

---

# Informed Asymmetric Actor-Critic: Leveraging Privileged Signals Beyond Full-State Access

---

Daniel Ebi<sup>1\*</sup> Damien Ernst<sup>2</sup> Klemens Böhm<sup>1</sup> Gaspard Lambrechts<sup>3,4</sup>

<sup>1</sup> Karlsruhe Institute of Technology <sup>2</sup> University of Liège

<sup>3</sup> McGill University <sup>4</sup> Mila - Québec AI Institute

## Abstract

Asymmetric reinforcement learning leverages privileged information available during training to improve learning under partial observability. Existing asymmetric actor-critic methods typically assume access to the full environment state to condition the critic during training, which is often unrealistic in practice. We introduce the informed asymmetric actor-critic framework that allows the critic to be conditioned on arbitrary state-dependent privileged signals, and show that any such signal yields unbiased policy gradient estimates. This substantially expands the set of admissible privileged information and raises the problem of selecting the most informative signals for learning. To this end, we propose two novel informativeness criteria: a dependence-based test that can be applied prior to training, and a test based on improvements in value prediction that can be applied post hoc. Experiments on partially observable benchmarks and synthetic environments demonstrate that carefully selected privileged signals can match or outperform full-state asymmetric baselines while relying on strictly less state information.

## 1 Introduction

Reinforcement learning (RL) has emerged as a powerful tool for optimizing control policies in various domains, including heating, ventilation, and air conditioning systems [1], energy systems [2, 3], autonomous driving [4], and robotics [5]. However, many real-world applications involve partial observability, where agents must make decisions based on incomplete and noisy observations. This setting is formalized by partially observable Markov decision processes (POMDPs) [6], where optimal actions depend on the history of past observations and actions.

To address this, RL methods for fully observable settings have been adapted by learning history-dependent policies, typically using recurrent neural networks (RNNs) that encode observation-action sequences [7–12]. Many approaches compress these sequences into compact latent representations by introducing auxiliary learning objectives [13–21]. While these methods can, in principle, learn optimal policies, they assume the same level of observability during training and execution, restricting policy learning to the limited information available at deployment. Yet, in practice, this assumption is unnecessarily restrictive and possibly suboptimal. Many training environments provide additional information that is unavailable or impractical to access during execution, including multi-view sensory input in robotics [22], simulator states available during training in simulation, expert policies, and external knowledge obtained from foundation models by querying representations conditioned on observations and eventual additional context. These signals do not necessarily correspond to the full state or satisfy the Markov property, yet they can provide useful information for improving learning.

Several approaches have explored the use of privileged information. For example, some methods train privileged expert policies conditioned on the true state and then imitate them [23]. However, these

---

\*Correspondence to: Daniel Ebi <daniel.ebi@kit.edu>.

methods often lack theoretical guarantees and may lead to suboptimal policies in POMDPs [24]. To mitigate this, Warrington et al. [24] propose constraining the expert policy to yield an optimal policy under partial observability. Another line of work exploits privileged information in model-based RL by constructing world models that summarize histories or integrate additional state signals. Examples include the Informed Dreamer [25], the Wasserstein Believer [26], and the Scaffolder [27].

Finally, asymmetric actor-critic methods offer a promising approach for leveraging privileged information by conditioning the actor on observable histories and the critic on privileged signals during training [28–31]. Performance improvements arise from better value estimation rather than richer actor inputs, as the actor does not access privileged signals. Pinto et al. [22] introduce an early asymmetric actor-critic method with a state-conditioned critic that achieves strong empirical performance. However, this formulation is generally ill-defined in POMDPs unless additional assumptions hold, such as state-decodability, where the belief over latent states collapses to a Dirac distribution given the history [32]. Baisero and Amato [32] address this issue by introducing the history-state value function to ensure unbiased gradients. Theoretical work has established convergence guarantees for policy gradient and actor-critic methods in fully observable settings [33, 34]. More recently, these analyses have been extended to partially observable settings, including symmetric recurrent natural actor-critic methods using RNNs [12] and asymmetric settings with fixed agent state and linear function approximators [35] or tabular belief-weighted formulations [36]. However, existing asymmetric actor-critic approaches often assume full-state access, leaving the case of privileged partial signals underexplored.

In this work, we generalize asymmetric actor-critic methods by introducing the informed asymmetric actor-critic framework, which allows the critic to be conditioned on arbitrary state-dependent privileged signals without requiring full-state access. We show that such state-conditioned signals can be leveraged while preserving unbiased policy gradients. This substantially broadens the class of admissible privileged training-time signals and extends the scope of asymmetric actor-critics, but also raises a natural question: if any state-conditioned signal is admissible, which ones should be selected?

Since there is no inherent criterion for preferring one admissible signal over another, the selection and evaluation of useful signals becomes a central challenge in asymmetric RL. We provide a first systematic answer to this question by introducing two informativeness criteria that assess the utility of additional signals w.r.t. value estimation: (i) a residual-based dependency measure, and (ii) a value-prediction-based measure. Related work in causal decision-making and bandit settings studies variable selection using explicit causal models to identify variables relevant for optimal policies [37–39]. In contrast, we do not assume access to such a structure and instead study privileged signals from a statistical perspective focused on their utility for value estimation.

We empirically evaluate our framework on benchmark environments and synthetic informed POMDPs. Our results show that leveraging informative privileged signals can significantly improve policy learning, challenging the assumption that full-state access is essential for effective asymmetric actor-critics. These findings underscore the importance of understanding how privileged information can be selected, evaluated, and integrated to facilitate learning in partially observable settings.

## 2 Background

In this section, we formally introduce partially observable Markov decision processes (POMDPs) and the symmetric actor-critic paradigm, and then describe the informed POMDP framework that motivates our informed asymmetric actor-critic framework.

### 2.1 Partially Observable Markov Decision Processes

A POMDP [6] models sequential decision-making under uncertainty as tuple  $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$ , where  $\mathcal{S}$ ,  $\mathcal{A}$ , and  $\mathcal{O}$  denote the state, action, and observation spaces. The transition probabilities  $T(s_{t+1}|s_t, a_t)$  describe the process dynamics. The environment emits observations via  $O(o_t|s_t)$  at time  $t$  and selects actions  $a_t$  based on the observable history  $h_t$ , defined as the sequence of past observations and actions, including the current observation. Specifically, the history can be written as  $h_t = (o_0, a_0, \dots, o_t)$ . We define the set of observable histories as  $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}_t$ , where  $\mathcal{H}_t \subseteq \mathcal{O} \times (\mathcal{A} \times \mathcal{O})^t$  is the set of histories of size  $t$ . The immediate reward of the agent

is given by  $r_t := R(s_t, a_t)$ , with  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ . We assume that there exists a constant  $r_{\max} > 0$  such that  $|r_t| \leq r_{\max}$  for all  $t$ .  $P$  specifies the initial state distribution. The objective is to maximize the expected return  $J(\pi) = \mathbb{E}^\pi [\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$ , where  $\pi(a_t|h_t)$  denotes a history-dependent policy and  $\gamma \in [0, 1]$  is a discount factor. The future return from time  $t$  is defined as  $G_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k})$ , which leads to the history-based  $Q$ -function, the expected return given the current history and action,

$$Q^\pi(h_t, a_t) = \mathbb{E}_{s_{t:\infty}, a_{t+1:\infty}}^\pi [G_t | h_t, a_t],$$

and the corresponding history value function

$$V^\pi(h_t) = \sum_{a_t \in \mathcal{A}} \pi(a_t | h_t) Q^\pi(h_t, a_t).$$

## 2.2 Symmetric Actor-Critic Paradigm

Given a history-dependent policy  $\pi_\theta$  parametrized by  $\theta$  (actor), the policy gradient in a POMDP [40, 41] is given by

$$\nabla_\theta J(\pi_\theta) = \mathbb{E} \left[ \sum_t \gamma^t Q^\pi(h_t, a_t) \nabla_\theta \log \pi_\theta(a_t | h_t) \right]. \quad (1)$$

In practice, this expectation is estimated via Monte Carlo by sampling histories and actions in the POMDP, which is known to yield high-variance estimates. It has been shown that subtracting any function of the history, referred to as a baseline, from the  $Q$ -function does not bias the gradient while potentially reducing variance. Commonly, the value function  $V^\pi(h_t)$  is used as this baseline, yielding the advantage formulation

$$\nabla_\theta J(\pi_\theta) = \mathbb{E} \left[ \sum_t \gamma^t A^\pi(h_t, a_t) \nabla_\theta \log \pi_\theta(a_t | h_t) \right], \quad (2)$$

where  $A^\pi(h_t, a_t) = Q^\pi(h_t, a_t) - V^\pi(h_t)$  is the so-called advantage function. Unfortunately, the exact forms in Equations 1-2 involve the unknown functions  $Q^\pi(\cdot)$  and  $V^\pi(\cdot)$ . The  $Q$ -function  $Q^\pi(h_t, a_t)$  can, in principle, be estimated using a single Monte Carlo sample of  $G_t$ , but  $V^\pi(h_t)$  cannot because the sampled  $G_t$  is not conditionally independent of  $a_t$  given  $h_t$ . Practitioners typically approximate  $Q^\pi(\cdot)$  or  $V^\pi(\cdot)$  with parametric functions  $Q_\theta^\pi$  or  $V_\theta^\pi$  (critic), using temporal difference (TD) learning. Since a baseline requires  $V^\pi(\cdot)$ , many algorithms avoid estimating  $Q^\pi$  directly and instead rely on an advantage estimator that uses only the critic,  $\hat{A}_t = r_t + \gamma V_\theta^\pi(h_{t+1}) - V_\theta^\pi(h_t)$ , where  $r_t$  is the observed reward at time  $t$ . This estimator of the advantage is unbiased if  $V_\theta^\pi$  perfectly approximates  $V^\pi$ .

## 2.3 Informed POMDPs

Modeling the training-execution asymmetry, in which the critic accesses additional state-dependent signals during training, but the policy relies only on observation-action histories, requires a framework that formalizes these signals without altering execution-time dynamics. The informed POMDP formalism [25] provides a principled approach by augmenting the standard POMDP with an information variable. Specifically, the informed POMDP introduces an information space  $\mathcal{I}$  and an information function  $I: \mathcal{S} \rightarrow \Delta(\mathcal{I})$ , which gives the probability to obtain information  $i_t \in \mathcal{I}$  in the true state  $s_t \in \mathcal{S}$ . Hence, the informed POMDP is defined by the 10-tuple  $(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, I, \tilde{\mathcal{O}}, R, P, \gamma)$ . In contrast to the standard POMDP, the observation function is defined as  $\tilde{\mathcal{O}}: \mathcal{I} \rightarrow \Delta(\mathcal{O})$  and denotes the probability to obtain  $o_t \in \mathcal{O}$  given  $i_t \in \mathcal{I}$ . The defining assumption of the informed POMDP is that the observation  $o_t$  is conditionally independent of the true state  $s_t$  given the information  $i_t$ , i.e.,  $p(o_t | i_t, s_t) = \tilde{\mathcal{O}}(o_t | i_t)$ .<sup>1</sup> Each informed POMDP induces an underlying execution POMDP defined as  $(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, P, \gamma)$ , with observation function  $O(o_t | s_t) = \mathbb{E}_{i_t | s_t} [\tilde{\mathcal{O}}(o_t | i_t)]$ , which governs the agent's interaction at deployment.

<sup>1</sup>This formulation is not restrictive. Suppose a POMDP with observation  $o_t$  and auxiliary observation  $o_t^+$  such that  $(o_t, o_t^+) \sim O^+(o_t, o_t^+ | s_t)$ . Then, one can always define an information  $i_t = (o_t, o_t^+)$  such that  $p(i_t, o_t | s_t) = I(i_t | s_t) \tilde{\mathcal{O}}(o_t | i_t)$  holds.

Thus, the conditional-independence assumption offers a convenient modeling abstraction for introducing an information variable that is at least as informative about the underlying state as the execution-time observation, while preserving full generality. As a result, the informed POMDP framework naturally subsumes both standard POMDPs and settings with additional training-time signals, making it particularly well-suited for modeling asymmetric learning scenarios.

### 3 Informed Asymmetric Actor-Critic

In this section, we develop an asymmetric actor-critic framework that leverages arbitrary state-conditioned information during training. Building on the informed POMDP paradigm, we introduce informed history-based  $Q$ - and value functions and derive an unbiased policy gradient estimator. Our formulation generalizes the history-state critic of Baisero and Amato [32]: instead of requiring full-state access, we show that any privileged signal  $i_t \sim I(i_t|s_t)$  suffices to obtain unbiased value estimates, without imposing additional assumptions on the POMDP. This expands the set of usable training-time signals in asymmetric actor-critics, motivating the question of how to select among them.

#### 3.1 Informed History-based Value Functions

Given an informed POMDP, we define the informed history-based reward as

$$R(h_t, i_t, a_t) = \mathbb{E}_{s_t} [R(s_t, a_t) | h_t, i_t]. \quad (3)$$

This reward function leverages additional state-conditioned information available during training while remaining unbiased w.r.t. the standard history-based reward when taking expectations over  $p(i_t|h_t)$  (cf. Lemma A.1 in Appendix A).

**Informed history  $Q$ -function.** We next introduce the informed history  $Q$ -function, which extends the history-state critic of Baisero and Amato [32] to arbitrary state-conditioned privileged signals:

$$Q^\pi(h_t, i_t, a_t) = \mathbb{E}_{s_{t:\infty}, a_{t+1:\infty}} [G_t | h_t, i_t, a_t].$$

This  $Q$ -function is an unbiased estimator of the standard history-based counterpart (cf. Lemma A.2), i.e.,  $\mathbb{E}_{i_t|h_t} [Q^\pi(h_t, i_t, a_t)] = Q^\pi(h_t, a_t)$ .

**Informed history value function.** Based on the proposed informed history  $Q$ -function, we can define the informed history value function:

$$V^\pi(h_t, i_t) = \mathbb{E}_{s_{t:\infty}, a_{t:\infty}} [G_t | h_t, i_t], \quad (4)$$

which admits the time-invariant policy-based decomposition

$$V^\pi(h, i) = \sum_{a \in \mathcal{A}} \pi(a|h) Q^\pi(h, i, a). \quad (5)$$

The corresponding Bellman equation is

$$Q^\pi(h, i, a) = R(h, i, a) + \gamma \mathbb{E}_{o', i' | h, i, a} [V^\pi(h', i')], \quad (6)$$

where  $h' = hao'$ . Analogously to the  $Q$ -function, the informed history value function is an unbiased estimator of the standard history value (cf. Lemma A.3), i.e.,  $\mathbb{E}_{i_t|h_t} [V^\pi(h_t, i_t)] = V^\pi(h_t)$ .

In our setting, the privileged signal  $i_t$  provides additional state-dependent context that can reduce the ambiguity about the underlying state given a fixed history. Formally, letting  $H(s_t|h_t)$  denote the conditional entropy of the state, standard information-theoretic results imply that conditioning on  $i_t$  never increases, and on average reduces, uncertainty about the true state, i.e.,  $E_{i_t|h_t} [H(s_t|h_t, i_t)] \leq H(s_t|h_t)$ . This does not guarantee that  $V^\pi(h_t, i_t)$  is inherently easier to approximate than  $V^\pi(h_t)$ ; rather, it is the conditional expectation of a return random variable whose variance may be lower. In particular, by the law of total variance applied to the return  $G_t$ ,  $E_{i_t|h_t} [\text{Var}(G_t|h_t, i_t)] \leq \text{Var}(G_t|h_t)$ , so an appropriate choice of  $i_t$  can reduce the variance of value targets. It mainly occurs in environments with high “value aliasing”, where distinct states yielding different returns may exist under the same observable history [35]. In such cases,  $i_t$  can help disambiguate the state, providing a cleaner target for the critic.

When the privileged signal coincides with the full state,  $i_t = s_t$ , our formulation reduces to the history-state value function of Baisero and Amato [32] (cf. Corollary B.1 in Appendix B), showing that their framework is a special case of our more general formalization.

