Adversarial Imitation Learning from Visual Observations using Latent Information

Vittorio Giammarino *vgiammar@bu.edu*
Division of Systems Engineering
Boston University

James Queeney *queeney@merl.com*
Mitsubishi Electric Research Laboratories

Ioannis Ch. Paschalidis *yannisp@bu.edu*
Department of Electrical and Computer Engineering
Division of Systems Engineering
Faculty of Computing & Data Sciences
Boston University

Reviewed on OpenReview: *https://openreview.net/forum?id=ydPHjgf6h0*

Abstract

We focus on the problem of imitation learning from visual observations, where the learning agent has access to videos of experts as its sole learning source. The challenges of this framework include the absence of expert actions and the partial observability of the environment, as the ground-truth states can only be inferred from pixels. To tackle this problem, we first conduct a theoretical analysis of imitation learning in partially observable environments. We establish upper bounds on the suboptimality of the learning agent with respect to the divergence between the expert and the agent latent state-transition distributions. Motivated by this analysis, we introduce an algorithm called Latent Adversarial Imitation from Observations, which combines off-policy adversarial imitation techniques with a learned latent representation of the agent's state from sequences of observations. In experiments on high-dimensional continuous robotic tasks, we show that our model-free approach in latent space matches state-of-the-art performance. Additionally, we show how our method can be used to improve the efficiency of reinforcement learning from pixels by leveraging expert videos. To ensure reproducibility, we provide free access to all the learning curves and open-source our code.

1 Introduction

Learning from videos represents a compelling opportunity for the future, as it offers a cost-effective and efficient way to teach autonomous agents new skills and behaviors. Compared to other methods, video recording is a faster and more flexible alternative for gathering data. Moreover, with the abundance of high-quality videos available on the internet, learning from videos has become increasingly accessible in recent years. However, despite the potential benefits, this approach remains challenging as it involves several technical problems that must be addressed simultaneously in order to succeed. These problems include representation learning, significant computational demands due to high-dimensional observation space, the partial observability of the decision process, and lack of expert actions. Our objective is to establish algorithms capable of overcoming all of these challenges, enabling the learning of complex robotics tasks directly from videos of experts.

Formally, our focus is on the problem of *Visual Imitation from Observations (V-IfO)*. In V-IfO, the learning agent does not have access to a pre-specified reward function, and instead has to learn by imitating an

Table 1: A summary of imitation from experts: Imitation Learning (IL), Imitation from Observations (IfO), Visual Imitation Learning (V-IL), and Visual Imitation from Observations (V-IfO).

	IL	IfO	V-IL	V-IfO
Fully observable environment	✓	✓	✗	✗
Access to expert actions	✓	✗	✓	✗

expert's behavior. Additionally, in V-IfO, expert actions are not accessible during the learning process, and the pixel-based observations we obtain from video frames result in partial observability. The absence of expert actions and the partial observability of the environment distinguish V-IfO from other types of imitation from experts. Specifically, we identify three other frameworks previously addressed in the literature: *Imitation Learning (IL)* (Atkeson & Schaal, 1997; Abbeel & Ng, 2004; Ross & Bagnell, 2010; Reske et al., 2021; Giammarino et al., 2023a) where states are fully observable and expert state-action pairs are accessible, *Visual Imitation Learning (V-IL)* (Rafailov et al., 2021) which explores the idea of imitating directly from pixels but still assumes that expert actions are provided to the learning agent, and *Imitation from Observations (IfO)* (Torabi et al., 2018b;a) which retains full observability but considers only the availability of expert states. Table 1 summarizes these frameworks.

In order to address the V-IfO problem, this paper introduces both theoretical and algorithmic contributions. First, we provide a theoretical analysis of the problem and demonstrate that the suboptimality of the learning agent can be upper bounded by the divergence between the expert and the agent latent state-transition distributions. Our analysis motivates the reduction of the V-IfO problem to two subproblems: (i) estimating a proper latent representation from sequences of observations and (ii) efficiently minimizing the divergence between expert and agent distributions in this latent space. Next, we propose practical solutions to these subproblems. By doing so, we formalize a novel algorithm called *Latent Adversarial Imitation from Observations (LAIfO)*, which tackles the divergence minimization step using off-policy adversarial imitation techniques (Ghasemipour et al., 2020) and recovers a latent representation of the ground-truth state by means of observations stacking (Mnih et al., 2013; 2015) and data augmentation (Laskin et al., 2020b; Kostrikov et al., 2020; Yarats et al., 2021). We evaluate our algorithm on the DeepMind Control Suite (Tassa et al., 2018), demonstrating that our model-free approach in latent space achieves state-of-the-art performance. We conclude by showing how LAIfO can be used on challenging environments, such as the humanoid from pixels (Tassa et al., 2018), to improve Reinforcement Learning (RL) efficiency by leveraging expert videos.

The remainder of the paper is organized as follows: Section 2 provides a summary of the most related works to this paper. Section 3 introduces notation and background on RL and IL. Section 4 provides a theoretical analysis of the V-IfO problem. Section 5 introduces our algorithm, LAIfO, and outlines how it can leverage expert videos to improve data efficiency of RL from pixels. Finally, Section 6 presents our experimental results and Section 7 concludes the paper providing a general discussion on our findings.

