(\pi) - J_{B_h}(\pi) + J_{B_h}(\pi) - J_{A_h}(\pi)\Big]. \tag{36}$$

We first compare A_{h-1} and B_h. Since both hybrids use $(\widehat{O}, \widehat{T})$ during the first $h-1$ stages, they induce the same law over the public history immediately before the step-h diagnostic and over the diagnostic action m_h. Conditional on this common history and on $m_h = m$, let $\boldsymbol{x}_h = \boldsymbol{x}$ and let $\boldsymbol{b}_h$ be the nominal pre-diagnostic belief, namely the conditional distribution of d_h under the common past.

The step-h diagnostic kernels induce the following two joint distributions of (d_h, y_h) on $\mathcal{D} \times \mathcal{Y}_m$:

$$\mu^\star_h(d, y) := \boldsymbol{b}_h(d)\, O^\star_{h,m}(y \mid d, \boldsymbol{x}), \qquad \widehat{\mu}_h(d, y) := \boldsymbol{b}_h(d)\, \widehat{O}_{h,m}(y \mid d, \boldsymbol{x}).$$

Their total-variation distance satisfies

$$\begin{aligned}
\mathrm{TV}(\mu^\star_h, \widehat{\mu}_h) &= \frac{1}{2}\sum_{d\in\mathcal{D}}\sum_{y\in\mathcal{Y}_m} \boldsymbol{b}_h(d)\left|O^\star_{h,m}(y \mid d, \boldsymbol{x}) - \widehat{O}_{h,m}(y \mid d, \boldsymbol{x})\right| \\
&= \sum_{d\in\mathcal{D}} \boldsymbol{b}_h(d)\ \mathrm{TV}\Big(O^\star_{h,m}(\cdot \mid d, \boldsymbol{x}), \widehat{O}_{h,m}(\cdot \mid d, \boldsymbol{x})\Big) \\
&\le \sum_{d\in\mathcal{D}} \boldsymbol{b}_h(d)\, \alpha(h, d, \boldsymbol{x}, m),
\end{aligned} \tag{37}$$

where the last inequality follows from Lemma 6.

Conditional on the common pre-diagnostic history, let $F^O_h(d, y)$ denote the expected cost from the step-h diagnostic decision through the end of the episode, given $d_h = d$ and $y_h = y$, under the common policy and continuation kernels used by the two hybrids after the diagnostic output is generated. The dependence of F^O_h on the conditioned public history is suppressed.

We include the step-h diagnostic cost in F^O_h, even though it is common to the two hybrids. This ensures that, under $0 \le c_{\mathrm{M}} + c_{\mathrm{I}} \le c_{\max}$,

$$0 \le F^O_h(d, y) \le L_{\max} + (H - h + 1)c_{\max} = W_h.$$

Once (d_h, y_h) is fixed, the two hybrids use the same intervention policy, the same step-h transition kernel $T^\star$, and the same subsequent kernels. Hence F^O_h is the same test function under $\mu^\star_h$ and $\widehat{\mu}_h$.

By Lemma 2 and equation 37,

$$\begin{aligned}\left|\mathbb{E}_{\mu_h^\star}\left[F_h^O(d_h, y_h)\right] - \mathbb{E}_{\widehat{\mu}_h}\left[F_h^O(d_h, y_h)\right]\right| &\le W_h \ \mathrm{TV}(\mu_h^\star, \widehat{\mu}_h) \\ &\le W_h \sum_{d\in\mathcal{D}} \boldsymbol{b}_h(d)\alpha(h, d, \boldsymbol{x}, m) \\ &= \Gamma_h^O(\boldsymbol{b}_h, \boldsymbol{x}, m). \end{aligned} \tag{38}$$

Taking expectation over the common pre-diagnostic history and diagnostic action therefore gives

$$\left|J_{A_{h-1}}(\pi) - J_{B_h}(\pi)\right| \le \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\Gamma_h^O(\boldsymbol{b}_h, \boldsymbol{x}_h, m_h)\right]. \tag{39}$$

The expectation may be written under the fully nominal model because the quantity inside the expectation is determined before the step-h diagnostic output, and the first $h-1$ stages are nominal in both hybrids.

We next compare B_h and A_h. These hybrids have the same law through the step-h diagnostic output: both use $(\widehat{O}, \widehat{T})$ during the first $h-1$ stages and use $\widehat{O}$ for the current diagnostic. Therefore, conditional on the common post-diagnostic public history, the latent state d_h has the nominal posterior $\bar{\boldsymbol{b}}_h$. They also induce the same intervention action i_h under the fixed public-history policy.

Conditional on this common history and on $i_h = i$, define the true and nominal belief-mixed laws of the step-h transition outcome (d_{h+1}, u_h) by

$$\nu_h^\star := \sum_{d\in\mathcal{D}} \bar{\boldsymbol{b}}_h(d)\, T_{h,i}^\star(\cdot \mid d, \boldsymbol{x}_h), \qquad \widehat{\nu}_h := \sum_{d\in\mathcal{D}} \bar{\boldsymbol{b}}_h(d)\, \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x}_h).$$

Let $F_h^T(d', u)$ denote the conditional expected cost from immediately after the step-h transition through the end of the episode, under the common continuation policy and kernels. Since the current diagnostic and intervention costs have already been incurred,

$$0 \le F_h^T(d', u) \le L_{\max} + (H-h)c_{\max} = W_{h+1}.$$

On the transition-confidence event, Lemma 1 gives

$$\mathrm{TV}(\nu_h^\star, \widehat{\nu}_h) \le \sum_{d\in\mathcal{D}} \bar{\boldsymbol{b}}_h(d)\, \beta_h(d, \boldsymbol{x}_h, i).$$

Applying Lemma 2 and then averaging over the common post-diagnostic history and intervention action yields

$$\begin{aligned}\left|J_{B_h}(\pi) - J_{A_h}(\pi)\right| &\le \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[W_{h+1}\sum_{d\in\mathcal{D}} \bar{\boldsymbol{b}}_h(d)\, \beta_h(d, \boldsymbol{x}_h, i_h)\right] \\ &= \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)\right]. \end{aligned} \tag{40}$$

Here the expectation is again under the fully nominal model because all random variables appearing in the integrand are determined before the step-h transition, and the two hybrids agree with the nominal model up to that point.