The careful reader will have noticed that all definitions and unbiasedness results remain valid when replacing  $i_t$  with an arbitrary auxiliary observation  $o_t^+$ . Indeed, since  $h_t$  contains  $o_t$  by definition, defining  $\tilde{i}_t = (o_t, o_t^+)$  implies that conditioning on  $(h_t, o_t^+)$  is equivalent to conditioning on  $(h_t, \tilde{i}_t)$ , so all results hold.

### 3.2 Informed Asymmetric Policy Gradient

Using the informed history-based  $Q$ -function, we define the informed asymmetric policy gradient

$$\nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right],$$

where the policy  $\pi_{\theta}$  depends only on the history  $h_t$ , while the critic may additionally condition on training-time information  $i_t$ .

Incorporating privileged state-conditioned information into the critic does not bias the policy gradient: for any choice of  $i_t$ , the informed asymmetric policy gradient coincides with the standard policy gradient.

**Theorem 3.1** (Informed asymmetric policy gradient). *Given an informed POMDP, the informed asymmetric policy gradient is equivalent to the standard policy gradient. Formally,*

$$\nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}) = \nabla_{\theta} J(\pi_{\theta}).$$

*Proof.* Given Equation 1 and Lemma A.2, we have

$$\begin{aligned} \nabla_{\theta} J(\pi_{\theta}) &= \mathbb{E} \left[ \sum_t \gamma^t Q^{\pi}(h_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &\stackrel{(a)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, a_t} \left[ Q^{\pi}(h_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &\stackrel{(b)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, a_t} \left[ \mathbb{E}_{i_t | h_t} \left[ Q^{\pi}(h_t, i_t, a_t) \right] \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &\stackrel{(c)}{=} \sum_t \gamma^t \mathbb{E}_{h_t, i_t, a_t} \left[ Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &\stackrel{(d)}{=} \mathbb{E} \left[ \sum_t \gamma^t Q^{\pi}(h_t, i_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t) \right] \\ &= \nabla_{\theta}^{\text{IAAC}} J(\pi_{\theta}). \end{aligned}$$

Here, (a) and (d) follow from linearity of expectation and the assumption that the infinite discounted series is absolutely integrable, allowing interchange of expectation and summation; (b) uses Lemma A.2, and (c) applies the law of total expectation. This concludes the proof.  $\square$

This result generalizes prior work on asymmetric actor-critic methods beyond state-based critics and recovers the asymmetric policy gradient of Baisero and Amato [32] as a special case when  $i_t = s_t$  (cf. Corollary B.2). Moreover, it extends straightforwardly to value functions that take an arbitrary auxiliary observation  $o_t^+$  in place of  $i_t$  as input.

Hence, Theorem 3.1 provides a theoretical justification for using arbitrary state-conditioned privileged signals during training. It shows that, under standard policy-gradient assumptions, symmetric and informed actor-critic methods optimize the same objective  $J(\pi_{\theta})$  and share the same first-order stationary condition  $\nabla_{\theta} J(\pi_{\theta}) = 0$ . Signal choice thus mainly affects optimization through changes in the variance of the value estimates, rather than asymptotic optimality.

In particular, our framework supports the use of a state-conditioned expert policy as a privileged signal, i.e.,  $i_t = a_t^* \sim \pi^*(\cdot|s_t)$ , relating to ideas in asymmetric imitation learning while avoiding gradient bias. Indeed,  $a_t^*$  depends on the environment state  $s_t$ , and Theorem 3.1 guarantees unbiased policy gradients for the history-dependent actor, even when the expert policy itself is not deployable under partial observability. This enables the critic to exploit oracle information for value estimation without requiring the actor to perform direct imitation, which was shown to be suboptimal in asymmetric settings [24].

**Informed history critic.** As in symmetric actor-critic methods, we learn a value function using TD learning, and we use the TD error to form a low-variance informed advantage estimate  $\hat{A}_t^{\text{IAAC}}$ . The informed history critic  $\hat{V}: \mathcal{H} \times \mathcal{I} \rightarrow \mathbb{R}$  additionally receives a privileged signal and approximates the informed history value  $V^\pi(h_t, i_t)$ . Combined with a history-dependent actor  $\hat{\pi}_\theta(a_t|h_t)$ , this yields the *informed asymmetric actor-critic (IAAC)*. The resulting policy gradient estimator is

$$\hat{\nabla}_\theta^{\text{IAAC}} J(\pi_\theta) = \mathbb{E} \left[ \sum_t \gamma^t \hat{A}_t^{\text{IAAC}} \nabla_\theta \log \hat{\pi}_\theta(a_t|h_t) \right].$$

## 4 Informativeness of Privileged Signals

While the informed asymmetric actor-critic framework guarantees unbiased policy gradients for any state-conditioned privileged signals, not all such signals are equally useful for learning in practice. The choice of  $i_t \in \mathcal{I}$  can affect the critic’s learning efficiency and stability, motivating the need for criteria to select the most informative signals.

Here, we focus on the informativeness of  $i_t$  from the perspective of predicting future returns, which is the critic’s primary role. To this end, we introduce two criteria that quantify the informativeness of privileged signals and enable systematic comparison across different choices of  $i_t$ .

### 4.1 Residual Informativeness

The utility of a privileged signal  $i_t$  in an informed critic depends on whether it provides information about future returns beyond what is already encoded in the observable history-action pair  $(h_t, a_t)$ . Ideally, this question can be formalized through the conditional independence (CI) hypothesis  $\mathbb{H}_0^{\text{CI}}: G_t \perp\!\!\!\perp i_t | h_t, a_t$ , which corresponds to the factorization

$$p(G_t, i_t, h_t, a_t) = p(h_t, a_t)p(G_t|h_t, a_t)p(i_t|h_t, a_t).$$

Rejecting  $\mathbb{H}_0^{\text{CI}}$  indicates that  $i_t$  contains additional, non-redundant information about future returns.

Such a test requires access to samples from the conditional independence null distribution corresponding to  $\mathbb{H}_0^{\text{CI}}$ . Unfortunately, directly obtaining samples from this distribution is generally infeasible. Instead, interaction with the environment yields samples from the observed joint distribution  $p(G_t, i_t, h_t, a_t) = p(h_t, a_t)p(G_t, i_t|h_t, a_t)$ , where conditional independence cannot be assumed. Sampling independently from  $p(G_t|h_t, a_t)$  would require learning this conditional distribution, which is challenging for high-dimensional histories, and may introduce additional estimation noise into any direct test.

To circumvent this, we first remove the components of  $G_t$  and  $i_t$  that are predictable from  $(h_t, a_t)$ , and then test whether their remaining components are dependent.

**Definition 4.1** ( $\alpha$ -residual informativeness). Let  $G_t$ ,  $i_t$ ,  $h_t$ , and  $a_t$  be random variables with finite second moments. Define the conditional-mean residuals

$$e_{G_t} := G_t - \mathbb{E}[G_t|h_t, a_t], \quad e_{i_t} := i_t - \mathbb{E}[i_t|h_t, a_t].$$

A privileged signal  $i_t$  is  $\alpha$ -residual-informative if a statistical test of independence rejects the null hypothesis

$$\mathbb{H}_0^{\text{res}}: e_{G_t} \perp\!\!\!\perp e_{i_t},$$

in favor of the alternative  $\mathbb{H}_1^{\text{res}}: e_{G_t} \not\perp\!\!\!\perp e_{i_t}$ , at significance level  $\alpha > 0$ .

Under  $\mathbb{H}_0^{\text{CI}}$ , the residuals satisfy  $e_{G_t} \perp\!\!\!\perp e_{i_t} | h_t, a_t$ ; see also Proposition C.1 in Appendix C. Thus, conditional independence implies conditional residual independence, although the converse does

---

**Algorithm 1**  $\alpha$ -Residual-Informativeness Test

---

- 1: **Input:** Dependency measure  $\rho$ , significance level  $\alpha$ , number of folds  $K$ , number of permutations  $B$ , and episodes  $\{(o_{n,t}, a_{n,t}, i_{n,t}, G_{n,t})_{t=0}^{T-1}\}_{n=0}^{N-1}$ .
  - 2: Encode histories with a recurrent net:  $z_t = f_{\text{RNN}}(h_t)$ .
  - 3: Partition data into  $K$  folds.
  - 4: **for** each fold  $k = 1$  to  $K$  **do**
  - 5:   Train regression models on  $K - 1$  folds:
  - 6:   Predict  $\hat{\mathbb{E}}[G_t|z_t, a_t]$  and  $\hat{\mathbb{E}}[i_t|z_t, a_t]$  on held-out fold.
  - 7:   Compute residuals  $\hat{e}_{G_t}$  and  $\hat{e}_{i_t}$  on held-out fold (Eq. 7).
  - 8: **end for**
  - 9: Aggregate cross-fitted residuals across held-out folds.
  - 10: Compute observed dependence  $\rho_{\text{obs}}$  using  $\rho(\cdot, \cdot)$ .
  - 11: **for**  $b = 1$  to  $B$  **do**
  - 12:   Permute residuals  $\hat{e}_{G_t}$  on episode-level.
  - 13:   Compute  $\rho(\hat{e}_{G_t}^{(b)}, \hat{e}_{i_t})$  for permuted residuals.
  - 14: **end for**
  - 15: Compute empirical  $p$ -value according to Eq. 8.
  - 16: **Output:** Declare  $i_t$   $\alpha$ -residual-informative if  $p < \alpha$ .
- 

not hold in general [42]. As with the CI test, directly testing conditional residual independence is infeasible because the conditional residual distributions  $p(e_{G_t}|h_t, a_t)$  and  $p(e_{i_t}|h_t, a_t)$  are unknown.

Instead, we test the unconditional residual independence hypothesis  $\mathbb{H}_0^{\text{res}}$ . This relaxation is motivated by our actual objective: detecting whether  $i_t$  explains variation in  $G_t$  that remains unexplained after accounting for  $(h_t, a_t)$ . Consequently, rejecting residual independence provides evidence that the privileged signal contributes predictive information beyond the observable history-action pair.

To approximate the null distribution required for this residual test, we construct a surrogate return whose predictable component matches that of  $G_t$ , while its residual component is independent of  $i_t$ . Specifically, we (i) sample  $\tilde{G}_t$  independently from the marginal distribution  $p(G_t)$ , which is easily feasible from the collected episodes, and (ii) shift these samples to preserve the estimated conditional mean  $\mathbb{E}[G_t|h_t, a_t]$ . Thus,  $\tilde{G}_t^{\text{null}} = \tilde{G}_t - \mathbb{E}[\tilde{G}_t] + \mathbb{E}[G_t|h_t, a_t]$ .

By construction, the surrogate matches the first conditional moment of the target distribution  $p(G_t|h_t, a_t)$ , while its higher-order moments follow those of the marginal distribution  $p(G_t)$ .

The corresponding null residual

$$e_{\tilde{G}_t}^{\text{null}} := \tilde{G}_t^{\text{null}} - \mathbb{E}[G_t|h_t, a_t] = \tilde{G}_t - \mathbb{E}[\tilde{G}_t]$$

is independent of  $e_{i_t}$ , and thus satisfies  $\mathbb{H}_0^{\text{res}}$ . Further details on the surrogate construction and its relation to conditional residual independence are given in Appendix C.

**Practical Implementation.** In practice, we approximate  $\mathbb{E}[G_t|h_t, a_t]$  and  $\mathbb{E}[i_t|h_t, a_t]$  using regression models capable of capturing nonlinear relationships (e.g., random forests or neural networks), since  $G_t$  and  $i_t$  typically exhibit nonlinear dependence. Histories are generally encoded using a recurrent network to produce a fixed-length representation  $z_t = f_{\text{RNN}}(h_t)$ , yielding residual estimates

$$\hat{e}_{G_t} := G_t - \hat{\mathbb{E}}[G_t|z_t, a_t], \quad \hat{e}_{i_t} := i_t - \hat{\mathbb{E}}[i_t|z_t, a_t], \quad (7)$$

and  $\hat{e}_{\tilde{G}_t}^{\text{null}} := \tilde{G}_t^{\text{null}} - \hat{\mathbb{E}}[G_t|z_t, a_t]$  for the null sample. We quantify the dependence between  $\hat{e}_{G_t}$  and  $\hat{e}_{i_t}$  using a general dependency measure  $\rho(\cdot, \cdot)$  (e.g., mutual information or Hilbert-Schmidt Independence Criterion). By design,  $\rho(\hat{e}_{G_t}, \hat{e}_{i_t}) = 0$  under independence and increases with stronger dependence. To account for finite-sample variability and temporal correlations, we approximate the surrogate null by episode-wise permutation of the observed residuals. Residuals  $\hat{e}_{G_t}$  are shuffled across independently collected episodes, while keeping the temporal ordering within each episode unchanged, thereby preserving within-episode dependence structure and the estimated conditional mean. Cross-fitting is used to prevent overfitting: the data set is split into  $K$  folds, with regressors trained on  $K - 1$  folds and evaluated on the held-out fold.

We compute the empirical  $p$ -value with  $B$  permutations as

$$p = \frac{1 + \sum_{b=1}^B \mathbf{1}\{\rho(\hat{e}_{G_t}^{(b)}, \hat{e}_{i_t}) \geq \rho(\hat{e}_{G_t}, \hat{e}_{i_t})\}}{B + 1}, \quad (8)$$

where  $\hat{e}_{G_t}^{(b)}$  denotes the residual under the  $b$ -th permutation and  $\mathbf{1}\{\cdot\}$  is the indicator function, which equals 1 if the condition inside the braces holds and 0 otherwise. The privileged signal  $i_t$  is declared  $\alpha$ -residual-informative if  $p < \alpha$ . Algorithm 1 summarizes the procedure.

Notably, the  $\alpha$ -residual-informativeness criterion can be evaluated on episodes collected under any exploratory policy, without requiring a trained actor or critic. In particular, a random policy can be used, although the resulting data may suffer from insufficient state-action coverage, which can limit the reliability of the estimated quantities. Nevertheless, the criterion allows estimating the informativeness of privileged signals prior to training and supports informed signal selection. To obtain meaningful residuals, the history encoding function  $f_{\text{RNN}}(\cdot)$  is typically trained in advance to produce effective representations of the interaction history. More generally, any encoding scheme (e.g., fixed-window or padding-based approaches) can be used, as long as it maps the history to a fixed-length representation.

As detailed in Appendix C.2, for Lipschitz-continuous dependency measures, perturbations in the residuals can be bounded in terms of the regression error. If this error does not decrease with increasing sample size, the resulting test may deviate from its nominal significance level and exhibit reduced statistical power. To address this, we use cross-fitting and random forests, which are typically less prone to overfitting than deep neural networks in low-data regimes.

## 4.2 Informativeness via Return Prediction Gain

Informativeness can also be assessed via the critic’s predictive accuracy: a privileged signal is informative if its inclusion improves value estimation. This post-hoc criterion provides a quantitative measure of a signal’s contribution to return prediction. It requires only episodes collected under any fixed policy and does not depend on policy performance.

Consider a symmetric critic  $\hat{Q}(h_t, a_t)$  and an informed asymmetric critic  $\hat{Q}(h_t, i_t, a_t)$ . For a set of  $N$  episodes  $\{\tau_j\}_{j=1}^N$ , we define the episode-level squared-error gain

$$L^{\tau_j} := \frac{1}{T_j} \sum_{t=0}^{T_j-1} ((\hat{Q}(h_t, a_t) - G_t)^2 - (\hat{Q}(h_t, i_t, a_t) - G_t)^2),$$

where  $T_j$  is the length of episode  $\tau_j$ .  $L^{\tau_j} > 0$  indicates that including  $i_t$  improves return prediction on that episode.

**Definition 4.2** ( $(\epsilon, \delta)$ -prediction informativeness). A privileged signal  $i_t$  is said to be  $(\epsilon, \delta)$ -prediction-informative with  $\epsilon \geq 0$  if a statistical test rejects the null hypothesis

$$\mathbb{H}_0 : \mathbb{E}[L^\tau] \leq \epsilon,$$

in favor of the alternative  $\mathbb{H}_1 : \mathbb{E}[L^\tau] > \epsilon$ , at significance level  $\delta$ , based on the set of episode-level gains  $\{L^{\tau_j}\}_{j=1}^N$ .