2 Related work

In recent years, several studies have focused on the IL problem (Atkeson & Schaal, 1997; Abbeel & Ng, 2004; Ross & Bagnell, 2010; Reddy et al., 2019; Reske et al., 2021; Giammarino et al., 2023a) and, in particular, on the generative adversarial IL framework (Ho & Ermon, 2016) which has emerged as one of the most promising approaches for IL. Adversarial IL builds upon a vast body of work on inverse RL (Russell, 1998; Ng et al., 2000; Abbeel & Ng, 2004; Syed & Schapire, 2007; Ziebart et al., 2008; Syed et al., 2008). The primary goal of inverse RL is to identify a reward function that enables expert trajectories (i.e., state-action pairs) to be optimal. The reward function obtained from the inverse RL step is then used to train agents in order to match the expert's expected reward. In the fully observable setting, adversarial IL was originally formalized in Ho & Ermon (2016) and Fu et al. (2017). It was later extended to the observation only setting in Torabi et al. (2018b) and to the visual setting in Karnan et al. (2022). Furthermore, the adversarial IfO problem has been theoretically analyzed in Yang et al. (2019) and Cheng et al. (2021). Note that all of these studies are built upon on-policy RL (Schulman et al., 2017), which provides good learning stability but is known

for poor sample efficiency. In recent works, this efficiency issue has been addressed by using off-policy RL algorithms in the adversarial optimization step (Haarnoja et al., 2018; Lillicrap et al., 2015). These include DAC (Kostrikov et al., 2018), SAM (Blondé & Kalousis, 2019), and ValueDICE (Kostrikov et al., 2019) for the adversarial IL problem, and OPOLO (Zhu et al., 2020) and MobILE (Kidambi et al., 2021) for the adversarial IfO problem. Another line of research has tackled IfO by directly estimating expert actions and subsequently deploying IL techniques on the estimated state-action pairs (Torabi et al., 2018a; Liu et al., 2018; Behbahani et al., 2019; Zhang & Ohn-Bar, 2021; Zhang et al., 2022; Shaw et al., 2023; Yang et al., 2023). Finally, recent studies have investigated offline alternatives to the adversarial IL framework (Dadashi et al., 2020).

All of the aforementioned works consider fully observable environments modeled as *Markov Decision Processes (MDPs)*. However, when dealing with pixels, individual observations alone are insufficient for determining optimal actions. As a result, recent works (Rafailov et al., 2021; Hu et al., 2022) have treated the V-IL problem as a *Partially Observable Markov Decision Process (POMDP)* (Astrom, 1965). In particular, Rafailov et al. (2021) addressed the V-IL problem by proposing a model-based extension (Hafner et al., 2019a;b) of generative adversarial IL called VMAIL. The work in Gangwani et al. (2020) also considered IL in a POMDP in order to handle missing information in the agent state, but did not directly focus on learning from pixels. The more difficult V-IfO problem, on the other hand, has received less attention in the literature. To the best of our knowledge, this problem has only been considered by the recent algorithm PatchAIL (Liu et al., 2023), where off-policy adversarial IL is performed directly on the pixel space. Different from Liu et al. (2023), we first study V-IfO from a theoretical perspective, which motivates an algorithm that performs imitation on a *latent representation of the agent state* rather than directly on the pixel space as in PatchAIL. This difference is crucial to ensure improved computational efficiency.

Our work is also related to the RL from pixels literature which tackles the challenge of maximizing an agent's expected return end-to-end, from pixels to actions. This approach has proven successful in playing Atari games (Mnih et al., 2013; 2015). Recently, RL from pixels has also been extended to tackle continuous action space tasks, such as robot locomotion, by leveraging either data augmentation techniques (Laskin et al., 2020a;b; Kostrikov et al., 2020; Lee et al., 2020; Raileanu et al., 2021; Yarats et al., 2021) or variational inference (Hafner et al., 2019a;b; Lee et al., 2020; Hafner et al., 2020).

Finally, another line of research has focused on the visual imitation problem in the presence of domain mismatch, also known as third-person imitation learning (Stadie et al., 2017; Okumura et al., 2020; Cetin & Celiktutan, 2021; Giammarino et al., 2023b). This paradigm relaxes the assumption that the agent and the expert are defined on the same decision process and represents a generalization of the imitation from experts frameworks introduced in Table 1.

3 Preliminaries

Unless indicated otherwise, we use uppercase letters (e.g., S_t) for random variables, lowercase letters (e.g., s_t) for values of random variables, script letters (e.g., $\mathcal{S}$) for sets, and bold lowercase letters (e.g., $\boldsymbol{\theta}$) for vectors. Let $[t_1 : t_2]$ be the set of integers t such that $t_1 \leq t \leq t_2$; we write S_t such that $t_1 \leq t \leq t_2$ as $S_{t_1:t_2}$. We denote with $\mathbb{E}[\cdot]$ expectation, with $\mathbb{P}(\cdot)$ probability, and with $\mathbb{D}_f(\cdot, \cdot)$ an f-divergence between two distributions of which the total variation (TV) distance, $\mathbb{D}_{\text{TV}}(\cdot, \cdot)$, and the Jensen-Shannon divergence, $\mathbb{D}_{\text{JS}}(\cdot||\cdot)$, are special cases.