Finally, applying the triangle inequality to equation 36, and summing equation (39) and equation (40) over h, gives

$$\begin{aligned}\left|J_{O^\star,T^\star}(\pi) - J_{\widehat{O},\widehat{T}}(\pi)\right| &\le \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\sum_{h=1}^{H} \Gamma_h^O(\boldsymbol{b}_h, \boldsymbol{x}_h, m_h)\right] \\ &\quad + \mathbb{E}_{\widehat{O},\widehat{T},\pi}\left[\sum_{h=1}^{H} \Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h)\right] \\ &= \mathcal{B}_O(\pi) + \mathcal{B}_T(\pi). \end{aligned}$$

By the absorbing-state convention, the sums over $h = 1, \ldots, H$ are equivalent to the corresponding sums up to the termination time. This proves equation (35). Since the argument holds for any fixed policy on the same confidence event, it also applies to the learned policy and every comparator without an additional union bound over policies.

Define $\mathcal{U}_{\widehat{O},\widehat{T}}(\pi) := J_{\widehat{O},\widehat{T}}(\pi) + \mathcal{B}_O(\pi) + \mathcal{B}_T(\pi)$. The diagnostic penalty is the conditional expectation of $W_h\alpha(h, d_h, \boldsymbol{x}_h, m_h)$ before the output is observed, and the transition penalty is the conditional expectation of $W_{h+1}\beta_h(d_h, \boldsymbol{x}_h, i_h)$ after the output is observed. The recursions in equation (30)– equation (34) are therefore Bellman recursions for this objective under $(\widehat{O}, \widehat{T})$, by the same backward-induction argument used in Step 2 of the proof of Theorem 1. Planning optimality gives

$$\mathcal{U}_{\widehat{O},\widehat{T}}(\widehat{\pi}^O) \le \inf_{\pi} \mathcal{U}_{\widehat{O},\widehat{T}}(\pi) + \epsilon_{\text{plan}}.$$

For every admissible comparator π°, apply equation (35) to the learned policy and the comparator, with planning optimality between the two applications:

$$\begin{aligned} J_{O^\star,T^\star}(\widehat{\pi}^O) &\le \mathcal{U}_{\widehat{O},\widehat{T}}(\widehat{\pi}^O) \\ &\le \mathcal{U}_{\widehat{O},\widehat{T}}(\pi^\circ) + \epsilon_{\text{plan}} \\ &\le J_{O^\star,T^\star}(\pi^\circ) + 2\mathcal{B}_O(\pi^\circ) + 2\mathcal{B}_T(\pi^\circ) + \epsilon_{\text{plan}}. \end{aligned}$$

Subtracting the comparator cost proves the claimed inequality. The two confidence events hold jointly with probability at least $1 - \delta - \delta_O$ by a union bound. Taking $\pi^\circ = \pi^\star_{OT}$ gives suboptimality. □

Theorem 5 explains how the known-diagnostic condition in Assumption 1 can be relaxed in the stated calibration setting. An error in a diagnostic model can change which output is seen and which intervention is selected afterward. The extra penalty therefore belongs before the output is observed. The final bound separates this diagnostic-model uncertainty from transition-model uncertainty, and weights both by the decisions made by the comparator under the estimated model. When all diagnostic kernels are known exactly, their radii can be set to zero. Then $\mathcal{B}_O = 0$ and the bound reduces to the main comparator guarantee.

The calibration labels and their confidence bounds remain essential premises of this result. The extension applies when the diagnostic confidence sets are obtained from an independent labeled calibration sample. Joint estimation of O and T from the same partially labeled trajectories would require additional identifiability analysis.

H Extension to Approximate Transition Sharing

The main model assumes that, conditional on the latent state and the chosen intervention, the environment transition is shared across diagnostic choices and outputs; see equation (1). This section relaxes that assumption by allowing the same intervention to induce slightly different transition dynamics across diagnostic branches. We measure this deviation relative to a shared reference kernel and incorporate it as an additional misspecification term in the pessimistic penalty. The resulting guarantee is conditional on a valid confidence bound for the reference kernel: closeness of the branch-specific kernels to a common reference alone does not justify the pooled confidence radius derived for the exactly shared model.

H.1 Reference dynamics and their uncertainty

Suppose the actual branch-specific transition is $T^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x})$. Assume there exists a shared reference kernel $\overline{T}^\star_{h,i}$ and known upper bounds $\zeta_h(d, \boldsymbol{x}, m, y, i)$ such that

$$\mathrm{TV}\big(T^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}), \overline{T}^\star_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \le \zeta_h(d, \boldsymbol{x}, m, y, i). \tag{41}$$

Assume the pooled estimator has confidence

$$\mathrm{TV}\big(\overline{T}^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \le \beta_h(d, \boldsymbol{x}, i).$$

The reference-confidence inequality is a separate premise, not a consequence of Lemma 1 under misspecification. Different reveal rates can weight branches differently in the Q and K estimates, so naive pooling need not concentrate around the chosen reference with the exact-model radii. In this section, $J_{T^\star}$ denotes cost under the actual branch-dependent kernels, while $J_{\widehat{T}}$ uses the shared nominal kernels.

H.2 Performance guarantee under approximate sharing

For a post-diagnostic branch (m, y), define

$$\Gamma_h^{\mathrm{tot}}(\bar{\boldsymbol{b}}, \boldsymbol{x}, m, y, i) = W_{h+1} \sum_d \bar{\boldsymbol{b}}(d)\big\{\beta_h(d, \boldsymbol{x}, i) + \zeta_h(d, \boldsymbol{x}, m, y, i)\big\}. \tag{42}$$

Let $\widehat{\pi}_\zeta$ be an ϵ_{plan}-optimal policy obtained by replacing Γ_h with Γ_h^{tot} in the intervention backup. Define

$$\mathcal{B}_{\mathrm{tot}}(\pi) := \mathbb{E}_{\widehat{T},\pi}\left[\sum_{h=1}^{\tau} \Gamma_h^{\mathrm{tot}}(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, m_h, y_h, i_h)\right].$$

Theorem 6 (Robustness to approximate transition sharing)**.** *Under equation (41) and the reference-kernel confidence event, every public-history comparator π° satisfies*

$$J_{T^\star}(\widehat{\pi}_\zeta) - J_{T^\star}(\pi^\circ) \le 2\mathcal{B}_{\mathrm{tot}}(\pi^\circ) + \epsilon_{\mathrm{plan}}.$$

In particular, the result applies to an optimal policy of the actual branch-dependent model.

Proof. Work on the stated confidence event for the reference kernel.