**Practical Implementation.** In practice, the test procedure depends on the number of episodes collected  $N$ . For small sample sizes, we apply a one-sided bootstrap test: we resample  $B > N$  episodes with replacement, compute bootstrap means  $\bar{L}_b^*$  for  $1 \leq b \leq B$ , and estimate the empirical  $p$ -value as  $p = \frac{1 + \sum_{b=1}^B \mathbf{1}\{\bar{L}_b^* \leq 0\}}{B + 1}$ , rejecting  $\mathbb{H}_0$  if  $p < \delta$ . For larger sample sizes (e.g.,  $N > 1,000$ ), we apply a one-sided  $t$ -test based on the sample mean  $\bar{L}$  and variance  $\hat{\sigma}^2$ . The test statistic is  $t = \frac{\bar{L} - \epsilon}{\hat{\sigma}/\sqrt{N}}$ , which under  $\mathbb{H}_0$  follows a Student- $t$  distribution with  $N - 1$  degrees of freedom. We reject  $\mathbb{H}_0$  if the corresponding one-sided  $p$ -value is below  $\delta$ .

## 4.3 Selecting Privileged Signal Generators

Typically, we consider a state-conditioned candidate feature vector  $c_t = (c_t^1, \dots, c_t^M)$ , with  $M \in \mathbb{Z}_{\geq 0}$ , where each component corresponds to a candidate feature extracted from the underlying state or auxiliary training-time information. A privileged signal is then formed by selecting a

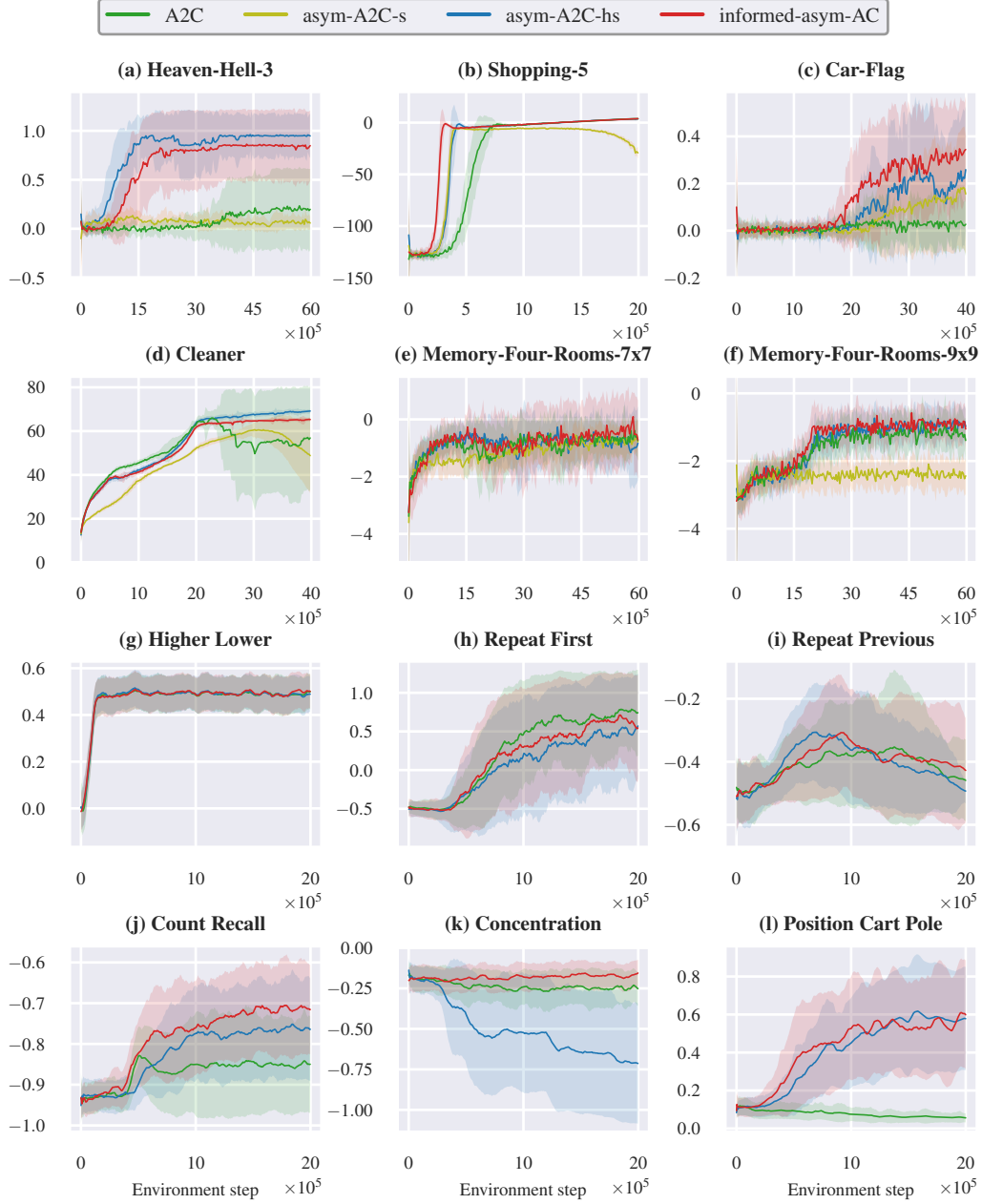


Figure 1: Learning performance on six benchmark navigation tasks (a-f) and six POPGym environments (g-l). Curves show mean episodic returns smoothed using a moving average over 100 episodes. Means and standard deviations are computed across 20 independent runs.

subset of components  $\mathcal{Z} \subseteq \{1, \dots, M\}$ , yielding the signal  $i_t := (c_t^m)_{m \in \mathcal{Z}}$ . Different subsets may encode distinct aspects of the environment state and are not assumed to be equally informative a priori.

Both proposed informativeness criteria support principled selection over such feature subsets via hypothesis testing. For all subsets, we evaluate the sequence  $\{(c_t^m)_{m \in \mathcal{Z}}\}_{t=0}^T$  and deem the feature subset informative if the corresponding null hypothesis is rejected. In the pre-training setting, we apply the  $\alpha$ -residual-informativeness test and retain subsets for which residual independence is

rejected at significance level  $\alpha$ . In the post-hoc setting, we apply the  $(\epsilon, \delta)$ -prediction-informativeness test and retain subsets that yield a statistically significant improvement in value prediction accuracy ( $\epsilon = 0$ ). Among informative candidates, one may select the feature subset with the largest effect size or smallest  $p$ -value, corresponding to the privileged signal that provides the strongest additional information about future returns under the respective criterion.

When only a single  $i_t$  is available or when  $c_t$  cannot be naturally decomposed into  $M$  components, the same procedure can still be used to compare different learned encodings of the signal, or of the observation in the degenerate case  $i_t = o_t$ . In all cases, Theorem 3.1 applies without assumptions on the dimensionality or structure of  $i_t$ .

## 5 Experiments

We evaluate our informed asymmetric actor-critic framework along two dimensions. First, we assess its performance on standard POMDP benchmarks against symmetric and asymmetric baselines. Second, we validate the proposed informativeness criteria in controlled informed POMDPs, where privileged signals can be systematically varied. We release our source code on GitHub<sup>2</sup> to ensure reproducibility and report wall-clock runtimes in Appendix F. Ablations and further results can be found in Appendix G.

### 5.1 Benchmark Performance

**Environments.** We use six benchmark navigation tasks from Baisero and Amato [32] and six environments from the POPGym suite [43]. The navigation tasks consist of *Heaven-Hell-3*, *Shopping-5*, *Car-Flag*, *Cleaner*, *Memory-Four-Rooms-7x7*, and *Memory-Four-Rooms-9x9*, while the POPGym suite includes *Higher Lower*, *Repeat First*, *Repeat Previous*, *Count Recall*, *Concentration*, and *Position Cart Pole*. For each environment, we define task-specific privileged signals available only to the critic; the actor receives the same inputs, except for the additional information. Detailed descriptions of all environments are provided in Appendix D.1.

**Baselines.** We compare our informed asymmetric actor-critic method, *informed-asym-A2C*, against three advantage actor-critic (A2C) variants: (1) *A2C*, a symmetric approach using a history-based critic  $\hat{V}(h)$ ; (2) *asym-A2C-hs*, an asymmetric variant with a history-state critic  $\hat{V}(h, s)$ ; and (3) *asym-A2C-s*, an asymmetric approach with a state-based critic  $\hat{V}(s)$ , evaluated only on the navigation tasks. We use model architectures and hyperparameters from Baisero and Amato [32] and Morad et al. [43], respectively; see Appendix E.1.

**Results.** Figure 1 reports the learning curves across all twelve benchmark tasks, showing episodic returns smoothed over 100 episodes and aggregated over 20 independent runs. For the POPGym environments, we present only the best-performing variant of *informed-asym-A2C*. Across navigation environments, our method consistently improves sample efficiency and training stability relative to A2C. In *Heaven-Hell-3*, *informed-asym-A2C* exhibits strong performance relative to A2C and *asym-A2C-s*, though it is slightly outperformed by *asym-A2C-hs*. In *Shopping-5*, the informed asymmetric actor-critic converges faster than *asym-A2C-hs* and achieves comparable asymptotic return. In *Car-Flag*, it outperforms all baselines in both convergence speed and asymptotic return. In *Cleaner*, *asym-A2C-hs* achieves marginally higher returns, but *informed-asym-A2C* converges at a similar rate with greater stability than A2C, which suffers a performance drop after 2.5 million steps. In the Memory-Four-Rooms tasks, *informed-asym-A2C* outperforms both asymmetric baselines. These findings are consistent with those in the POPGym environments. *Informed-asym-A2C* achieves the highest final returns in most of the environments. In *Higher Lower*, however, all methods perform similarly. On *Repeat First*, A2C performs best, while on *Position Cart Pole* both asymmetric variants clearly outperform A2C, which fails to converge within 2 million environment steps. For *Concentration*, the full-state A2C variant appears to struggle with convergence, potentially due to the high-dimensional state representation containing many less relevant or irrelevant features.

Overall, these results demonstrate that asymmetric critics leveraging appropriately structured privileged signals can match or outperform full-state critics while relying on strictly less information.

<sup>2</sup><https://github.com/EbiDa/informed-asymmetric-a2c>

Table 1: Comparison of candidate privileged signals using the residual-based and post-hoc prediction informativeness criteria. We provide mean  $\pm$  standard deviation across ten independent runs on a randomly sampled informed POMDP with reward weights  $w_r = [0.0001, 0.0001, -0.0001, -1.0, 1.0]$ .

| Privileged signal                    | Residual Informativeness            |                             | Prediction Informativeness  |                             |
|--------------------------------------|-------------------------------------|-----------------------------|-----------------------------|-----------------------------|
|                                      | $\rho_{\text{obs}}$ ( $\uparrow$ )  | $p$ -value ( $\downarrow$ ) | $L_\tau$ ( $\uparrow$ )     | $p$ -value ( $\downarrow$ ) |
| $i_t = [s_t^1, s_t^2]$               | $3.0\text{e-}05 \pm 8.1\text{e-}06$ | $0.438 \pm 0.290$           | $-1.4\text{e-}02 \pm 0.014$ | $0.801 \pm 0.330$           |
| $i_t = [s_t^1, s_t^2, s_t^3]$        | $5.5\text{e-}05 \pm 1.9\text{e-}05$ | $0.170 \pm 0.169$           | $0.007 \pm 0.016$           | $0.397 \pm 0.338$           |
| $i_t = [s_t^1, s_t^2, s_t^4]$        | $7.6\text{e-}05 \pm 3.2\text{e-}05$ | $0.119 \pm 0.262$           | $0.035 \pm 0.018$           | $0.043 \pm 0.077$           |
| $i_t = [s_t^1, s_t^2, s_t^5]$        | $5.0\text{e-}05 \pm 1.5\text{e-}05$ | $0.191 \pm 0.228$           | $0.018 \pm 0.019$           | $0.232 \pm 0.248$           |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$ | $7.4\text{e-}05 \pm 2.4\text{e-}05$ | $0.068 \pm 0.116$           | $0.046 \pm 0.018$           | $0.009 \pm 0.016$           |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$ | $5.8\text{e-}05 \pm 1.5\text{e-}05$ | $0.106 \pm 0.141$           | $0.026 \pm 0.019$           | $0.120 \pm 0.152$           |
| $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$ | $7.6\text{e-}05 \pm 2.1\text{e-}05$ | $0.035 \pm 0.051$           | $0.064 \pm 0.020$           | $0.003 \pm 0.006$           |
| $i_t = s_t$                          | $7.0\text{e-}05 \pm 1.8\text{e-}05$ | $0.038 \pm 0.051$           | $0.056 \pm 0.032$           | $0.072 \pm 0.131$           |

## 5.2 Validation of Informativeness Criteria

**Environments.** To empirically validate the proposed informativeness criteria in a controlled setting, we consider a distribution of synthetic informed POMDPs with finite state space ( $|\mathcal{S}| = 20$ ), discrete action space ( $|\mathcal{A}| = 4$ ), and continuous observation and privileged information spaces. Transition dynamics are generated following François-Lavet et al. [44] by sparsifying state-action transitions and normalizing to obtain valid probability distributions. Each state is associated with a latent Gaussian feature vector ( $s_t \in \mathbb{R}^5$ ), and rewards are defined as linear functions of these features with weights  $w_r$  sampled uniformly from  $[-1, 1]$ . Privileged signals  $i_t$  are constructed by masking subsets of state features, while observations  $o_t$  correspond to noisy feature subsets of  $i_t$ ; see Appendix D.3 for details.

**Informativeness criteria.** For the residual-based criterion, we estimate dependence between return-prediction residuals and privileged-signal residuals via the Hilbert-Schmidt Independence Criterion (HSIC) [45] with Gaussian RBF kernels, where bandwidths are chosen via the median heuristic and a Nystrom approximation [46] with 512 landmarks is applied for efficiency. We collect 250 episodes of length 25 under a random policy and perform episode-level permutation testing with  $B = 1,000$  resamples. Conditional expectations are estimated via cross-fitting with a 100-tree random forest using 5-fold cross-validation. Observation-action histories are encoded with an RNN of width 64, trained on 100 training episodes. For the post-hoc criterion, we train symmetric and informed recurrent critics via TD learning using 5-fold cross-validation on the same environments for 2,500 episodes of length 25. In the informed critic, the 64-dimensional representation of the history is concatenated with the privileged signal and passed through a feedforward network to output a single value estimate. We evaluate episode-level gains in value prediction accuracy across held-out folds, with  $\epsilon = 0$ .

**Results.** Table 1 reports test statistics and  $p$ -values for different candidate privileged signals under both criteria across ten independent runs on a randomly sampled environment. In this setting, the agent observes noiseless  $o_t = (s_t^1, s_t^2)$ , while the reward weight vector  $w_r$  assigns most of its weight to state components  $s_t^4$  and  $s_t^5$ , which therefore primarily determine the reward. Privileged signals that include these components consistently exhibit stronger residual dependence and higher predictive gains. In contrast, signals containing only redundant or irrelevant components (e.g.,  $s_t^3$ ) yield weaker statistical evidence under both criteria. Notably, full-state access does not always achieve the best scores, indicating that additional features can be uninformative when misaligned with return prediction. Overall, the results show that both criteria effectively identify informative signals and that selecting return-relevant features leads to improved performance (cf. AUC reported in Table 2).

## 5.3 Discussion

Our findings suggest that learning performance depends primarily on whether the input contains reward-relevant information, rather than on the amount of information provided. In particular, adding

Table 2: Area under the mean episodic return curve (AUC) for the informed asymmetric actor-critic using different privileged signals on a randomly sampled informed POMDP with reward weights  $w_r = [0.0001, 0.0001, -0.0001, -1.0, 1.0]$ . AUCs are computed from 50 test episodes evaluated every 50 gradient steps during training. We report mean  $\pm$  standard deviation across ten independent runs. The symmetric A2C baseline achieves an AUC of  $1.06 \times 10^5 \pm 1.61 \times 10^4$ .

| Privileged signal                                 | AUC (sum of returns) ( $\uparrow$ )   |
|---------------------------------------------------|---------------------------------------|
| $i_t = \emptyset$                                 | $1.06\text{e}+05 \pm 1.61\text{e}+04$ |
| $i_t = [s_t^1, s_t^2]$                            | $1.07\text{e}+05 \pm 1.58\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^3]$                     | $1.08\text{e}+05 \pm 1.56\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^4]$                     | $1.08\text{e}+05 \pm 1.57\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^5]$                     | $1.10\text{e}+05 \pm 1.69\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$              | $1.07\text{e}+05 \pm 1.56\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$              | $1.11\text{e}+05 \pm 1.97\text{e}+04$ |
| $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$              | $1.23\text{e}+05 \pm 1.76\text{e}+04$ |
| $i_t = s_t = [s_t^1, s_t^2, s_t^3, s_t^4, s_t^5]$ | $1.19\text{e}+05 \pm 1.83\text{e}+04$ |

state features that are unrelated or only weakly related to future returns can degrade value estimation by introducing structured noise, even when using high-capacity function approximators. This effect is especially pronounced when inputs include rapidly varying components unrelated to the return signal, akin to a “noisy TV” scenario.

