Active Measuring in Reinforcement Learning With Delayed Negative Effects

Daiqi Gao
Harvard University

Ziping Xu
University of North Carolina at Chapel Hill

Aseel Rawashdeh
Harvard University

Predrag Klasnja
University of Michigan

Susan A. Murphy
Harvard University

Abstract

Measuring states in reinforcement learning (RL) can be costly in real-world settings and may negatively influence future outcomes. We introduce the Actively Observable Markov Decision Process (AOMDP), where an agent not only selects control actions but also decides whether to measure the latent state. The measurement action reveals the true latent state but may have a negative delayed effect on the environment. We show that this reduced uncertainty enables sample-efficient learning and may increase the value of the optimal policy despite these costs. We formulate an AOMDP as a periodic partially observable MDP and propose an online RL algorithm based on belief states. To approximate the belief states, we further propose a sequential Monte Carlo method to jointly approximate the posterior of unknown static environment parameters and unobserved latent states. We evaluate the proposed algorithm in a digital health application, where the agent decides when to deliver digital interventions and when to assess users' psychological status through surveys.

1 INTRODUCTION

Reinforcement learning (RL) in domains such as games often assumes that the states and rewards are fully observable. In many real-world applications, however, measuring the states and rewards can be costly and may affect state transitions. For example, in digital health, an RL algorithm decides when to send interventions to help users alleviate depression. To adapt these interventions effectively, the algorithm needs to measure users' emotions through ecological momentary assessment (EMA) (Targum et al., 2021). Yet frequent assessments impose a burden on users and may reduce user engagement and intervention effectiveness in the longer term. Similar trade-offs arise in other behavioral interventions, such as physical activity suggestions (Choi et al., 2025) and dyadic interventions (Xu et al., 2025), where querying latent psychological states improves decision quality but can also reduce engagement. In robotics, a robot exploring an unknown map may need to activate energy-intensive sensors to improve its understanding of the environment, but doing so drains battery power quickly and may limit the exploration range (Choudhury et al., 2020).

In such problems, actions naturally have two components. Control actions (e.g., sending digital interventions or robot moves) affect environment transitions and are optimized to maximize the cumulative rewards as in the standard RL setting. Measurement actions (e.g., sending surveys or activating sensors) reveal the latent state for better control action decisions, but may negatively affect future states.

We model this class of problems as an Actively Observable Markov Decision Process (AOMDP), which can be viewed as an extension of a Partially Observable Markov Decision Process (POMDP) with an explicit measurement decision. In each decision cycle, the agent selects both a control action and a measurement action. When a measurement action is taken, the emission includes the true latent state, though this can introduce delayed negative effects on future states. The reward in our formulation is a deterministic func-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

tion of the next latent and observed states, so it can be latent when the latent state is not directly observed. This differs slightly from the most common POMDP definitions, where rewards are directly observed. In this setting, learning remains possible since the reward can be inferred through belief updates from emissions, observed states, and measurements. Our framework highlights the tradeoff between the immediate benefits of collapsing state uncertainty and the potential delayed negative impact of measurement.

Main Contributions. First, we propose the AOMDP framework to formalize problems where measuring fully resolves state uncertainty but may negatively affect future states. Under certain assumptions, we show that tabular AOMDPs are learnable with polynomial samples, in contrast to general POMDPs that may require exponential sample complexity. Further, we carefully characterize the trade-off between benefits of state uncertainty reduction and potential negative effects into future states.

Second, we formulate AOMDP as a periodic POMDP with period length two to address the different state and action spaces when choosing the measurement action and the control action. We propose an online, lightweight RL algorithm based on the corresponding periodic belief MDPs for settings with limited data like digital health. It builds on Randomized Least-Squares Value Iteration (RLSVI), separating measurement and control updates and tailoring target constructions for the two actions.

Third, to obtain the belief state in an unknown environment, we develop a sequential Monte Carlo (SMC) method (Del Moral et al., 2006) to approximate the joint posterior of unknown static environment parameters and unobserved stochastic latent states. A key insight is that the observed state can be viewed an emission of the previous latent state, thus helping update the weights of the particle trajectory.

Finally, we apply the proposed algorithm in a digital health application for promoting physical activity, where an RL agent decides when to send interventions and when to query users about their latent commitment to being active.

2 RELATED WORK

Active learning in RL strategically identifies and queries an oracle for information (typically the rewards) at the most informative states to improve policy learning. These methods generally aim to maximize learning efficiency and performance within a finite number of queries, implicitly assuming uniform cost over all queries. Within this field, active reward learning minimizes the number of reward queries while ensuring a near-optimal policy (Daniel et al., 2014; Kong and Yang, 2022). Information-directed reward learning selects a query to provide to the expert at each query time defined by a fixed schedule to maximize the return (Lindner et al., 2021). Active queries in RL from human feedback (RLHF) select the conversations or experts to query in order to increase query efficiency (Das et al., 2024; Ji et al., 2024; Liu et al., 2024).

Active Feature Acquisition (AFA) in RL focuses on dynamically selecting a subset of features to observe, seeking to maximize task performance while minimizing the explicit costs associated with each feature acquisition. For example, Yin et al. (2020) proposed an AFA-POMDP framework that simultaneously learns a task policy and a feature acquisition policy by penalizing the reward with a fixed, predefined cost for every feature queried at each decision time. Alternatively, Li and Oliva (2021) and Yu et al. (2023) utilized MDP frameworks to sequentially select individual features, incurring explicit costs at each step until terminating to make a final prediction. While AFA paradigms typically select a subset of features under immediate explicit costs without affecting state transitions, our framework evaluates whether to query the latent state at each decision time, where the cost is reflected in delayed negative effects on future states.

The works most closely aligned with our problem setting focus on frameworks that explicitly incorporate measurement costs into the RL objective. In these settings, rewards and costs are typically expressed in the same units, enabling joint optimization to balance the reward of the control action and the cost of the measurement action. Although the reward in RL can be viewed as a deterministic function of the next state, there is a difference between actively measuring latent states and measuring latent rewards. When the reward is latent, the probability of measuring will always converge to zero as the estimation of the expected reward becomes more accurate. However, when the state is latent, the probability of measuring may not converge to zero even if the transition model is known or well learned. Due to the stochasticity in state transitions, it is not possible to accurately predict the state needed for selecting the action.

To actively measure the latent state, Nam et al. (2021) formally proposed the Action-Contingent Noiselessly Observable MDPs (ACNO-MDPs) framework, which formulated the problem as a special case of a POMDP. The cost of their measurement action was fixed and observed along with the reward. However, in AOMDP, the negative effect is delayed and incorporated into future states. Further, an ACNO-MDP did not al-

low unobserved rewards or always-passively-observed states and emissions. Nam et al. (2021) proposed algorithms for both tabular and continuous settings, but their deep RL algorithm for continuous settings was not feasible for problems with limited data, e.g., in many digital health applications. Moreover, the estimated transition parameters and latent states were not guaranteed to be drawn from their posterior distributions (see a detailed discussion in Appendix E). Krale et al. (2023, 2024); Avalos et al. (2024) proposed lightweight algorithms for solving ACNO-MDPs, but they only considered tabular settings. In addition to the works on latent states, several works focused on deciding when to measure the latent rewards (Krueger et al., 2016; Schulze and Evans, 2018; Tucker et al., 2023; Parisi et al., 2024). However, we consider the problem with both latent states and latent rewards, with latent rewards modeled as part of the next latent state.

In a mixed observability MDP (MOMDP) (Ong et al., 2010), part of the state is always observed, while the rest is always latent. Our setting is related but differs in that measurement actions can reveal the latent state, and that rewards can be latent since they depend on latent states. Sinha et al. (2024) developed a periodic policy for a POMDP, but their underlying environment is stationary.

3 PROBLEM SETUP

In active measuring, the agent interacts with the environment following the dynamic depicted in Figure 1. At each time step t, the agent observes state Z_t and emission O_t (not the latent state U_t), and decides the measurement action $I_{t,1}$. Taking $I_{t,1} = 1$ will reveal the true latent state U_t, while taking $I_{t,1} = 0$ means that U_t remains latent. We view both O_t and $I_{t,1}U_t$ as emissions of U_t. Then, the agent makes a decision about the control action $A_{t,2}$, and the environment generates the next state $(Z_{t+1}, U_{t+1}) \sim \mathbb{T}(\cdot \mid Z_t, U_t, I_{t,1}, A_{t,2})$ and the next emission $O_{t+1} \sim \mathbb{O}(\cdot \mid U_{t+1})$. The definition of $\mathbb{T}$ guarantees that the transition to the next states Z_{t+1}, U_{t+1} does not depend on the history prior to time t. The reward after taking $A_{t,2}$ is $R_t = r(Z_{t+1}, U_{t+1})$, where r is a known deterministic function. Since the reward depends on the latent state U_t, the reward may also be latent. Note that there is no instantaneous reward at time $(t, 1)$ for $I_{t,1}$ (between taking $I_{t,1}$ and $A_{t,2}$). Thus the effect of $I_{t,1}$ on rewards is only via future states Z_{t+1} and U_{t+1}. While the control action $A_{t,2}$ only affects the transition $\mathbb{T}$, the measurement action $I_{t,1}$ affects both the transition $\mathbb{T}$ and the emission $I_{t,1}U_t$. The observed history before choosing $I_{t,1}$ is $H_{t,1} = \{Z_1, O_1, I_{1,1}, I_{1,1}U_1, A_{1,2}, \ldots, Z_t, O_t\}$.

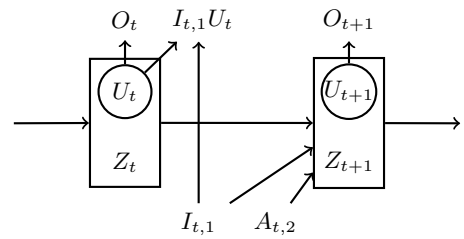

Figure 1: The diagram showing the environment of an AOMDP. A directed edge connected to a square indicates edges to each node within the square. Edges pointing to the control (A_t) and measurement actions (I_t) are omitted.

The observed history before choosing $A_{t,2}$ is $H_{t,2} = H_{t,1} \cup \{I_{t,1}, I_{t,1}U_t\}$.

Definition 1 formally defines active measuring with delayed effects on the environment as an MDP with mixed observable states and special emission structures. For a set $\mathcal{S}$, let $\Delta(\mathcal{S})$ be the set of probability measures on the measurable space $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, where $\mathcal{B}(\mathcal{S})$ is the Borel σ-algebra on $\mathcal{S}$.

Definition 1 (AOMDP)**.** An *Actively Observable MDP* is a tuple $(\mathcal{Z}, \mathcal{U}, \mathcal{O}, \mathcal{A}, \mathcal{I}, \mathbb{T}, \mathbb{O}, r, \gamma)$, where $\mathcal{A}$ is the control action space, $\mathcal{I} = \{0, 1\}$ is the measurement action space, $\mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ is the space of observed states, $\mathcal{U} \subseteq \mathbb{R}^{d_U}$ is the space of latent states, and $\mathcal{O} \subseteq \mathbb{R}^{d_O}$ is the space of emissions of $\mathcal{U}$. The emission function is $\mathbb{O} : \mathcal{U} \mapsto \Delta(\mathcal{O})$, with $\mathbb{O}(o_t \mid u_t)$ being the probability density function (p.d.f.) of O_t given $U_t = u_t$ (overloading notation). The transition function for Z_t and U_t is $\mathbb{T} : \mathcal{Z} \times \mathcal{U} \times \mathcal{I} \times \mathcal{A} \mapsto \Delta(\mathcal{Z} \times \mathcal{U})$, with $\mathbb{T}(z_{t+1}, u_{t+1} \mid z_t, u_t, i_{t,1}, a_{t,2})$ being the p.d.f. of Z_{t+1}, U_{t+1} given $Z_t = z_t, U_t = u_t, I_{t,1} = i_{t,1}, A_{t,2} = a_{t,2}$. The reward function $r : \mathcal{Z} \times \mathcal{U} \mapsto \mathbb{R}$ is a known deterministic function of the next observed and latent states. The discount factor $\gamma \in [0, 1)$ is a constant. We assume that Z_t and O_t are observed before choosing $I_{t,1}$, and that $I_{t,1}U_t$ is observed before choosing $A_{t,2}$.

The goal is to find a policy $\boldsymbol{\pi}$ that selects $I_{t,1}$ and $A_{t,2}$ to maximize the expected discounted sum of rewards, i.e. $\boldsymbol{\pi}^* = \operatorname{argmax}_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}\{\sum_{t=1}^{\infty} \gamma^{t-1} R_t\}$. The positive effect of taking the measurement action $I_{t,1} = 1$ is twofold: (1) learning: it helps learn the transition and emission functions, either directly via SMC or indirectly in a model-free RL algorithm; and (2) optimization: it provides an accurate state for choosing $A_{t,2}$. This second benefit exists even when the environment is known. A setting with truly no reward information (neither directly observed rewards nor inferable latent states) is impossible to solve, since an agent would receive no learning signal. However, in our setting, the

emissions O_t and $I_{t,1}U_t$ together with the next observed state Z_{t+1} help infer the latent state U_{t+1} and thus the latent reward R_t.

3.1 AOMDP and Periodic POMDP

One major difference between AOMDPs and stationary POMDPs is that the state, action, and emission spaces are time-inhomogeneous at the step $(t,1)$ and step $(t,2)$, while being periodic on a higher level index t. This structure is a special case of a periodic POMDP (with possibly latent rewards), which allows non-stationarity within a period but assumes the period structure is homogeneous over time. For conciseness of notation, we use the index $(t, K+1)$ below to refer to the index $(t+1,1)$, and use $(t,0)$ to refer to $(t-1,K)$.

Definition 2 (K-Periodic POMDP)**.** A *K-periodic POMDP* is a tuple $(\mathcal{S}_{1:K}, \mathcal{O}_{1:K}, \mathcal{A}_{1:K}, \mathbb{T}_{1:K}, \mathbb{O}_{1:K}, r_{1:K}, \gamma)$, where $\mathcal{S}_k \subset \mathbb{R}^{d_{S_k}}$ is the latent state space, $\mathcal{O}_k \subset \mathbb{R}^{d_{O_k}}$ is the emission space for $\mathcal{S}_k$, $\mathcal{A}_k$ is the action space, and γ is the discount factor. The transition function for the latent state is $\mathbb{T}_k : \mathcal{S}_{k-1} \times \mathcal{A}_{k-1} \mapsto \Delta(\mathcal{S}_k)$, with $\mathbb{T}_k(s_{t,k} \mid s_{t,k-1}, a_{t,k-1})$ being the p.d.f. of $S_{t,k}$ given $S_{t,k-1} = s_{t,k-1}$ and $A_{t,k-1} = a_{t,k-1}$. The emission function is $\mathbb{O}_k : \mathcal{S}_k \mapsto \Delta(\mathcal{O}_k)$, with $\mathbb{O}_k(o_{t,k} \mid s_{t,k})$ being the p.d.f. of $O_{t,k}$ given $S_{t,k} = s_{t,k}$. The reward function $r_k : \mathcal{S}_{k+1} \mapsto \mathbb{R}$ is a known deterministic function of the next latent state.

The following lemma describes how an AOMDP can be modeled as a Periodic POMDP.

Lemma 1. *An AOMDP is a special case of a periodic POMDP with period length $K = 2$. Further, at time $k = 1$, the emission of the state $S^I_{t,1} = [Z_t, U_t]$ is $O^I_{t,1} = [Z_t, O_t]$, and the reward is zero. At time $k = 2$, the emission of the state $S^A_{t,2} = [Z_t, U_t, I_{t,1}]$ is $O^A_{t,2} = [Z_t, O_t, I_{t,1}U_t, I_{t,1}]$, and the reward is R_t. The discount factors are $\gamma_1 = \gamma_2 = \sqrt{\gamma}$.*

A periodic POMDP can be viewed as a stationary POMDP by augmenting the state with the time index k, similar as how a periodic MDP is viewed as a stationary MDP (Riis, 1965; Sinha et al., 2024). As discussed in Kaelbling et al. (1998), a POMDP can be solved using a belief MDP, whose optimal policy is a Markov stationary policy based on the belief state. Here, a belief state is a probability measure that represents the posterior distribution of the latent state given the observed history, and can be viewed as a sufficient statistic of the history. Incorporating periodicity, the optimal policy of a K-periodic POMDP is a sequence of K Markov policies based on the belief state, and can be solved using a periodic belief MDP.

Definition 3 (K-Periodic Belief MDP)**.** A *K-periodic belief MDP* is a tuple $(\mathcal{B}_{1:K}, \mathcal{A}_{1:K}, \mathbb{T}_{1:K}, r_{1:K}, \gamma)$, where $\mathcal{B}_k = \Delta(\mathcal{S}_k)$ is the set of belief states over the latent state space $\mathcal{S}_k$ at time k in a period, $\mathcal{A}_k$ is the action space, and γ is the discount factor. The transition function for the belief states is $\mathbb{T}_k : \mathcal{B}_{k-1} \times \mathcal{A}_{k-1} \mapsto \mathcal{B}_k$. With $r_k : \mathcal{S}_{k+1} \mapsto \mathbb{R}$ being a known reward function of the next latent state, the reward of the belief state is $r_k(b^S_{t,k+1}) = \int r_k(s) b^S_{t,k+1}(s) ds$.