For each latent state, the triangle inequality gives

$$\begin{aligned}
&\mathrm{TV}\big(T^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \\
&\quad \le \mathrm{TV}\big(T^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}), \overline{T}^\star_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \\
&\qquad + \mathrm{TV}\big(\overline{T}^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\big) \\
&\quad \le \zeta_h(d, \boldsymbol{x}, m, y, i) + \beta_h(d, \boldsymbol{x}, i).
\end{aligned}$$

Applying Lemma 3 to any post-diagnostic belief $\bar{\boldsymbol{b}}$ therefore gives

$$\begin{aligned}
&\mathrm{TV}\left(\sum_d \bar{\boldsymbol{b}}(d) T^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}), \sum_d \bar{\boldsymbol{b}}(d) \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\right) \\
&\quad \le \sum_d \bar{\boldsymbol{b}}(d)\big\{\zeta_h(d, \boldsymbol{x}, m, y, i) + \beta_h(d, \boldsymbol{x}, i)\big\}.
\end{aligned} \tag{43}$$

Fix a public-history policy π. Repeat the transition-replacement argument in Step 1 of the proof of Theorem 1. The two processes being compared now differ at step h between the actual branch-specific kernel and the estimated shared kernel. They agree on the history, diagnostic branch (m_h, y_h), and intervention i_h before that transition. Their common past uses the estimated kernels, so the conditional damage distribution is the nominal posterior $\bar{\boldsymbol{b}}_h$. The common continuation cost is bounded by W_{h+1}. Combining equation (43) with Lemma 2, the absolute step-h cost difference is at most

$$\mathbb{E}_{\widehat{T},\pi}\big[\Gamma_h^{\mathrm{tot}}(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, m_h, y_h, i_h)\big].$$

Summing the differences over the horizon gives

$$\big|J_{T^\star}(\pi) - J_{\widehat{T}}(\pi)\big| \le \mathcal{B}_{\mathrm{tot}}(\pi).$$

The modified backup minimizes $\mathcal{U}_{\mathrm{tot}}(\pi) := J_{\widehat{T}}(\pi) + \mathcal{B}_{\mathrm{tot}}(\pi)$. Indeed, the total penalty in equation (42) is the conditional expectation of the additional latent cost $W_{h+1}\{\beta_h(d_h, \boldsymbol{x}_h, i_h) + \zeta_h(d_h, \boldsymbol{x}_h, m_h, y_h, i_h)\}$. Using Step 2 for the learned policy and the comparator gives

$$\begin{aligned}
J_{T^\star}(\widehat{\pi}_\zeta) &\le \mathcal{U}_{\mathrm{tot}}(\widehat{\pi}_\zeta) \\
&\le \mathcal{U}_{\mathrm{tot}}(\pi^\circ) + \epsilon_{\mathrm{plan}} \\
&\le J_{T^\star}(\pi^\circ) + 2\mathcal{B}_{\mathrm{tot}}(\pi^\circ) + \epsilon_{\mathrm{plan}}.
\end{aligned}$$

Subtracting the comparator cost proves the theorem. □

Theorem 6 allows the same intervention to have slightly different effects after different diagnostics, provided those differences are bounded and the reference kernel has a valid confidence bound. The penalty then covers both the error in estimating the reference and its deviation from the actual diagnostic branch. The comparison is with policies in the actual branch-dependent environment, not only with policies that are optimal in the reference model.

H.3 Separating statistical error from misspecification

A useful special case is a uniform misspecification bound

$$\zeta_h(d, \boldsymbol{x}, m, y, i) \le \zeta.$$

Since at most H interventions are executed,

$$\mathcal{B}_{\mathrm{tot}}(\pi) \le \mathcal{B}_{\widehat{T}}(\pi) + \zeta \sum_{h=1}^{H} W_{h+1}.$$

Consequently, the comparator bound is at most $2\mathcal{B}_{\widehat{T}}(\pi^\circ) + 2\zeta \sum_{h=1}^{H} W_{h+1} + \epsilon_{\mathrm{plan}}$. The second term is a misspecification contribution; it vanishes with ζ, not merely with more logged data.

This separates two reasons for a large performance bound. More informative labeled data can reduce the estimation radii β_h, but cannot by itself reduce a fixed discrepancy ζ between the branch dynamics and the reference. If the reference estimator has shrinking confidence radii on the rows used by the comparator, the statistical contribution can decrease while the displayed misspecification contribution remains.

I Extension to a Hard Resource Budget

The deployment objective in the main model accounts for diagnostic and intervention costs additively, but does not impose a pathwise resource constraint on individual trajectories. This section extends the formulation by requiring every realized diagnostic–intervention sequence to satisfy a finite hard budget. The remaining budget becomes an additional observable planning-state variable, while the unknown environmental transition kernels are unchanged. Consequently, the budget constraint modifies the feasible policy class and the planning recursion without introducing new transition parameters or statistical uncertainty.

I.1 Budget accounting and feasible decisions

Let B_h denote the remaining budget immediately before the step-h diagnostic decision, with known initial value B_1. Suppose that completing the diagnostic/intervention branch (m_h, y_h, i_h) consumes the known nonnegative expenditure

$$g_h(\boldsymbol{x}_h, m_h, y_h, i_h).$$

The branch is feasible only if

$$g_h(\boldsymbol{x}_h, m_h, y_h, i_h) \le B_h.$$

On a nonterminal transition, the remaining budget evolves deterministically according to

$$B_{h+1} = B_h - g_h(\boldsymbol{x}_h, m_h, y_h, i_h). \tag{44}$$

The expenditure is also charged on a terminal branch, although no subsequent budget state is then needed.

We assume that B_1 and all expenditures lie on a finite grid, and let $\mathcal{B}$ be the resulting finite set of reachable nonnegative remaining budgets. The budget is a harness-side resource variable: it affects which harness actions are admissible but does not alter the diagnostic or environment dynamics. In particular,

$$y_h \sim O_{h,m_h}(\cdot \mid d_h, \boldsymbol{x}_h)$$

and, after the intervention is chosen,

$$(d_{h+1}, \boldsymbol{x}_{h+1}, e_h, z_h) \sim T^{\star}_{h,i_h}(\cdot \mid d_h, \boldsymbol{x}_h),$$

exactly as in the main model. Diagnostics therefore remain read-only with respect to the environment and damage state, even though selecting a diagnostic may consume harness-side budget through g_h.