Notably, selecting informative training-time signals is inherently task-dependent, similar in spirit to the design of observation spaces. Our framework extends this perspective by treating training-time information as an additional degree of freedom, allowing auxiliary variables to serve as candidate signals. Our criteria automatically identify informative subsets from a large set of candidates in a fully data-driven manner, without requiring manual feature engineering. This motivates principled signal selection as a complementary factor to architecture design and optimization in asymmetric actor-critic methods.

## 6 Conclusion

We introduced the informed asymmetric actor-critic framework, which enables the critic to leverage arbitrary state-dependent privileged signals during training while preserving unbiased policy gradients. This generalizes asymmetric actor-critic methods beyond full-state critics and shifts the focus from access to state information towards the choice of informative training-time signals. We further proposed two complementary criteria to assess signal informativeness, and demonstrated empirically that appropriately selected signals can improve value estimation and learning efficiency.

Future work may extend these criteria to account for signal complexity or direct policy effects, enabling better trade-offs between informativeness and model capacity, integrate signal selection end-to-end within the learning loop, and explore asymmetric critics that condition on histories of privileged signals.

## Acknowledgments

This work was supported by the *Helmholtz Association Initiative and Networking Fund* on the HAICORE@KIT partition. Daniel Ebi acknowledges financial support by the *German Research Foundation (DFG)* as part of the Research Training Group GRK 2153: “Energy Status Data – Informatics Methods for its Collection, Analysis and Exploitation”. This work was carried out while Gaspard Lambrechts was a postdoctoral fellow of the Fund for Scientific Research (FNRS) at the University of Liège, supported by the *Wallonia-Brussels Federation* in Belgium.

## References

- [1] Khalil al Sayed, Boodi Abhinandana, Roozbeh Sadeghian Broujeny, and Karim Beddiar. Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review. *Journal of Building Engineering*, 95, 2024.
- [2] Vincent François-Lavet, David Taralla, Damien Ernst, and Raphaël Fonteneau. Deep reinforcement learning solutions for energy microgrids management. In *Thirteenth European Workshop on Reinforcement Learning*, 2016.
- [3] Daniel Ebi, Edouard Fouché, Marco Heyden, and Klemens Böhm. MicroPPO: Safe power flow management in decentralized micro-grids with proximal policy optimization. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, 2024.
- [4] Ahmad Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017.
- [5] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27), 2025.
- [6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2), 1998.
- [7] Bram Bakker. Reinforcement learning with long short-term memory. In *Advances in Neural Information Processing Systems*, 2001.
- [8] Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Recurrent policy gradients. *Logic Journal of the IGPL*, 18(5), 2010.
- [9] Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *2015 AAAI fall symposium series*, 2015.
- [10] Marvin Zhang, Zoe McCarthy, Chelsea Finn, Sergey Levine, and Pieter Abbeel. Learning deep neural network policies with continuous memory states. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016.
- [11] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for POMDPs. *CoRR*, abs/1704.07978, 2017.
- [12] Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for POMDPs. *Transactions on Machine Learning Research*, 2025.
- [13] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.
- [14] Lars Buesing, Theophane Weber, Sébastien Racanière, S. M. Ali Eslami, Danilo Jimenez Rezende, David P. Reichert, Fabio Viola, Frederic Besse, Karol Gregor, Demis Hassabis, and Daan Wierstra. Learning and querying fast generative models for reinforcement learning. *CoRR*, abs/1802.03006, 2018.
- [15] Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A. Pires, Toby Pohlen, and Rémi Munos. Neural predictive belief representations. *CoRR*, abs/1811.06407, 2018.
- [16] Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for RL. *Advances in Neural Information Processing Systems*, 32, 2019.
- [17] Dongqi Han, Kenji Doya, and Jun Tani. Variational recurrent models for solving partially observable control tasks. In *International Conference on Learning Representations*, 2020.

- [18] Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Alché, Remi Munos, and Mohammad Gheshlaghi Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020.
- [19] Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [20] Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12), 2022.
- [21] Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive RL. In *The Twelfth International Conference on Learning Representations*, 2024.
- [22] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *Robotics: Science and Systems*, 2018.
- [23] Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-driven planning via imitation learning. *The International Journal of Robotics Research*, 37(13-14), 2018.
- [24] Andrew Warrington, Jonathan W Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in pomdps. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.
- [25] Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging additional information in model-based RL. *Reinforcement Learning Journal*, 2024.
- [26] Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. The Wasserstein Believer: Learning belief updates for partially observable environments through reliable latent space models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [27] Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [28] Jonas Degraeve, Federico Felici, Jonas Buchli, Michael Neunert, Brendan D. Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin A. Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897), 2022.
- [29] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976), 2023.
- [30] Miguel Vasco, Takuma Seno, Kenta Kawamoto, Kaushik Subramanian, Peter R Wurman, and Peter Stone. A super-human vision-based reinforcement learning agent for autonomous racing in Gran Turismo. *Reinforcement Learning Journal*, 2024.
- [31] Peter Dür, Mireille Gheche, Guilherme Maeda, Nobuhiko Mukai, Naoya Takahashi, Stefan Heusser, Hamdi Sahloul, Yamen Saraiji, Pavel Adodin, Yin Bi, Sam Blakeman, Christian Conti, Dunai Hitos, Yunpu Hu, Farshad Khadivar, Raphaela Kreiser, Luz Martinez, Fabian Schilling, Ricardo Morales, and Michael Spranger. Outplaying elite table tennis players with an autonomous robot. *Nature*, 652(8111), 2026.

- [32] Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022.
- [33] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *The Eighth International Conference on Learning Representations*, 2020.
- [34] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98), 2021.
- [35] Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Forty-second International Conference on Machine Learning*, 2025.
- [36] Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [37] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.
- [38] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 2020.
- [39] Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [40] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, 1999.
- [41] Daan Wierstra, Alexander Foerster, Jan Peters, and Jürgen Schmidhuber. Solving deep memory POMDPs with recurrent policy gradients. In *Artificial Neural Networks – ICANN 2007*, 2007.
- [42] Hao Zhang, Yewei Xia, Kun Zhang, Shuigeng Zhou, and Jihong Guan. Conditional independence test based on residual similarity. *ACM Transactions on Knowledge Discovery from Data*, 17(8), 2023.
- [43] Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POP-Gym: Benchmarking partially observable reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [44] Vincent François-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *Journal of Artificial Intelligence Research*, 65(1), 2019.
- [45] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt Norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory, ALT’05*, 2005.
- [46] Florian Kalinke and Zoltán Szabó. Nyström  $m$ -Hilbert-Schmidt independence criterion. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 2023.
- [47] Hector Geffner and Blai Bonet. Solving large POMDPs using real time dynamic programming. In *Working Notes Fall AAAI Symposium on POMDPs*, 1998.
- [48] Andrea Baisero. gym-pomdps: Gym environments from POMDP files. <https://github.com/abaisero/gym-pomdps>, 2019. Accessed: 2026-01-28.

- [49] Hai Nguyen. POMDP robot domains. <https://github.com/hai-h-nguyen/pomdp-domains>, 2021. Accessed: 2026-01-28.
- [50] Shuo Jiang and Christopher Amato. Multi-agent reinforcement learning with directed exploration and selective memory reuse. In *Proceedings of the ACM Symposium on Applied Computing*, 2021.
- [51] Andrea Baisero and Sammie Katt. gym-gridverse: Gridworld domains for fully and partially observable settings. <https://github.com/abaisero/gym-gridverse>, 2021. Accessed: 2026-01-28.
- [52] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983.
- [53] Andrea Baisero and Sammie Katt. asym-porl: Asymmetric methods for partially observable reinforcement learning. <https://github.com/abaisero/asym-rlpo>, 2021. Accessed: 2026-01-28.

## A Unbiasedness Results

This section establishes unbiasedness properties of the informed history-based reward (Lemma A.1),  $Q$ - (Lemma A.2), and value functions (Lemma A.3), which are central to the informed asymmetric actor-critic framework.

**Lemma A.1** (Unbiasedness of the informed history-based reward). *In an informed POMDP, the informed history-based reward function  $R(h_t, i_t, a_t)$  satisfies*

$$\mathbb{E}_{i_t|h_t} [R(h_t, i_t, a_t)] = R(h_t, a_t),$$

for all  $h_t \in \mathcal{H}$  and  $a_t \in \mathcal{A}$ , where the expectation is taken under the belief  $p(i_t|h_t)$ .

*Proof.* Using the definition of the standard history-based reward function, i.e.,

$$R(h_t, a_t) = \mathbb{E}_{s_t|h_t} [R(s_t, a_t)] = \sum_{s_t \in \mathcal{S}} R(s_t, a_t) p(s_t|h_t),$$

and applying the law of total probability, we obtain:

$$\begin{aligned} R(h_t, a_t) &= \sum_{s_t \in \mathcal{S}} p(s_t|h_t) R(s_t, a_t) \\ &= \sum_{s_t \in \mathcal{S}} \left( \sum_{i_t \in \mathcal{I}} p(s_t|h_t, i_t) p(i_t|h_t) \right) R(s_t, a_t) \\ &= \sum_{i_t \in \mathcal{I}} \left( \sum_{s_t \in \mathcal{S}} R(s_t, a_t) p(s_t|h_t, i_t) \right) p(i_t|h_t) \\ &= \mathbb{E}_{i_t|h_t} [\mathbb{E}_{s_t|h_t, i_t} [R(s_t, a_t)]] = \mathbb{E}_{i_t|h_t} [R(h_t, i_t, a_t)]. \end{aligned}$$

This concludes the proof.  $\square$

**Lemma A.2** (Unbiasedness of the informed history  $Q$ -function). *In an informed POMDP, the informed history  $Q$ -function  $Q^\pi(h_t, i_t, a_t)$  satisfies*

$$\mathbb{E}_{i_t|h_t} [Q^\pi(h_t, i_t, a_t)] = Q^\pi(h_t, a_t),$$

for all  $h_t \in \mathcal{H}$  and  $a_t \in \mathcal{A}$ .

*Proof.* Starting with the definition of the history  $Q$ -function and using the law of total expectation, we have:

$$\begin{aligned} Q^\pi(h_t, a_t) &= \mathbb{E}_{s_{t:\infty}, a_{t+1:\infty}} \left[ \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \middle| h_t, a_t \right] \\ &= \mathbb{E}_{i_t|h_t} \left[ \mathbb{E}_{s_{t:\infty}, a_{t+1:\infty}} \left[ \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \middle| h_t, i_t, a_t \right] \right] \\ &= \mathbb{E}_{i_t|h_t} \left[ \mathbb{E}_{s_{t:\infty}, a_{t+1:\infty}} [G_t | h_t, i_t, a_t] \right] = \mathbb{E}_{i_t|h_t} [Q^\pi(h_t, i_t, a_t)]. \end{aligned}$$

This concludes the proof.  $\square$

**Lemma A.3** (Unbiasedness of the informed history value function). *In an informed POMDP, the informed history value function  $V^\pi(h_t, i_t)$  satisfies*

$$\mathbb{E}_{i_t|h_t} [V^\pi(h_t, i_t)] = V^\pi(h_t)$$

for all  $h_t \in \mathcal{H}$ .

*Proof.* Given the definition of the history value function, i.e.,

$$V^\pi(h_t) = \mathbb{E}_{s_{t:\infty}, a_{t:\infty}} \left[ \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \middle| h_t \right],$$

and using the law of total expectation, we have:

$$\begin{aligned}
V^\pi(h_t) &= \mathbb{E}_{s_{t:\infty}, a_{t:\infty}}^\pi \left[ \sum_k^\infty \gamma^k R(s_{k+t}, a_{k+t}) \middle| h_t \right] \\
&= \mathbb{E}_{i_t | h_t} \left[ \mathbb{E}_{s_{t:\infty}, a_{t:\infty}}^\pi \left[ \sum_k^\infty \gamma^k R(s_{k+t}, a_{k+t}) \middle| h_t, i_t \right] \right] \\
&= \mathbb{E}_{i_t | h_t} \left[ \mathbb{E}_{s_{t:\infty}, a_{t:\infty}}^\pi [G_t | h_t, i_t] \right] = \mathbb{E}_{i_t | h_t} [V^\pi(h_t, i_t)].
\end{aligned}$$

This concludes the proof.  $\square$

## B Auxiliary Results

This section collects our auxiliary results.

**Corollary B.1** (Relation of  $V^\pi(h, i)$  to the history-state value function of Baisero and Amato [32]). *The informed history value function  $V^\pi(h, i)$  reduces to the history-state value function for  $i = s$ , where  $s \in \mathcal{S}$  denotes the true environment state. In particular,*

$$V^\pi(h, s) = \sum_{a \in \mathcal{A}} \pi(a|h) Q^\pi(h, s, a),$$

where the history-state action-value function is defined as

$$Q^\pi(h, s, a) = R(s, a) + \gamma E_{s', o' | s, a} [V^\pi(h', s')],$$

with  $s' \sim p(s'|s, a)$ ,  $o' \sim \tilde{O}(o'|s')$ ,  $i' = s'$ , and  $h' = ha o'$ .

By Lemma A.3, this formulation provides an alternative unbiased estimator of the history value function, i.e.,

$$V^\pi(h) = E_{s|h} [V^\pi(h, s)],$$

as previously established by Baisero and Amato [32].

**Corollary B.2** (Relation of  $\nabla_\theta^{\text{IAAC}} J(\pi_\theta)$  to the asymmetric policy gradient of Baisero and Amato [32]). *The informed asymmetric policy gradient  $\nabla_\theta^{\text{IAAC}} J(\pi_\theta)$  reduces to the asymmetric policy gradient introduced by Baisero and Amato [32] for  $i = s$ , where  $s \in \mathcal{S}$  denotes the true environment state. In particular,*

$$\nabla_\theta^{\text{AAC}} J(\pi_\theta) = E \left[ \sum_{t=0}^\infty \gamma^t Q^\pi(h_t, s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | h_t) \right].$$

Following Lemma A.1 and A.3, this formulation recovers an alternative asymmetric policy gradient estimator that is equivalent to the standard policy gradient, i.e.,

$$\nabla_\theta^{\text{AAC}} J(\pi_\theta) = \nabla_\theta J(\pi_\theta),$$

as established by Baisero and Amato [32].

## C Analysis of the $\alpha$ -Residual-Informativeness Test

In this section, we describe the construction of the proposed  $\alpha$ -residual-informativeness test and show how the regression error affects the population dependency measure.

### C.1 Test Construction

The goal of the test is to determine whether the privileged signal  $i_t$  contains information about future returns  $G_t$  beyond what is already captured by the history-action pair  $(h_t, a_t)$ . Ideally, this corresponds to testing the conditional independence (CI) hypothesis

$$\mathbb{H}_0^{\text{CI}} : G_t \perp\!\!\!\perp i_t \mid h_t, a_t.$$

Let  $G_t, i_t, h_t$ , and  $a_t$  be random variables with joint distribution  $p(G_t, i_t, h_t, a_t)$  and finite second moments. Define the conditional-mean residuals

$$e_{G_t} := G_t - \mathbb{E}[G_t | h_t, a_t], \quad e_{i_t} := i_t - \mathbb{E}[i_t | h_t, a_t].$$

**Proposition C.1.** If  $\mathbb{H}_0^{\text{CI}}$  holds, then

$$e_{G_t} \perp\!\!\!\perp e_{i_t} \mid h_t, a_t.$$

*Proof.* Under  $\mathbb{H}_0^{\text{CI}}$ , the conditional joint distribution factorizes as

$$p(G_t, i_t | h_t, a_t) = p(G_t | h_t, a_t) p(i_t | h_t, a_t).$$

Subtracting measurable functions of  $(h_t, a_t)$  from each variable does not affect conditional independence given  $(h_t, a_t)$ , so that

$$p(e_{G_t}, e_{i_t} | h_t, a_t) = p(e_{G_t} | h_t, a_t) p(e_{i_t} | h_t, a_t).$$

This concludes the proof.  $\square$

The converse does not hold in general [42]. Thus, conditional residual independence is a necessary but not sufficient consequence of conditional independence.