Concretely, for the measurement action in an AOMDP, the belief state of a latent state $S^I_{t,1} = [Z_t, U_t] \in \mathcal{S}^I$ given the history $H_{t,1}$ is $b^{S^I}_{t,1} \in \mathcal{B}^I$, where $\mathcal{B}^I = \Delta(\mathcal{S}^I)$ is the set of belief states over $\mathcal{S}^I$. In other words, $b^{S^I}_{t,1}(s^I) = p(s^I \mid H_{t,1})$ is the p.d.f. of the posterior distribution of $S^I_{t,1}$ given $H_{t,1}$. When the observed state is $Z_t = z_t$, we have $b^{S^I}_{t,1} = \delta_{z_t} \otimes b^U_{t,1}$, where δ is the Dirac measure, $b^U_{t,1}$ is the belief state of U_t at time $(t,1)$, and $\mu_1 \otimes \mu_2$ denotes the product measure of μ_1 and μ_2. Similarly, for the control action, the belief state of a latent state $S^A_{t,2} = [Z_t, U_t, I_{t,1}] \in \mathcal{S}^A$ given the history $H_{t,2}$ is $b^{S^A}_{t,2} \in \mathcal{B}^A$, where $\mathcal{B}^A = \Delta(\mathcal{S}^A)$ and $b^{S^A}_{t,2}(s^A) = p(s^A \mid H_{t,2})$. When the observed state is $Z_t = z_t$ and the measurement action is $I_{t,1} = i_{t,1}$, we have $b^{S^A}_{t,2} = \delta_{z_t} \otimes b^U_{t,2} \otimes \delta_{i_{t,1}}$, where $b^U_{t,2}$ is the updated belief state of U_t after observing $I_{t,1}U_t$. We will discuss how to estimate the belief state $b^U_{t,1:2}$ of U_t when $\mathbb{T}$ is unknown in Section 4.1. The instantaneous reward after taking $A_{t,2}$ is $r(Z_{t+1}, b^U_{t+1,1}) = r(b^{S^I}_{t+1,1}) = \int r(s) b^{S^I}_{t+1,1}(s) ds$. The understanding of the reward may be updated after taking $I_{t+1,1}$ and observing $I_{t+1,1}U_{t+1}$.

Lemma 1 implies that the AOMDP can be solved as a periodic belief MDP with $K = 2$. Augmenting the state with the time index (Riis, 1965; Sinha et al., 2024), we have that the optimal policy $\boldsymbol{\pi}^*$ of an AOMDP is in the class of periodic Markov policies $\Pi = \{\boldsymbol{\pi} = \{\pi^I, \pi^A\} | \pi^I : \mathcal{B}^I \mapsto \mathcal{I}, \pi^A : \mathcal{B}^A \mapsto \mathcal{A}\}$. The Q-functions of $\boldsymbol{\pi} \in \Pi$ for A and I are defined as

$$\mathcal{Q}^{I\boldsymbol{\pi}}(b^{S^I}_{t,1}, i_{t,1}) := \mathbb{E}^{\boldsymbol{\pi}}\Big\{ \sum_{l=t}^{\infty} \gamma^{l-t} r(b^{S^I}_{l+1,1}) \,\Big|\, b^{S^I}_{t,1}, i_{t,1} \Big\},$$

$$\mathcal{Q}^{A\boldsymbol{\pi}}(b^{S^A}_{t,2}, a_{t,2}) := \mathbb{E}^{\boldsymbol{\pi}}\Big\{ \sum_{l=t}^{\infty} \gamma^{l-t-\frac{1}{2}} r(b^{S^I}_{l+1,1}) \,\Big|\, b^{S^A}_{t,2}, a_{t,2} \Big\}.$$

The Bellman optimality equations for the AOMDP are

$$\mathcal{Q}^{I*}(b^{S^I}_{t,1}, i_{t,1}) = \mathbb{E}\Big\{ \sqrt{\gamma} \max_{a \in \mathcal{A}} \mathcal{Q}^{A*}(b^{S^A}_{t,2}, a) \,\Big|\, b^{S^I}_{t,1}, i_{t,1} \Big\}, \tag{1}$$

$$\begin{aligned} \mathcal{Q}^{A*}(b^{S^A}_{t,2}, a_{t,2}) = \mathbb{E}\Big\{ & r(b^{S^I}_{t+1,1}) + \\ & \sqrt{\gamma} \max_{i \in \mathcal{I}} \mathcal{Q}^{I*}(b^{S^I}_{t+1,1}, i) \,\Big|\, b^{S^A}_{t,2}, a_{t,2} \Big\}, \end{aligned} \tag{2}$$

which is extended from results of periodic belief MDPs.

Algorithm 1 Estimating Belief State $b^U_{t,1}$

Input: history $h_{t,1}$, J particles $\{\widehat{u}^{(j)}_{1:t-1,1:2}\}_{j=1}^J$ with weights $\{w^{(j)}_{t-1,2}\}_{j=1}^J$, and the prior of $\boldsymbol{\theta}$.

1: **for** $j \in \{1:J\}$ **do**
2: Draw $\widehat{\boldsymbol{\theta}}^{(j)}_t \sim p(\boldsymbol{\theta} \mid \widehat{u}^{(j)}_{1:t-1,2}, h_{t-1,2})$, $\widetilde{u}^{(j)}_{t,1} \sim p(u_t \mid z_t, \widehat{\boldsymbol{\theta}}^{(j)}_t, \widehat{u}^{(j)}_{t-1,2}, z_{t-1}, i_{t-1,1}, a_{t-1,2})$.
3: Particle weight: $\widetilde{w}^{(j)}_{t,1} \propto w^{(j)}_{t-1,2} p(z_t \mid \widehat{\boldsymbol{\theta}}^{(j)}_t, \widetilde{u}^{(j)}_{t-1,2}, z_{t-1}, i_{t-1,1}, a_{t-1,2}) p(o_t \mid \widehat{\boldsymbol{\theta}}^{(j)}_t, \widetilde{u}^{(j)}_{t,1})$.
4: Calculate the effective sample size ESS $:= [\sum_{j=1}^J (\widetilde{w}^{(j)}_{t,1})^2]^{-1}$. If ESS $< 0.5J$, resample from $\{\widehat{u}^{(j)}_{1:t-1,1:2}, \widetilde{u}^{(j)}_{t,1}\}_{j=1}^J$ with weights $\{\widetilde{w}^{(j)}_{t,1}\}_{j=1}^J$ to obtain J new particles $\{\widehat{u}^{(j)}_{1:t-1,1:2}, \widehat{u}^{(j)}_{t,1}\}_{j=1}^J$ with weights $w^{(j)}_{t,1} = 1/J$ for all j. Otherwise, set $\widehat{u}^{(j)}_{t,1} = \widetilde{u}^{(j)}_{t,1}$ and $w^{(j)}_{t,1} = \widetilde{w}^{(j)}_{t,1}$.
5: **end for**
6: The estimated belief state is $\widehat{b}^U_{t,1}(u) = \sum_{j=1}^J w^{(j)}_{t,1} \delta(u - \widehat{u}^{(j)}_{t,1})$.

4 METHODOLOGY

To learn the optimal Q-function $\mathcal{Q}^*_k$ online, we adapt the RLSVI algorithm (Osband et al., 2016) to the periodic belief MDP based on the Bellman optimality equations (1) and (2). RLSVI is a model-free algorithm that encourages exploration via posterior sampling (see details in Appendix C.2). In order to obtain the belief states $b^{S^I}_{t,1}$ and $b^{S^A}_{t,2}$, we assume parametric transition and emission models in Section 4.1. We decouple the learning of the Q-function from the transition and emission models to enhance robustness against model misspecification, as identifying the sign of the advantage function for binary actions is often more robust than correctly specifying full conditional distributions (see simulations with misspecified transitions in Appendix D.8).

We approximate each optimal Q-function by a linear function of the basis function ϕ, i.e.

$$\begin{aligned} \mathcal{Q}^{I*}(b^{S^I}_{t,1}, I_{t,1}) &= \phi^I(b^{S^I}_{t,1}, I_{t,1})^\top \boldsymbol{\beta}^I, \\ \mathcal{Q}^{A*}(b^{S^A}_{t,2}, A_{t,2}) &= \phi^A(b^{S^A}_{t,2}, A_{t,2})^\top \boldsymbol{\beta}^A, \end{aligned} \tag{3}$$

where ϕ^I and ϕ^A are the basis functions, and $\boldsymbol{\beta}^I$ and $\boldsymbol{\beta}^A$ are the parameters to be learned. For example, the basis can be of linear, polynomial, or Gaussian functions of the state and action. For this choice of basis functions, RLSVI fits a Bayesian linear regression (BLR) model to the target, which is the estimated optimal Q-function.

Algorithm 2 Estimating Belief State $b^U_{t,2}$

Input: history $h_{t,2}$, J particles $\{\widehat{u}^{(j)}_{1:t-1,2}, \widehat{u}^{(j)}_{t,1}\}_{j=1}^J$ with weights $\{w^{(j)}_{t-1}\}_{j=1}^J$.

1: **if** $i_{t,1} = 1$ **then**
2: When $I_{t,1}U_t = u_t$, set $\widehat{u}^{(j)}_{t,2} = u_t$ for all $j \in \{1:J\}$. Update the particle weight $w^{(j)}_{t,2} = w^{(j)}_{t-1,2} p(z_t, u_t \mid \widehat{\boldsymbol{\theta}}^{(j)}_t, \widehat{u}^{(j)}_{t-1,2}, z_{t-1}, i_{t-1,1}, a_{t-1,2})$.
3: **else**
4: Set $\widehat{u}^{(j)}_{t,2} = \widehat{u}^{(j)}_{t,1}$ for each $j \in \{1:J\}$. Update the particle weight $w^{(j)}_{t,2} = w^{(j)}_{t,1}$.
5: **end if**
6: The estimated belief state is $\widehat{b}^U_{t,2}(u) = \sum_{j=1}^J w^{(j)}_{t,2} \delta(u - \widehat{u}^{(j)}_{t,2})$.

4.1 Approximating the Belief State

In this section, we generalize the Particle Belief MDP (PB-MDP) approximation (Lim et al., 2023) to the AOMDP. The PB-MDP approximates the belief states by using SMC to maintain a set of J particles.

Generalizing the PB-MDP approximation first requires addressing the challenge that neither the transition function $\mathbb{T}$ nor the emission function $\mathbb{O}$ is known. To address this, we use parametric models; denote $\boldsymbol{\theta}$ as the joint parameters of the transition and emission functions $\mathbb{T}, \mathbb{O}$. The second challenge is that augmenting the latent state with the static parameter of the environment, $\boldsymbol{\theta}$, in standard particle filtering fails, since the parameter space is only explored in the initial step and its posterior distribution degenerates as time increases (Kantas et al., 2015). We use ideas from particle learning (Storvik, 2002; Carvalho et al., 2010), which samples a new $\boldsymbol{\theta}$ at every step. Our solution enables efficient online SMC with static parameter estimation since, under some working models (detailed in Section D.2), the posterior of $\boldsymbol{\theta}$ given the history and fixed values of the latent states is closed form. Finally, the belief state of U_t is updated at both $(t,1)$ and $(t,2)$.

In the AOMDP, we only need to derive the belief states $b^U_{t,1:2}$ for U_t to construct $b^{S^I}_{t,1}$ and $b^{S^A}_{t,2}$ (discussed in Section 3.1). Following particle learning, at each time t we first draw the parameter $\widehat{\boldsymbol{\theta}}^{(j)}_t$ from its posterior given the observed history $h_{t,k}$ and the value of one particle $\widehat{u}^{(j)}_{1:t-1,2}$ up to time $t-1$, before drawing the new state $U^{(j)}_t$ given $\widehat{\boldsymbol{\theta}}^{(j)}_t$, $h_{t,k}$, and $\widehat{u}^{(j)}_{1:t-1,2}$. In SMC, each particle represents a possible trajectory of latent states $U_{1:t}$ up to the current time t. The marginal posterior of the current latent state U_t given the history $H_{t,1}$ or $H_{t,2}$ is approximated by the empirical distri-

bution of the last state in each particle $\{U_t^{(j)}\}_{j=1}^J$. Detailed implementations are provided in Algorithms 1 and 2, with the full derivation of the sampling scheme in Appendix B.5. Therein, we used $p(\cdot|\cdot)$ to denote the general p.d.f. associated with the conditional distributions defined by the AOMDP transition and emission functions. Denote the approximate posterior distributions from Algorithms 1 and 2 by $\widehat{b}_{t,1}^U$ and $\widehat{b}_{t,2}^U$ respectively. The belief states of $S_{t,1}^I$ and $S_{t,2}^A$ are approximated by $\widehat{b}_{t,1}^{S^I} = \delta_{z_t} \otimes \widehat{b}_{t,1}^U$ and $\widehat{b}_{t,2}^{S^A} = \delta_{z_t} \otimes \widehat{b}_{t,2}^U \otimes \delta_{i_{t,1}}$.

Notice that $p(z_t \mid \boldsymbol{\theta}, u_{t-1}^{(j)}, z_{t-1}, i_{t-1,1}, a_{t-1,2})$ is used to approximate the belief state $\widehat{b}_{t,1}^U$ in Algorithm 1, even though it does not involve U_t. This is because each particle is a draw from the posterior of the latent state trajectory. Z_t acts as an emission of the latent state trajectory. Indeed a small likelihood of Z_t indicates that the previous latent state value $U_{t-1}^{(j)}$ is less likely to be the true value, and this trajectory should therefore be down-weighted. Furthermore, in Algorithm 2, when $i_{t,1} = 1$, the particle weight $w_{t,2}^{(j)}$ is updated directly from the weight $w_{t-1,2}^{(j)}$ of the previous period. Specifically, we calculate the joint likelihood of Z_t and U_t given $[U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}] = [\widehat{u}_{t-1,2}^{(j)}, z_{t-1}, i_{t-1,1}, a_{t-1,2}]$ and $\widehat{\boldsymbol{\theta}}_t^{(j)}$. This prevents the weights $\{w_{t,2}^{(j)}\}_{j=1}^J$ from collapsing to zero if the true value u_t is not contained in the support of $\widehat{b}_{t,1}^U$ at time $(t, 1)$. Lastly, Algorithm 1 resamples the whole trajectory when the ESS is small for better numerical stability (Liu and Chen, 1998; Del Moral et al., 2006).

4.2 Constructing the Target

Motivated by the fact that the control and measurement actions have different effects on the environment, we construct the targets for them differently.

For the measurement action I, when $b_{t,1}^{S^I} = \delta_{z_t} \otimes b_{t,1}^U$, the target in Equation (1) can be rewritten as

$$\mathcal{Q}^{I*}(b_{t,1}^{S^I}, 1) = \sqrt{\gamma} \int \max_{a \in \mathcal{A}} Q^{A*}(\delta_{z_t} \otimes \delta_{u_t} \otimes \delta_1, a) b_{t,1}^U(u_t) du_t, \tag{4}$$

$$\mathcal{Q}^{I*}(b_{t,1}^{S^I}, 0) = \sqrt{\gamma} \max_{a \in \mathcal{A}} Q^{A*}(\delta_{z_t} \otimes b_{t,1}^U \otimes \delta_0, a). \tag{5}$$

Note that (4) is an integration over the distribution of emission $I_{t,1}U_t$. This is derived based on the known transition function from $b_{t,1}^U$ to $b_{t,2}^U$ to increase numerical stability (see the proof in Appendix B.4). When $b_{t,1}^{S^I} = \delta_{z_t} \otimes b_{t,1}^U$ and $I_{t,1} = 1$, the emission $I_{t,1}U_t$ has a p.d.f. $b_{t,1}^U$. Given $I_{t,1}U_t = u_t$, we have $b_{t,2}^{S^A}(s^A) = \delta_{z_t} \otimes \delta_{u_t} \otimes \delta_1$. When $b_{t,1}^{S^I} = \delta_{z_t} \otimes b_{t,1}^U$ and $I_{t,1} = 0$, we have $P(I_{t,1}U_t = 0) = 1$ and $b_{t,2}^{S^A}(s^A) = \delta_{z_t} \otimes b_{t,1}^U \otimes \delta_0$. This is different from the standard RLSVI (Osband et al., 2016), which approximates the target with a single observation of the next state to increase computational efficiency when the transition is unknown. In addition, we use double Q-learning to alleviate the maximization bias in the targets. This is essential for the active measuring target. Notice that if there were no delayed effect of measurement actions, the second benefit of measuring (obtaining an accurate state) comes exactly from the difference between (4) and (5) (see details in Proposition 3), which could be significantly overestimated due to maximization bias. Based on the above discussion, we define the target in BLR for the measurement action as follows. When $I_{l,1} = 1$, define

$$Y_l^I = \sqrt{\gamma} \sum_{j=1}^J w_{t,1}^{(j)} [\phi^A(\delta_{z_t} \otimes \delta_{\widehat{u}_{t,1}^{(j)}} \otimes \delta_1, a^{(j)})^\top \widetilde{\boldsymbol{\beta}}_{t-1}^A],$$
$$\text{where } a^{(j)} = \operatorname*{argmax}_{a \in \mathcal{A}} [\phi^A(\delta_{z_t} \otimes \delta_{\widehat{u}_{t,1}^{(j)}} \otimes \delta_1, a)^\top \widetilde{\boldsymbol{\beta}}_{t^-}^A],$$

since $\widehat{b}_{t,1}^U(u) = \sum_{j=1}^J w_{t,1}^{(j)} \delta(u - \widehat{u}_{t,1}^{(j)})$. When $I_{l,1} = 0$, define

$$Y_l^I = \sqrt{\gamma} \phi^A(\delta_{z_t} \otimes \widehat{b}_{t,1}^U \otimes \delta_0, a')^\top \widetilde{\boldsymbol{\beta}}_{t-1}^A,$$
$$\text{where } a' = \operatorname*{argmax}_{a \in \mathcal{A}} [\phi^A(\delta_{z_t} \otimes \widehat{b}_{t,1}^U \otimes \delta_0, a)^\top \widetilde{\boldsymbol{\beta}}_{t^-}^A].$$

Here, $\widetilde{\boldsymbol{\beta}}_{t-1}^A$ is the estimated parameter at time $t-1$, and $\widetilde{\boldsymbol{\beta}}_{t^-}^A$ is copied from $\widetilde{\boldsymbol{\beta}}_t^A$ every C steps.