Because the diagnostic is selected before its output is observed, feasibility must be imposed at both decision nodes. Define the set of feasible interventions after observing (m, y) by

$$\mathcal{I}_h^{\mathrm{bud}}(\boldsymbol{x}, B, m, y) := \{i \in \mathcal{I} : g_h(\boldsymbol{x}, m, y, i) \le B\}. \tag{45}$$

Let

$$\mathcal{Y}_{h,m}^{\mathrm{poss}}(\boldsymbol{x}) := \{y \in \mathcal{Y}_m : O_{h,m}(y \mid d, \boldsymbol{x}) > 0 \text{ for some } d \in \mathcal{D}\}$$

be the set of physically possible outputs of diagnostic m. A diagnostic is admissible only if every possible output leaves at least one feasible intervention:

$$\mathcal{M}_h^{\mathrm{bud}}(\boldsymbol{x}, B) := \left\{ m \in \mathcal{M} : \mathcal{I}_h^{\mathrm{bud}}(\boldsymbol{x}, B, m, y) \neq \varnothing \text{ for every } \; y \in \mathcal{Y}_{h,m}^{\mathrm{poss}}(\boldsymbol{x}) \right\}. \tag{46}$$

These sets enforce feasibility even after diagnostic outputs assigned zero probability by the nominal belief. We require $\mathcal{M}_h^{\mathrm{bud}}(\boldsymbol{x}, B)$ to be nonempty at every planning state reachable under the true, nominal, or hybrid models used below. A zero-expenditure fallback available at every state is a sufficient condition. Without such a completion condition, assigning infinite value only to nominal infeasible states is not enough: an unmodeled transition could still lead to an infeasible state.

The use of all physically possible outputs is deliberately independent of the current belief. Thus the comparator class below is the class defined by these feasible sets, not every policy that happens to satisfy a budget only on true-model histories of positive probability.

I.2 Planning with the remaining budget

Let Π_{bud} be the class of public-history policies that choose only actions in the feasible sets above. The budget-aware planner uses planning states of the form

$$(\boldsymbol{b}_h, \boldsymbol{x}_h, B_h),$$

while retaining the original latent transition rows

$$r = (h, d, \boldsymbol{x}, i).$$

In particular, the budget is not added to the row used for estimating $T_{h,i}^{\star}$.

For a feasible branch (m, y, i) at remaining budget B, write

$$B^+ := B - g_h(\boldsymbol{x}, m, y, i).$$

Starting from

$$\widehat{V}_{H+1}^{\mathrm{bud}} = 0,$$

define, in the same context as the main intervention backup,

$$\widehat{G}_{h+1}^{\mathrm{bud}}(u; B^+) := \begin{cases} \ell(z), & u = (\boldsymbol{x}', 1, z), \\ \widehat{V}_{h+1}^{\mathrm{bud}}\left(\widehat{\Phi}_{h,i}(\bar{\boldsymbol{b}}, \boldsymbol{x}, u), \boldsymbol{x}', B^+\right), & u = (\boldsymbol{x}', 0, \perp). \end{cases}$$

The budget-aware intervention backup is

$$\begin{aligned} \widehat{Q}_h^{I,\mathrm{bud}}(\bar{\boldsymbol{b}}, \boldsymbol{x}, B, m, y, i) := c_{\mathrm{I}}(\boldsymbol{x}, m, y, i) + \Gamma_h(\bar{\boldsymbol{b}}, \boldsymbol{x}, i) \\ + \sum_{u \in \mathcal{U}} \widehat{p}_{h,i}(u \mid \bar{\boldsymbol{b}}, \boldsymbol{x})\, \widehat{G}_{h+1}^{\mathrm{bud}}(u; B^+), \end{aligned} \tag{47}$$

for

$$i \in \mathcal{I}_h^{\mathrm{bud}}(\boldsymbol{x}, B, m, y).$$

The diagnostic backup is

$$\begin{aligned} \widehat{Q}_h^{M,\mathrm{bud}}(\boldsymbol{b}, \boldsymbol{x}, B, m) := c_{\mathrm{M}}(\boldsymbol{x}, m) \\ + \sum_{y \in \mathcal{Y}_{h,m}^{\mathrm{poss}}(\boldsymbol{x})} p_{h,m}(y \mid \boldsymbol{b}, \boldsymbol{x}) \min_{i \in \mathcal{I}_h^{\mathrm{bud}}(\boldsymbol{x}, B, m, y)} \widehat{Q}_h^{I,\mathrm{bud}}\left(\bar{\boldsymbol{b}}^{m,y}, \boldsymbol{x}, B, m, y, i\right), \end{aligned} \tag{48}$$

and

$$\widehat{V}_h^{\mathrm{bud}}(\boldsymbol{b}, \boldsymbol{x}, B) = \min_{m \in \mathcal{M}_h^{\mathrm{bud}}(\boldsymbol{x}, B)} \widehat{Q}_h^{M,\mathrm{bud}}(\boldsymbol{b}, \boldsymbol{x}, B, m). \tag{49}$$

Let $\widehat{\pi}^{\mathrm{bud}}$ be an ϵ_{plan}-optimal policy for this recursion.

I.3 Preservation of the Statistical Guarantee Under Budget Augmentation

The next budget is a known deterministic function of the current branch. Appending that coordinate leaves total-variation transition error unchanged, so the same confidence radii and comparator argument apply.

Proposition 4 (Budget-feasible policy guarantee)**.** *Suppose the feasible-completion condition above holds, the reachable budget set $\mathcal{B}$ is finite, the expenditure function g_h is known, and the diagnostic and transition kernels do not depend on the remaining budget except through the harness actions selected. Then the transition-confidence results of the main analysis remain unchanged. Moreover, on the transition-confidence event, every comparator $\pi^\circ \in \Pi_{\mathrm{bud}}$ satisfies*

$$J_{T^\star}(\widehat{\pi}^{\mathrm{bud}}) - J_{T^\star}(\pi^\circ) \le 2\mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}},$$

where $\mathcal{B}_{\widehat{T}}$ is the same nominal transition-uncertainty functional as in the main text.

Consequently, if

$$\pi^\star_{\mathrm{bud}} \in \arg\min_{\pi \in \Pi_{\mathrm{bud}}} J_{T^\star}(\pi),$$

then

$$J_{T^\star}(\widehat{\pi}^{\mathrm{bud}}) - J_{T^\star}(\pi^\star_{\mathrm{bud}}) \le 2\mathcal{B}_{\widehat{T}}(\pi^\star_{\mathrm{bud}}) + \epsilon_{\mathrm{plan}}.$$

The realized-count bound also applies. The population rate additionally requires independent episodes and finite latent coverage of an optimal budget-feasible comparator, as in Corollary 1.

Proof. We first verify that adding the budget preserves the transition-error bound, then apply the cost comparison and planning argument to feasible policies.

The remaining budget is observable and is a deterministic function of the initial budget and the previously selected diagnostic/intervention branches. It therefore introduces no additional latent variable.