We model the decision process as an infinite-horizon discounted POMDP described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathcal{T}, \mathcal{U}, \mathcal{R}, \rho_0, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, and $\mathcal{X}$ is the set of observations. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$ is the transition probability function where $P(\mathcal{S})$ denotes the space of probability distributions over $\mathcal{S}$, $\mathcal{U} : \mathcal{S} \to P(\mathcal{X})$ is the observation probability function, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function which maps state-action pairs to scalar rewards. Alternatively, the reward function can also be expressed as $\mathcal{R} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ mapping state-transition pairs to scalar rewards rather than state-action pairs. Finally, $\rho_0 \in P(\mathcal{S})$ is the initial state distribution and $\gamma \in [0, 1)$ the discount factor. The true environment state $s \in \mathcal{S}$ is unobserved by the agent. Given an action $a \in \mathcal{A}$, the next state is sampled such that $s' \sim \mathcal{T}(\cdot|s, a)$, an observation is generated as $x' \sim \mathcal{U}(\cdot|s')$, and a reward $\mathcal{R}(s, a)$ or $\mathcal{R}(s, s')$ is computed. Note that an MDP is a special case of a POMDP where the underlying state s is directly observed.

Reinforcement learning Given an MDP and a stationary policy $\pi : \mathcal{S} \to P(\mathcal{A})$, the RL objective is to maximize the expected total discounted return $J(\pi) = \mathbb{E}_{\tau_\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)]$ where $\tau_\pi = (s_0, a_0, s_1, a_1, \dots)$. A stationary policy π induces a normalized discounted state visitation distribution defined as $d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s|\rho_0, \pi, \mathcal{T})$, and we define the corresponding normalized discounted state-action visitation distribution as $\rho_\pi(s, a) = d_\pi(s)\pi(a|s)$. Finally, we denote the state value function of π as $V^\pi(s) = \mathbb{E}_{\tau_\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)|S_0 = s]$ and the state-action value function as $Q^\pi(s, a) = \mathbb{E}_{\tau_\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)|S_0 = s, A_0 = a]$. When a function is parameterized with parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$ we write $\pi_{\boldsymbol{\theta}}$.

Generative adversarial imitation learning Assume we have a set of expert demonstrations $\tau_E = (s_{0:T}, a_{0:T})$ generated by the expert policy π_E, a set of trajectories $\tau_{\boldsymbol{\theta}}$ generated by the policy $\pi_{\boldsymbol{\theta}}$, and a discriminator network $D_{\boldsymbol{\chi}} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ parameterized by $\boldsymbol{\chi}$. Generative adversarial IL (Ho & Ermon, 2016) optimizes the min-max objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\chi}} \; \mathbb{E}_{\tau_E}[\log(D_{\boldsymbol{\chi}}(s, a))] + \mathbb{E}_{\tau_{\boldsymbol{\theta}}}[\log(1 - D_{\boldsymbol{\chi}}(s, a))]. \tag{1}$$

Maximizing (1) with respect to $\boldsymbol{\chi}$ is effectively an inverse RL step where a reward function, in the form of the discriminator $D_{\boldsymbol{\chi}}$, is inferred by leveraging τ_E and $\tau_{\boldsymbol{\theta}}$. On the other hand, minimizing (1) with respect to $\boldsymbol{\theta}$ can be interpreted as an RL step, where the agent aims to minimize its expected cost. It has been demonstrated that optimizing the min-max objective in (1) is equivalent to minimizing $\mathbb{D}_{\text{JS}}(\rho_{\pi_{\boldsymbol{\theta}}}(s, a)||\rho_{\pi_E}(s, a))$, so we are recovering the expert state-action visitation distribution (Ghasemipour et al., 2020).

Latent representation in POMDP When dealing with a POMDP, a policy $\pi_{\boldsymbol{\theta}}(x_t)$ that selects an action a_t based on a single observation $x_t \in \mathcal{X}$ is likely to perform poorly since x_t lacks enough information about the unobservable true state s_t. It is therefore beneficial to estimate a distribution of the true state from the full history of prior experiences. To that end, we introduce a latent variable $z_t \in \mathcal{Z}$ such that $z_t = \phi(x_{\leq t}, a_{<t})$, where ϕ maps the history of observations and actions to $\mathcal{Z}$. Alternatively, when actions are not observable, we have $z_t = \phi(x_{\leq t})$. The latent variable z_t should be estimated such that $\mathbb{P}(s_t|x_{\leq t}, a_{<t}) \approx \mathbb{P}(s_t|z_t)$, meaning that z_t represents a sufficient statistic of the history for estimating a distribution of the unobservable true state s_t. It is important to clarify that this does not imply $\mathcal{Z} \equiv \mathcal{S}$.