**Unconditional residual independence test.** However, directly testing conditional residual independence is infeasible because it would require samples from

$$p(e_{G_t}, e_{i_t}, h_t, a_t) = p(h_t, a_t) p(e_{G_t} | h_t, a_t) p(e_{i_t} | h_t, a_t),$$

where, in particular, the conditional residual distribution  $p(e_{G_t} | h_t, a_t)$  is unknown.

Instead, our proposed test evaluates the weaker unconditional residual independence hypothesis

$$\mathbb{H}_0^{\text{res}} : e_{G_t} \perp\!\!\!\perp e_{i_t}.$$

This relaxation is motivated by the objective of detecting whether the privileged signal explains variation in  $G_t$  beyond what is captured by  $(h_t, a_t)$ . Thus, residual dependence provides evidence that  $i_t$  contains additional predictive information beyond the observable history-action pair.

**First-moment-matching surrogate return.** To perform the residual independence test, we require samples where the residual component of  $G_t$  is independent of  $i_t$ . Since sampling directly from  $p(G_t | h_t, a_t)$  is generally infeasible, we construct a surrogate return with the same conditional mean structure but an independent residual component.

**Assumption C.2** (Existence of a surrogate). There exists a random variable  $\tilde{G}_t$  defined on the same probability space such that

- (i)  $\tilde{G}_t \stackrel{d}{=} G_t$  (same marginal distribution),
- (ii)  $\tilde{G}_t \perp\!\!\!\perp i_t, h_t, a_t$  (independence from the privileged signal, history, and action),
- (iii)  $\mathbb{E}[\tilde{G}_t^2] < \infty$  (finite second moment).

Under Assumption C.2, the surrogate null  $\tilde{G}_t^{\text{null}}$  is then obtained as

$$\tilde{G}_t^{\text{null}} := \tilde{G}_t - \mathbb{E}[\tilde{G}_t] + \mathbb{E}[G_t | h_t, a_t]. \quad (9)$$

This construction preserves the conditional mean of  $G_t$ , i.e.,  $\mathbb{E}[G_t | h_t, a_t]$ , while replacing the unexplained component with an independent sample from the marginal return distribution.

The corresponding surrogate null residual is

$$e_{\tilde{G}_t}^{\text{null}} := \tilde{G}_t^{\text{null}} - \mathbb{E}[G_t | h_t, a_t] = \tilde{G}_t - \mathbb{E}[\tilde{G}_t].$$

**Proposition C.3** (Properties of the surrogate null residual). *Under Assumption C.2,*

$$\mathbb{E}[e_{\tilde{G}_t}^{\text{null}} | h_t, a_t] = 0, \quad \text{and} \quad e_{\tilde{G}_t}^{\text{null}} \perp\!\!\!\perp e_{i_t},$$

*almost surely.*

*Proof.* First, we prove zero conditional mean. By definition,

$$e_{\tilde{G}_t}^{\text{null}} = \tilde{G}_t - \mathbb{E}[\tilde{G}_t],$$

where  $\mathbb{E}[\tilde{G}_t]$  is a constant. Taking the conditional expectation given  $(h_t, a_t)$ , yields

$$\mathbb{E}[e_{\tilde{G}_t}^{\text{null}} | h_t, a_t] = \mathbb{E}[\tilde{G}_t | h_t, a_t] - \mathbb{E}[\tilde{G}_t].$$

Since  $\tilde{G}_t$  is independent of  $(h_t, a_t)$  by Assumption C.2 (ii),

$$\mathbb{E}[\tilde{G}_t | h_t, a_t] = \mathbb{E}[\tilde{G}_t].$$

Therefore,

$$\mathbb{E}[e_{\tilde{G}_t}^{\text{null}} | h_t, a_t] = 0.$$

Second, we show independence from  $e_{i_t}$ . By Assumption C.2 (ii),

$$\tilde{G}_t \perp\!\!\!\perp i_t, h_t, a_t.$$

Moreover, the surrogate residual  $e_{\tilde{G}_t}^{\text{null}} = \tilde{G}_t - \mathbb{E}[\tilde{G}_t]$  is a measurable function  $f$  of  $\tilde{G}_t$ , i.e.,

$$e_{\tilde{G}_t}^{\text{null}} = f(\tilde{G}_t),$$

while  $e_{i_t}$  is a measurable function  $g$  of  $(i_t, h_t, a_t)$ , i.e.,

$$e_{i_t} = g(i_t, h_t, a_t) = i_t - \mathbb{E}[i_t | h_t, a_t].$$

Since measurable transformations preserve independence, it follows that

$$e_{\tilde{G}_t}^{\text{null}} = f(\tilde{G}_t) \perp\!\!\!\perp g(i_t, h_t, a_t) = e_{i_t}.$$

This concludes the proof.  $\square$

Therefore, the surrogate construction generates residual samples satisfying the null hypothesis  $\mathbb{H}_0^{\text{res}}$ . The surrogate matches the first conditional moment of  $p(G_t | h_t, a_t)$ , while its higher-order moments follow those of the marginal distribution of  $G_t$ . Consequently, the proposed permutation test operates under a first-moment-matching surrogate null rather than the exact conditional independence null.

The proposed criterion should therefore be interpreted as a practical test for residual informativeness rather than a test of conditional independence itself. While conditional independence implies conditional residual independence, the converse does not hold in general. Furthermore, our test evaluates unconditional residual dependence, which provides a tractable relaxation of conditional residual independence. Thus, failure to reject  $\mathbb{H}_0^{\text{res}}$  does not imply conditional independence; however, rejection indicates that the privileged signal contains information about the unexplained variation of  $G_t$  beyond what is captured by  $(h_t, a_t)$ .

Hence, this criterion is aligned with value-function learning, where the central objective is to identify whether additional signals improve the prediction of future returns beyond the information already available in the observable history-action pair.

## C.2 Effect of Regression Error on Residual Dependence Estimation

This subsection discusses how errors in estimating the conditional means affect the population dependency measure underlying the residual informativeness test.

**Regression error and residual perturbation.** Let  $f_G(h_t, a_t) = \mathbb{E}[G_t | h_t, a_t]$  and  $f_i(h_t, a_t) = \mathbb{E}[i_t | h_t, a_t]$  denote the true conditional expectations, and let  $\hat{f}_G(h_t, a_t), \hat{f}_i(h_t, a_t)$  be learned estimators. Define the regression errors

$$\ell_{G_t} = f_G(h_t, a_t) - \hat{f}_G(h_t, a_t), \tag{10}$$

$$\ell_{i_t} = f_i(h_t, a_t) - \hat{f}_i(h_t, a_t). \tag{11}$$

We assume that the regression errors have finite second moments, i.e.,  $\mathbb{E}[\ell_{G_t}^2] < \infty$ , and  $\mathbb{E}[\ell_{i_t}^2] < \infty$ .

The population residuals are

$$e_{G_t} = G_t - f_G(h_t, a_t), \quad e_{i_t} = i_t - f_i(h_t, a_t),$$

while the estimated residuals used in practice satisfy

$$\hat{e}_{G_t} = G_t - \hat{f}_G(h_t, a_t) = e_{G_t} + \ell_{G_t}, \quad \hat{e}_{i_t} = i_t - \hat{f}_i(h_t, a_t) = e_{i_t} + \ell_{i_t}.$$

Throughout this section, we assume sample splitting, following the procedure described in Algorithm 1, such that the regression estimation procedure is independent of the samples used for the dependence test.

Let  $\rho(X, Y)$  denote a population dependency measure between square-integrable random variables  $X$  and  $Y$ . We assume the following stability property.

**Assumption C.4** (Lipschitz continuity in  $L^2$ ). There exists a constant  $L_\rho > 0$  such that for all random variables  $X, X', Y, Y'$  with finite second moments,

$$|\rho(X, Y) - \rho(X', Y')| \leq L_\rho (\|X - X'\|_2 + \|Y - Y'\|_2),$$

where  $\|Z\|_2 := \sqrt{\mathbb{E}[Z^2]}$ .

This assumption holds for several commonly used dependency measures under suitable regularity conditions.

Applying Assumption C.4 with

$$(X, Y) = (e_{G_t}, e_{i_t}), \quad (X', Y') = (\hat{e}_{G_t}, \hat{e}_{i_t}),$$

yields

$$|\rho(e_{G_t}, e_{i_t}) - \rho(\hat{e}_{G_t}, \hat{e}_{i_t})| \leq L_\rho (\|\ell_{G_t}\|_2 + \|\ell_{i_t}\|_2). \quad (12)$$

Thus, estimation errors in the conditional means perturb the population residual dependency measure by at most an additive margin proportional to the  $L^2$  regression errors of the conditional expectations. Consequently, accurate estimation of the conditional means ensures that the estimated residual dependence measure remains close to the population residual dependence criterion.

## D Environments

In this section, we detail the partially observable environments used in our experiments.

### D.1 Navigation Environments

This subsection presents the partially observable benchmark navigation environments used to assess the learning performance of the proposed informed asymmetric actor-critic method.

**Heaven-Hell-3.** The Heaven-Hell task [47, 48] is a partially observable navigation problem in a grid-world environment with a corridor-like structure that forks into three distinct branches. Two of these branches correspond to terminal exits: one leading to a positive outcome (heaven) and the other to a negative outcome (hell). The third branch leads to a non-terminal location where the agent can interact with an oracle (priest) that provides information needed to disambiguate the exits. The agent is initially unaware of which terminal exit corresponds to heaven.

The underlying state includes both the agent’s position and the true location of the heaven exit. As observation, however, the agent either perceives its own location or, when visiting the priest, receives an observation that reveals the location of heaven. We construct the privileged information by combining the agent’s location with its Manhattan distance to the heaven terminal.

At each time step, the agent selects an action from the discrete set {NORTH, SOUTH, EAST, WEST}. The environment is deterministic, and movement is constrained by the grid-world layout. To solve the task optimally, the agent must first visit the priest to acquire the necessary information about the correct exit, then return to the fork and proceed to the identified heaven location.

The agent receives sparse feedback in the form of a terminal reward: a reward of 1.0 for exiting to heaven, and a reward of -1.0 for exiting to hell.

**Shopping-5.** The Shopping-5 environment [48] models another grid-world navigation task in which an agent must buy an item from a store. The environment is modeled as a two-dimensional grid-world of size  $5 \times 5$ , with the item placed randomly at one of the grid cells. The agent begins at an arbitrary location and must first locate and then buy the item. While the agent’s position is fully observable, the item’s position is hidden and must be explicitly queried.

Hence, the full state encodes both the agent’s position and the item’s location, represented compactly as integers. Observations are similarly encoded, but are partial: at each time step, the agent observes either its own position or, upon executing a query, the position of the item. Similar to the Heaven-Hell task, we construct the privileged information by computing the current Manhattan distance between the agent and the item.

At each time step, the agent selects an action from the discrete set {UP, DOWN, LEFT, RIGHT, QUERY, BUY}. The four movement actions update the agent’s position deterministically within the bounds of the grid. Executing the QUERY action returns the location of the item, but is subject to a cost. The BUY action attempts to purchase the item at the agent’s current position; if executed in the correct cell, it completes the task successfully.

The environment provides a dense reward signal to encourage efficient behavior: a reward of -1.0 for moving, a reward of -2.0 for querying the item’s location, a reward of -5.0 for a BUY action in the wrong cell, and a reward of +10.0 for a BUY action in the correct cell.

Optimal behavior requires the agent to query the item’s location once, retain that information internally, and efficiently navigate to the target cell before executing a successful BUY action.

**Car-Flag.** The Car-Flag environment [49] models a continuous control task where an agent controls a car moving along a one-dimensional track via discrete force-control actions. At the two ends of the track are terminal flags: one corresponding to a positive outcome (good flag) and the other to a negative outcome (bad flag). Reaching either flag terminates the episode. Additionally, an intermediate information flag is placed along the track; when reached, it reveals the position of the good flag. While the task is conceptually similar to Heaven-Hell, key differences are the force-control and the position of the information flag.

Both the state and observation spaces are represented as three-dimensional real-valued vectors. The state includes the agent’s position, velocity, and the position of the good flag. The observation mirrors the state structure, but the third component (i.e., the good flag’s position) is masked, i.e., set to zero, when the agent is outside the observation range of the information flag; and the agent’s velocity is always hidden. In the informed setting, we provide the agent its velocity as a privileged partial signal.

At each time step, the agent selects an action from a discrete set of six force-control inputs, i.e., {LEFT\_HIGH, LEFT\_MEDIUM, LEFT\_LOW, RIGHT\_LOW, RIGHT\_MEDIUM, RIGHT\_HIGH}, which apply varying levels of acceleration to the left or right, or the action NONE, which applies no force.

The environment provides a sparse, terminal reward signal: a reward of 1.0 for reaching the good flag, and a reward of -1.0 for reaching the bad flag.

Optimal behavior requires the agent to first locate the information flag to identify the correct goal, then apply appropriate force controls to reach the good flag while avoiding the bad one.

**Cleaner.** Originally designed as a two-agent cooperative task, the Cleaner environment [50] is adapted in this work to a single-agent control problem via fully centralized training and execution. In this formulation, the joint actions and observations are constructed via the Cartesian product of the corresponding spaces of the two individual agents. The environment is a maze-like  $13 \times 13$  grid-world in which two robots must collectively cover and clean the entire area. The task is considered complete once every non-wall cell has been visited by at least one of the agents.

The full environment state is represented as a binary tensor of shape  $13 \times 13 \times 5$ , where each channel encodes the presence of: (i) a wall, (ii) a dirty cell, (iii) a cleaned cell, (iv) the first agent, and (v) the second agent. Each agent’s local observation is a  $3 \times 3 \times 3$  binary tensor that captures the immediate neighborhood centered around the agent, including information about walls, dirty cells, and clean cells. As privileged input, the critic receives a  $13 \times 13 \times 5$  tensor encoding the agent’s own position within the grid world, while masking out the position of the other agent by setting its corresponding cells to zero.

Each agent independently selects from four movement actions: UP, DOWN, LEFT, and RIGHT. In the centralized setting, where both agents are controlled jointly, the action space is the Cartesian product of the individual action sets, yielding a total of 16 composite actions.

At each time step, the agent receives a reward proportional to the number of new cells cleaned during that step. The possible reward values are: a reward of 0.0 if no new cells are cleaned, a reward of 1.0 if one agent cleaned a new cell, and a reward of 2.0 if both agents cleaned a new cell.

**Memory-Four-Rooms.** The so-called Gridverse suite [51] defines a collection of partially observable environments in which agents interact within structured grid-worlds. In this work, we consider the  $7 \times 7$ -Memory-Four-Room and  $9 \times 9$ -Memory-Four-Room environments. While actions are encoded as categorical indices, both states and observations are structured representations comprising multiple semantically meaningful components. Importantly, these components differ between state and observation, and some are only available in the state representation. The key components are:

- **Grid component:** A tensor of shape  $3 \times 7 \times 7$  for  $7 \times 7$ -Memory-Four-Room or  $3 \times 9 \times 9$  for  $9 \times 9$ -Memory-Four-Room, where each channel encodes a semantic property of the environment (e.g., cell type, cell color, or status). The observation includes a rotated, agent-centric  $3 \times 2 \times 3$ -view of this grid rendered from the first-person perspective of the agent. Cells obstructed by walls are occluded in the observation.
- **Agent-ID-Grid component:** A binary matrix of size  $7 \times 7$  or  $9 \times 9$ , respectively, indicating the agent’s absolute position. This component is included only in the state.
- **Agent component:** A three-dimensional categorical array encoding the agent’s position and orientation. In the state, this is expressed in absolute coordinates, while in the observation, it is provided relative to the agent’s perspective and is thus constant, and not necessary for control.