Fix a pre-transition branch $(h, d, \boldsymbol{x}, B, m, y, i)$ and let

$$B^+ = B - g_h(\boldsymbol{x}, m, y, i).$$

The true transition law of the environmental outcome together with the next budget can be written as

$$\widetilde{T}^\star_{h,m,y,i}(d', u, B' \mid d, \boldsymbol{x}, B) := T^\star_{h,i}(d', u \mid d, \boldsymbol{x})\, \mathbf{1}\{B' = B^+\}, \tag{50}$$

and its nominal counterpart is

$$\widehat{\widetilde{T}}_{h,m,y,i}(d', u, B' \mid d, \boldsymbol{x}, B) := \widehat{T}_{h,i}(d', u \mid d, \boldsymbol{x})\, \mathbf{1}\{B' = B^+\}. \tag{51}$$

Although these complete kernels depend on (m, y) through the known deterministic budget coordinate, their unknown environmental component remains the shared intervention kernel. Since the true and nominal models use the same deterministic update,

$$\begin{aligned} &\mathrm{TV}\left(\widetilde{T}^\star_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}, B), \widehat{\widetilde{T}}_{h,m,y,i}(\cdot \mid d, \boldsymbol{x}, B)\right) \\ &\quad = \mathrm{TV}\left(T^\star_{h,i}(\cdot \mid d, \boldsymbol{x}), \widehat{T}_{h,i}(\cdot \mid d, \boldsymbol{x})\right). \end{aligned} \tag{52}$$

Therefore the original radius $\beta_h(d, \boldsymbol{x}, i)$ simultaneously controls every feasible budget and diagnostic branch. No additional confidence radius or union bound over $\mathcal{B}$, m, or y is required.

The Q/K factorization and the selective-reveal concentration argument also remain unchanged. The quantities estimated from the offline data are still the environmental kernels

$$Q^\star_{h,i}(u \mid d, \boldsymbol{x}) \quad \text{and} \quad K^\star_{h,i}(d' \mid d, \boldsymbol{x}, u),$$

and samples may still be pooled across diagnostics, outputs, and budget values. The budget is measurable from the public history and hence adds no information beyond the history already conditioned upon in the conditional-representativeness assumption.

Fix a budget-feasible public-history policy π. In the transition-replacement argument from Step 1 of the proof of Theorem 1, the two processes being compared have the same public history, remaining budget, diagnostic output, and selected intervention before the transition being replaced. Hence they also have the same deterministic next budget B^+. Their only difference is the environmental transition law $T^\star_{h,i}$ versus $\widehat{T}_{h,i}$. Conditional on the common branch, the continuation value may depend on B^+, but its range is still bounded by W_{h+1}. Equation 52 therefore gives the same local bound

$$\Gamma_h(\bar{\boldsymbol{b}}_h, \boldsymbol{x}_h, i_h).$$

Summing the hybrid differences yields

$$\left|J_{T^\star}(\pi) - J_{\widehat{T}}(\pi)\right| \le \mathcal{B}_{\widehat{T}}(\pi) \qquad \text{for every } \pi \in \Pi_{\mathrm{bud}}.$$

The recursion in equation (47)–equation (49) is the Bellman recursion for

$$J_{\widehat{T}}(\pi) + \mathcal{B}_{\widehat{T}}(\pi)$$

minimized over Π_{bud}. Indeed, the budget update is known and deterministic, and the feasible action sets enforce equation 44 pathwise. Consequently,

$$J_{\widehat{T}}(\widehat{\pi}^{\mathrm{bud}}) + \mathcal{B}_{\widehat{T}}(\widehat{\pi}^{\mathrm{bud}}) \le J_{\widehat{T}}(\pi^\circ) + \mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}}$$

for every $\pi^\circ \in \Pi_{\mathrm{bud}}$. Applying the policy-wise simulation bound first to $\widehat{\pi}^{\mathrm{bud}}$ and then to π° gives

$$J_{T^\star}(\widehat{\pi}^{\mathrm{bud}}) - J_{T^\star}(\pi^\circ) \le 2\mathcal{B}_{\widehat{T}}(\pi^\circ) + \epsilon_{\mathrm{plan}}.$$

Taking $\pi^\circ = \pi^\star_{\mathrm{bud}}$ proves the second claim.

Finally, because the statistical row remains $r = (h, d, \boldsymbol{x}, i)$, the realized counts, confidence radii, and coverage coefficients are exactly those of the main analysis, evaluated under the budget-feasible comparator. Only the planning state space grows by the finite factor $|\mathcal{B}|$. For each fixed $(\boldsymbol{x}, B)$, the alpha-vector argument also remains valid, so the budget-aware value function is piecewise linear and concave in the belief. □

Proposition 4 shows that enforcing a hard budget does not require a separate learned transition model for each remaining budget. For the same $(h, d, \boldsymbol{x}, i)$, the unknown environmental transition is unchanged; only the known resource balance and the permitted actions differ. The planner can therefore reuse the same labeled transitions and confidence radii while tracking the remaining budget. Feasibility must hold after every physically possible diagnostic output, including outputs assigned zero probability by the estimated belief.

The comparison is with the best policy in the specified budget-feasible class. For a fixed log, the counts and radii are unchanged. The budget-feasible comparator can visit different latent rows, however, so its coverage coefficients and the resulting numerical bound may differ. The conclusion also relies on a known budget update, feasible completion at every relevant state, and environmental and diagnostic kernels that do not otherwise depend on the remaining budget.

J Synthetic Experiment Details

This section specifies the environment, offline logs, and evaluation procedures for the synthetic results in the main text. We first describe the common workflow and then explain the settings used for the overall comparison, sample-size study, persistent-damage study, and sharing comparison. The final subsection adds catastrophe statistics that are not shown in the main-text tables and figure.

J.1 Canonical workflow

State and episode structure. An episode has at most six decision steps. The observed state records workflow progress and the current proposal type, which is either safe or risky. The hidden damage state is $d_h \in \{C, R, K\}$, denoting clean, recoverable, and critical states. Progress and damage are separate variables: an operation can advance the task while leaving the environment damaged. Each episode starts at progress zero with $d_1 = C$. At each live step, a new proposal is risky with probability 0.25 and safe otherwise.

The harness first chooses a diagnostic from Skip, Cheap, and Reveal, and then an intervention from Execute, Block, Repair, and Rollback. Execute increases progress by one. Block and Repair leave progress unchanged; Rollback reduces it by one, with a minimum of zero. An Execute that reaches progress three completes the task. A catastrophic outcome ends the episode immediately, and an episode still active after the sixth step ends in timeout.