For the control action A, the target is constructed based on a *2-step TD prediction* (Sutton and Barto, 2018, Chapter 7). This allows the target of the control policy to be updated based on itself rather than the measurement policy, which we find improves numerical stability as the instantaneous reward for the measurement action is zero. Define the target as

$$Y_l^A = r(Z_{l+1}, \widehat{b}_{l+1,1}^U) + \gamma \phi^A(\widehat{b}_{l+1,2}^{S^A}, a')^\top \widetilde{\boldsymbol{\beta}}_{t-1}^A,$$
$$\text{where } a' = \operatorname*{argmax}_{i \in \mathcal{I}} \phi^A(\widehat{b}_{l+1,2}^{S^A}, i')^\top \widetilde{\boldsymbol{\beta}}_{t^-}^A.$$

Similarly, $\widetilde{\boldsymbol{\beta}}_{t^-}^A$ is copied from $\widetilde{\boldsymbol{\beta}}_t^A$ every C steps. The reward can be estimated as $r(Z_{t+1}, \widehat{b}_{t+1,1}^U) = \sum_{j=1}^J w_{t+1,1}^{(j)} r(Z_{t+1}, \widehat{u}_{t+1,1}^{(j)})$ by definition.

Let $\boldsymbol{X}_l^I = \phi^I(\widehat{b}_{l,1}^{S^I}, I_{l,1})$ and $\boldsymbol{X}_l^A = \phi^A(\widehat{b}_{l,2}^{S^A}, A_{l,2})$ be the covariates in the BLR for $l \in \{1 : t\}$. The construction of ϕ^I and ϕ^A is problem-dependent. Appendix D.1 describes how ϕ^I and ϕ^A are constructed for the application in Section 6. For example, when the belief state is approximately normal, we can use the weighted mean and standard deviation of the particles $\{\widehat{u}_{t,k}^{(j)}\}_{j=1}^J$, $k = 1$ or 2, along with the observed state Z_t to construct the basis functions.

To update the parameter $\boldsymbol{\beta}^I$, we fit a BLR on $\mathbf{Y}^I := [Y^I_{1:t}]^\top$ using $\mathbf{X}^I := [\boldsymbol{X}^I_{1:t}]^\top$. Similarly, for $\boldsymbol{\beta}^A$, we fit a BLR on $\mathbf{Y}^A = [Y^A_{1:t}]^\top$ using $\mathbf{X}^A = [\boldsymbol{X}^A_{1:t}]^\top$. Following Osband et al. (2016), Gaussian priors are chosen for computationally tractable posterior sampling, which also guarantees stochastically optimistic Q-values. The posterior of $\boldsymbol{\beta}^I_t$ is $N(\boldsymbol{\mu}^I_t, \boldsymbol{\Sigma}^I_t)$, and the posterior of $\boldsymbol{\beta}^A_t$ is $N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$, where

$$\begin{aligned}
\boldsymbol{\Sigma}^I_t &= [(\mathbf{X}^I_t)^\top \mathbf{X}^I_t/(\sigma^I)^2 + \lambda^I \mathbf{I}]^{-1}, \\
\boldsymbol{\mu}^I_t &= \boldsymbol{\Sigma}^I_t[(\mathbf{X}^I_t)^\top \mathbf{Y}^I_t/(\sigma^I)^2], \\
\boldsymbol{\Sigma}^A_t &= [(\mathbf{X}^A_t)^\top \mathbf{X}^A_t/(\sigma^A)^2 + \lambda^A \mathbf{I}]^{-1}, \\
\boldsymbol{\mu}^A_t &= \boldsymbol{\Sigma}^A_t[(\mathbf{X}^A_t)^\top \mathbf{Y}^A_t/(\sigma^A)^2],
\end{aligned} \tag{6}$$

and $\mathbf{I}$ is the identity matrix. An estimate of $\boldsymbol{\beta}^I_t$ is then obtained by drawing $\widetilde{\boldsymbol{\beta}}^I_t \sim N(\boldsymbol{\mu}^I_t, \boldsymbol{\Sigma}^I_t)$ from the posterior distribution. Similarly, $\widetilde{\boldsymbol{\beta}}^A_t \sim N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$ is drawn. Finally, $I_{t,1}$ and $A_{t,2}$ are selected by maximizing the estimated Q-functions based on $\widetilde{\boldsymbol{\beta}}^I_t$ and $\widetilde{\boldsymbol{\beta}}^A_t$. See Algorithm 3 in Appendix C.

5 BENEFITS OF MEASURING

In this section, we discuss the benefits of measurement actions from two aspects—sample complexity benefits and policy improvement benefits—even though they may have negative delayed effects on future cumulative rewards. First, measuring may reduce the sample complexity of learning, leading to a more identifiable environment, as it reveals the latent state and thus removes a major source of uncertainty. Second, measuring can directly increase the value of the optimal policy by providing state information that enables better decisions in subsequent steps.

5.1 Sample Complexity Improvement

We first consider the impact of measuring on the number of samples required to learn an ϵ-optimal policy in an unknown environment. In general, learning in POMDPs can be fundamentally challenging: without further assumptions, the sample complexity may grow exponentially with the horizon H. In finite-horizon tabular POMDPs with always observed rewards, Liu et al. (2022) introduced the notion of an m-step α-weakly revealing POMDP (Definition 4), which requires that the latent states can be distinguished through m-step observations and actions, and thus the system becomes strongly identifiable. They showed that under the m-step α-weakly revealing condition, there exists an algorithm that learns an ϵ-optimal policy with $\text{poly}(S, A^m, H)$ samples, where S is the number of latent states, A is the action space size, and H is the horizon length.

Definition 4 (m-step α-weakly revealing condition)**.** There exist $m \in \mathbb{N}$ and $\alpha > 0$ such that $\sigma_S(M) \geq \alpha$, where for all $(\mathbf{a}, \mathbf{o}) \in \mathcal{A}^{m-1} \times \mathcal{O}^m$ and $s \in \mathcal{S}$,

$$[M_t]_{(\mathbf{a},\mathbf{o}),s} := \mathbb{P}\left(o_{t:t+m-1} = \mathbf{o} \mid s_t = s, a_{t:t+m-2} = \mathbf{a}\right).$$

Here, $\sigma_S(M_t)$ denotes the S-th singular value of the m-step emission matrix M_t.

The following proposition shows that the presence of measurements guarantees a strong form of this condition in the special case of a tabular AOMDP with observed rewards.

Proposition 2. *Any finite-horizon tabular AOMDP with the reward $r(Z_{t+1})$ depending only on the observed state satisfies the 2-step 1-weakly revealing condition.*

This result guarantees that there always exists an algorithm that learns an ϵ-optimal policy with polynomial sample complexity for AOMDPs. Any AOMDP can be learned with sample efficiency, in contrast to general POMDPs without measurement actions.

5.2 Optimal Policy Value Improvement

Beyond the learning efficiency benefits, we consider the case where the environment dynamics are fully known. Even in this setting, measurement actions can strictly increase the value of the optimal policy by reducing uncertainty about the latent state U_t, thereby enabling more informed control decisions. Specifically, a measurement collapses the belief distribution $b^U_{t,1}$ (the posterior over U_t before measuring) to a Dirac measure at the true latent state, while not measuring forces the agent to act under uncertainty.

To clearly characterize the improvement, we define $\mathcal{V}^{A*}_{z,i}(b^U) := \max_a \mathcal{Q}^{A*}(\delta_z \otimes b^U \otimes \delta_i, a)$ for $i \in \{0, 1\}$ as the optimal value function of the control action under the measurement action being 0 and 1, respectively.

Proposition 3. *At belief state $b^{S^I} := \delta_z \otimes b^U$, the advantage of measuring is*

$$\begin{aligned}
&\mathcal{Q}^{I*}(b^{S^I}, 1) - \mathcal{Q}^{I*}(b^{S^I}, 0) \\
=&\underbrace{\mathbb{E}_{b^U}[\mathcal{V}^{A*}_{z,1}(\delta_u) - \mathcal{V}^{A*}_{z,0}(\delta_u)]}_{\text{①Delayed effect}} + \underbrace{\mathbb{E}_{b^U}[\mathcal{V}^{A*}_{z,0}(\delta_u)] - \mathcal{V}^{A*}_{z,0}(b^U)}_{\text{②Immediate effect}}.
\end{aligned}$$

Proposition 3 decomposes the advantage of measuring into two components: the immediate effect (②), which arises from removing uncertainty in the current decision and is always nonnegative by Jensen's inequality; and the delayed effect (①), which reflects how measuring affects the distribution of the next-step state and may negatively impact future rewards. Further suboptimality bounds would require problem-dependent assumptions about the positive and negative effects of measurement actions.

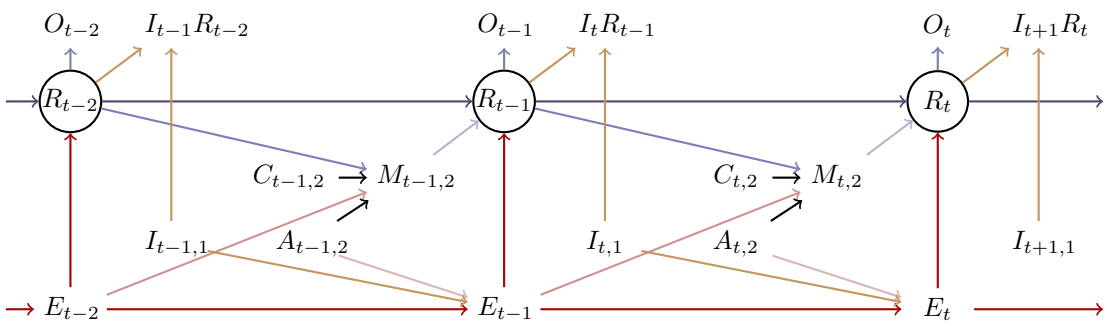

Figure 2: Causal DAG of HeartSteps. Arrows pointing to the actions are omitted.

In the special case where $\mathcal{V}_{z,1}^{A*} = \mathcal{V}_{z,0}^{A*}$ (that is, measuring does not affect the environment), the delayed effect vanishes and the advantage is always nonnegative. In this regime, measuring strictly improves the policy value.

6 APPLICATION

We apply the proposed algorithm to HeartSteps, a digital intervention designed to help users increase and maintain physical activity (PA) levels by increasing users' commitment to being active (Choi et al., 2025). Figure 2 shows a simplified causal directed acyclic graph (DAG) for HeartSteps, which represents the strongest causal relations among the variables. Each time t represents a day. The reward R_t is the user's commitment to being active on day t, which is measured through user surveys and can be unobserved when no survey is sent. The measurement action I_t is whether to send a survey to query the user about R_{t-1}, trading off information gain about the reward and the decrease in engagement. The control action A_t is whether to send a walking-suggestion notification during day t. The engagement E_t captures the negative effects of the two actions. Excessive notifications or surveys reduce E_t and thereby reduce the effectiveness of interventions A_t on R_t. The context C_t represents the evolving needs of the user, e.g., the logarithm of the prior 30-minute step count before the intervention time. The proximal outcome M_t is a mediator between A_t and R_t, e.g., the logarithm of the post 30-minute step count. The emission O_t can be the number of daily unprompted bouts of PA. Key factors that affect long-term behavior change are R_t and E_t. We utilize the public simulation testbed developed by Gao et al. (2025). It contains 42 different environments, constructed from the data on each of the 42 participants in HeartSteps. We can show that Figure 2 is a special case of an AOMDP (see Appendix D.1). Implementation details are provided in Appendices D.2-D.4.

We compare the proposed active-measure algorithm with the always-measure and never-measure algorithms, which always take $I_{t,1}$ to be one or zero and choose $A_{t,2}$ with RLSVI. It is also compared against Dyna-ATMQ (Krale et al., 2023), which focuses on discrete states and assumes a fixed measurement cost observed together with the reward. ATMQ learns the negative effect only from the observed cost. Therefore, when implementing ATMQ, we discretize the states and treat the cost as a tuning parameter. The details of the baseline algorithms are provided in Appendices D.6 and D.7.

We consider different scenarios and report the average cumulative reward $\sum_{t=1}^{T} R_t$ across 42 users (environments), subtracting the average cumulative reward of the zero policy that takes $I_{t,1} = A_{t,2} = 0$ for all t. The experiment is repeated 50 times for each method in each scenario. Figure 3 shows the average and 95% confidence intervals across the 50 replications for the average cumulative reward. The first row represents the scenarios with the minimal positive effect of the interaction $A_{t,2}R_{t-1}$ on the next reward R_t. The influence of $A_{t,2}R_{t-1}$ indicates how important it is to accurately obtain the value of R_{t-1} through measurement when choosing $A_{t,2}$. The second and third rows correspond to scenarios with small and medium positive effect sizes of $A_{t,2}R_{t-1}$ on R_t. The first column represents the scenarios with no negative effect of the measurement action $I_{t,1}$ on the next reward R_t. The second and third columns correspond to scenarios with minimal and small negative effect sizes of $I_{t,1}$ on R_t. The last column has the same settings as the second column but sets the strength of $R_t \to O_t$ to zero, i.e., the passively collected emission O_t no longer helps infer R_t. Details of testbed variants and their effect sizes are provided in Appendix D.5.

When there is no negative effect of the measurement action in Subfigures 3(a), 3(e), and 3(i), the always-measure algorithm performs the best as it does not need to learn to measure, while the active-measure al-

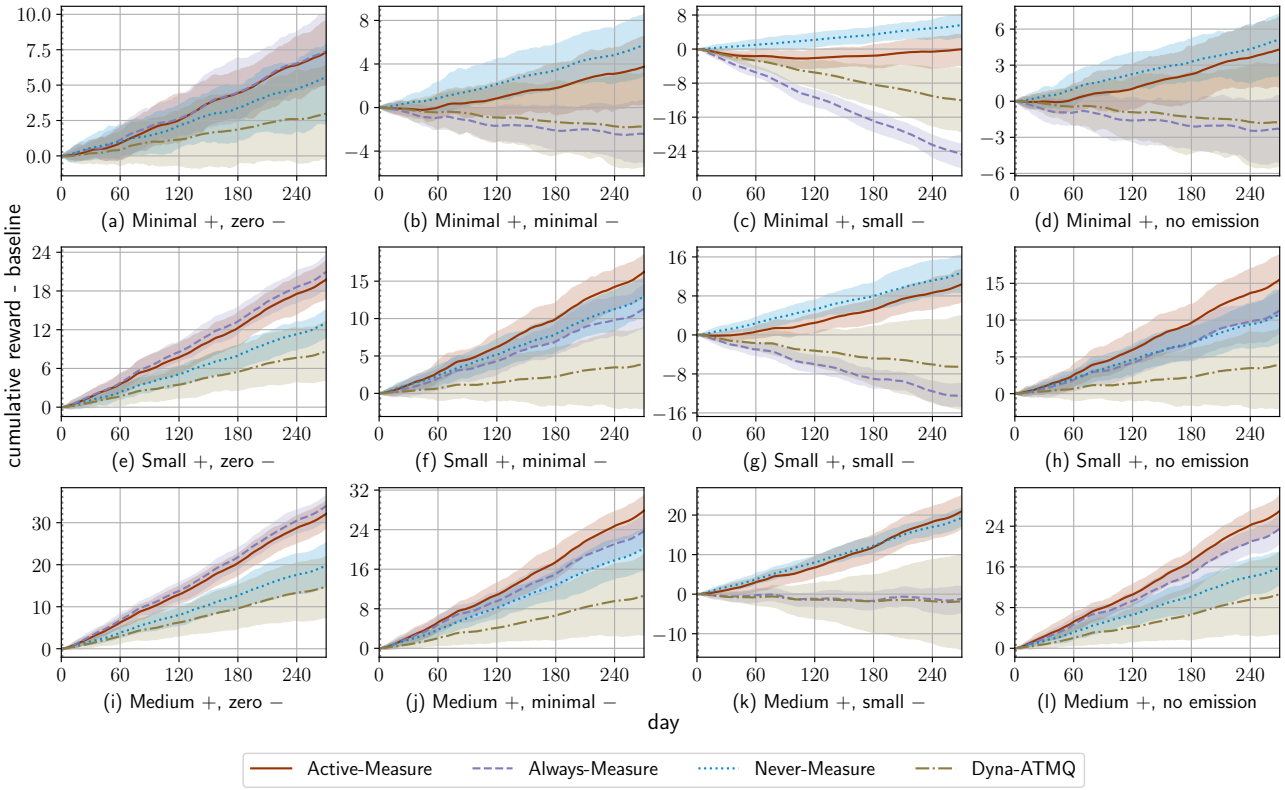

Figure 3: The average cumulative reward, subtracting the average cumulative reward of the zero policy. In each subfigure title, "+" denotes the positive effect of $A_{t,2}R_{t-1} \to R_t$ (rows: minimal, small, medium), and "-" denotes the negative effect of $I_{t,1} \to R_t$ (columns 1–3: zero, minimal, small). The last column ("zero O_t") sets $R_t \to O_t$ to zero, so the passive emission O_t is uninformative about R_t.

gorithm follows closely. When the measurement action has a small negative effect on the reward in Subfigures 3(c), 3(g), and 3(k), the never-measure algorithm performs best as it does not need to learn the negative effect. Active-measure has lower cumulative reward at the beginning, but it gradually picks up the negative signal while learning the transition and emission functions, and its cumulative reward increases faster than never-measure later on. When the measurement action has a small negative effect, with a small or medium positive effect of the control action in Subfigures 3(f) and 3(j), active-measure performs significantly better than the baseline methods. Recall that the environment for Subfigures 3(b) and 3(c) has a minimal effect of $A_{t,2}R_{t-1}$ on the reward. Under these scenarios, the observed states E_{t-1} and $C_{t,2}$ may dominate the decision policy, as there is no need to learn the latent state R_{t-1} perfectly. When a negative effect of $I_{t,1}$ exists, never-measure performs better. An increased effect of $A_{t,2}R_{t-1}$ means that different values of the latent state R_{t-1} may flip the sign of the optimal action $A_{t,2}$, and thus the benefit of measuring is more significant in Subfigures 3(i) and 3(j). Comparing Subfigures 3(j) and 3(l), we see that without an informative emission O_t, the cumulative reward of never-measure decreases significantly. Dyna-ATMQ generally achieves lower cumulative rewards since it requires discretizing the states and cannot detect small changes in the continuous rewards. See Appendix D.8 for additional results.