The environment contains a good exit, a bad exit, and a beacon, each placed randomly at the start of each episode. The beacon shares its color with the good exit, and successful task completion requires the agent to first locate the beacon, memorize its color, and then navigate to the exit of matching color while avoiding the bad exit.

As privileged input, the critic is provided with an agent-centered  $3 \times 3 \times 5$  tensor, offering an expanded view of the agent’s local surroundings.

At each time step, the agent selects from the following discrete action set {MOVE\_FORWARD, MOVE\_BACKWARD, MOVE\_LEFT, MOVE\_RIGHT, TURN\_LEFT, TURN\_RIGHT, PICK\_N\_DROP, ACTUATE}. The MOVE\_ actions are interpreted relative to the agent’s orientation, while TURN\_ actions modify the orientation itself. Although the action set also includes PICK\_N\_DROP and ACTUATE for generality, these actions have no effect in the Memory-Four-Rooms tasks, as there are no doors or pickable objects.

The reward signal is composed of the following terms: a living reward of -0.05 per time step, a reward of +5.0 for reaching the good exit, and a reward of -5.0 for reaching the bad exit.

## D.2 POPGym Environments

This subsection presents the partially observable environments from the POPGym suite [43] used to benchmark our informed asymmetric actor-critic method.

**Higher Lower.** The Higher Lower task is a partially observable card game, in which the agent must predict whether the next drawn card has a higher or lower rank than the current card. At each step, the agent observes the current card and selects one of two discrete actions, HIGHER or LOWER, corresponding to the prediction that the next card will be of higher or lower rank, respectively.

After each action, the next card is revealed and becomes the reference card for the next decision. As the agent progresses through the deck, it must remember the previous card’s rank to predict whether the next card will be higher or lower.

The reward is scaled by the deck size: correct predictions yield a positive reward, incorrect predictions a negative reward, and ties (identical ranks) result in zero reward.

We consider four variants of privileged information for this environment. The *full-state* signal provides the complete internal card-count representation of the environment state. The *expert* corresponds to an expert policy revealing whether the next card is higher or lower than the current card, effectively encoding the optimal prediction. The *previous-card* signal provides access to the previously observed card rank, while the *both-cards* signal additionally includes the current observed card together with the previous one.

**Repeat First.** Repeat First is a partially observable card-game-based memory task. At the beginning of each episode, the agent observes an initial card suit that must be memorized throughout the episode. Subsequently, the agent receives a sequence of additional card suits and, at each step, must output the suit of the initial card.

The environment therefore requires the agent to retain information over long time horizons while processing distracting intermediate observations. At each time step, the agent selects one of four discrete actions corresponding to the card suits SPADES, HEARTS, DIAMONDS, and CLUBS.

The reward is scaled by the deck size: correctly predicting the initial card yields a positive reward, whereas incorrect predictions yield a negative reward.

We consider four variants of privileged information for this environment. The *hand* signal reveals the suit of the current card in hand. The *dealt* signal provides statistics over already dealt cards (i.e., frequency of observed suits). The *full-state* signal combines both sources of information. Finally, the *first-card* signal provides access to the suit of the first observed card in the sequence.

**Repeat Previous.** Repeat Previous is a partially observable card-game-based memory task. The agent observes a sequence of card suits and, at each time step, must predict the suit observed  $k$  steps earlier in the sequence. In our experiments, we use  $k = 4$ , requiring the agent to maintain a memory of past observations while processing continuously arriving inputs.

At each step, the agent selects one of four discrete actions corresponding to the four standard card suits: SPADES, HEARTS, DIAMONDS, and CLUBS. Rewards are scaled by the deck size: correctly recalling the target suit yields a positive reward, whereas incorrect predictions yield a negative reward.

We consider the same four variants of privileged information as in the Repeat First task, with the only difference that the *first-card* signal is replaced by the *previous-card* signal.

**Count Recall.** Count Recall is a partially observable card-game-based memory task, designed to test order-agnostic memory and counting ability. At each time step, the agent observes a pair of cards corresponding to a currently drawn card and a query card. The task requires the agent to predict how many times the queried card has appeared in the history of previously observed cards.

At each step, the agent selects an integer-valued action corresponding to the predicted count. Rewards are scaled by the deck size: correct predictions of the queried count yield a positive reward, while incorrect predictions yield a negative reward.

We consider five variants of privileged information for this environment. The *freq-cards-seen* signal provides the full histogram of observed card frequencies. The *freq-cards-queried* signal provides the histogram of queried card frequencies. The *last-obs* signal provides the most recent observation pair, and the *expert* signal reveals the true count of the queried card at the current step. Finally, the *full-state* signal corresponds to the full environment state, combining the frequencies of observed and queried cards with the most recent observation pair.

**Concentration.** Concentration is a partially observable card-matching task, inspired by the classic memory game of the same name. A full deck of cards is initially placed face-down, and the agent interacts with the environment by selecting cards to flip.

At each time step, the agent chooses a card to reveal. If two consecutively selected cards match (according to the chosen matching criterion, e.g., in rank or color), the pair is removed from the game and the agent receives a positive reward. Otherwise, the selection is penalized, and the cards are flipped back face-down.

We consider six variants of privileged information for this environment. The *all-values* signal provides access to the underlying identities of all cards in the deck. The *is-face-up* signal indicates which

cards are currently revealed. The *first-card* and *second-card* signals provide the identities of the most recently selected cards. The *flipped-cards* signal provides both of the last two selected cards jointly. Finally, the *full-state* signal combines card identities, visibility information, and the most recent selection history.

**Position Cart Pole.** The position-only cartpole task is a partially observable variant of the classic control benchmark [52]. The environment implements the standard cartpole mechanics, but the agent does not directly observe the cart’s horizontal velocity or the pole’s angular velocity. Instead, it receives only the cart’s position along the horizontal axis and the pole angle, requiring it to infer the missing velocity information over time.

At each time step, the agent selects a discrete action corresponding to the standard cartpole control inputs: PUSH-LEFT or PUSH-RIGHT. Successful balancing yields a small positive reward scaled by the episode length, while failure results in a negative reward.

We consider four variants of privileged information for this environment. The *x-velocity* signal provides access to the cart’s horizontal velocity. The *angle-velocity* signal provides access to the angular velocity of the pole. The *both-velocities* signal provides both velocity components jointly. Finally, the *full-state* signal reveals the full environment state.

### D.3 Synthetic Informed POMDPs

We generate a distribution of synthetic informed POMDP instances with a finite state space of size  $|\mathcal{S}|$ , a discrete action space of size  $|\mathcal{A}|$ , and continuous observation and information spaces. Following the methodology of François-Lavet et al. [44], transition probabilities are randomly assigned by setting each  $(s_t, a_t, s_{t+1})$ -entry to zero with probability 0.75, and sampling uniformly from  $[0, 1]$  otherwise. To ensure valid transitions, we assign a non-zero probability to a randomly chosen next state whenever all transitions from a given state-action pair are initially zero. We then normalize the probabilities so that they sum to one. Each state is associated with a Gaussian feature vector  $s_t \in \mathbb{R}^{d_s}$  with  $d_s \in \mathbb{Z}_{>0}$ :  $s_t \sim \mathcal{N}(\mathbf{0}, \sigma_s^2 \mathbf{I}_{d_s})$ , where  $\mathbf{0}$  is the  $d_s$ -dimensional zero vector,  $\mathbf{I}_{d_s}$  the  $d_s \times d_s$  identity matrix, and  $\sigma_s^2$  the variance. Rewards are linear functions of the state features, with reward weights  $w_r \in \mathbb{R}^{d_s}$  sampled uniformly from  $[-1, 1]$  at initialization.

We first construct the privileged information  $i_t \in \mathbb{R}^{d_i}$  by selecting a subset of  $1 \leq d_i \leq d_s$  state features using a binary masking vector  $x_i \in \{0, 1\}^{d_s}$  and applying a selection matrix  $W_i \in \{0, 1\}^{d_i \times d_s}$ :

$$i_t = W_i(x_i \odot s_t)$$

where  $\odot$  denotes element-wise multiplication. Observations  $o_t \in \mathbb{R}^{d_o}$  are generated by masking a subset of  $1 \leq d_o \leq d_i$  features from  $i_t$  using a binary masking vector  $x_o \in \{0, 1\}^{d_i}$ , applying a selection matrix  $W_o \in \{0, 1\}^{d_o \times d_i}$  to the masked privileged signal, and adding Gaussian noise:

$$o_t = W_o(x_o \odot i_t) + \beta_o \epsilon_o, \quad \epsilon_o \sim \mathcal{N}(\mathbf{0}, \sigma_o^2 \mathbf{I}_{d_o}),$$

where  $\beta_o \geq 0$  modulates the observation noise and  $\sigma_o^2$  is the variance.

## E Model Architectures and Hyperparameters

In this section, we describe the actor and critic architectures as well as the hyperparameters used in our experiments.

### E.1 Navigation Tasks

We use the implementation of environments [53, 48, 51, 49] and actor-critic methods provided by Baisero and Amato [32], extending them to the informed setting.

In each task, a 128-dimensional single-layer gated recurrent unit (GRU) encodes the concatenated action and observation features into a history representation. While the actor and critic networks share this architectural component, their parameters are maintained separately. The subsequent actor and critic network components differ across environments as follows:

- For the Heaven-Hell-3 and Shopping-5 tasks, we employ a 64-dimensional embedding model to represent states, actions, and observations. Both the actor and critic networks consist of two-layer feedforward neural networks with 512 and 256 units, respectively, using ReLU activations in the hidden layers and a linear output layer.
- For the Car-Flag and Cleaner environments, actions are represented as one-hot encodings of their respective categorical indices. As the state and observation representations provided by these environments are already flattened and structurally simple, no additional embedding is applied. The actor’s and critic’s subsequent networks adopt the same architecture used for the Heaven-Hell-3 and Shopping-5 tasks.
- For the Memory-Four-Room tasks, the  $3 \times 2 \times 3$  observation tensors are initially processed by an embedding layer that maps each categorical value to an 8-dimensional vector. The resulting embedded tensor is then flattened into a 144-dimensional feature vector, which serves as the observation input to both the actor and critic networks. Actions, provided as categorical indices, are represented using one-dimensional embedding layers. For the states, the grid component is first embedded and then concatenated with the agent-ID grid. A three-layer convolutional network subsequently processes this combined input. The output of the convolutional network is concatenated with the agent components. The actor and critic networks each consist of a hidden layer with 512 units using ReLU activation, followed by a linear output layer.

We encode the privileged information analogously to the observations. The embedded privileged information is then concatenated with the latent history representation before being passed to the task-specific feedforward neural network.

Table 3: Hyperparameters for the benchmark navigation environments.

| Environment                    | $\eta_\pi$ | $\eta_{\hat{V}}$ | $\lambda_0$ |
|--------------------------------|------------|------------------|-------------|
| Heaven-Hell-3                  | 0.001      | 0.001            | 0.1         |
| Shopping-5                     | 0.001      | 0.0003           | 3.0         |
| Car-Flag                       | 0.001      | 0.001            | 0.03        |
| Cleaner                        | 0.001      | 0.001            | 1.0         |
| $7 \times 7$ -Memory-Four-Room | 0.0003     | 0.001            | 0.1         |
| $9 \times 9$ -Memory-Four-Room | 0.001      | 0.0003           | 0.3         |

For each environment and method, we use the hyperparameter values recommended by Baisero and Amato [32] to ensure comparability with prior work. Table 3 summarizes the actor learning rate  $\eta_\pi$ , critic learning rate  $\eta_{\hat{V}}$ , and the initial negative-entropy weight  $\lambda_0$  selected for each environment. Additionally, the following model hyperparameters are applied across all environments: discount factor is set to  $\gamma = 0.99$ , episodes are automatically terminated if they exceed 100 time steps; two episodes are sampled per gradient update; a frozen target network is used to stabilize critic training, with target parameters updated every 10,000 time steps; and the negative-entropy weight  $\lambda$  decays linearly over 2 million time steps to a final value equal to one-tenth of  $\lambda_0$ .

## E.2 POPGym Environments

We use the environment implementations of Morad et al. [43], and adopt their recommended actor and critic architectures, extending them to the informed setting.

For each environment and method, a 256-dimensional single-layer GRU encodes the interaction history into a fixed-length representation. Before being processed by the GRU, observations are projected into a 128-dimensional zero-mean, unit-variance representation using a linear layer followed by layer normalization and a LeakyReLU activation. The GRU hidden state is projected onto a 128-dimensional feature vector, which is fed into separate actor and critic heads. Both heads are implemented as two-layer feedforward networks with 128 hidden units and LeakyReLU activations. For the A2C variants with an informed or full-state critic, we encode the privileged signal analogously to the observations. The resulting embedding is concatenated with the history representation before being passed to the critic head.

We use the hyperparameter values recommended by Morad et al. [43] for the actor-critic architecture to ensure comparability with prior work. We set the learning rate to  $\eta = 5 \times 10^{-4}$ , use backpropagation-through-time truncation length of 1,024, a zero entropy regularization weight  $\lambda = 0$ , and a discount factor of  $\gamma = 0.99$ . For the encoding of the privileged signal, we use an embedding size of 64 for the card-game tasks and an embedding size of 128 for the position-only cartpole environment.

### E.3 Synthetic Informed POMDPs

For the synthetic informed POMDP environments, the actor is implemented as a single-layer GRU with hidden size 64, followed by a linear readout layer. Both the symmetric and informed history critics use a separate single-layer GRU with hidden size 64 to produce fixed-length representations of the interaction history, followed by a linear readout layer that outputs the value estimate. For the informed history critic, the GRU hidden state is concatenated with the privileged signal before being passed to the linear layer. The learning rate for both the actor and the critics is set to  $\eta = 1 \times 10^{-4}$ , and we use a fixed discount factor of  $\gamma = 0.99$  across all environments.

## F Wall-Clock Runtimes

All experiments were conducted on a cluster node equipped with 96 cores running at 3.0 GHz and 93.75 GB of RAM allocated per task. Tables 4 and 5 report the wall-clock training times for different A2C variants and privileged signals across the navigation and PopGym environments, respectively. Table 6 summarizes the wall-clock runtimes for the experiments in Section 5.2.