Damage transitions. Table 3 gives the fixed simulator transitions. Each row is a probability distribution over the next damage outcome, including catastrophe. The rows depend on the current damage and intervention, and on the proposal type for Execute. They do not depend on which diagnostic was used. This is the transition-sharing structure tested in the main text. The terms "safe" and "risky" distinguish proposal classes in this simulator; a safe proposal is not a guarantee of a damage-free transition.

Table 3: Fixed damage-transition probabilities in the canonical workflow. The last four columns give the distribution of the next damage outcome.

Intervention / proposal	Current damage	C	R	K	Catastrophe
Execute / safe	C	0.97	0.03	0.00	0.00
	R	0.00	0.88	0.10	0.02
	K	0.00	0.00	0.85	0.15
Execute / risky	C	0.65	0.30	0.05	0.00
	R	0.00	0.50	0.40	0.10
	K	0.00	0.00	0.55	0.45
Block	C	1.00	0.00	0.00	0.00
	R	0.00	1.00	0.00	0.00
	K	0.00	0.00	1.00	0.00
Repair	C	1.00	0.00	0.00	0.00
	R	0.85	0.15	0.00	0.00
	K	0.00	0.65	0.30	0.05
Rollback	C	1.00	0.00	0.00	0.00
	R	0.95	0.05	0.00	0.00
	K	0.60	0.30	0.08	0.02

Diagnostic observations and costs. Skip returns no information. Cheap returns a binary pass/fail result, with failure probabilities 0.10, 0.65, and 0.90 in states C, R, and K, respectively. Reveal returns the current damage label exactly. These observation rules are fixed during an experiment. Table 4 records the diagnostic charges, intervention charges, and terminal losses. Completion is therefore not automatically a good outcome: completing in a damaged state can still incur a substantial loss.

Table 4: Costs and terminal losses in the canonical workflow. Entries in the last column follow the order in the middle column.

Component	Choice or final damage state	Cost / loss
Diagnostic	Skip, Cheap, Reveal	$0,\ 0.5,\ 1.5$
Intervention	Execute, Block, Repair, Rollback	$0,\ 0.5,\ 2,\ 2.5$
Completion	C, R, K	$0,\ 20,\ 70$
Timeout	C, R, K	$40,\ 55,\ 80$
Catastrophe	Catastrophic termination	100

We measure suboptimality relative to the true-model optimum under the corresponding evaluation protocol. All simulator parameters are held fixed when varying the amount of offline data or the availability of latent-state labels.

J.2 Offline logs and selective damage labels

Logging policy. The logger chooses a diagnostic using the proposal-dependent probabilities in Table 5. After observing its output, it draws an intervention from Table 6. These probabilities specify the data-collection policy, not the learned policy. All reported canonical studies use the fixed simulator above; their differences concern data size, label retention, or the estimator and planner being compared.

Table 5: Base diagnostic probabilities used to collect canonical logs. The consecutive-Reveal rule described below can override this draw.

Proposal	Skip	Cheap	Reveal
Safe	0.65	0.25	0.10
Risky	0.35	0.40	0.25

Table 6: Intervention probabilities for the canonical logging policy, conditional on the diagnostic output.

Diagnostic output	Execute	Block	Repair	Rollback
No result or pass	0.90	0.10	0.00	0.00
Fail	0.15	0.25	0.45	0.15
Revealed C	0.90	0.10	0.00	0.00
Revealed R	0.15	0.15	0.50	0.20
Revealed K	0.05	0.15	0.30	0.50

Which labels are recorded. Reveal supplies the current damage label. A destination label is available when the next diagnostic reveals the state, or when a noncatastrophic terminal outcome records the final damage state. After a nonterminal Reveal, the logger forces another Reveal at the next step with probability 0.5. This creates consecutive source–destination labels without labeling every transition. The episode begins in the known clean state.

The number N counts independently simulated episodes. It is not a count of labeled transitions: early completion and catastrophe change the number of live steps, and selective diagnostics change how many of those steps are labeled. Each dataset replicate has its own deterministic random seed. Unless stated otherwise, synthetic results are averaged over 20 independently generated offline datasets, and $\pm$ denotes the standard deviation across these dataset replicates.

Reveal-label masking. For the cross-diagnostic sharing experiment, we first generate a fixed offline dataset using the canonical logging policy described above. This logging policy determines when the Reveal diagnostic is selected and, consequently, which latent damage labels are recorded in the dataset. We then vary label availability by applying a nested post-processing mask to these recorded labels: each available label is retained with probability p_{label}. Thus, p_{label} controls the fraction of already recorded damage labels exposed to the transition estimators; it does not change how often Reveal is selected during data collection. In particular, $p_{\mathrm{label}} = 1$ means that all labels recorded in the original dataset are retained. The underlying trajectories, actions, transition outcomes, costs, and public histories are identical across all values of p_{label}.

J.3 Baselines

Oracle. The Oracle is a true-model belief planner used to define the zero-suboptimality reference. It knows the true transition and diagnostic kernels but receives the same observations as the learned policies: the current latent damage state is not provided for free, and informative diagnostics still incur their usual costs. We solve the resulting finite-horizon POMDP exactly with alpha-vector dynamic programming.

Plug-in planner. The plug-in planner is the main structured non-pessimistic comparison. For each proposal–intervention pair, transition counts are pooled across diagnostic choices and outputs but not

across different proposals or interventions. The estimator uses an empirical-Bayes pseudo-count of total strength one, with its distribution obtained by pooling the available damage rows within the same proposal–intervention primitive. The fitted transition center is then used directly by the exact alpha-vector planner, without an uncertainty penalty.

RA-PEVI. RA-PEVI uses the same latent-state representation, estimated transition center, observation model, horizon, and alpha-vector planner as the plug-in baseline. The difference is the uncertainty-aware planning penalty. Thus, the RA-PEVI versus plug-in comparison keeps the estimated center fixed and changes how transition uncertainty enters planning.

Myopic router. The myopic router is a deterministic policy designed to isolate the importance of persistent historical damage. On a safe proposal it skips diagnosis and executes immediately. On a risky proposal it selects Reveal and then executes in state C, repairs in state R, and rolls back in state K. The policy therefore responds to the immediate risk of the current proposal but does not carry unresolved damage information forward when the current proposal is safe.