7 DISCUSSION

The proposed algorithm can be naturally extended to problems with multiple measurement actions or control actions. For example, in HeartSteps, there can be two possible digital interventions per day. Such problems can still fit into the periodic POMDP framework. For the n-step TD prediction used to construct the target of the control action in RLSVI, n now depends on the time of the next nonzero instantaneous reward. In addition, while we focus on settings where the measurement action has no instantaneous reward, the proposed method can incorporate it in the target of the measurement action in RLSVI.

Acknowledgements

This research was funded by NIH grants P50DA054039, P41EB028242, R01HL125440-06A1, UH3DE028723, and P30AG073107-03 GY3 Pilots. Susan Murphy holds concurrent appointments at Harvard University and as an Amazon Scholar. This paper describes work performed at Harvard University and is not associated with Amazon.

References

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Raphaël Avalos, Eugenio Bargiacchi, Ann Nowe, Diederik Roijers, and Frans A Oliehoek. Online planning in pomdps with state-requests. In *Reinforcement Learning Conference*, 2024.

Colin Bellinger, Rory Coles, Mark Crowley, and Isaac Tamblyn. Active measure reinforcement learning for observation cost minimization. In *Canadian Conference on Artificial Intelligence*. Canadian Artificial Intelligence Association (CAIAC), 2021.

Colin Bellinger, Andriy Drozdyuk, Mark Crowley, and Isaac Tamblyn. Balancing information with observation costs in deep reinforcement learning. In *Canadian Conference on Artificial Intelligence*, 2022.

Carlos M Carvalho, Michael S Johannes, Hedibert F Lopes, and Nicholas G Polson. Particle learning and smoothing. *Statistical Science*, 25(1):88–106, 2010.

Soo Ji Serisse Choi, Pei-Yao Hung, Mengyun Liu, Walter Dempsey, Mark W Newman, and Predrag Klasnja. Improving affective associations with physical activity via a message-based mhealth intervention (walktojoy): Proof-of-concept study. *Journal of medical Internet research*, 27:e75792, 2025.

Shushman Choudhury, Nate Gruver, and Mykel J Kochenderfer. Adaptive informative path planning with multimodal sensing. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, pages 57–65, 2020.

Christian Daniel, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. Active reward learning. In *Robotics: Science and systems*, volume 98, 2014.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *International Conference on Machine Learning Workshop on Theoretical Foundations of Foundation Models*, 2024.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

Daiqi Gao, Hsin-Yu Lai, Predrag Klasnja, and Susan Murphy. Harnessing causality in reinforcement learning with bagged decision times. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2025.

Samuel Holt, Alihan Hüyük, and Mihaela van der Schaar. Active observing in continuous-time control. *Advances in Neural Information Processing Systems*, 36:46054–46092, 2023.

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *arXiv preprint arXiv:2402.09401*, 2024.

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.

Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, and Nicolas Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30(3):328–351, 2015.

Dingwen Kong and Lin Yang. Provably feedback-efficient reinforcement learning via active reward learning. *Advances in Neural Information Processing Systems*, 35:11063–11078, 2022.

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, pages 212–220, 2023.

Merlijn Krale, Thiago D Simão, Jana Tumova, and Nils Jansen. Robust active measuring under model uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21276–21284, 2024.

David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost. In *Advances in Neural Information Processing Systems FILM Workshop*, 2016.

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

Michael H Lim, Tyler J Becker, Mykel J Kochenderfer, Claire J Tomlin, and Zachary N Sunberg. Optimality guarantees for particle belief approximation of pomdps. *Journal of Artificial Intelligence Research*, 77:1591–1636, 2023.

David Lindner, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause. Information directed reward learning for reinforcement

learning. *Advances in Neural Information Processing Systems*, 34:3850–3862, 2021.

Jun S Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

HyunJi Alex Nam, Scott Fleming, and Emma Brunskill. Reinforcement learning with state observation costs in action-contingent noiselessly observable markov decision processes. *Advances in Neural Information Processing Systems*, 34:15650–15666, 2021.

Sylvie CW Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, 29(8):1053–1068, 2010.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

Simone Parisi, Alireza Kazemipour, and Michael Bowling. Beyond optimism: Exploration with partially observable rewards. *Advances in Neural Information Processing Systems*, 37:65415–65444, 2024.

Jens Ove Riis. Discounted markov programming in a periodic process. *Operations Research*, 13(6):920–929, 1965.

Sebastian Schulze and Owain Evans. Active reinforcement learning with monte-carlo tree search. *arXiv preprint arXiv:1803.04926*, 2018.

Amit Sinha, Matthieu Geist, and Aditya Mahajan. Periodic agent-state based q-learning for pomdps. *Advances in Neural Information Processing Systems*, 37:62123–62159, 2024.

Geir Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on signal Processing*, 50(2):281–289, 2002.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. Cambridge: MIT press, 2018.

Steven D Targum, Colin Sauder, Miriam Evans, John N Saber, and Philip D Harvey. Ecological momentary assessment as a measurement tool in depression trials. *Journal of psychiatric research*, 136: 256–264, 2021.

Aaron D Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Uncertainty in Artificial Intelligence*, pages 2147–2156. PMLR, 2023.

Ziping Xu, Hinal Jajal, Sung Won Choi, Inbal Nahum-Shani, Guy Shani, Alexandra M Psihogios, Pei-Yao Hung, and Susan A Murphy. Reinforcement learning on dyads to enhance medication adherence. In *International Conference on Artificial Intelligence in Medicine*, pages 490–499. Springer, 2025.

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *Management Science*, 70(6):3841–3855, 2024.

Haiyan Yin, Yingzhen Li, Sinno Jialin Pan, Cheng Zhang, and Sebastian Tschiatschek. Reinforcement learning with efficient active feature acquisition. In *Advances in Neural Information Processing Systems Learning Meets Combinatorial Algorithms Workshop*, 2020.

Zheng Yu, Yikuan Li, Joseph Chahn Kim, Kaixuan Huang, Yuan Luo, and Mengdi Wang. Deep reinforcement learning for cost-effective medical diagnosis. In *International Conference on Learning Representations*, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, see Sections 3 and 4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, see Appendix D.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, see the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, see Section 5.
   (b) Complete proofs of all theoretical results. Yes, see Appendix B.
   (c) Clear explanations of any assumptions. Yes, see Section 5.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, see the supplementary material.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, see Appendix D.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, see Section 6 and Appendix D.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, see Appendix D.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. Yes, see Section 6.

   (b) The license information of the assets, if applicable. Not Applicable.

   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

   (d) Information about consent from data providers/curators. Not Applicable.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. Not Applicable.

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Active Measuring With Delayed Negative Effects in Reinforcement Learning: Supplementary Materials

A Table of Notations

In this section, we summarize the definitions of notation.

Table A.1: Table of Notations.

Notation	Definition
$U_t,\ Z_t,\ O_t$	Latent state, observed state, and emission at time t.
$I_{t,1},\ A_{t,2}$	Measurement action (phase $k=1$) and control action (phase $k=2$).
$r,\ R_t$	Deterministic reward function and realized reward $R_t = r(Z_{t+1}, U_{t+1})$.
$S^I_{t,1},\ S^A_{t,2}$	Latent states for the measurement and control actions.
$H_{t,1},\ H_{t,2}$	Observed histories before selecting $I_{t,1}$ and $A_{t,2}$, respectively.
$b^{S^I}_{t,1} \in \mathcal{B}^I,\ b^{S^A}_{t,2} \in \mathcal{B}^A$	Belief states for $I_{t,1}$ and $A_{t,2}$, with $\mathcal{B}^I = \Delta(\mathcal{S}^I)$ and $\mathcal{B}^A = \Delta(\mathcal{S}^A)$.
$b^U_{t,1},\ b^U_{t,2}$	Belief states of the latent state U_t before and after choosing the measurement action.
$\mathcal{Z}, \mathcal{U}, \mathcal{O}$	Observed-state, latent-state, and emission spaces.
$\mathcal{A}, \mathcal{I} = \{0, 1\}$	Control-action and measurement-action spaces.
$\mathcal{S}^I,\ \mathcal{S}^A$	Latent-state spaces for $S^I_{t,1}$ and $S^A_{t,2}$.
$\mathcal{B}(\mathcal{S})$	Borel σ-algebra on $\mathcal{S}$.
$\Delta(\mathcal{S})$	Set of probability measures on the measurable space $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$.
$\mathbb{T}$	Transition kernel for (Z_{t+1}, U_{t+1}) given $(Z_t, U_t, I_{t,1}, A_{t,2})$.
$\mathbb{O}$	Emission kernel for O_t given U_t.
$d_Z,\ d_U,\ d_O$	Dimensions of observed state, latent state, and emission spaces.
δ_x	Dirac measure concentrated at point x.
$\mu_1 \otimes \mu_2$	Product measure of μ_1 and μ_2.
$\boldsymbol{\pi} = \{\pi^I, \pi^A\}$	Policy for measurement and control actions.
J	Number of particles used in SMC.
$\boldsymbol{\theta}^{(j)}$	The jth particle of the static parameter.
$\widehat{u}^{(j)}_{t,k},\ w^{(j)}_{t,k}$	The jth particle of the latent state and its weight at time t, k.
$\phi^I(\cdot),\ \phi^A(\cdot)$	Basis functions for the Q-functions of I and A.
$\boldsymbol{\beta}^I,\ \boldsymbol{\beta}^A$	BLR coefficients for measurement and control Q-functions.
$\sigma^I,\ \sigma^A$	BLR noise hyperparameters for measurement and control actions.
$\lambda^I,\ \lambda^A$	BLR prior hyperparameters for measurement and control actions.

B Definitions and Proofs

In this section, we define periodic POMDPs and periodic belief MDPs and provide proofs of the theoretical results.

B.1 Periodic POMDP and Periodic Belief MDP

A periodic policy is denoted by $\boldsymbol{\pi} := \{\pi_{1:K}\}$, where each $\pi_k : \mathcal{B}_k \mapsto \mathcal{A}_k$ is a Markov policy. The Q-function of a periodic belief MDP under policy $\boldsymbol{\pi}$ is

$$\mathcal{Q}_k^{\boldsymbol{\pi}}(b_{t,k}^S, a_{t,k}) := \mathbb{E}^{\boldsymbol{\pi}}\left\{ \sum_{(i,j):(i,j)\geq(t,k)} \gamma^{K(i-t)+j-k} \cdot r_j(b_{i,j+1}^S) \,\middle|\, b_{t,k}^S, A_{t,k} = a_{t,k} \right\},$$

where $(i,j) > (t,k)$ means $i = t$ and $j > k$, or $i > t$, and $(i,j) = (t,k)$ means $i = t$ and $j = k$. The value function is defined as $\mathcal{V}_k(b_{t,k}^S) = \mathbb{E}^{\boldsymbol{\pi}}\{\mathcal{Q}_k(b_{t,k}^S, A_{t,k}) \mid b_{t,k}^S\}$. Extending the results from the periodic MDP to the periodic belief MDP, the Bellman optimality equation for the belief state is

$$\mathcal{Q}_k^*(b_{t,k}^S, a_{t,k}) = \mathbb{E}\left\{ r_k(b_{t,k+1}^S) + \gamma_k \max_{a_{t,k+1}\in\mathcal{A}_{k+1}} \mathcal{Q}_{k+1}^*(b_{t,k+1}^S, a_{t,k+1}) \,\middle|\, b_{t,k}^S, A_{t,k} = a_{t,k} \right\} \tag{7}$$

for $k \in \{1:K\}$. The optimal value function is then $\mathcal{V}_k^*(b_{t,k}^S) = \max_{a_{t,k}\in\mathcal{A}_k} \mathcal{Q}_k^*(b_{t,k}^S, a_{t,k})$.

Theoretically, the belief state can be updated as $b_{t,k+1}^S(s_{t,k+1}) = \mathbb{P}(s_{t,k+1}, o_{t,k+1} \mid b_{t,k}^S, a_{t,k})/\mathbb{P}(o_{t,k+1} \mid b_{t,k}^S, a_{t,k})$, where $\mathbb{P}(s_{t,k+1}, o_{t,k+1} \mid b_{t,k}^S, a_{t,k}) = \int_{\mathcal{S}_k} \mathbb{T}_{k+1}(s_{t,k+1} \mid s, a_{t,k})\mathbb{O}_{k+1}(o_{t,k+1} \mid s_{t,k+1}) b_{t,k}^S(s)\, ds$, and $\mathbb{P}(o_{t,k+1} \mid b_{t,k}^S, a_{t,k}) = \int_{\mathcal{S}_{k+1}} \mathbb{P}(s', o_{t,k+1} \mid b_{t,k}^S, a_{t,k})\, ds'$.

Note that the right-hand side of Equation (7) is an expectation over $O_{t,k+1}$, i.e.,

$$\int \left[r_k(b_{t,k+1}^S) + \gamma_k \max_{a_{t,k+1}\in\mathcal{A}_{k+1}} \mathcal{Q}_{k+1}^*(b_{t,k+1}^S, a_{t,k+1}) \right] p(o_{t,k+1} \mid b_{t,k}^S, a_{t,k})\, do_{t,k+1},$$

where $b_{t,k+1}^S$ is a function of $a_{t,k}$, $o_{t,k+1}$, and $b_{t,k}^S$. That is, $b_{t,k+1}^S(s_{t,k+1}) = p(s_{t,k+1} \mid b_{t,k}^S, a_{t,k}, o_{t,k+1})$. The maximization is taken separately for each $o_{t,k+1}$.

B.2 AOMDP as a Periodic POMDP

In this subsection, we prove Lemma 1, which formulates the AOMDP as a 2-periodic POMDP.

Proof. With some overload of notation, recall that the components $(\mathcal{Z}, \mathcal{U}, \mathcal{O}, \mathcal{A}, \mathcal{I}, \mathbb{T}, \mathbb{O}, r, \gamma)$ of an AOMDP are not indexed, whereas the components $(\mathcal{S}_{1:K}, \mathcal{O}_{1:K}, \mathcal{A}_{1:K}, \mathbb{T}_{1:K}, \mathbb{O}_{1:K}, r_{1:K}, \gamma_{1:K})$ of a K-periodic POMDP are indexed by the time step k. In an AOMDP, the variables Z_t, U_t, O_t are indexed only by the period number t, with the actions $I_{t,1}$ and $A_{t,2}$ indexed by both t and k. In contrast, in a K-periodic POMDP, the variables $S_{t,k}, O_{t,k}, A_{t,k}$ are all indexed by both t and k.

Action space. In an AOMDP, there are two decision times in each period t. The measure action $I_{t,1}$ and the control action $A_{t,2}$ are the two actions in a period. The action spaces in the 2-periodic POMDP are $\mathcal{A}_1 = \mathcal{I} = \{0,1\}$ and $\mathcal{A}_2 = \mathcal{A} = \{0,1\}$.