4 Theoretical analysis

Recall that we consider the V-IfO problem where expert actions are not available and the ground-truth states $s \in \mathcal{S}$ are not observable (see Table 1). As a result, a latent representation $z \in \mathcal{Z}$ is inferred from the history of observations and used by the learning agent to make decisions.

Throughout the paper we make the following assumptions: (i) the expert and the agent act on the same POMDP and (ii) the latent variable z can be estimated from the history of observations as $z_t = \phi(x_{\leq t})$ such that $\mathbb{P}(s_t|z_t, a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Assumption (i) is instrumental for both our derivations and experiments. Relaxing this assumption would lead to dynamics mismatch (Gangwani et al., 2022) and visual domain adaptation problems (Giammarino et al., 2023b), which represent interesting extensions for future work. On the other hand, assumption (ii) explicitly states the characteristics required by the latent variable z; i.e., z_t can be successfully estimated from the history of observations $x_{\leq t}$ in order to approximate a sufficient statistic of the history. Note that this is a common assumption in the IL literature for POMDPs (Gangwani et al., 2020; Rafailov et al., 2021), and estimating such a variable is a non-trivial problem that we address in the next section. We further discuss the importance of this assumption from a theoretical perspective in Appendix B (Remark 1).

On the latent space $\mathcal{Z}$, we can define the normalized discounted latent state visitation distribution as $d_{\pi_{\boldsymbol{\theta}}}(z) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(z_t = z|\rho_0, \pi_{\boldsymbol{\theta}}, \mathcal{T}, \mathcal{U})$ and the normalized discounted latent state-action visitation distribution as $\rho_{\pi_{\boldsymbol{\theta}}}(z, a) = d_{\pi_{\boldsymbol{\theta}}}(z)\pi_{\boldsymbol{\theta}}(a|z)$. Further, we define the latent state-transition visitation distribution as $\rho_{\pi_{\boldsymbol{\theta}}}(z, z') = d_{\pi_{\boldsymbol{\theta}}}(z)\int_{\mathcal{A}} \mathbb{P}(z'|z, \bar{a})\pi_{\boldsymbol{\theta}}(\bar{a}|z)d\bar{a}$ and the normalized discounted joint distribution as $\rho_{\pi_{\boldsymbol{\theta}}}(z, a, z') = \rho_{\pi_{\boldsymbol{\theta}}}(z, a)\mathbb{P}(z'|z, a)$, where

$$\mathbb{P}(z'|z, a) = \int_{\mathcal{S}}\int_{\mathcal{S}}\int_{\mathcal{X}} \mathbb{P}(z'|x', a, z)\mathcal{U}(x'|s')\mathcal{T}(s'|s, a)\mathbb{P}(s|z)dx'ds'ds. \tag{2}$$

Finally, we obtain $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z')$ as

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') = \frac{\mathbb{P}(z'|z,a)\pi_{\boldsymbol{\theta}}(a|z)}{\int_{\mathcal{A}} \mathbb{P}(z'|z,\bar{a})\pi_{\boldsymbol{\theta}}(\bar{a}|z)d\bar{a}}.$$

Note that we write $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}$, with $\pi_{\boldsymbol{\theta}}$ as subscript, in order to explicitly denote the dependency on the policy and omit the subscript, as in (2), when such probability depends only on the environment.

We start by considering the case in which $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $J(\pi) = \mathbb{E}_{\tau_\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)]$. The following Theorem shows how the suboptimality of $\pi_{\boldsymbol{\theta}}$ can be upper bounded by the TV distance between latent state-transition visitation distributions, reducing the V-IfO problem to a divergence minimization problem in the latent space $\mathcal{Z}$.

Theorem 1. *Consider a POMDP, and let $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $z_t = \phi(x_{\leq t})$ such that $\mathbb{P}(s_t|z_t, a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Then, the following inequality holds:*

$$\big|J(\pi_E) - J(\pi_{\boldsymbol{\theta}})\big| \leq \frac{2R_{\max}}{1-\gamma} \mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big) + C,$$

where $R_{\max} = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\mathcal{R}(s,a)|$ and

$$C = \frac{2R_{\max}}{1-\gamma} \mathbb{E}_{\rho_{\pi_{\boldsymbol{\theta}}}(z,z')}\big[\mathbb{D}_{\mathrm{TV}}\big(\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z'), \mathbb{P}_{\pi_E}(a|z,z')\big)\big]. \tag{3}$$

Proof. Using the definition of $J(\pi_{\boldsymbol{\theta}})$, we first upper bound the performance difference between expert and agent by $\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(s,a), \rho_{\pi_E}(s,a)\big)$. Next, we bound the latter divergence by $\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,a), \rho_{\pi_E}(z,a)\big)$ using the assumption $\mathbb{P}(s_t|z_t,a_t) = \mathbb{P}(s_t|z_t)$ and noticing that $\mathbb{P}(s_t|z_t)$ is policy independent. Finally, we bound this last divergence in terms of $\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big)$ (Lemma 3 in Appendix B). We provide the full derivations in Appendix C. □