Observable-branch learner. The observable-branch baseline fits dynamics directly at the level of the complete observable branch, indexed by the stage, public workflow state, proposal type, diagnostic choice, diagnostic output, and intervention. It therefore estimates different transition models for different diagnostic branches rather than sharing a latent intervention row across them. Planning uses finite-horizon recursion over this fitted observable model. An unobserved branch–intervention pair is assigned cost 100, matching the largest terminal-loss scale in the canonical workflow. No additional count-dependent pessimistic correction is applied to observed branches.

Recurrent IQL. Recurrent IQL is a model-free offline RL baseline that receives the complete public interaction history but no simulator-private damage labels. Before the current diagnostic result is observed, its input is a 17-dimensional feature vector containing normalized stage and progress, the proposal type, and one-hot encodings of the previous diagnostic, diagnostic output, and intervention. After the current diagnostic result is observed, a 9-dimensional encoding of the current diagnostic and output is appended.

The model uses a one-layer GRU with hidden width 64. It has separate diagnostic and intervention Q heads, separate value heads for the two decision stages, and separate diagnostic and intervention policy heads. Training combines TD regression for both Q functions, expectile regression for the value functions, and advantage-weighted behavior cloning. Because the objective is cost minimization, the actor weight is proportional to $\exp((V - Q)/\tau)$. We use expectile 0.3, temperature 0.1, cost scale 100, Adam with learning rate 3×10^{-3}, batch size 256, 250 behavior-cloning warm-up updates, and 2,000 total updates. Evaluation selects the actor argmax action.

Decision Transformer. Decision Transformer is a history-based sequence-model baseline that also uses only public information. Each workflow step is represented by two causal tokens: a pre-diagnostic token that predicts one of the three diagnostics and a post-diagnostic token that predicts one of the four interventions. The 15-dimensional state encoding contains normalized stage and progress, proposal type, a pre/post decision indicator, the selected diagnostic, and its output. State, return-to-go, previous-action, and positional embeddings are combined before the Transformer.

The model has hidden width 128, three causal Transformer layers, four attention heads, feed-forward width 512, GELU activations, pre-normalization, dropout 0.1, and a maximum sequence length of 12 tokens. A single seven-action output head represents the three diagnostic and four intervention choices. We train with masked action cross-entropy using AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256, and 3,000 updates, with gradient clipping at 0.25. At evaluation, the return-to-go conditioning target is fixed to the true-model Oracle expected cost and is identical across training seeds.

Branch-complete PEVI. For the cross-diagnostic sharing experiment, we use Branch-complete PEVI as the information-matched nonsharing comparison. It receives exactly the same source–destination-labeled transitions as Shared PEVI and uses the same prior strength and pessimistic planning rule. The only difference is the transition parameterization: Branch-complete retains a separate latent transition row for each proposal, diagnostic, output, and intervention branch, whereas

Shared PEVI pools compatible rows across diagnostic branches. Neither method infers missing latent labels from partially observed sequences.

J.4 SETTINGS FOR THE MAIN RESULTS

Overall comparison and sample-size study. Table 1 reports the main comparison at $N = 5,000$. Structured methods use exact alpha-vector planning, and policy performance is evaluated by exact recursion under the true simulator. Figure 1(a) uses the same canonical workflow and varies the number of logged episodes over $N \in \{250, 500, 1000, 2000, 5000, 10000\}$. The sample-size sweep uses a separate set of dataset seeds from the main comparison.

Persistent-damage study. Figure 1(b)–(c) uses a separate $N = 5000$ experiment on the same six-step workflow. We evaluate decisions for which the current proposal is safe while unresolved damage from an earlier interaction is still present. The true latent state is used only to define this evaluation subset and is not provided to the learned policy. The protection rate measures how often the policy takes a protective response in these states, while the unresolved-completion rate measures how often an episode completes with damage still present.

Cross-diagnostic sharing comparison. Table 2 isolates the effect of sharing transition information across diagnostic branches. We fix $N = 2000$ and vary the retained label probability over $p_{\text{label}} \in \{1, 0.5, 0.2, 0.1, 0.05\}$. The logged trajectories, actions, transition outcomes, costs, and public histories are held fixed, and the retained label sets are nested across values of p_{label}. Shared PEVI and Branch-complete PEVI receive the same labeled source–destination transitions, use the same prior strength, and use the same pessimistic planner. The only difference is the transition parameterization: Shared PEVI pools a latent intervention row across diagnostic-output branches, whereas Branch-complete PEVI estimates a separate row for each branch.

This sharing experiment is evaluated on a belief grid with spacing 0.05. It is separate from the canonical sample-size sweep. Branch-complete PEVI is also different from the observable-branch learner in Table 1: Branch-complete PEVI retains the latent-state model but removes cross-diagnostic row sharing, while the observable-branch learner models dynamics directly at the observable branch level.

J.5 ADDITIONAL OUTCOMES

The main figure reports protective behavior and unresolved damage at completion. Table 7 additionally reports catastrophe probability from the same persistent-damage study. RA-PEVI has a lower catastrophe probability than the myopic router, complementing the persistent-damage comparison in Figure 1. The Oracle uses the true simulator and is included only as a reference.

Table 7: **Catastrophic outcomes in the persistent-damage study.** The experiment uses $N = 5000$ logged episodes. Lower probability is better.

Method	Catastrophe probability
Oracle	0.00187
RA-PEVI	0.00312
Myopic router	0.00891

K APPWORLD EXPERIMENT DETAILS

This section gives the full experimental protocol for the AppWorld study in Section 7.

K.1 HARNESS PROTOCOL AND APPLICATION STATE

Tool proposals. The evaluation uses AppWorld as the execution environment for a controlled sequence of file-system operations. Each episode uses a separate directory namespace and contains

eight target positions. Before interaction begins, each position is assigned a binary target-existence state. The harness never observes this state directly.

Execute commits the corresponding directory operation through the AppWorld file-system API. If the target was absent, the execution is treated as an erroneous operation and contributes one unit of unresolved damage. Block suppresses the operation. Repair removes the most recent unresolved erroneous effect through the same native file-system interface. Environment changes are committed within the episode and persist across subsequent decisions.

Diagnose–then–intervene decisions. At every proposal position, the harness makes a two-stage decision. It first chooses between Skip and Diagnose. Diagnose performs a read-only native query of whether the current target directory exists. Its binary result is passed through a controlled noise channel that flips the answer with probability 0.2. After observing the diagnostic result, the policy chooses the intervention to commit. Under Skip, the intervention is chosen directly without an additional query. Execute commits the proposed AppWorld operation. Block suppresses the proposal and leaves the target unchanged. Repair resolves the most recent unresolved effect and consumes the current proposal position.