State space. Define the latent state for $I_{t,1}$ as $S_{t,1}^I = [Z_t, U_t]$, and the latent state for $A_{t,2}$ as $S_{t,2}^A = [Z_t, U_t, I_{t,1}]$. The state spaces in the 2-periodic POMDP are $\mathcal{S}_1 = \mathbb{R}^{d_Z+d_U}$ and $\mathcal{S}_2 = \mathbb{R}^{d_Z+d_U} \times \{0,1\}$.

Emission space. Since $O_t \in \mathcal{O}$ and $I_{t,1}U_t \in \mathbb{R}$ is the emission of U_t in an AOMDP, the emissions in the 2-periodic POMDP can be defined as $O_{t,1}^I = [Z_t, O_t]$ and $O_{t,2}^A = [Z_t, O_t, I_{t,1}U_t, I_{t,1}]$. The corresponding emission spaces are $\mathcal{O}_1 = \mathbb{R}^{d_Z} \times \mathcal{O}$ and $\mathcal{O}_2 = \mathbb{R}^{d_Z} \times \mathcal{O} \times \mathbb{R}^{d_U} \times \mathcal{I}$.

Discount factor. Since the instantaneous reward of $I_{t,1}$ is zero and the discount factor on $R_t = r(Z_{t+1}, U_{t+1})$ is γ, we define $\gamma_1 = \gamma_2 = \sqrt{\gamma}$.

Transition function. For the state $S_{t-1,2}^A = s_{t-1,2}^A := [z_{t-1}, u_{t-1}, i_{t-1,1}]$, the action $A_{t-1,2} = a_{t-1,2}$, and the next state $S_{t,1}^I = s^I := [z, u]$, the transition function at time $k = 1$ is defined as

$$\mathbb{T}_1(s^I \mid s_{t-1,2}^A, a_{t-1,2}) = \mathbb{T}(z, u \mid z_{t-1}, u_{t-1}, i_{t-1,1}, a_{t-1,2}).$$

For the state $S^I_{t,1} = s^I_{t,1} := [z_t, u_t]$, the action $I_{t,1} = i_{t,1}$, and the next state $S^A_{t,2} = s^A := [z, u, i]$, the transition function at time $k = 2$ is defined as

$$\mathbb{T}_2(s^A \mid s^I_{t,1}, i_{t,1}) = \delta(z - z_t)\,\delta(u - u_t)\,\mathbb{1}(i = i_t).$$

Emission function. For the state $S^I_{t,1} = s^I_{t,1} := [z_t, u_t]$ and emission $O^I_{t,1} = o^I := [z, o]$, the emission function is

$$\mathbb{O}_1(o^I \mid s^I_{t,1}) = \delta(z - z_t)\,\mathbb{O}(o \mid u_t).$$

Similarly, for the state $S^A_{t,2} = s^A_{t,2} := [z_t, u_t, i_{t,1}]$ and emission $O^A_{t,2} = o^A := [z, o, u, i]$, the emission function is

$$\mathbb{O}_2(o^A \mid s^A_{t,2}) = \delta(z - z_t)\,\mathbb{O}(o \mid u_t)\left[\mathbb{1}(i_{t,1} = 1)\delta(u - u_t) + \mathbb{1}(i_{t,1} = 0)\delta(u - 0)\right]\mathbb{1}(i = i_{t,1}).$$

Reward function. Since the instantaneous reward of $I_{t,1}$ is zero, the reward function at time $k = 1$ is

$$r_1(s^A_{t,2}) = 0 \quad \text{for all } s^A_{t,2} \in \mathcal{S}_2.$$

The reward function at time $k = 2$ is

$$r_2(s^I_{t+1,1}) = r(z_{t+1}, u_{t+1}) \quad \text{for } s^I_{t+1,1} := [z_{t+1}, u_{t+1}].$$

□

B.3 Sample Complexity Improvement

In this subsection, we prove Proposition 2.

Proof. To utilize the conclusion in Liu et al. (2022), we redefine the action as $\widetilde{A}_t = [I_{t,1}, A_{t,2}]$, the state as $\widetilde{S}_t = [Z_{t-1}, U_{t-1}, I_{t-1,1}, A_{t-1,2}, Z_t, U_t]$, and the emission as $\widetilde{O}_t = [Z_{t-1}, O_{t-1}, I_{t-1,1}U_{t-1}, I_{t-1,1}, A_{t-1,2}, Z_t, O_t]$, so that $\widetilde{O}_t$ depends only on $\widetilde{S}_t$. This can be viewed as a special case of solving an AOMDP as a periodic POMDP.

For the measurement action $I_{t,1}$, the state $\widetilde{S}_t$ is equivalent to $S^I_{t,1} = [Z_t, U_t]$, since $Z_{t-1}, U_{t-1}, I_{t-1,1}, A_{t-1,2} \perp\!\!\!\perp Z_{t+1}, U_{t+1} \mid b^{S^I}_{t,1}, I_{t,1}$. For the control action $A_{t,2}$, the state $\widetilde{S}_t$ is a special case of $S^A_{t,2} = [Z_t, U_t, I_{t,1}]$, since $Z_{t-1}, U_{t-1}, I_{t-1,1}, A_{t-1,2} \perp\!\!\!\perp Z_{t+1}, U_{t+1} \mid b^{S^I}_{t,1}, I_{t,1}$ and $\{Z_t, U_t\} \subsetneq S^A_{t,2}$. Selecting $A_{t,2}$ based on $\widetilde{S}_t$ therefore provides a lower bound for the sample complexity of selecting $A_{t,2}$ based on $S^A_{t,2}$.

Now we investigate the submatrix $M_{t,\widetilde{a}}$, corresponding to the rows of M_t where the action is $\widetilde{a}$, i.e., $(M_{t,\widetilde{a}})_{o,s} = [M_t]_{(\widetilde{a},o),s}$. Suppose the sizes of the observed state, latent state, emission, measurement action, and control action are Z, U, O, I, and A, respectively. Then the size of the state space is $\widetilde{S} = Z \times U \times I \times A \times Z \times U$, and the size of the emission space is $\widetilde{O} = Z \times O \times U \times I \times A \times Z \times O$. Thus, the submatrix $M_{t,\widetilde{a}}$ has dimension $\widetilde{O}^2 \times \widetilde{S}$.

When we take $\widetilde{a} = [1, a]$ for any $a \in \mathcal{A}$ (i.e., $I_{t,1} = 1$), we will show that $M_{t,\widetilde{a}}$ contains a square submatrix of size $\widetilde{S} \times \widetilde{S}$ that has full rank. Specifically, denote the row value as $[z_0, o_0, iu_0, i_0, a_0, z_1, o_1, z_1, o_1, iu_1, i_1, a_1, z_2, o_2]$ (the concatenation of emissions $\widetilde{O}_t$ and $\widetilde{O}_{t+1}$), and the column value as $[z'_0, u'_0, i'_0, a'_0, z'_1, u'_1]$ (the state $\widetilde{S}_t$). Consider the submatrix obtained by fixing values of o_0, o_1, o_2, and z_2. Then the size of this submatrix is $\widetilde{S} \times \widetilde{S}$, as only z_0, iu_0, i_0, a_0, z_1, and iu_1 are allowed to vary. This submatrix is diagonal, with diagonal entries equal to one for indices $z_0 = z'_0$, $iu_0 = u'_0$, $i_0 = i'_0$, $a_0 = a'_0$, $z_1 = z'_1$, and $iu_1 = u'_1$. All off-diagonal entries are zero. Therefore, $\sigma_{\widetilde{S}}(M_{t,\widetilde{a}}) = 1$.

According to Proposition 3 of Liu et al. (2022), since $\max_i \sigma_{\widetilde{S}}(M_{t,\widetilde{a}}) \geq 1$ with the maximizer $\widetilde{a} = [1, a]$, the 2-step 1-weakly revealing condition is satisfied. □

B.4 Bellman Optimality Equation for the Measure Action

In this subsection, we prove Proposition 3.

Proof. To derive the target for the measurement action I in RLSVI, first note that the emission distribution of $I_{t,1}U_t$ is

$$p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1}) = \begin{cases} b^U_{t,1}(u_t), & \text{if } I_{t,1} = 1, \\ \delta(u_t - 0), & \text{if } I_{t,1} = 0, \end{cases}$$

where $\delta(\cdot)$ is the Dirac delta function.

When $I_{t,1} = 1$, the p.d.f. of the belief state $b^{S^A}_{t,2}$ at value $s^A = [z, u, i]$ given the previous belief state $b^{S^I}_{t,1} = \delta_{z_t} \otimes b^U_{t,1}$, the last action $I_{t,1} = 1$, and the emission $I_{t,1}U_t = u_t$ is

$$\begin{aligned}
&b^{S^A}_{t,2}(s^A) = p(s^A \mid b^{S^I}_{t,1}, I_{t,1} = 1, I_{t,1}U_t = u_t) \\
=&p_{Z_t}(z \mid b^{S^I}_{t,1}, I_{t,1} = 1, I_{t,1}U_t = u_t) \cdot P(I_{t,1} = i \mid b^{S^I}_{t,1}, I_{t,1} = 1, I_{t,1}U_t = u_t, Z_t = z) \\
&\cdot p_{U_t}(u \mid \delta_{z_t} \otimes b^U_{t,1}, I_{t,1} = 1, I_{t,1}U_t = u_t, Z_t = z, I_{t,1} = i) \\
=&p_{Z_t}(z \mid \delta_{z_t}) \cdot P(I_{t,1} = i \mid I_{t,1} = 1) \cdot \frac{p_{U_t}(u \mid b^U_{t,1}, I_{t,1} = 1)\, p_{I_{t,1}U_t}(u_t \mid U_t = u, I_{t,1} = 1)}{p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1} = 1)} \\
=&\delta(z - z_t)\, \delta(u - u_t)\, \mathbb{1}(i = 1).
\end{aligned}$$

The last equality holds since $p_{I_{t,1}U_t}(u_t \mid U_t = u, I_{t,1} = 1) = \delta(u - u_t)$. Therefore, the transition function from $b^{S^I}_{t,1}$ to $b^{S^A}_{t,2}$ is

$$\begin{aligned}
\mathcal{T}(b^{S^A}_{t,2} \mid b^{S^I}_{t,1}, I_{t,1} = 1) &= \int \delta_{\delta_{z_t} \otimes \delta_{u_t} \otimes \delta_1}\, p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1})\, du_t \\
&= \int \delta_{\delta_{z_t} \otimes \delta_{u_t} \otimes \delta_1}\, b^U_{t,1}(u_t)\, du_t,
\end{aligned}$$

since $p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1}) = b^U_{t,1}(u_t)$ when $I_{t,1} = 1$. Then, based on the Bellman optimality equation (1),

$$\begin{aligned}
\mathcal{Q}^{I*}(b^{S^I}_{t,1}, I_{t,1}) &= \int \left[\sqrt{\gamma} \max_{a_{t,2} \in \mathcal{A}} \mathcal{Q}^{A*}(b^{S^A}_{t,2}, a_{t,2})\right] p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1})\, du_t \\
&= \sqrt{\gamma} \int \max_{a_{t,2} \in \mathcal{A}} Q^{A*}(\delta_{z_t} \otimes \delta_{u_t} \otimes \delta_1, a_{t,2})\, b^U_{t,1}(u_t)\, du_t,
\end{aligned} \tag{8}$$

since the instantaneous reward of $I_{t,1}$ is zero.

On the other hand, when $I_{t,1} = 0$,

$$\begin{aligned}
&b^{S^A}_{t,2}(s^A) = p(s^A \mid b^{S^I}_{t,1}, I_{t,1} = 0, I_{t,1}U_t = u_t) \\
=&p_{Z_t}(z \mid \delta_{z_t}) \cdot P(I_{t,1} = i \mid I_{t,1} = 0) \cdot \frac{p_{U_t}(u \mid b^U_{t,1}, I_{t,1} = 0)\, p_{I_{t,1}U_t}(u_t \mid U_t = u, I_{t,1} = 0)}{p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1} = 0)} \\
=&\delta(z - z_t)\, b^U_{t,1}(u)\, \mathbb{1}(i = 0),
\end{aligned}$$

The last equality holds since $p_{I_{t,1}U_t}(u_t \mid U_t = u, I_{t,1} = 0) = p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1} = 0) = \delta(u_t - 0)$ and $p_{U_t}(u \mid b^U_{t,1}, I_{t,1} = 0) = b^U_{t,1}(u)$. Therefore, the transition function from $b^{S^I}_{t,1}$ to $b^{S^A}_{t,2}$ is

$$\begin{aligned}
\mathcal{T}(b^{S^A}_{t,2} \mid b^{S^I}_{t,1}, I_{t,1}) &= \int \delta_{\delta_{z_t} \otimes b^U_{t,1} \otimes \delta_0}\, p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1})\, du_t \\
&= \delta_{\delta_{z_t} \otimes b^U_{t,1} \otimes \delta_0},
\end{aligned}$$

since $p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1}) = \delta(u_t - 0)$ when $I_{t,1} = 0$. Similarly, based on the Bellman optimality equation (7),

$$\begin{aligned}
\mathcal{Q}^{I*}(b^{S^I}_{t,1}, I_{t,1}) &= \int \left[\sqrt{\gamma} \max_{a_{t,2} \in \mathcal{A}} \mathcal{Q}^{A*}(b^{S^A}_{t,2}, a_{t,2})\right] p_{I_{t,1}U_t}(u_t \mid b^U_{t,1}, I_{t,1})\, du_t \\
&= \sqrt{\gamma} \max_{a_{t,2} \in \mathcal{A}} Q^{A*}(\delta_{z_t} \otimes b^U_{t,1} \otimes \delta_0, a_{t,2}).
\end{aligned} \tag{9}$$

Comparing equations (8) and (9), we see that in (8) the optimal control action $a_{t,2}$ is selected based on each possible value of the emission $I_{t,1}U_t$, while in (9) it is selected based on the entire belief state $b_{t,2}^{S^A}$. Further, we have

$$\begin{aligned}\mathcal{Q}^{I*}(b_{t,1}^{S^I},1)-\mathcal{Q}^{I*}(b_{t,1}^{S^I},0)=&\int \max_{a_{t,2}\in\mathcal{A}} Q^{A*}(\delta_{z_t}\otimes\delta_{u_t},a_{t,2})\, b_{t,1}^{U}(u_t)\, du_t\\ &-\max_{a_{t,2}\in\mathcal{A}} Q^{A*}(\delta_{z_t}\otimes b_{t,1}^{U},a_{t,2})\\ =&\mathbb{E}[\mathcal{V}_1^{A*}(\delta_u)]-\mathcal{V}_0^{A*}(b_{t,1}^{U})\\ =&\mathbb{E}[\mathcal{V}_1^{A*}(\delta_u)]-\mathbb{E}[\mathcal{V}_0^{A*}(\delta_u)]+\mathbb{E}[\mathcal{V}_0^{A*}(\delta_u)]-\mathcal{V}_0^{A*}(b_{t,1}^{U}).\end{aligned}$$

□

Remark. When the measurement action does not affect the environment, $I_{t,1}$ is not part of the state $S_{t,2}^A$. In this case, we have $b_{t,2}^{S^A}=\delta_{z_t}\otimes\delta_{u_t}$ when $I_{t,1}=1$ and $b_{t,2}^{S^A}=\delta_{z_t}\otimes b_{t,1}^{U}$ when $I_{t,1}=0$. By Jensen's inequality, it follows that

$$\mathcal{Q}^{I*}(b_{t,1}^{S^I},1)=\int \max_{a_{t,2}\in\mathcal{A}} Q^{A*}(\delta_{z_t}\otimes\delta_{u_t},a_{t,2})\, b_{t,1}^{U}(u_t)\, du_t \geq \max_{a_{t,2}\in\mathcal{A}} Q^{A*}(\delta_{z_t}\otimes b_{t,1}^{U},a_{t,2})=\mathcal{Q}^{I*}(b_{t,1}^{S^I},0).$$

However, this inequality may not hold when $I_{t,1}$ does affect the environment.