Theorem 1 addresses the challenge of considering rewards that depend on actions without the ability to observe expert actions. Consequently, in our setting, we cannot compute C in (3). Similar to the MDP case (Yang et al., 2019), a sufficient condition for $C = 0$ is the injectivity of $\mathbb{P}(z'|z,a)$ in (2) with respect to a, indicating that there is only one action corresponding to a given latent state transition. This property ensures that $\mathbb{P}(a|z,z')$ remains unaffected by different executed policies, ultimately reducing C to zero. For the sake of completeness, we formally state this result in Appendix C. However, in our setting, it is difficult to guarantee the injectivity of $\mathbb{P}(z'|z,a)$ due to its dependence on both the environment through $\mathcal{U}(x'|s')$ and $\mathcal{T}(s'|s,a)$, and the latent variable estimation method through $\mathbb{P}(z'|x',a,z)$ and $\mathbb{P}(s|z)$. Instead, we demonstrate in Theorem 2 how redefining the reward function as $\mathcal{R}: \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, which is commonly observed in robotics learning, allows us to reformulate the result in Theorem 1 without the additive term C in (3).

Theorem 2. *Consider a POMDP, and let $\mathcal{R}: \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and $z_t = \phi(x_{\leq t})$ such that $\mathbb{P}(s_t|z_t, a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Then, the following inequality holds:*

$$\big|J(\pi_E) - J(\pi_{\boldsymbol{\theta}})\big| \leq \frac{2R_{\max}}{1-\gamma} \mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big),$$

where $R_{\max} = \max_{(s,s')\in\mathcal{S}\times\mathcal{S}} |\mathcal{R}(s,s')|$.

Proof. The proof proceeds similarly to the one for Theorem 1, by using that $\mathbb{P}(s,s'|z,z')$ is not characterized by the policy but only by the environment. We show the full proof in Appendix C. □

In summary, Theorems 1 and 2 show that, assuming we have a latent space $\mathcal{Z}$ that can effectively approximate a sufficient statistic of the history, the imitation problem can be performed entirely on this latent space. Note that this is in contrast with the existing literature (Liu et al., 2023), where imitation is performed on the observation space $\mathcal{X}$. As a result of this analysis, our algorithm is characterized by two main ingredients: a practical method to estimate $z \in \mathcal{Z}$ from sequences of observations, and an efficient optimization pipeline to minimize $\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big)$.

5 Latent Adversarial Imitation from Observations

In the following, we introduce the main components of our algorithm LAIfO. Motivated by our theoretical analysis in the previous section, our algorithm combines techniques for adversarial imitation from observations and latent variable estimation. First, we outline our adversarial imitation pipeline in the latent space $\mathcal{Z}$, which leverages off-policy adversarial imitation from observations (Kostrikov et al., 2018; Blondé & Kalousis, 2019; Zhu et al., 2020) in order to minimize the divergence between the latent state-transition visitation distributions of the agent and expert. Then, we describe a simple and effective approach for estimating the latent state z that makes use of observations stacking (Mnih et al., 2013; 2015) and data augmentation (Laskin et al., 2020b; Kostrikov et al., 2020; Yarats et al., 2021). Finally, we show how LAIfO can leverage expert videos to enhance the efficiency of RL from pixels in a number of highly challenging tasks.

Off-policy adversarial imitation from observations Based on the results in Section 4, given a latent variable z that captures a sufficient statistic of the history, we can minimize the suboptimality of the policy $\pi_{\boldsymbol{\theta}}$ by solving the minimization problem

$$\min_{\boldsymbol{\theta}} \quad \mathbb{D}_{\text{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z, z'), \rho_{\pi_E}(z, z')\big). \tag{4}$$

We propose to optimize the objective in (4) using off-policy adversarial IfO. We initialize two replay buffers $\mathcal{B}_E$ and $\mathcal{B}$ to respectively store the sequences of observations generated by the expert and the agent policies, from which we infer the latent state-transitions (z, z'). Note that we write $(z, z') \sim \mathcal{B}$ to streamline the notation. Then, given a discriminator $D_{\boldsymbol{\chi}} : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$, we write

$$\max_{\boldsymbol{\chi}} \quad \mathbb{E}_{(z,z')\sim\mathcal{B}_E}[\log(D_{\boldsymbol{\chi}}(z, z'))] + \mathbb{E}_{(z,z')\sim\mathcal{B}}[\log(1 - D_{\boldsymbol{\chi}}(z, z'))] + g\big(\nabla_{\boldsymbol{\chi}} D_{\boldsymbol{\chi}}\big), \tag{5}$$

where $g(\cdot)$ is defined in (6). As mentioned, alternating the maximization of the loss in (5) with an RL step leads to the minimization of $\mathbb{D}_{\text{JS}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z, z') || \rho_{\pi_E}(z, z')\big)$ (Goodfellow et al., 2020). Since $\mathbb{D}_{\text{JS}}(\cdot||\cdot)$ can be used to upper bound $\mathbb{D}_{\text{TV}}(\cdot, \cdot)$ (cf. Lemma 1 in Appendix B), this approach effectively minimizes the loss in (4). In order to stabilize the adversarial training process, it is important to ensure local Lipschitz-continuity of the learned reward function (Blondé et al., 2022). Therefore, as proposed in Gulrajani et al. (2017), we include in (5) the gradient penalty term