Table 4: Wall-clock training time across the six benchmark navigation tasks for different A2C variants. We report absolute training time (in seconds) and relative mean runtime w.r.t. the mean runtime of the symmetric A2C (100%). Means and standard deviations are computed across 20 independent runs.

| Environment           | Algorithm         | Wall-Clock Time (s)                   | Relative Time (%) |
|-----------------------|-------------------|---------------------------------------|-------------------|
| Memory-Four-Rooms-7x7 | a2c               | $2.18\text{e}+04 \pm 1.52\text{e}+03$ | 100.00            |
|                       | informed-asym-a2c | $2.32\text{e}+04 \pm 1.66\text{e}+03$ | 106.04            |
|                       | asym-a2c-s        | $2.88\text{e}+04 \pm 2.29\text{e}+03$ | 132.03            |
|                       | asym-a2c-hs       | $3.45\text{e}+04 \pm 2.28\text{e}+03$ | 158.02            |
| Memory-Four-Rooms-9x9 | a2c               | $2.13\text{e}+04 \pm 1.70\text{e}+03$ | 100.00            |
|                       | informed-asym-a2c | $2.38\text{e}+04 \pm 1.19\text{e}+03$ | 111.62            |
|                       | asym-a2c-s        | $2.76\text{e}+04 \pm 1.54\text{e}+03$ | 129.48            |
|                       | asym-a2c-hs       | $3.53\text{e}+04 \pm 1.10\text{e}+03$ | 165.93            |
| Heaven-Hell-3         | asym-a2c-s        | $5.81\text{e}+03 \pm 1.23\text{e}+03$ | 72.75             |
|                       | a2c               | $7.99\text{e}+03 \pm 1.47\text{e}+03$ | 100.00            |
|                       | informed-asym-a2c | $8.97\text{e}+03 \pm 1.66\text{e}+03$ | 112.27            |
|                       | asym-a2c-hs       | $9.15\text{e}+03 \pm 1.34\text{e}+03$ | 114.45            |
| Shopping-5            | asym-a2c-s        | $2.16\text{e}+03 \pm 2.94\text{e}+02$ | 56.71             |
|                       | a2c               | $3.81\text{e}+03 \pm 7.73\text{e}+02$ | 100.00            |
|                       | asym-a2c-hs       | $4.28\text{e}+03 \pm 5.81\text{e}+02$ | 112.27            |
|                       | informed-asym-a2c | $4.35\text{e}+03 \pm 3.08\text{e}+02$ | 113.96            |
| Car-Flag              | asym-a2c-s        | $4.38\text{e}+03 \pm 1.08\text{e}+03$ | 72.26             |
|                       | informed-asym-a2c | $5.94\text{e}+03 \pm 1.44\text{e}+03$ | 97.93             |
|                       | a2c               | $6.07\text{e}+03 \pm 1.67\text{e}+03$ | 100.00            |
|                       | asym-a2c-hs       | $6.20\text{e}+03 \pm 1.74\text{e}+03$ | 102.19            |
| Cleaner               | asym-a2c-s        | $5.39\text{e}+03 \pm 3.42\text{e}+02$ | 90.66             |
|                       | a2c               | $5.95\text{e}+03 \pm 4.46\text{e}+02$ | 100.00            |
|                       | asym-a2c-hs       | $6.20\text{e}+03 \pm 2.46\text{e}+02$ | 104.26            |
|                       | informed-asym-a2c | $6.23\text{e}+03 \pm 3.29\text{e}+02$ | 104.68            |

Table 5: Wall-clock training time across the six POPGym environments for different privileged signals. We report absolute training time (in seconds) and relative mean runtime w.r.t. the mean runtime of symmetric A2C (100%). Means and standard deviations are computed across 20 independent runs.

| Environment        | Privileged Signal  | Wall-Clock Time (s)                   | Relative Time (%) |
|--------------------|--------------------|---------------------------------------|-------------------|
| Higher Lower       | none               | $8.76\text{e}+03 \pm 1.36\text{e}+02$ | 100.00            |
|                    | full-state         | $9.46\text{e}+03 \pm 1.04\text{e}+02$ | 107.93            |
|                    | previous-card      | $9.51\text{e}+03 \pm 7.47\text{e}+01$ | 108.51            |
|                    | both-cards         | $9.52\text{e}+03 \pm 6.90\text{e}+01$ | 108.66            |
|                    | expert             | $9.53\text{e}+03 \pm 7.82\text{e}+01$ | 108.80            |
| Repeat Previous    | none               | $9.12\text{e}+03 \pm 8.92\text{e}+02$ | 100.00            |
|                    | hand               | $9.66\text{e}+03 \pm 1.00\text{e}+03$ | 105.90            |
|                    | dealt              | $9.76\text{e}+03 \pm 1.04\text{e}+03$ | 106.92            |
|                    | full-state         | $9.91\text{e}+03 \pm 8.90\text{e}+02$ | 108.62            |
|                    | previous-card      | $9.97\text{e}+03 \pm 8.70\text{e}+02$ | 109.30            |
| Repeat First       | none               | $9.14\text{e}+03 \pm 7.97\text{e}+02$ | 100.00            |
|                    | dealt              | $9.70\text{e}+03 \pm 9.52\text{e}+02$ | 106.10            |
|                    | hand               | $9.76\text{e}+03 \pm 8.91\text{e}+02$ | 106.85            |
|                    | full-state         | $9.80\text{e}+03 \pm 9.87\text{e}+02$ | 107.19            |
|                    | first-card         | $9.95\text{e}+03 \pm 9.31\text{e}+02$ | 108.86            |
| Count Recall       | none               | $1.18\text{e}+04 \pm 7.38\text{e}+02$ | 100.00            |
|                    | expert             | $1.22\text{e}+04 \pm 4.24\text{e}+02$ | 103.88            |
|                    | freq-cards-seen    | $1.23\text{e}+04 \pm 4.95\text{e}+02$ | 104.79            |
|                    | freq-cards-queried | $1.24\text{e}+04 \pm 4.88\text{e}+02$ | 105.12            |
|                    | full-state         | $1.24\text{e}+04 \pm 4.61\text{e}+02$ | 105.26            |
|                    | last-obs           | $1.26\text{e}+04 \pm 4.53\text{e}+02$ | 107.47            |
| Concentration      | none               | $6.67\text{e}+03 \pm 4.80\text{e}+02$ | 100.00            |
|                    | flipped-cards      | $6.90\text{e}+03 \pm 4.40\text{e}+02$ | 103.48            |
|                    | second-card        | $6.97\text{e}+03 \pm 3.95\text{e}+02$ | 104.50            |
|                    | first-card         | $7.04\text{e}+03 \pm 4.43\text{e}+02$ | 105.46            |
|                    | all-values         | $7.15\text{e}+03 \pm 4.49\text{e}+02$ | 107.17            |
|                    | full-state         | $7.17\text{e}+03 \pm 3.58\text{e}+02$ | 107.44            |
|                    | is-face-up         | $7.25\text{e}+03 \pm 2.94\text{e}+02$ | 108.73            |
| Position Cart Pole | none               | $8.87\text{e}+03 \pm 7.90\text{e}+01$ | 100.00            |
|                    | full-state         | $9.06\text{e}+03 \pm 1.02\text{e}+02$ | 102.14            |
|                    | both-velocities    | $1.01\text{e}+04 \pm 2.51\text{e}+02$ | 113.34            |
|                    | x-velocity         | $1.02\text{e}+04 \pm 6.57\text{e}+01$ | 115.48            |
|                    | angle-velocity     | $1.10\text{e}+04 \pm 1.83\text{e}+01$ | 123.62            |

Table 6: Wall-clock runtime comparison of the two informativeness tests across different privileged signals  $i_t$ . Means and standard deviations are computed across 10 independent runs.

| Privileged Signal                           | Wall-Clock Runtime (s)        |                                 |
|---------------------------------------------|-------------------------------|---------------------------------|
|                                             | Residual Informativeness Test | Prediction Informativeness Test |
| $i_t = [s_t^1, s_t^2]$                      | $939.70 \pm 43.76$            | $898.24 \pm 65.57$              |
| $i_t = [s_t^1, s_t^2, s_t^3]$               | $936.96 \pm 42.71$            | $906.89 \pm 56.61$              |
| $i_t = [s_t^1, s_t^2, s_t^4]$               | $964.76 \pm 39.17$            | $910.51 \pm 48.68$              |
| $i_t = [s_t^1, s_t^2, s_t^5]$               | $950.20 \pm 31.40$            | $913.10 \pm 69.74$              |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$        | $939.95 \pm 46.23$            | $890.51 \pm 75.44$              |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$        | $931.06 \pm 30.45$            | $889.61 \pm 82.82$              |
| $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$        | $919.11 \pm 50.37$            | $905.27 \pm 55.26$              |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4, s_t^5]$ | $946.33 \pm 21.69$            | $896.10 \pm 48.37$              |

## G Additional Experiments

This section summarizes our additional empirical results.

### G.1 Learning Performance on POPGym Environments

Table 7 reports the final performance (mean episodic return) and area under the learning curve (AUC) after 2 million steps for actor-critic variants with access to different privileged signals on the POPGym environments.

Table 7: Learning performance of actor-critic variants with different privileged signals on the POPGym environments, computed as mean  $\pm$  standard deviation over 20 independent runs.

| Environment        | Privileged Signal  | Final Performance ( $\uparrow$ ) | AUC ( $\uparrow$ )      |
|--------------------|--------------------|----------------------------------|-------------------------|
| Higher Lower       | full-state         | 4.90e-01 $\pm$ 6.91e-02          | 9.59e+05 $\pm$ 1.99e+05 |
|                    | expert             | 4.92e-01 $\pm$ 9.53e-02          | 9.59e+05 $\pm$ 1.99e+05 |
|                    | both-cards         | 4.81e-01 $\pm$ 9.23e-02          | 9.58e+05 $\pm$ 1.97e+05 |
|                    | previous-card      | 5.02e-01 $\pm$ 1.20e-01          | 9.55e+05 $\pm$ 1.98e+05 |
|                    | none               | 4.83e-01 $\pm$ 7.82e-02          | 9.38e+05 $\pm$ 1.26e+05 |
| Repeat Previous    | previous-card      | 3.79e-01 $\pm$ 3.02e-01          | 6.81e+05 $\pm$ 5.40e+05 |
|                    | dealt              | 4.77e-01 $\pm$ 7.46e-02          | 8.28e+05 $\pm$ 2.88e+05 |
|                    | none               | 4.60e-01 $\pm$ 1.20e-01          | 8.36e+05 $\pm$ 2.64e+05 |
|                    | full-state         | 5.06e-01 $\pm$ 7.44e-02          | 8.75e+05 $\pm$ 2.58e+05 |
|                    | hand               | 5.02e-01 $\pm$ 3.70e-02          | 9.14e+05 $\pm$ 1.91e+05 |
| Repeat First       | none               | 8.29e-01 $\pm$ 5.34e-01          | 6.63e+05 $\pm$ 1.18e+06 |
|                    | hand               | 5.51e-01 $\pm$ 8.15e-01          | 5.38e+05 $\pm$ 1.34e+06 |
|                    | first-card         | 3.45e-01 $\pm$ 9.20e-01          | 2.86e+05 $\pm$ 1.44e+06 |
|                    | dealt              | 8.76e-01 $\pm$ 4.52e-01          | 1.66e+05 $\pm$ 1.41e+06 |
|                    | full-state         | 9.00e-01 $\pm$ 4.47e-01          | 1.96e+04 $\pm$ 1.47e+06 |
| Count Recall       | freq-cards-queried | 7.18e-01 $\pm$ 1.41e-01          | 1.57e+06 $\pm$ 2.19e+05 |
|                    | last-obs           | 7.75e-01 $\pm$ 1.19e-01          | 1.61e+06 $\pm$ 2.27e+05 |
|                    | full-state         | 7.53e-01 $\pm$ 1.10e-01          | 1.63e+06 $\pm$ 1.93e+05 |
|                    | freq-cards-seen    | 7.69e-01 $\pm$ 1.32e-01          | 1.66e+06 $\pm$ 2.07e+05 |
|                    | none               | 8.45e-01 $\pm$ 1.13e-01          | 1.76e+06 $\pm$ 1.68e+05 |
| Concentration      | expert             | 8.84e-01 $\pm$ 1.01e-01          | 1.82e+06 $\pm$ 1.29e+05 |
|                    | all-values         | 1.45e-01 $\pm$ 7.59e-02          | 3.49e+05 $\pm$ 2.13e+05 |
|                    | second-card        | 1.57e-01 $\pm$ 1.11e-01          | 3.79e+05 $\pm$ 2.24e+05 |
|                    | first-card         | 1.29e-01 $\pm$ 9.48e-02          | 3.89e+05 $\pm$ 2.43e+05 |
|                    | none               | 2.38e-01 $\pm$ 1.39e-01          | 4.22e+05 $\pm$ 2.32e+05 |
|                    | is-face-up         | 2.20e-01 $\pm$ 1.06e-01          | 4.71e+05 $\pm$ 2.73e+05 |
|                    | flipped-cards      | 2.59e-01 $\pm$ 2.79e-01          | 4.91e+05 $\pm$ 4.69e+05 |
| Position Cart Pole | full-state         | 4.24e-01 $\pm$ 4.03e-01          | 7.00e+05 $\pm$ 5.60e+05 |
|                    | angle-velocity     | 5.82e-01 $\pm$ 2.95e-01          | 8.68e+05 $\pm$ 4.64e+05 |
|                    | both-velocities    | 5.37e-01 $\pm$ 3.45e-01          | 8.48e+05 $\pm$ 4.78e+05 |
|                    | full-state         | 5.69e-01 $\pm$ 2.61e-01          | 8.26e+05 $\pm$ 4.20e+05 |
|                    | x-velocity         | 1.65e-01 $\pm$ 1.05e-01          | 2.57e+05 $\pm$ 1.49e+05 |
|                    | none               | 5.55e-02 $\pm$ 2.09e-02          | 1.53e+05 $\pm$ 7.50e+04 |

### G.2 Effect of Privileged Signal Choice on Policy Performance

We study how different privileged signal generators affect the quality of the learned policy. To this end, we train a symmetric actor-critic baseline and multiple informed asymmetric actor-critic (IAAC) agents in the same fixed randomly sampled environment, each IAAC variant using a different privileged signal as described in Section 5.2. Each model is trained for 15,000 gradient steps, where

every step uses a batch of 16 episodes. Every 50 gradient steps, we evaluate the current policy by estimating the mean episodic return over 50 evaluation episodes. All experiments are repeated with 10 random seeds. Figure 2 presents the actor loss, critic loss, and episodic return on evaluation episodes during training for selected actor-critic variants, smoothed using a moving average over 500 episodes.

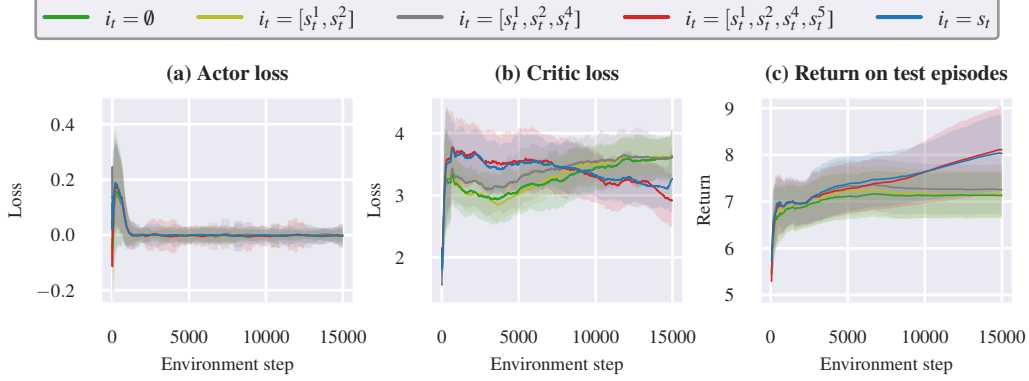


Figure 2: Learning performance of selected actor-critic variants on a randomly sampled informed POMDP. The curves depict (a) the actor loss, (b) the critic loss, and (c) episodic returns evaluated on 50 test episodes. The results are smoothed using a moving average over 500 environment steps, with the means and standard deviations computed across 10 independent runs.

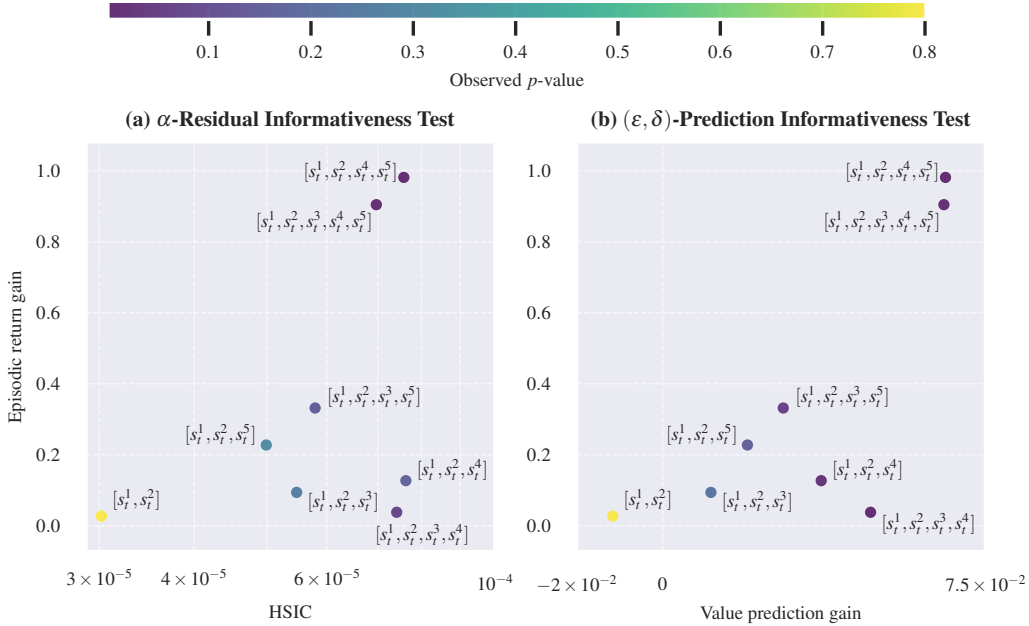


Figure 3: Relationship between estimated signal informativeness and downstream policy performance. Each point corresponds to a distinct privileged signal configuration. The x-axis shows the effect size of the (a) residual-based informativeness measure (observed HSIC) and (b) post-hoc prediction informativeness measure (value prediction gain), while the y-axis shows the average gain in mean episodic return relative to a symmetric actor-critic baseline, aggregated over all evaluation episodes and averaged across 10 runs. Color of the point indicates the corresponding  $p$ -value of the respective informativeness test.