B.5 Belief Propagation With Sequential Monte Carlo

Belief state $b_{t,1}^U$. To obtain the belief state $b_{t,1}^U$ for U_t, note that the joint posterior distribution of $U_{1:t}$ and $\boldsymbol{\theta}$ given the observed history $H_{t,1}=\{Z_1,O_1,I_{1,1},I_{1,1}U_1,A_{1,2},\ldots,Z_t,O_t\}$ is

$$\begin{aligned}p(U_{1:t},\boldsymbol{\theta}\mid H_{t,1})=&p(U_{1:t},\boldsymbol{\theta}\mid H_{t-1,2},A_{t-1,2},Z_t,O_t)\\ =&p(U_{1:t-1}\mid H_{t-1,2},A_{t-1,2})\cdot\\ &p(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2},A_{t-1,2})\cdot\\ &p(Z_t\mid\boldsymbol{\theta},U_{1:t-1},H_{t-1,2},A_{t-1,2})\cdot\\ &p(U_t\mid Z_t,\boldsymbol{\theta},U_{1:t-1},H_{t-1,2},A_{t-1,2})\cdot\\ &p(O_t\mid U_t,Z_t,\boldsymbol{\theta},U_{1:t-1},H_{t-1,2},A_{t-1,2})\cdot\\ &[p(Z_t,O_t\mid H_{t-1,2},A_{t-1,2})]^{-1}\\ \propto&p(U_{1:t-1}\mid H_{t-1,2})\cdot p(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})\cdot\\ &p(Z_t\mid\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})\cdot\\ &p(U_t\mid Z_t,\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})\cdot p(O_t\mid\boldsymbol{\theta},U_t).\end{aligned}$$

Here, $p(U_{1:t-1}\mid H_{t-1,2},A_{t-1,2})=p(U_{1:t-1}\mid H_{t-1,2})$ since $A_{t-1,2}$ is chosen only based on $H_{t-1,2}$, i.e., $A_{t-1,2}\perp\!\!\!\perp U_{1:t-1}\mid H_{t-1,2}$. Suppose the proposal distribution can be factorized as

$$q(U_{1:t},\boldsymbol{\theta}\mid H_{t,1})=q(U_t\mid\boldsymbol{\theta},U_{1:t-1},H_{t,1})\,q(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})\,q(U_{1:t-1}\mid H_{t-1,2}).$$

Then the importance weight can be written as

$$\begin{aligned}\widetilde{W}_{t,1}=&\frac{p(U_{1:t},\boldsymbol{\theta}\mid H_{t,1})}{q(U_{1:t},\boldsymbol{\theta}\mid H_{t,1})}\\ \propto& p(U_{1:t-1}\mid H_{t-1,2})\cdot p(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})\cdot p(Z_t\mid\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})\cdot\\ &p(U_t\mid Z_t,\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})\cdot p(O_t\mid\boldsymbol{\theta},U_t)\cdot\\ &[q(U_{1:t-1}\mid H_{t-1,2})\,q(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})\,q(U_t\mid\boldsymbol{\theta},U_{1:t-1},H_{t,1})]^{-1}\\ =& W_{t-1,2}\cdot\frac{p(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})}{q(\boldsymbol{\theta}\mid U_{1:t-1},H_{t-1,2})}\cdot\frac{p(U_t\mid Z_t,\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})}{q(U_t\mid\boldsymbol{\theta},U_{1:t-1},H_{t,1})}\cdot\\ &p(Z_t\mid\boldsymbol{\theta},U_{t-1},Z_{t-1},I_{t-1,1},A_{t-1,2})\cdot p(O_t\mid\boldsymbol{\theta},U_t),\end{aligned}$$

where

$$W_{t-1,2} = \frac{p(U_{1:t-1} \mid H_{t-1,2})}{q(U_{1:t-1} \mid H_{t-1,2})}$$

is the weight for the marginal posterior distribution of $U_{1:t-1}$ given the history $H_{t-1,2}$. If we take

$$\begin{aligned} q(\boldsymbol{\theta} \mid U_{1:t-1}, H_{t-1,2}) &= p(\boldsymbol{\theta} \mid U_{1:t-1}, H_{t-1,2}), \\ q(U_t \mid \boldsymbol{\theta}, U_{1:t-1}, H_{t,1}) &= p(U_t \mid Z_t, \boldsymbol{\theta}, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}), \end{aligned}$$

then the importance weight $\widetilde{W}_{t,1}$ can be updated as

$$\widetilde{W}_{t,1} \propto W_{t-1,2} \cdot p(Z_t \mid \boldsymbol{\theta}, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}) \cdot p(O_t \mid \boldsymbol{\theta}, U_t).$$

Belief state $b^U_{t,2}$. To obtain the belief state $b^U_{t,1}$ for U_t, we break down the posterior distribution $p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,2})$ separately for the cases $I_t = 0$ and $I_t = 1$.

When $I_t = 0$, the joint posterior distribution of $U_{1:t}$ and $\boldsymbol{\theta}$ given the observed history $H_{t,2} = H_{t,1} \cup \{I_{t,1}, I_{t,1}U_t\}$ is

$$\begin{aligned} p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,2}) =& p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}, I_{t,1}, I_{t,1}U_t) \\ =& p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}, I_{t,1}) \cdot \\ & p(I_{t,1}U_t \mid \boldsymbol{\theta}, U_{1:t}, H_{t,1}, I_{t,1}) \\ & [p(I_{t,1}U_t \mid H_{t,1}, I_{t,1})]^{-1} \\ \propto& p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}) \cdot p(I_{t,1}U_t \mid U_t, I_{t,1}). \end{aligned}$$

Here, $p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}, I_{t,1}) = p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1})$ since $I_{t,1}$ is chosen only based on $H_{t,1}$, i.e., $I_{t,1} \perp\!\!\!\perp (U_{1:t}, \boldsymbol{\theta}) \mid H_{t,1}$. Take the proposal distribution as

$$q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,2}) = q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}),$$

which is the same as the previous proposal distribution. Then the importance weight can be written as

$$\begin{aligned} \widetilde{W}_{t,2} =& \frac{p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,2})}{q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,2})} \\ \propto& p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1}) \cdot p(I_{t,1}U_t \mid U_t, I_{t,1}) \cdot [q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1})]^{-1}. \end{aligned}$$

For any latent state $U_t = u_t$, we have $p(I_{t,1}U_t = 0 \mid U_t = u_t, I_{t,1} = 0) = 1.$. Then we have

$$\widetilde{W}_{t,2} \propto \frac{p(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1})}{q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1})} \cdot p(I_{t,1}U_t \mid U_t, I_{t,1}) = W_{t,1},$$

where

$$W_{t,1} = \frac{p(U_{1:t} \mid H_{t,1})}{q(U_{1:t} \mid H_{t,1})}$$

is the weight for the marginal posterior distribution of $U_{1:t}$ given the history $H_{t,1}$. Thus, the proposal distribution and the particle weights remain the same as at time $(t, 1)$.

When $I_t = 1$, we do not need the parameter $\boldsymbol{\theta}$ to approximate the belief state of the latent state. The posterior distribution $p(U_{1:t} \mid H_{t,2})$ can be expressed as

$$\begin{aligned} p(U_{1:t} \mid H_{t,2}) =& p(U_{1:t} \mid H_{t-1,2}, A_{t-1,2}, Z_t, O_t, I_{t,1}, I_{t,1}U_t) \\ =& p(U_{1:t-1} \mid H_{t-1,2}, A_{t-1,2}, I_{t,1}) \cdot \\ & p(Z_t \mid U_{1:t-1}, H_{t-1,2}, A_{t-1,2}, I_{t,1}) \cdot \\ & p(I_{t,1}U_t \mid Z_t, U_{1:t-1}, H_{t-1,2}, A_{t-1,2}, I_{t,1}) \cdot \\ & p(U_t \mid I_{t,1}U_t, Z_t, U_{1:t-1}, H_{t-1,2}, A_{t-1,2}, I_{t,1}) \cdot \\ & p(O_t \mid U_t, Z_t, I_{t,1}U_t, U_{1:t-1}, H_{t-1,2}, A_{t-1,2}, I_{t,1}) \cdot \\ & [p(Z_t, O_t, I_{t,1}U_t \mid H_{t-1,2}, A_{t-1,2}, I_{t,1})]^{-1} \\ \propto& p(U_{1:t-1} \mid H_{t-1,2}) \cdot p(Z_t \mid U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}) \cdot \\ & p(I_{t,1}U_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}) \cdot \\ & p(U_t \mid I_{t,1}U_t, Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}) \cdot p(O_t \mid U_t). \end{aligned}$$

When the true latent state $U_t = u_t$, we have

$$p(I_{t,1}U_t = u \mid U_t = u_t, I_{t,1} = 1) = \delta(u - u_t),$$

according to the definition of the emission model. Therefore,

$$\begin{aligned}
&p(U_t = u_t \mid I_{t,1}U_t = u, Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1} = 1) \\
=&\frac{p(I_{t,1}U_t = u \mid U_t = u_t, I_{t,1} = 1)\, p(U_t = u_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1} = 1)}{p(I_{t,1}U_t = u \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1} = 1)} \\
=&\delta(u - u_t).
\end{aligned}$$

Suppose the proposal distribution can be factorized as

$$q(U_{1:t} \mid H_{t,2}) = q(U_{1:t-1} \mid H_{t-1,2})\, q(U_t \mid U_{1:t-1}, H_{t,2}).$$

Now if we take

$$q(U_t \mid U_{1:t-1}, H_{t,2}) = p(U_t \mid I_{t,1}U_t, Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}),$$

the importance weight can be written as

$$\begin{aligned}
\widetilde{W}_{t,2} \propto&\frac{p(U_{1:t-1} \mid H_{t-1,2})}{q(U_{1:t-1} \mid H_{t-1,2})} \cdot \frac{p(U_t \mid I_{t,1}U_t, Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1})}{q(U_t \mid U_{1:t-1}, H_{t,2})} \cdot \\
&p(Z_t \mid U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}) \cdot p(I_{t,1}U_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}) \cdot p(O_t \mid U_t) \\
\propto&W_{t-1,2} \cdot p(Z_t \mid U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}) \cdot p(I_{t,1}U_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}) \cdot p(O_t \mid U_t).
\end{aligned}$$

In this case, the proposal distribution $q(U_t \mid U_{1:t-1}, H_{t,2})$ simplifies to

$$q(U_t = u \mid U_{1:t-1}, H_{t,1}, I_{t,1} = 1, I_{t,1}U_t = u_t) = \delta(u - u_t),$$

which places all mass on $U_t = u_t$. This indicates that when $I_{t,1} = 1$, we should always draw $\widehat{U}_{t,2}^{(j)} = u_t$. Then, since $\widehat{U}_{t,2}^{(j)}$ has the same value for all particles j, the likelihood term $p(O_t \mid \widehat{U}_{t,2}^{(j)})$ is identical across all particles in the importance weight. Now we have

$$\begin{aligned}
&p(I_{t,1}U_t = u_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1}) \\
=&\int p(I_{t,1}U_t = u' \mid U_t = u, Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1} = 1)\, p(U_t = u \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}, I_{t,1})\, du \\
=&\int \delta(u_t - u)\, p(U_t = u \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2})\, du \\
=&p(U_t = u_t \mid Z_t, U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}).
\end{aligned}$$

Therefore, after normalization, we obtain

$$\widetilde{W}_{t,2}^{(j)} \propto W_{t-1,2}^{(j)} \cdot p(Z_t = z_t, U_t = u_t \mid U_{t-1}, Z_{t-1}, I_{t-1,1}, A_{t-1,2}),$$

when we observe $Z_t = z_t$ and $I_{t,1}U_t = u_t$.

Note that the decomposition of $p(U_{1:t} \mid H_{t,2})$ from $H_{t-1,2}$ instead of $H_{t,1}$ allows us to resample the particles of U_t. Otherwise, if the true value u_t were not drawn from the proposal distribution $q(U_{1:t}, \boldsymbol{\theta} \mid H_{t,1})$ at time $(t, 1)$, then all particle weights $\widetilde{W}_{t,2}^{(j)}$ would be zero.

C Additional Algorithm Details

In this section, we provide implementation details of the active-measure algorithm.

C.1 Active-Measure

Algorithm 3 provides a full description of the Active-Measure procedure.

Algorithm 3 Active-Measure

Input: Hyperparameters $\lambda^I, (\sigma^I)^2, \lambda^A, (\sigma^A)^2, C$.

Observe Z_1 and O_1. Construct the belief state $b^U_{1,1}$ for U_1 using Algorithm 1.
for $t \geq 1$ **do**
  Set $I_{t,1} = 1$ if $t = 1$; otherwise set $I_{t,1} = \text{argmax}_{i \in \mathcal{I}} \phi^I(b^{S^I}_{t,1}, i)^\top \widetilde{\boldsymbol{\beta}}^I_t$.
  Observe $I_{t,1}U_t$. Update the belief state $b^U_{t,2}$ for U_t using Algorithm 2.
  Draw $\widetilde{\boldsymbol{\beta}}^I_t \sim N(\boldsymbol{\mu}^I_t, \boldsymbol{\Sigma}^I_t)$ using (6).
  Set $A_{t,2} \sim \text{Bernoulli}(0.5)$ if $t = 1$; otherwise set $A_{t,2} = \text{argmax}_{a \in \mathcal{A}} \phi^A(b^{S^A}_{t,2}, a)^\top \widetilde{\boldsymbol{\beta}}^A_t$.
  Observe Z_{t+1} and O_{t+1}. Construct the belief state $b^U_{t+1,1}$ for U_{t+1} using Algorithm 1.
  Draw $\widetilde{\boldsymbol{\beta}}^A_t \sim N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$ using (6).
end for

C.2 RLSVI

For completeness, we describe the standard RLSVI algorithm (Osband et al., 2016) in Algorithm 4. Here we assume a stationary MDP setting, where the state is S_t, the action is A_t, and the reward is R_t.

Define

$$\boldsymbol{X}_l = \phi(s_l, a_l),$$
$$Y_l = r_l + \gamma \max_a \phi(s_{l+1}, a)^\top \widetilde{\boldsymbol{\beta}}_{t-1},$$

for $l \in \{1 : t\}$, where ϕ is the basis function. Let $\mathbf{X} := [\boldsymbol{X}_{1:t}]^\top$ and $\mathbf{Y} := [Y_{1:t}]^\top$. We fit a BLR model for $\mathbf{Y}$ given $\widetilde{\boldsymbol{\beta}}_{t-1}$. The posterior of $\boldsymbol{\beta}_t$ is $N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where

$$\boldsymbol{\Sigma}_t = \left[(\mathbf{X}_t)^\top \mathbf{X}_t/\sigma^2 + \lambda \mathbf{I}\right]^{-1},$$
$$\boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t \left[(\mathbf{X}_t)^\top \mathbf{Y}_t/\sigma^2\right].$$

An estimate of $\boldsymbol{\beta}_t$ is then obtained by drawing $\widetilde{\boldsymbol{\beta}}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ from the posterior distribution.

Algorithm 4 Stationary RLSVI

Input: Hyperparameters λ, σ^2, and initialization $\widetilde{\boldsymbol{\beta}}_0 = \mathbf{0}$.

Observe the initial state s_1.
for $t \geq 1$ **do**
  Draw $\widetilde{\boldsymbol{\beta}}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ based on the previous parameter estimated $\widetilde{\boldsymbol{\beta}}_{t-1}$.
  Select the action $a_t = \text{argmax}_a \phi(s_t, a)^\top \widetilde{\boldsymbol{\beta}}_t$.
  Observe the reward r_t and the next state s_{t+1}.
end for

D Application Details

In this section, we provide details for the HeartSteps application. The experiment code and HeartSteps environment can be found here: github.com/StatisticalReinforcementLearningLab/Active-Measuring.

D.1 HeartSteps as a Special Case of an AOMDP

Definition. In HeartSteps, the reward R_t corresponds to the next latent state U_{t+1}, and the proximal outcome and engagement $[M_{t,2}, E_t]$ form the observed state Z_{t+1}. The emission O_t of the latent state U_t is indexed as the emission O_{t-1} of the latent reward R_{t-1}. The reward $r(Z_{t+1}, U_{t+1})$ of the AOMDP is defined as $r(M_{t,2}, E_t, R_t) = R_t$.

State Construction. Lemma 1 states that the state of $I_{t,1}$ is $S^I_{t,1} = [Z_t, U_t]$, which includes $[M_{t-1,2}, E_{t-1}, R_{t-1}]$. However, the causal DAG suggests that $M_{t-1,2}$ is independent of future rewards and states given E_{t-1}, R_{t-1}, and thus can be omitted from the state without affecting the optimal value function (Gao et al., 2025). Notice that a context variable $C_{t,2}$ is observed between $I_{t,1}$ and $A_{t,2}$, which does not exist in the general AOMDP framework. However, since $C_{t,2}$ is exogenous and independent of other variables given $M_{t,2}$, it does not affect the belief propagation and only becomes part of the state $S^A_{t,2}$ of $A_{t,2}$. Therefore, the optimal state for I_t is $S^I_{t,1} = [E_{t-1}, R_{t-1}]$, and the optimal state for A_t is $S^A_{t,2} = [E_{t-1}, R_{t-1}, C_{t,2}, I_{t,1}]$.

SMC. In Algorithms 1 and 2, the probability $p(z_t \mid \widehat{\boldsymbol{\theta}}^{(j)}, \widetilde{u}^{(j)}_{t-1,2}, z_{t-1}, i_{t-1,1}, a_{t-1,2})$ is used to update the particle weight. However, the causal DAG implies that $E_{t-1} \perp\!\!\!\perp R_{t-2} \mid E_{t-2}, I_{t-1,1}, A_{t-1,2}$. Therefore, the belief states $b^U_{t,1}$ and $b^U_{t,2}$ do not depend on E_{t-1}. Only $M_{t-1,2}$ is needed to update the particle weight (see details in Appendix D.2).