$$g\big(\nabla_{\boldsymbol{\chi}} D_{\boldsymbol{\chi}}\big) = \lambda \mathbb{E}_{(\hat{z},\hat{z}')\sim\mathbb{P}_{(\hat{z},\hat{z}')}}[(||\nabla_{\boldsymbol{\chi}} D_{\boldsymbol{\chi}}(\hat{z}, \hat{z}')||_2 - 1)^2], \tag{6}$$

where λ is a hyperparameter, and $\mathbb{P}_{(\hat{z},\hat{z}')}$ is defined such that $(\hat{z}, \hat{z}')$ are sampled uniformly along straight lines between pairs of transitions respectively sampled from $\mathcal{B}_E$ and $\mathcal{B}$. For additional details on the importance of the term in (6) for improved stability, refer to our ablation experiments in Appendix E and to Gulrajani et al. (2017). Finally, from a theoretical standpoint, note that we should perform importance sampling correction in order to account for the effect of off-policy data when sampling from $\mathcal{B}$ (Queeney et al., 2021; 2022). However, neglecting off-policy correction works well in practice and does not compromise the stability of the algorithm (Kostrikov et al., 2018).

Latent variable estimation from observations Note that the problem in (4) is defined on the latent space $\mathcal{Z}$. Therefore, we now present a simple and effective method to estimate the latent variable z from sequences of observations. Inspired by the model-free RL from pixels literature, we propose to combine the successful approaches of observations stacking (Mnih et al., 2013; 2015) and data augmentation (Laskin et al., 2020b; Kostrikov et al., 2020; Yarats et al., 2021). We stack together the most recent $d \in \mathbb{N}$ observations, and provide this stack as an input to a feature extractor which is trained during the RL step. More specifically, we define a feature extractor $\phi_{\boldsymbol{\delta}} : \mathcal{X}^d \to \mathcal{Z}$ such that $z = \phi_{\boldsymbol{\delta}}(x_{t^-:t})$ where $t - t^- + 1 = d$. When learning from pixels, we also apply data augmentation to the observations stack to improve the quality of the extracted features as in Kostrikov et al. (2020). We write $\text{aug}(x_{t^-:t})$ to define the augmented stack of observations. The latent representations z and z' are then computed respectively as $z = \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t^-:t}))$ and $z' = \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t^-+1:t+1}))$. We train the feature extractor $\phi_{\boldsymbol{\delta}}$ with the critic networks $Q_{\boldsymbol{\psi}_k}$ $(k = 1, 2)$ in order

to minimize the loss function

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\delta},\boldsymbol{\psi}_k}(\mathcal{B}) &= \mathbb{E}_{(z,a,z')\sim\mathcal{B}}[(Q_{\boldsymbol{\psi}_k}(z,a)-y)^2], \\ y &= r_{\boldsymbol{\chi}}(z,z') + \gamma \min_{k=1,2} Q_{\bar{\boldsymbol{\psi}}_k}(z',a'), \end{aligned} \tag{7}$$

$$r_{\boldsymbol{\chi}}(z,z') = D_{\boldsymbol{\chi}}\big(z,z'\big), \tag{8}$$

where $D_{\boldsymbol{\chi}}\big(z,z'\big)$ is the discriminator optimized in (5). In (7), a is an action stored in $\mathcal{B}$ used by the agent to interact with the environment, while $a' = \pi_{\boldsymbol{\theta}}(z') + \epsilon$ where $\epsilon \sim \text{clip}(\mathcal{N}(0,\sigma^2), -c, c)$ is a clipped exploration noise with c the clipping parameter and $\mathcal{N}(0,\sigma^2)$ a univariate normal distribution with zero mean and σ standard deviation. The reward function $r_{\boldsymbol{\chi}}(z,z')$ is defined as in (8), and $\bar{\boldsymbol{\psi}}_1$ and $\bar{\boldsymbol{\psi}}_2$ are the slow moving weights for the target Q networks. We provide more implementation details and the complete pseudo-code for our algorithm in Appendix D.

Note that the feature extractor $\phi_{\boldsymbol{\delta}}$ is shared by both the critics $Q_{\boldsymbol{\psi}_k}$, the policy $\pi_{\boldsymbol{\theta}}$, and the discriminator $D_{\boldsymbol{\chi}}$. However, we stop the backpropagation of the gradient from $\pi_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\chi}}$ into $\phi_{\boldsymbol{\delta}}$. The logic of this choice involves obtaining a latent variable z that is not biased towards any of the players in the adversarial IfO game in (5), but only provides the information necessary to determine the expert and agent expected performance. This design is motivated by our theoretical analysis which shows how, provided $\phi_{\boldsymbol{\delta}} : \mathcal{X}^d \to \mathcal{Z}$ where $\mathcal{Z}$ approximates a sufficient statistic of the history, all the networks in the adversarial IfO game can be directly defined on $\mathcal{Z}$ as $D_{\boldsymbol{\chi}} : \mathcal{Z}\times\mathcal{Z} \to [0,1]$, $Q_{\boldsymbol{\psi}_k} : \mathcal{Z}\times\mathcal{A} \to \mathbb{R}$ and $\pi_{\boldsymbol{\theta}} : \mathcal{Z} \to P(\mathcal{A})$. This is in contrast with the current literature on V-IfO where all the networks are defined on the observation space $\mathcal{X}$ as $D_{\boldsymbol{\chi}} : \mathcal{X}\times\mathcal{X} \to [0,1]$, $Q_{\boldsymbol{\psi}_k} : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ and $\pi_{\boldsymbol{\theta}} : \mathcal{X} \to P(\mathcal{A})$.