For each privileged signal generator, we compute the average gain in mean episodic return relative to the symmetric critic across all evaluation episodes during training, averaged over the 10 runs. In Figure 3, we plot this performance gain against the corresponding effect sizes of the two informativeness measures (observed HSIC and value prediction gain) to illustrate the relationship between signal informativeness and policy performance. The informativeness tests are conducted offline using separate episodes collected with a random policy, allowing us to assess how well the proposed criteria serve as a proxy for downstream policy improvement.

The upper-right quadrant of the scatter plots highlights the most relevant privileged signal generators, combining high estimated informativeness with strong gains in mean episodic return. Signals such as  $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$  and  $i_t = s_t$  are consistently identified as informative (low  $p$ -values: 0.05-residual informativeness and (0, 0.05)- or (0, 0.1)-prediction informativeness) and yield the largest performance improvements. In contrast, some signals (e.g.,  $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$  or  $i_t = [s_t^1, s_t^2, s_t^4]$ ) show strong effect sizes but smaller gains in episodic returns, while  $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$  provides higher episodic gains despite lower measured informativeness. Signals identical to the observations ( $i_t = o_t = [s_t^1, s_t^2]$ ) remain uninformative, in line with their limited contribution to policy improvement.

These results show that, while the proposed informativeness criteria are designed to quantify the predictive value of privileged signals for return estimation, they also serve as a reasonable proxy for downstream policy improvement. Notably, the tests are performed on episodes collected with a random policy, yet signals identified as informative tend to yield higher gains in episodic return during policy training. This suggests that value-based informativeness captures essential aspects of signals that facilitate control, while also highlighting the need for future research to develop criteria that directly assess a signal’s potential to enhance policy performance beyond value prediction. Interestingly, the full-state signal  $i_t = s_t$  does not always yield the highest episodic return gains, further highlighting the practical relevance of the informed asymmetric actor-critic framework, which can exploit any state-dependent privileged information beyond full-state access.

### G.3 Noisy observations

In addition to the noiseless observation setting studied in Section 5.2, we assess our informativeness criteria under noisy observation configurations. Recall that we generate observations by masking a subset of features from  $i_t$  and adding Gaussian noise modulated by  $\beta_o \geq 0$ .

Tables 8 and 9 report results for multiple privileged signal generators under noise levels  $\beta_o = 0.1$  and  $\beta_o = 0.5$ , averaged over ten independent runs, using the  $\alpha$ -residual-informativeness criterion and the post-hoc (0,  $\delta$ )-prediction-informativeness measure, respectively. Increasing observation noise from  $\beta_o = 0.0$  to  $\beta_o = 0.1$  or  $\beta_o = 0.5$  has only a minor impact on the relative ranking of privileged signals under both informativeness criteria. Signals that include the most return-relevant state components (notably those containing  $s_t^4$ ) remain consistently identified as informative, with low  $p$ -values in both the residual-based and prediction tests. In contrast, signals lacking these components (e.g.,  $[s_t^1, s_t^2]$ ) continue to appear uninformative across noise levels. While effect sizes tend to increase slightly with higher noise for the truly informative signals, the statistical conclusions remain generally unchanged.

This trend is intuitive. As observation noise increases, the history  $h_t$  becomes a less accurate proxy for the latent state, making the history-conditioned belief over states more uncertain. Privileged signals that provide additional state-conditioned information can therefore explain a larger fraction of the remaining return variance, increasing their relative informativeness compared to the history alone. An exception is  $i_t = [s_t^1, s_t^2, s_t^3]$  under the residual-based criterion: its mean  $p$ -value increases noticeably with higher noise, making it less likely to be deemed as informative. In contrast, the post-hoc prediction criterion shows, on average, an increase in effect size for this signal at higher noise, indicating that it still provides useful predictive information even when the residual test becomes more conservative.

### G.4 Pure-noise privileged signal

To evaluate whether the residual-informativeness criterion correctly rejects uninformative privileged signals, we perform a robustness ablation using artificially constructed noise signals. Specifically, we use independent Gaussian noise as privileged signals, i.e.,  $i_t \sim \mathcal{N}(0, \mathbf{I}_{d_i})$ , where  $\mathbf{I}_{d_i}$  denotes

the  $d_i$ -dimensional identity matrix. Since these signals contain no information about future returns beyond the observable history-action pair, the residual-informativeness test should not identify them as informative.

We evaluate Gaussian noise signals with dimensionalities ranging from  $d_i = 1$  to  $d_i = 5$ . For each dimensionality, we report the resulting test statistics and the corresponding  $p$ -values in Table 10. The residual-based test consistently assigns high  $p$ -values to these purely noisy privileged signals, remaining well above the significance threshold across all tested dimensionalities. This confirms that the proposed criterion does not falsely identify arbitrary additional features as informative and can effectively filter out uninformative privileged signals.

## G.5 Alternative feature subset

To assess the robustness of our results with respect to the choice of feature subsets used as observations, we repeat the informativeness tests from Section 5.2 using an alternative feature subset, with  $\beta_o = 0.0$ , while keeping the reward weights unchanged. This change in configuration also induces different signal generators, as we employ privileged signals of the form  $i_t = (o_t, o_t^+)$ . Comparing results across feature subsets allows us to evaluate the extent to which our findings depend on the particular choice of observable features.

Table 11 summarizes the performance of different privileged signal generators under both informativeness criteria, averaged over ten independent runs. For this analysis, the agent is provided with observations  $o_t = [s_t^1, s_t^3]$ . In line with the previous experiments, signals that add little beyond the observation (e.g.,  $i_t = [s_t^1, s_t^3]$ ) are identified as uninformative by both criteria, with near-zero effect sizes and large  $p$ -values. Incorporating return-relevant components increases effect sizes. Under the post-hoc prediction criterion, signals containing strongly reward-relevant features (i.e.,  $s_t^4$  and combinations including  $s_t^2$  and  $s_t^4$ ) yield clear and statistically significant gains, reflected in larger positive  $L_\tau$  and small  $p$ -values. This supports earlier findings that signals which include the most reward-informative state dimensions are consistently ranked highest. The residual-based criterion remains more conservative, as in the original observation setting. Although the dependency measure  $\rho_{\text{obs}}$  generally increases when informative features are added, statistical significance at level  $\alpha = 0.1$  is typically reached only for larger feature subsets (e.g., those containing both  $s_t^2$  and  $s_t^4$ , or the full state  $s_t$ ).

Overall, the relative ordering of privileged signal generators is largely preserved across observation choices: signals containing the most reward-relevant state components remain the most informative under both criteria, although in some cases, higher significance levels are required to reach statistical significance.

Table 8: Comparison of candidate privileged signals using the residual-based informativeness criterion across ten independent runs on a randomly sampled informed POMDP environment with noisy observations.

| Privileged Signal                    | $\beta_o = 0.1$                     |                               | $\beta_o = 0.5$                     |                               |
|--------------------------------------|-------------------------------------|-------------------------------|-------------------------------------|-------------------------------|
|                                      | $\rho_{\text{obs}} (\uparrow)$      | $p\text{-value} (\downarrow)$ | $\rho_{\text{obs}} (\uparrow)$      | $p\text{-value} (\downarrow)$ |
| $i_t = [s_t^1, s_t^2]$               | $3.2\text{e-}05 \pm 1.2\text{e-}05$ | $0.406 \pm 0.280$             | $3.8\text{e-}05 \pm 1.8\text{e-}05$ | $0.380 \pm 0.321$             |
| $i_t = [s_t^1, s_t^2, s_t^3]$        | $8.4\text{e-}05 \pm 3.5\text{e-}05$ | $0.069 \pm 0.120$             | $8.7\text{e-}05 \pm 3.9\text{e-}05$ | $0.105 \pm 0.223$             |
| $i_t = [s_t^1, s_t^2, s_t^4]$        | $1.2\text{e-}04 \pm 3.7\text{e-}05$ | $0.011 \pm 0.003$             | $1.3\text{e-}04 \pm 3.8\text{e-}05$ | $0.010 \pm 0.000$             |
| $i_t = [s_t^1, s_t^2, s_t^5]$        | $4.2\text{e-}05 \pm 1.5\text{e-}05$ | $0.262 \pm 0.303$             | $7.0\text{e-}05 \pm 4.0\text{e-}05$ | $0.088 \pm 0.087$             |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$ | $1.1\text{e-}04 \pm 3.7\text{e-}05$ | $0.018 \pm 0.017$             | $1.2\text{e-}04 \pm 3.7\text{e-}05$ | $0.012 \pm 0.006$             |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$ | $7.2\text{e-}05 \pm 2.7\text{e-}05$ | $0.061 \pm 0.065$             | $8.2\text{e-}05 \pm 3.5\text{e-}05$ | $0.043 \pm 0.062$             |
| $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$ | $9.8\text{e-}05 \pm 2.8\text{e-}05$ | $0.015 \pm 0.013$             | $1.2\text{e-}04 \pm 4.0\text{e-}05$ | $0.012 \pm 0.006$             |
| $i_t = s_t$                          | $9.5\text{e-}05 \pm 3.0\text{e-}05$ | $0.013 \pm 0.009$             | $1.1\text{e-}04 \pm 3.5\text{e-}05$ | $0.013 \pm 0.007$             |

Table 9: Comparison of candidate privileged signals using the post-hoc prediction informativeness criterion across ten independent runs on a randomly sampled informed POMDP environment with noisy observations.

| Privileged Signal                    | $\beta_o = 0.1$         |                             | $\beta_o = 0.5$         |                                     |
|--------------------------------------|-------------------------|-----------------------------|-------------------------|-------------------------------------|
|                                      | $L_\tau$ ( $\uparrow$ ) | $p$ -value ( $\downarrow$ ) | $L_\tau$ ( $\uparrow$ ) | $p$ -value ( $\downarrow$ )         |
| $i_t = [s_t^1, s_t^2]$               | $-0.013 \pm 0.014$      | $0.788 \pm 0.343$           | $0.004 \pm 0.012$       | $0.456 \pm 0.298$                   |
| $i_t = [s_t^1, s_t^2, s_t^3]$        | $0.007 \pm 0.016$       | $0.384 \pm 0.330$           | $0.021 \pm 0.014$       | $0.081 \pm 0.084$                   |
| $i_t = [s_t^1, s_t^2, s_t^4]$        | $0.035 \pm 0.018$       | $0.041 \pm 0.071$           | $0.049 \pm 0.017$       | $0.001 \pm 0.002$                   |
| $i_t = [s_t^1, s_t^2, s_t^5]$        | $0.019 \pm 0.019$       | $0.219 \pm 0.231$           | $0.036 \pm 0.017$       | $0.030 \pm 0.031$                   |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$ | $0.046 \pm 0.018$       | $0.009 \pm 0.016$           | $0.057 \pm 0.017$       | $3.2\text{e-}04 \pm 5.9\text{e-}04$ |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$ | $0.027 \pm 0.019$       | $0.110 \pm 0.137$           | $0.043 \pm 0.017$       | $0.010 \pm 0.011$                   |
| $i_t = [s_t^1, s_t^2, s_t^4, s_t^5]$ | $0.064 \pm 0.020$       | $0.003 \pm 0.006$           | $0.077 \pm 0.018$       | $1.5\text{e-}04 \pm 3.3\text{e-}04$ |
| $i_t = s_t$                          | $0.056 \pm 0.032$       | $0.073 \pm 0.131$           | $0.068 \pm 0.030$       | $0.022 \pm 0.041$                   |

Table 10: Residual-informativeness test on Gaussian-noise privileged signals  $i_t \sim \mathcal{N}(0, \mathbf{I}_{d_i})$ , where  $\mathbf{I}_{d_i}$  denotes the  $d_i$ -dimensional identity matrix. Means and standard deviations are computed across 10 independent runs.

| Privileged Signal                       | $\rho_{\text{obs}}$ ( $\uparrow$ )  | $p$ -value ( $\downarrow$ ) |
|-----------------------------------------|-------------------------------------|-----------------------------|
| $i_t \sim \mathcal{N}(0, \mathbf{I}_1)$ | $3.1\text{e-}05 \pm 1.3\text{e-}05$ | $0.482 \pm 0.263$           |
| $i_t \sim \mathcal{N}(0, \mathbf{I}_2)$ | $3.7\text{e-}05 \pm 9.5\text{e-}06$ | $0.279 \pm 0.193$           |
| $i_t \sim \mathcal{N}(0, \mathbf{I}_3)$ | $2.9\text{e-}05 \pm 6.7\text{e-}06$ | $0.464 \pm 0.239$           |
| $i_t \sim \mathcal{N}(0, \mathbf{I}_4)$ | $2.9\text{e-}05 \pm 6.7\text{e-}06$ | $0.527 \pm 0.271$           |
| $i_t \sim \mathcal{N}(0, \mathbf{I}_5)$ | $3.3\text{e-}05 \pm 1.5\text{e-}05$ | $0.430 \pm 0.294$           |

Table 11: Comparison of candidate privileged signals using residual-based and post-hoc prediction informativeness criteria across ten independent runs with observations  $o_t = [s_t^1, s_t^3]$ .

| Privileged Signal                    | Residual Informativeness            |                             | Prediction Informativeness  |                             |
|--------------------------------------|-------------------------------------|-----------------------------|-----------------------------|-----------------------------|
|                                      | $\rho_{\text{obs}}$ ( $\uparrow$ )  | $p$ -value ( $\downarrow$ ) | $L_\tau$ ( $\uparrow$ )     | $p$ -value ( $\downarrow$ ) |
| $i_t = [s_t^1, s_t^3]$               | $2.4\text{e-}05 \pm 6.6\text{e-}06$ | $0.673 \pm 0.273$           | $-1.9\text{e-}02 \pm 0.010$ | $0.973 \pm 0.041$           |
| $i_t = [s_t^1, s_t^2, s_t^3]$        | $4.9\text{e-}05 \pm 1.8\text{e-}05$ | $0.240 \pm 0.315$           | $0.023 \pm 0.010$           | $0.052 \pm 0.099$           |
| $i_t = [s_t^1, s_t^3, s_t^4]$        | $4.5\text{e-}05 \pm 1.6\text{e-}05$ | $0.314 \pm 0.329$           | $0.008 \pm 0.015$           | $0.307 \pm 0.345$           |
| $i_t = [s_t^1, s_t^3, s_t^5]$        | $5.8\text{e-}05 \pm 4.7\text{e-}05$ | $0.307 \pm 0.275$           | $0.004 \pm 0.014$           | $0.390 \pm 0.342$           |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^4]$ | $6.2\text{e-}05 \pm 1.6\text{e-}05$ | $0.057 \pm 0.082$           | $0.057 \pm 0.012$           | $0.003 \pm 0.008$           |
| $i_t = [s_t^1, s_t^2, s_t^3, s_t^5]$ | $5.4\text{e-}05 \pm 3.3\text{e-}05$ | $0.240 \pm 0.279$           | $0.039 \pm 0.012$           | $0.016 \pm 0.031$           |
| $i_t = [s_t^1, s_t^3, s_t^4, s_t^5]$ | $6.8\text{e-}05 \pm 3.8\text{e-}05$ | $0.080 \pm 0.076$           | $0.031 \pm 0.018$           | $0.107 \pm 0.192$           |
| $i_t = s_t$                          | $6.3\text{e-}05 \pm 3.0\text{e-}05$ | $0.071 \pm 0.102$           | $0.068 \pm 0.019$           | $0.006 \pm 0.013$           |