Basis Functions and Reward Functions. When a belief state is approximately normal, the mean and standard deviation of the particles serve as sufficient statistics of the normal distribution. Let $\{\widehat{r}^{(j)}_{t,k}\}_{j=1}^J$ be the particles and $\widehat{b}^R_{t,k}(u) = \sum_{j=1}^J w^{(j)}_{t,k}\delta(r - \widehat{r}^{(j)}_{t,k})$ be the estimated belief state of the latent reward R_t at time (t,k) for $k = 1, 2$. We thus define

$$\bar{b}^R_{t,k} := \mathbb{E}_{\widehat{b}^R_{t,k}}(s) = \sum_{j=1}^J w^{(j)}_{t,k}\widehat{r}^{(j)}_{t,k},$$
$$\widetilde{b}^R_{t,k} := \mathrm{Std}_{\widehat{b}^R_{t,k}}(s) = \left\{\sum_{j=1}^J w^{(j)}_{t,k}\big(\widehat{r}^{(j)}_{t,k} - \bar{b}^R_{t,k}\big)^2\right\}^{1/2}$$

as the expectation and standard deviation under the belief state. The basis functions are constructed as

$$\begin{aligned}\phi^I(b^{S^I}_{t,1}, I_{t,1}) =& [1, E_{t-1}, \bar{b}^R_{t-1,1}, \widetilde{b}^R_{t-1,1}, I_{t,1}, I_{t,1}E_{t-1}, I_{t-1,1}\bar{b}^R_{t-1,1}],\\ \phi^A(b^{S^A}_{t,2}, A_{t,2}) =& [1, E_{t-1}, \bar{b}^R_{t-1,2}, C_{t,2}, I_{t,1}, A_{t,2}, A_{t,2}E_{t-1}, A_{t,2}\bar{b}^R_{t-1,2}, A_{t,2}C_{t,2}].\end{aligned}$$

The standard deviation $\widetilde{b}^R_{t-1,2}$ is not included in ϕ^A since it is highly correlated with $I_{t,1}$.

Remember that the reward is defined as $r(Z_{t+1}, U_{t+1,1}) = R_t$, and the target for the control action in RLSVI is $r(Z_{t+1}, b^U_{t+1,1}) = \int r(Z_{t+1}, u)b^U_{t+1,1}(u)du$. Then it can be estimated as $r(Z_{t+1}, \widehat{b}^U_{t+1,1}) = \sum_{j=1}^J w^{(j)}_{t,1}\widehat{r}^{(j)}_{t,2} = \bar{b}^R_{t,2}$.

D.2 Update the Posterior of Unknown Parameters

Particle learning (Storvik, 2002; Carvalho et al., 2010) can be inefficient in a general POMDP where the posterior distribution of the parameters $\boldsymbol{\theta}$ is intractable. However, with certain working models, the posterior distribution of $\boldsymbol{\theta}$ admits a closed form. For example, models in the exponential family with conjugate priors yield closed-form posteriors, including linear models with Gaussian noise, binomial models, multinomial models, and Poisson models.

Here, we use a working model with linear mean and Gaussian noise. Specifically, suppose the mean of each variable is a linear function of its parents in the causal DAG, and the noise follows a Gaussian distribution. That is,

$$\begin{aligned}M_{t,2} &= \theta^M_0 + \theta^M_1 E_{t-1} + \theta^M_2 R_{t-1} + \theta^M_3 C_{t,2} + A_{t,2}(\theta^M_4 + \theta^M_5 E_{t-1} + \theta^M_6 R_{t-1} + \theta^M_7 C_{t,2}) + \epsilon^M_{t,2},\\ R_t &= \theta^R_0 + \theta^R_1 M_{t,2} + \theta^R_2 E_t + \theta^R_3 R_{t-1} + \epsilon^R_t,\\ O_t &= \theta^O_0 + \theta^O_1 R_t + \epsilon^O_t,\end{aligned} \tag{10}$$

where $\boldsymbol{\theta}^M = \theta^M_{0:7}$, $\boldsymbol{\theta}^R = \theta^R_{0:3}$, and $\boldsymbol{\theta}^O = \theta^O_{0:1}$. Let $\boldsymbol{\theta} := \{\boldsymbol{\theta}^M, \boldsymbol{\theta}^R, \boldsymbol{\theta}^O\}$ denote all the parameters used in SMC. The noise terms $\epsilon^M_{t,2}$, ϵ^R_t, and ϵ^O_t follow Gaussian distributions with mean zero and fixed variances σ^{2M}, σ^{2R}, and σ^{2O}, respectively. Suppose the prior of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^V \sim N(\boldsymbol{\nu}^V_0, \boldsymbol{\Lambda}^V_0)$ for $V \in \{M, R, O\}$.

Given the history

$$h_{t,2} = [E_0, O_0, I_{1,1}, I_{1,1}R_0, C_{1,2}, A_{1,2}, M_{1,2}, E_1, O_1, \ldots, E_{t-1}, O_{t-1}, I_{t,1}, I_{t,1}R_{t-1}],$$

and a particle value $\widehat{u}^{(j)}_{1:t,2} = \widehat{r}^{(j)}_{0:t-1,2}$, define

$$\begin{aligned}
\mathbf{X}_t^M :=& [X_{1:t-1}^M]^\top, \quad \text{with } X_l^M := [1, E_{l-1}, \widehat{r}^{(j)}_{l-1,2}, C_{l,2}, A_{l,2}, A_{l,2}E_{l-1}, A_{l,2}\widehat{r}^{(j)}_{l-1,2}, A_{l,2}C_{l,2}], \text{ for } l = 1, \ldots, t-1,\\
\mathbf{Y}_t^M :=& [M_{1:t-1,2}]^\top,\\
\mathbf{X}_t^R :=& [X_{1:t-1}^R]^\top, \quad \text{with } X_l^R := [1, M_{l,2}, E_t, \widehat{r}^{(j)}_{l-1,2}], \text{ for } l = 1, \ldots, t-1,\\
\mathbf{Y}_t^R :=& [\widehat{r}^{(j)}_{1:t-1,2}]^\top,\\
\mathbf{X}_t^O :=& [X_{1:t-1}^O]^\top, \quad \text{with } X_l^O := [1, \widehat{r}^{(j)}_{l,2}], \text{ for } l = 1, \ldots, t-1,\\
\mathbf{Y}_t^O :=& [O_{1:t-1}]^\top.
\end{aligned}$$

Under the working model (10), the posterior distribution of $\boldsymbol{\theta}^V$ is $\boldsymbol{\theta}^V \mid \widehat{u}^{(j)}_{1:t,2}, h_{t,2} \sim N(\boldsymbol{\nu}_t^V, \boldsymbol{\Lambda}_t^V)$, for $V \in \{M, R, O\}$, where

$$\begin{aligned}
\boldsymbol{\Lambda}_t^M &= \left\{ \frac{1}{\sigma_R^2} (\mathbf{X}_t^M)^\top \mathbf{X}_t^M + (\boldsymbol{\Lambda}_0^M)^{-1} \right\}^{-1},\\
\boldsymbol{\nu}_t^M &= \boldsymbol{\Lambda}_t^M \left\{ \frac{1}{\sigma_R^2} (\mathbf{X}_t^M)^\top \mathbf{Y}_t^M + (\boldsymbol{\Lambda}_0^M)^{-1} \boldsymbol{\nu}_0^M \right\}.
\end{aligned}$$

Then, a particle $\widehat{\boldsymbol{\theta}}_t^{(j)} = [\widehat{\boldsymbol{\theta}}_t^{M(j)}, \widehat{\boldsymbol{\theta}}_t^{R(j)}, \widehat{\boldsymbol{\theta}}_t^{O(j)}]$, is drawn with

$$\widehat{\boldsymbol{\theta}}_t^{V(j)} \mid \widehat{u}^{(j)}_{1:t-1,2}, h_{t-1,2} \sim N(\boldsymbol{\nu}_{t-1}^V, \boldsymbol{\Lambda}_{t-1}^V),$$

for $V \in \{M, R, O\}$.

As discussed in Appendix D.1, E_{t-1} is not needed when updating the belief particles. Given $\widehat{\boldsymbol{\theta}}_t^{(j)}$, we draw

$$\widetilde{r}^{(j)}_{t-1,1} \sim p(r_{t-1} \mid \widehat{\boldsymbol{\theta}}_t^{R(j)}, m_{t-1,2}, e_{t-1}, \widehat{r}^{(j)}_{t-2,2}),$$

and update the particle weight as

$$\widetilde{w}^{(j)}_{t,1} \propto w^{(j)}_{t-1,2}\, p(m_{t-1,2} \mid \widehat{\boldsymbol{\theta}}_t^{M(j)}, \widetilde{r}^{(j)}_{t-2,2}, e_{t-2}, c_{t-1,2}, a_{t-1,2})\, p(o_{t-1} \mid \widehat{\boldsymbol{\theta}}_t^{O(j)}, \widetilde{r}^{(j)}_{t-1,1}).$$

D.3 Reward Design

When the effects of actions on rewards are mediated by the proximal outcome $M_{t,2}$ and the engagement E_t, the causal effects become harder to detect. Fortunately, the mediators can be leveraged to construct improved rewards in the RL algorithm by following the idea of the surrogate index (Athey et al., 2019; Yang et al., 2024). The conditional mean of the reward given the mediators has smaller variance than the original reward. Based on the working model (10), the mean of the reward R_t can be estimated as $\widehat{R}_t := [1, M_{t,2}, E_t, \bar{b}^R_{t-1,2}]\, \widehat{\boldsymbol{\theta}}_t^{R(j)}$. We then use $\widehat{R}_t$ as the reward in Algorithm 3.

D.4 Hyperparameters

The prior mean $\boldsymbol{\nu}_0^V$ for $V \in \{M, R, O\}$ is estimated by pooling data across all users in HeartSteps V3, following the same procedure as in Gao et al. (2025). We obtain $\boldsymbol{\nu}_0^M = [-0.043, -0.026, 0.062, 0.418, 0.001, 0.003, -0.035, 0.011]$, $\boldsymbol{\nu}_0^R = [-0.005, 0.029, 0.012, 0.861]$, and $\boldsymbol{\nu}_0^O = [0.034, 0.534]$. Here, the second coordinate of $\boldsymbol{\nu}_0^R$ is computed by summing over the five original estimated parameters θ_k^R for the 5 interventions in HeartSteps V3. The prior covariance $\boldsymbol{\Lambda}_0^V$ is set to a diagonal matrix $0.01\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The noise variances σ^{2M}, σ^{2R}, and σ^{2O} are also estimated from HeartSteps

V3 as the variances of the residuals obtained by fitting linear regressions for M, R, and O, respectively. We have $\sigma^{2M} = 0.972$, $\sigma^{2R} = 0.240$, and $\sigma^{2O} = 0.637$. The same priors and noise variances are used for all users in our simulation experiments.

The discount factor is chosen as $\gamma = 0.9$ to balance discount regularization and the modeling of long-term effects. In HeartSteps, the optimal action not only leads to a high instantaneous reward but also places the user in a promising state that yields higher rewards in the future. In behavioral science, this process is referred to as habit formation.

The number of particles J is set to 50, which is sufficient to approximate the belief state under the working model (10). The numerical experiments in Lim et al. (2023) also demonstrate that between 10^1 and 10^2 particles already yield good performance in simpler problems. The parameters of the target $\widetilde{\boldsymbol{\beta}}^A_{t^-}$ are copied from $\widetilde{\boldsymbol{\beta}}^A_t$ every $C = 10$ steps.

When selecting the hyperparameters $\lambda^I, (\sigma^I)^2, \lambda^A, (\sigma^A)^2$, note that $\lambda \cdot \sigma^2$ is equivalent to the tuning parameter of an L_2 penalty. Therefore, we select $\lambda^I \cdot (\sigma^I)^2$ from $\{0.2, 0.5\}$, $(\sigma^I)^2$ from $\{0.02, 0.1\}$, $\lambda^A \cdot (\sigma^A)^2$ from $\{5, 20\}$, and $(\sigma^A)^2$ from $\{0.02, 0.1\}$.

D.5 Simulation Testbed

We construct our simulation testbed based on the public testbed developed by Gao et al. (2025). The original testbed includes five decision times per day. To focus on the discussion of active measuring, we consider a simplified setting with only one control action $A_{t,2}$ per day. Adapting the original testbed to the problem described in Figure 2, we aggregate the effects of the five actions $A_{t,1:5}$ on both the reward and engagement, effectively treating all five actions as identical. Furthermore, we introduce an additional effect of $I_{t,1}$ on the engagement E_t. Specifically, we modify equation (38) in Gao et al. (2025) as follows:

$$
\begin{aligned}
C_{t,2} &= \theta_0^C + \epsilon_{t,2}^C, \\
M_{t,2} &= \theta_0^M + \theta_1^M E_{t-1} + \theta_2^M R_{t-1} + \theta_3^M C_{t,2} + A_{t,2}(\theta_4^M + \theta_5^M E_{t-1} + \theta_6^M R_{t-1} + \theta_7^M C_{t,2}) + \epsilon_{t,2}^M, \\
E_t &= \theta_0^E + \theta_1^E E_{t-1} + \Big(\sum_{k=1}^{5} \theta_{k+1}^E\Big) A_{t,2} + \Big(\sum_{k=1}^{5} \theta_{k+6}^E\Big) A_{t,2} E_{t-1} + \theta_I^E I_{t,1} + \theta_{IE}^E I_{t,1} E_{t-1} + \epsilon_t^E, \\
R_t &= \theta_0^R + \Big(\sum_{k=1}^{5} \theta_k^R\Big) M_{t,2} + \theta_6^R E_t + \theta_7^R R_{t-1} + \epsilon_t^R, \\
O_t &= \theta_0^O + \theta_1^O R_t + \epsilon_t^O,
\end{aligned}
$$

where θ_I^E and θ_{IE}^E are manually set, while all other parameters are taken from the HeartSteps testbed. The noise terms $\epsilon_{t,2}^C$, $\epsilon_{t,2}^M$, ϵ_t^E, ϵ_t^R, and ϵ_t^O have mean zero and variances σ^{2C}, σ^{2M}, σ^{2E}, σ^{2R}, and σ^{2O}, respectively. In addition, to ensure that engagement has a positive effect on the reward, we clip θ_6^R as $\max\{\theta_6^R, 0.02\}$.

The effect size of the positive effect from $A_{t,2}R_{t-1}$ to R_t through $M_{t,2}$ is $\theta_6^M(\sum_{k=1}^5 \theta_k^R)/\sqrt{\sigma^{2R} + \sigma^{2M}(\sum_{k=1}^5 \theta_k^R)^2}$. The vanilla testbed, constructed directly from the HeartSteps dataset, has the minimal effect size. To examine the performance of the proposed algorithm across different testbed variants, we modify the parameter θ_6^M to 0.5 or 0.8 to achieve small and medium effect sizes. The average effect sizes across all users for the minimal, small, and medium positive effects are 0.026, 0.119, and 0.191, respectively.

The effect size of the negative effect from $I_{t,1}$ to R_t through E_t is $\theta_I^E \theta_6^R / \sqrt{\sigma^{2R} + \sigma^{2E}(\theta_6^R)^2}$. Since the measure action was not taken in HeartSteps V2, the vanilla testbed does not contain this negative effect. To create testbed variants with zero, minimal and small effect sizes, we set the parameters θ_I^E and θ_{IE}^E to $[0, 0]$, $[-0.08, 0.01]$, and $[-0.3, 0.02]$, respectively. The average effect sizes across all users for the zero, minimal, and small negative effects are 0, 0.010, and 0.039, respectively.

D.6 Details of Always-Measure and Never-Measure

Always-measure and never-measure algorithms set $I_{t,1}$ to one or zero with probability one and choose $A_{t,2}$ using RLSVI. Specifically, the state is $S^A_{t,2} = [E_{t-1}, R_{t-1}, C_{t,2}]$, since $I_{t,1}$ is a constant. Define the basis function

$$\phi^A(b^{S^A}_{t,2}, A_{t,2}) = [1, E_{t-1}, \bar{b}^U_{t-1,2}, C_{t,2}, A_{t,2}, A_{t,2}E_{t-1}, A_{t,2}\bar{b}^U_{t-1,2}, A_{t,2}C_{t,2}].$$

Similar to the active-measure algorithm, define $\boldsymbol{X}^A_l = \phi^A(b^{S^A}_{l,2}, A_{l,2})$ and

$$Y^A_l = r(Z_{l+1}, b^U_{l+1,1}) + \gamma\phi^I(b^{S^A}_{l+1,2}, a')^\top \widetilde{\boldsymbol{\beta}}^A_{t-1}, \quad \text{where} \quad a' = \operatorname*{argmax}_{i\in\mathcal{I}} \phi^I(b^{S^A}_{l+1,2}, i')^\top \widetilde{\boldsymbol{\beta}}^A_{t^-}.$$

Here, $\widetilde{\boldsymbol{\beta}}^A_{t^-}$ is copied from $\widetilde{\boldsymbol{\beta}}^A_t$ every C steps. We fit a BLR on $\mathbf{Y}^A = [Y^A_{1:t}]^\top$ using $\mathbf{X}^A = [\boldsymbol{X}^A_{1:t}]^\top$ to obtain the posterior distribution $N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$ of $\boldsymbol{\beta}^A_t$, where

$$\begin{aligned} \boldsymbol{\Sigma}^A_t &= \left[(\mathbf{X}^A_t)^\top \mathbf{X}^A_t/(\sigma^A)^2 + \lambda^A \mathbf{I}\right]^{-1}, \\ \boldsymbol{\mu}^A_t &= \boldsymbol{\Sigma}^A_t \left[(\mathbf{X}^A_t)^\top \mathbf{Y}^A_t/(\sigma^A)^2\right]. \end{aligned} \tag{11}$$

An estimate of $\boldsymbol{\beta}^A_t$ is then obtained by drawing $\widetilde{\boldsymbol{\beta}}^A_t \sim N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$ from the posterior distribution.