Finally, note that latent variable estimation is an active research area, and it is possible to apply other techniques such as variational inference (Lee et al., 2020) and contrastive learning (Chen et al., 2020; Grill et al., 2020). However, we will show in our experiments that the straightforward approach of observations stacking and data augmentation leads to strong performance in practice, without the need for more complicated estimation procedures. We include an ablation study on the importance of data augmentation in Appendix E.

Improving RL from pixels using expert videos We have so far considered the pure imitation setting where a reward function can only be estimated from expert data. However, for many real-world tasks a simple objective can often be provided to the learning agent. Assuming that videos of experts are also available, we show how we can use LAIfO to accelerate the RL learning process.

We combine the standard RL objective with our V-IfO objective in (4), leading to the combined problem

$$\max_{\boldsymbol{\theta}} \quad \mathbb{E}_{\tau_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}(s_t,a_t)\right] - \mathbb{D}_{\text{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big). \tag{9}$$

Using the adversarial IfO pipeline presented in (5), we can rewrite (9) as

$$\max_{\boldsymbol{\theta}} \quad \mathbb{E}_{\tau_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t \Big(\mathcal{R}(s_t,a_t) + r_{\boldsymbol{\chi}}\big(z_t,z_{t+1}\big)\Big)\right], \tag{10}$$

with $r_{\boldsymbol{\chi}}$ in (8). By learning $r_{\boldsymbol{\chi}}$ with LAIfO and optimizing the problem in (10) throughout training, we will show that we are able to significantly improve sample efficiency on challenging humanoid from pixels tasks (Tassa et al., 2018) compared to state-of-the-art RL from pixels algorithms (Yarats et al., 2021).

6 Experiments

In this section, we conduct experiments that aim to answer the following questions:

(1) For the V-IfO problem, how does LAIfO compare to PatchAIL (Liu et al., 2023), a state-of-the-art approach for V-IfO, in terms of asymptotic performance and computational efficiency?

(2) How does the V-IL version of LAIfO with access to expert actions, named *Latent Adversarial Imitation Learning (LAIL)*, compare to VMAIL (Rafailov et al., 2021), a state-of-the-art approach for V-IL?

Table 2: Experimental results for V-IfO (i.e., imitation from experts with partial observability and without access to expert actions). We use DDPG to train experts in a fully observable setting and collect 100 episodes of expert data. All of the expert policies can be downloaded by following the instructions in our code repository. BC is trained offline using expert observation-action pairs for 10^4 gradient steps. All the other algorithms are trained for 3×10^6 frames in walker run, hopper hop, cheetah run, quadruped run, and quadruped walk, and 10^6 frames for the other tasks. We evaluate the learned policies using average performance over 10 episodes. We run each experiment for 6 seeds. In the third, fourth and fifth columns, we report mean and standard deviation of final performance over seeds. In the last column, we report the ratio of wall-clock times between LAIfO and PatchAIL to achieve 75% of expert performance. For each task, we **highlight** the highest asymptotic performance between LAIfO and PatchAIL.

	Expert	BC	LAIfO (our)	PatchAIL-W (Liu et al., 2023)	Wall-clock time ratio to 75% expert performance (LAIfO (our) / PatchAIL-W)
Cup Catch	980	971 ± 9.7	$\mathbf{967 \pm 7.6}$	804 ± 357	0.69
Finger Spin	932	542 ± 219	$\mathbf{926 \pm 10.7}$	885 ± 30.8	0.67
Cartpole Swingup	881	329 ± 25.0	$\mathbf{873 \pm 3.6}$	842 ± 6.7	0.63
Cartpole Balance	990	648 ± 36.4	878 ± 239	$\mathbf{966 \pm 5.5}$	0.61
Pendulum Swingup	845	427 ± 142	786 ± 70.4	$\mathbf{829 \pm 23.7}$	0.90
Walker Walk	960	723 ± 137	$\mathbf{960 \pm 2.2}$	955 ± 7.0	0.15
Walker Stand	980	871 ± 77.7	961 ± 20.0	$\mathbf{971 \pm 10.5}$	0.27
Walker Run	640	133 ± 27.8	$\mathbf{618 \pm 4.6}$	569 ± 53.2	0.22
Hopper Stand	920	398 ± 96.4	800 ± 46.7	$\mathbf{867 \pm 33.9}$	0.16
Hopper Hop	217	45.9 ± 22.1	$\mathbf{206 \pm 8.5}$	191 ± 13.0	0.16
Quadruped Walk	970	337 ± 50.5	$\mathbf{594 \pm 92.9}$	263 ± 242	NA*
Quadruped Run	950	340 ± 75.2	$\mathbf{516 \pm 132}$	471 ± 219	NA*
Cheetah Run	900	106 ± 26.3	$\mathbf{773 \pm 41.2}$	695 ± 312	0.46

* 75% of the expert performance was not achieved by any algorithm.