Hidden application state. At each proposal position, the relevant hidden condition is whether the proposal target still exists. Executing a proposal whose target no longer exists creates one erroneous operation and increases the unresolved-damage count by one; executing a proposal with an existing target does not create new damage. Unresolved errors persist across later proposal positions until they are repaired. Diagnose provides a noisy observation of target existence, while Repair removes one previously unresolved error.

Diagnostic channel and costs. Diagnose calls the target-existence query and receives a binary observation. To make information acquisition nontrivial, we pass this observation through a fixed noisy channel that flips the binary answer with probability 0.2. The noise is part of the experimental protocol and is applied uniformly throughout data collection and evaluation. Diagnose incurs cost 0.02. Execute has zero intervention cost, while Block and Repair cost 0.1 and 0.5, respectively. Each unit of unresolved damage remaining at the end of the episode incurs terminal cost 20.

K.2 Offline data collection

Training tasks and trajectories. The offline corpus contains 720 episodes collected from nine AppWorld training tasks, with 80 episodes per task and eight proposal positions per episode. Seven tasks, containing 560 episodes, form the fitting split for the data-driven baselines, while two tasks, containing 160 episodes, are used for internal hyperparameter selection. The three AppWorld tasks used for policy evaluation are separate from these nine data-collection tasks.

Behavior-policy mixture. We generate the offline trajectories using a mixture of three scripted behavior profiles: Direct, Cautious, and Mixed. Their mixture weights are 0.4, 0.3, and 0.3, respectively. The three profiles are designed to produce different levels of intervention and diagnostic coverage. Direct strongly favors immediate execution, Cautious queries the environment much more often, and Mixed provides broader coverage of the available choices.

For clarity, Table 8 expresses the logging policies using the same two-stage decision structure as the learned harness. The Diagnose column gives the probability of requesting a diagnostic query. The remaining columns give the intervention distribution conditional on skipping diagnosis.

Table 8: **AppWorld logging-policy mixture.** Intervention probabilities are conditional on skipping diagnosis.

Controller	Weight	Diagnose	Execute | Skip	Block | Skip	Repair | Skip
Direct	0.40	0.05	0.789	0.158	0.053
Cautious	0.30	0.55	0.333	0.556	0.111
Mixed	0.30	0.25	0.533	0.333	0.133

After Diagnose, the logging policy commits Execute when the noisy diagnostic outcome indicates that the target is present and Block when it indicates that the target is absent. If Repair is unavailable, it is removed from the feasible intervention set and the remaining logging probabilities are renormalized.

Coverage in the offline data. The 560 fitting episodes contain 4,480 proposal positions. Among them, the logging policies choose Skip 3,230 times and Diagnose 1,250 times. The subsequent committed interventions contain 3,197 Execute decisions, 1,259 Block decisions, and 24 Repair decisions. The dataset therefore contains substantial coverage of both diagnostic choices and the common interventions, while direct Repair observations are comparatively sparse.

K.3 Baselines

Logging policy. We report the scripted data-collection policy as a reference. It follows the behavior mixture above and does not learn from the offline dataset.

Myopic. The myopic baseline estimates the immediate hazard of Execute from the offline data using a support-conditioned Beta–binomial model. At each proposal position, it compares the immediate costs of the available diagnostic and intervention choices, using a recovery-adjusted risk term, but does not model the effect of the current choice on future proposal positions. We use smoothing parameter 0.25 and recovery weight 0.75, selected using the internal-validation tasks.

Plug-in planner. The plug-in baseline is a model-based sequential planner that uses the center estimate of the uncertain execution model. It performs horizon-eight dynamic programming over accumulated unresolved damage and chooses among Execute, Block, and available Repair, without an uncertainty correction.

RA-PEVI. RA-PEVI uses the shared latent-state execution model together with uncertainty-aware sequential planning over diagnostic and intervention decisions. It jointly chooses whether to request Diagnose and, after observing its noisy binary outcome, which intervention to commit. The diagnostic model, costs, and action-feasibility rules are shared with the plug-in planner.

Recurrent IQL. Recurrent IQL provides a model-free history-based comparison. Each proposal position is represented as two successive decisions: a diagnostic decision followed by a committed intervention. Diagnostic and intervention costs are assigned to their corresponding decision events.

The model embeds the public input into a 64-dimensional representation and processes the interaction history with a two-layer GRU of hidden width 128. Separate Q and policy heads are used for the diagnostic and intervention decisions, together with a shared value head. Training uses TD regression for the Q functions, expectile regression for the value function, and advantage-weighted behavior cloning for the policy. A Polyak target network uses coefficient 0.005. Hyperparameter selection gives expectile 0.7, advantage temperature 1, and learning rate 10^{-4}. Training uses 5,000 gradient updates, and unavailable Repair actions are masked at evaluation.

K.4 Evaluation protocol and metrics

Evaluation scenarios. We evaluate on three held-out AppWorld tasks under two target-existence prevalence regimes. In the two regimes, the probability that a proposal target is present is 0.98 and 0.9995, respectively. For each task and regime, we sample 20 fixed binary masks specifying the target-existence condition across the eight proposal positions. The same mask is reused for all methods, giving $3 \times 2 \times 20 = 120$ matched evaluation scenarios.

Episode cost. If N_{diag}, N_{block}, and N_{repair} denote the number of diagnostic queries and committed Block and Repair interventions in an episode, and D_T is its final unresolved-damage count, then the episode cost is $C = 0.02N_{\mathrm{diag}} + 0.1N_{\mathrm{block}} + 0.5N_{\mathrm{repair}} + 20D_T$. Block and Repair counts include interventions selected after a diagnostic observation.

Table 9: **Additional safety statistics.** Harmful Execute is the mean number of harmful executions per episode, and Damage-free is the fraction of episodes ending with zero unresolved damage.

Method	Harmful Execute	Damage-free
Logging policy	0.058	95.0%
Recurrent IQL	0.062	94.7%
Myopic hazard	0.058	99.2%
Plug-in	0.058	99.2%
RA-PEVI	0.042	100.0%

Metrics. Mean cost averages total episode cost across the evaluation scenarios. Worst-10% cost averages the 12 highest episode costs. Terminal damage is the mean amount of unresolved damage remaining at the end of an episode.

We additionally report harmful Execute, the number of executed proposals that produce a harmful intermediate effect; Damage-free, the fraction of episodes ending with no unresolved damage. Harmful Execute can be nonzero even when final damage is zero because later Repair interventions can remove an earlier effect.

As shown in Table 9, RA-PEVI has the lowest harmful-Execute rate among the learned methods, and all evaluated episodes end without unresolved damage.