See Algorithm 5 for a full description of the always-measure and never-measure algorithms. The always-measure algorithm takes $P_0 = 1$, while the never-measure algorithm takes $P_0 = 0$. All other hyperparameters—λ^A, $(\sigma^A)^2$, C, J, and γ—and priors are set as described in Appendix D.4.

Algorithm 5 Always-Measure or Never-Measure

Input: Hyperparameters $\lambda^A, (\sigma^A)^2, C, P_0$.

Observe Z_1 and O_1. Construct the belief state $b^U_{1,1}$ for U_1 using Algorithm 1.
for $t \geq 1$ **do**
  Set $I_{t,1} = 1$ with probability P_0.
  Observe $I_{t,1}U_t$. Update the belief state $b^U_{t,2}$ for U_t using Algorithm 2.
  Take $A_{t,2} \sim$ Bernoulli(0.5) if $t = 1$; otherwise set $A_{t,2} = \operatorname{argmax}_{a\in\mathcal{A}} \phi^A(b^{S^A}_{t,2}, a)^\top \widetilde{\boldsymbol{\beta}}^A_{t-1}$.
  Observe Z_{t+1} and O_{t+1}.
  Draw $\widetilde{\boldsymbol{\beta}}^A_t \sim N(\boldsymbol{\mu}^A_t, \boldsymbol{\Sigma}^A_t)$ using (11).
end for

D.7 Details of Dyna-ATMQ

In implementing the Dyna-ATMQ algorithm, we adapted the open-source BAM-QMDP implementation by Krale et al. (2023) to accommodate our testbed structure. We retained the core ATM loop, which selects a control action from a belief-weighted Q-table and separately evaluates whether to measure based on the predicted measuring value and cost. The Dirichlet-distribution-based transition function, the Dyna-framework, and the decoupled action–measurement decision rule were preserved from the original BAM-QMDP code, while several components were modified to fit our setting.

We defined a 64-state discrete representation of the environment by discretizing $C_{t,2}$, E_{t-1}, and observed R_{t-1}. Each variable was discretized into four bins corresponding to $x < -\sigma$, $-\sigma \leq x < 0$, $0 \leq x < \sigma$, and $x \geq \sigma$, where x is the value of a continuous variable and σ is its standard deviation. Since all variables in the testbed were standardized, the cutoff values were -1, 0, and 1.

In addition, since Dyna-ATMQ only picks up the signal from an observed fixed cost, we need to treat the cost as a tuning parameter. We select the cost from values of 0.0, 0.01, 0.02, and 0.05, and report the performance of the one with the highest cumulative reward.

For other hyperparameters, the number of particles is set to 100, the number of offline training steps is 5, and the discount factor is $\gamma = 0.9$. Based on our simulation results, the average cumulative reward of Dyna-ATMQ can be improved by adding an exploration phase at the beginning. Therefore, we include a warm-up period of 20 decision times, during which Dyna-ATMQ takes the control and measurement actions with probability 0.5.

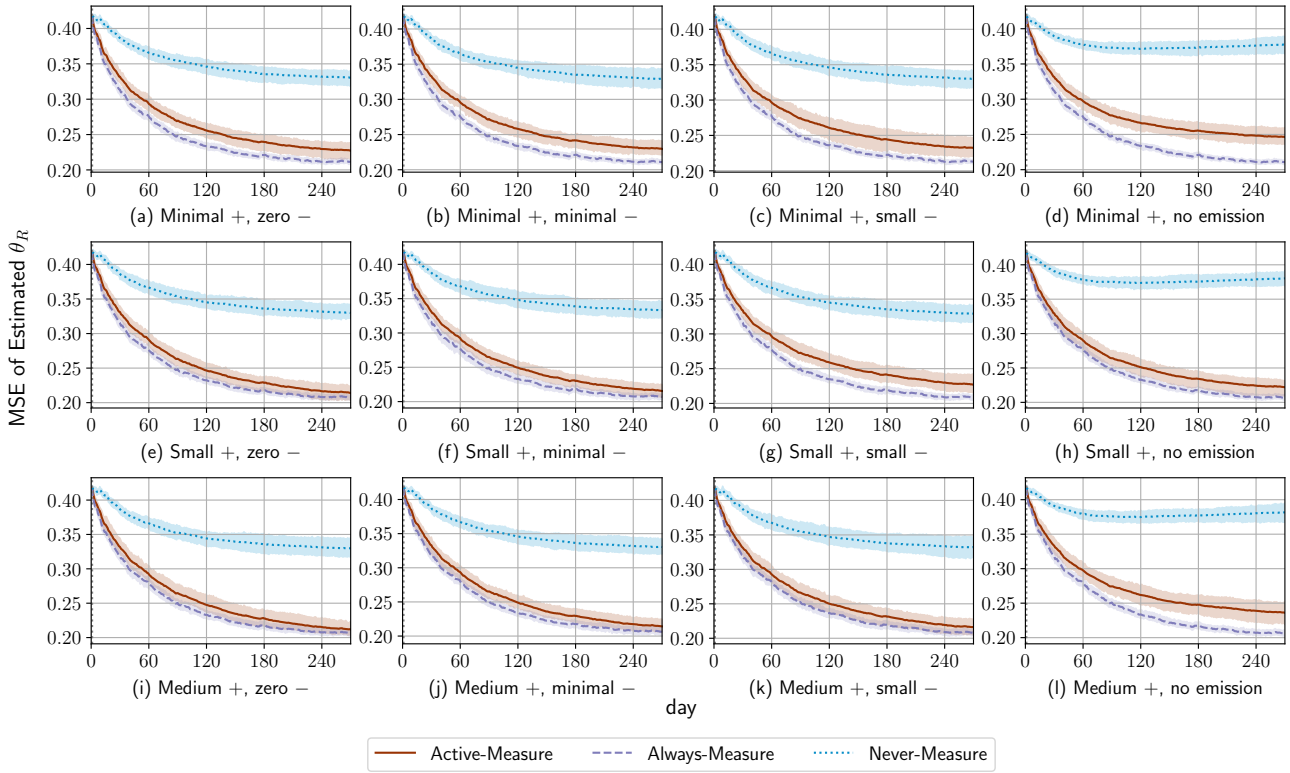

Figure D.1: The measurement rate. In each subfigure title, "+" denotes the positive effect of $A_{t,2}R_{t-1} \to R_t$ (rows: minimal, small, medium), and "-" denotes the negative effect of $I_{t,1} \to R_t$ (columns 1–3: zero, minimal, small). The last column ("zero O_t") sets $R_t \to O_t$ to zero, so the passive emission O_t is uninformative about R_t.

D.8 Additional Simulation Results

We report the measurement rate $\frac{1}{42}\sum_{i=1}^{42} \mathbb{1}(I_{t,i} = 1)$, averaged across all users at each time t. Figure D.1 shows its mean and 95% confidence interval over 50 replications. The experimental scenarios are the same as those in Figure 3. We observe that the measurement rate decreases as the negative effect increases. Moreover, under the same negative effect, the measurement rate increases as the positive effect increases, indicating that the benefits of measurement become greater.

In addition, we report the mean squared error (MSE) of $\widehat{\boldsymbol{\theta}}_t^{R(j)}$, averaged over J particles and 42 users, i.e., $\frac{1}{42J}\sum_{i=1}^{42}\sum_{j=1}^{J}\|\widehat{\boldsymbol{\theta}}_{t,i}^{R(j)} - \boldsymbol{\theta}_i^R\|$. Figure D.2 presents its mean and 95% confidence interval over 50 replications. The figure shows that the MSE of active-measure is very close to that of always-measure, while the MSE of never-measure is significantly larger. This suggests that a small number of measurements is sufficient to obtain a near-optimal estimate of the transition function. Furthermore, when the emission O_t is uninformative about the latent reward R_t, the MSE of never-measure increases substantially compared to when O_t is informative.

On average, active-measure takes about 38 minutes to complete one replication of the simulation for 42 users sequentially on a single CPU core of a cloud server, whereas always-measure and never-measure take about 34 minutes, and Dyna-ATMQ takes about 2 minutes.

Robustness of Active-Measure We conduct additional experiments using a misspecified transition model to evaluate the robustness of our proposed method. Specifically, we consider a general transition model from the

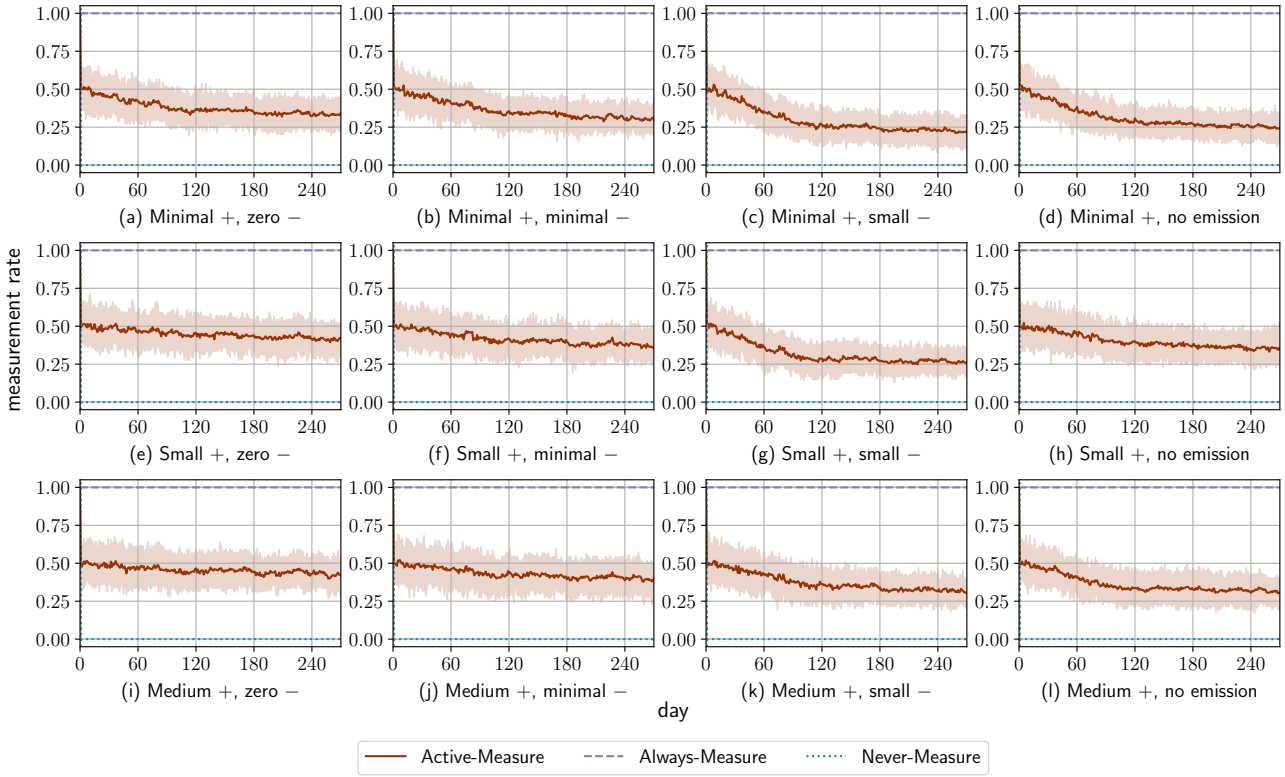

Figure D.2: The MSE of $\widehat{\boldsymbol{\theta}}_t^{R(j)}$. In each subfigure title, "+" denotes the positive effect of $A_{t,2}R_{t-1} \to R_t$ (rows: minimal, small, medium), and "-" denotes the negative effect of $I_{t,1} \to R_t$ (columns 1–3: zero, minimal, small). The last column ("zero O_t") sets $R_t \to O_t$ to zero, so the passive emission O_t is uninformative about R_t.

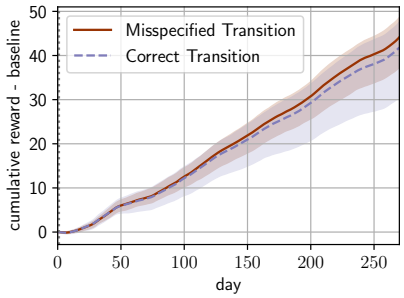

Figure D.3: Comparing the average cumulative reward under misspecified and correctly specified transition models, subtracting the average cumulative rewards of the zero policy.

current state $[E_{t-1}, R_{t-1}, C_t]$ to the next state $[M_{t,2}, E_t, R_t, O_t]$ given the actions $[I_{t,1}, A_{t,2}]$. That is,

$$\begin{aligned}
C_{t,2} &= \theta_0^C + \epsilon_{t,2}^C, \\
M_{t,2} &= \theta_0^M + \theta_1^M E_{t-1} + \theta_2^M R_{t-1} + \theta_3^M C_{t,2} + A_{t,2}(\theta_4^M + \theta_5^M E_{t-1} + \theta_6^M R_{t-1} + \theta_7^M C_{t,2}) + \epsilon_{t,2}^M, \\
E_t &= \theta_0^E + \theta_1^E E_{t-1} + \theta_2^E R_{t-1} + \theta_3^E C_{t,2} + A_{t,2}(\theta_4^E + \theta_5^E E_{t-1} + \theta_6^E R_{t-1} + \theta_7^E C_{t,2}) + \theta_I^E I_{t,1} + \theta_{IE}^E I_{t,1} E_{t-1} + \epsilon_t^E, \\
R_t &= \theta_0^R + \theta_1^R E_{t-1} + \theta_2^R R_{t-1} + \theta_3^R C_{t,2} + A_{t,2}(\theta_4^R + \theta_5^R E_{t-1} + \theta_6^R R_{t-1} + \theta_7^R C_{t,2}) + \epsilon_t^R, \\
O_t &= \theta_0^O + \theta_1^O E_{t-1} + \theta_2^O R_{t-1} + \theta_3^O C_{t,2} + A_{t,2}(\theta_4^O + \theta_5^O E_{t-1} + \theta_6^O R_{t-1} + \theta_7^O C_{t,2}) + \epsilon_t^O,
\end{aligned} \tag{12}$$

where θ_I^E and θ_{IE}^E are manually set to -0.1 and 0.01, respectively. Since HeartSteps V2 includes $K = 5$ decision times per day, the general model was first fitted to HeartSteps with five actions and contexts. To construct the testbed with only one control action, we then aggregate the effects of the five contexts $C_{t,1:5}$ and five actions $A_{t,1:5}$ on R_t, E_t, and O_t, as described in Appendix D.5.

Under this general testbed model, the working transition model described in Appendix D.2 for the proposed active-measure algorithm is misspecified. We compare it against the transition model specified as the true model in (12). Only the emission model is misspecified as $O_t = \theta_0^O + \theta_1^O R_t + \epsilon_t^O$, since the emission must depend only on the current latent reward in the AOMDP framework. Furthermore, because E_t is a function of the previous latent reward R_{t-1}, it is also used to update the particle weight, similar to $M_{t,2}$. In addition, since the mediational structure no longer exists, reward design is not applied in this setting. The model-free action selection algorithm remains the same as that described in Algorithm 3 for both transition models.

We compare the cumulative rewards—after subtracting the average cumulative rewards of the zero policy—in Figure D.3. We observe that the cumulative rewards are nearly identical under the misspecified and correctly specified transition models, with the misspecified model even exhibiting smaller variance. This demonstrates the advantage of using a parsimonious model in data-scarce settings: reducing the number of parameters may increase bias but can substantially reduce variance.

E Additional Related Work

The AOMDP extends the ACNO-MDP framework by allowing states Z_t and emissions O_t or $I_{t,1}R_t$ to be observed between control and measure actions, thereby providing more accurate latent-state estimation. While $I_{t,1}$ measures U_t under our indexing, in ACNO-MDP the measurement action I_{t-1} at time $t-1$ measures U_t. Our current indexing implicitly assumes that U_t and O_t occur before $I_{t,1}$. The definition of $\mathbb{O}$, which depends only on the latent variable U_t, follows that of Liu et al. (2022).

Bellinger et al. (2021) simultaneously chose the optimal control–measure action pair in tabular settings, but the learned policy always converged to non-measuring (see details in Krale et al., 2023). Bellinger et al. (2022) applied off-the-shelf deep RL algorithms to select the action pair in continuous settings based on the last measured state with a stale observation flag. Nam et al. (2021) proposed a heuristic for estimating latent states in continuous settings. They updated the transition function by maximizing the log-likelihood of the emission given the encoded

history and action, and then drew particles from the estimated transition functions. However, the estimated unknown parameters and unobserved latent states are not necessarily drawn from their posterior distributions given the observed history. First, the transition parameters were estimated via maximum likelihood rather than by constructing a posterior. Moreover, the particle weights were not updated according to the SMC framework. Under SMC, the weights remain unchanged when the state is unmeasured and become proportional to the likelihood of the latent-state value given the history when the state is measured.

Among the algorithms proposed for tabular settings, those introduced by Krale et al. (2023, 2024) were heuristic methods that partially ignored future state uncertainty. Avalos et al. (2024) also distinguished between the states of the two actions and separated the two decision steps, but they assumed a pre-known model and focused on planning. In continuous-time RL, Holt et al. (2023) assumed noisy emissions and proposed an offline, continuous-time, model predictive control (MPC) planner.