(3) What is the impact on performance due to partial observability and the absence of expert actions?

(4) Can LAIfO leverage expert videos to improve the efficiency of RL from pixels in high-dimensional continuous robotic tasks?

For more details about the hardware used to carry out these experiments, all the learning curves, additional ablation studies, and other implementation details, refer to Appendix E and to our code.

Visual Imitation from Observations In order to address Question (1), we evaluate LAIfO and PatchAIL (Liu et al., 2023), in its weight regularized version denoted by PatchAIL-W, on 13 different tasks from the DeepMind Control Suite (Tassa et al., 2018). We also compare these algorithms to Behavioral Cloning (BC) (Pomerleau, 1988) for reference. Note that BC uses observation-action pairs. The results are summarized in Table 2, Figure 1, and Figure 2. Table 2 includes the asymptotic performance of each algorithm, as well as the ratio of wall-clock times between LAIfO and PatchAIL to achieve 75% of expert performance. Figure 1 depicts the average return per episode throughout training as a function of wall-clock time. Moreover, we include in Figure 2 plots showing the average return per episode as a function of training steps. These results demonstrate that LAIfO can successfully solve the V-IfO problem, achieving asymptotic performance comparable to the state-of-the-art baseline PatchAIL. Importantly, *LAIfO is significantly more computationally efficient than PatchAIL*. This is well highlighted both in Table 2 and in Figure 1, where we show that *LAIfO always converges faster than PatchAIL in terms of wall-clock time*. This improved computational efficiency is the result of performing imitation on the latent space $\mathcal{Z}$, instead of directly on the high-dimensional observation space $\mathcal{X}$ (i.e., pixel space) as in PatchAIL. Finally, in Table 4 we examine the impact of the amount of expert data on performance. Throughout these experiments, LAIfO does not exhibit a tangible drop in performance due to the decrease of available expert data, and it consistently outperforms PatchAIL.

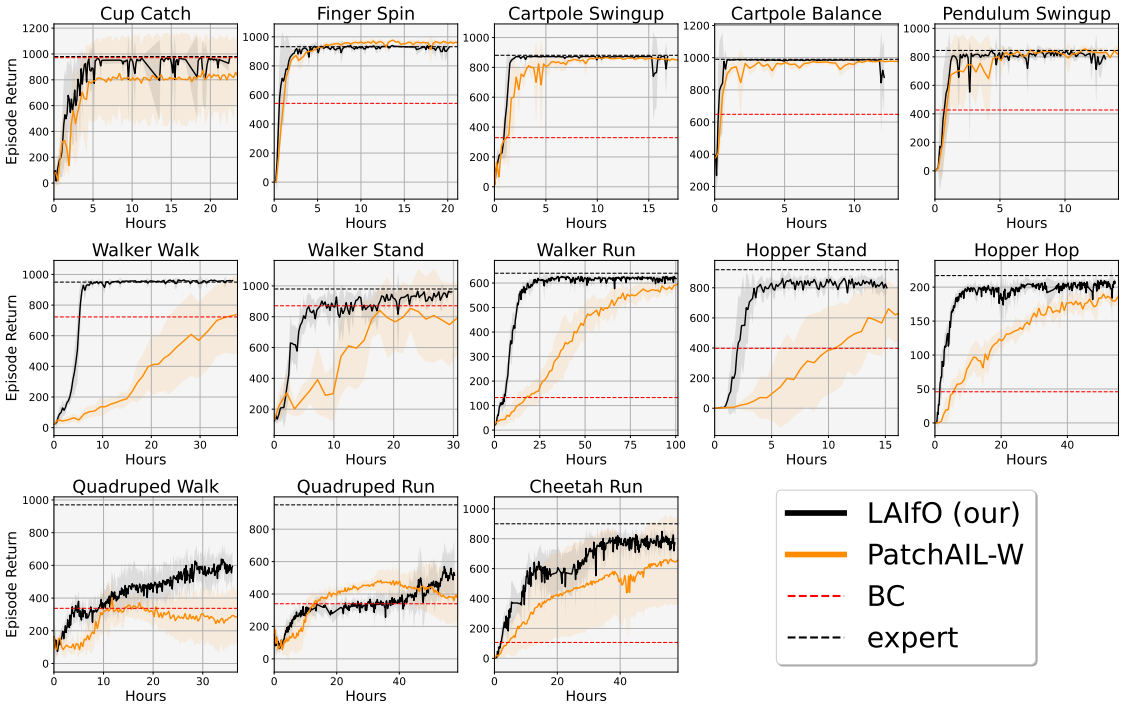

Figure 1: Learning curves for the V-IfO results in Table 2. Plots show the average return per episode as a function of wall-clock time. Our algorithm LAIfO achieves state-of-the-art asymptotic performance, and significantly reduces computation time compared to PatchAIL.

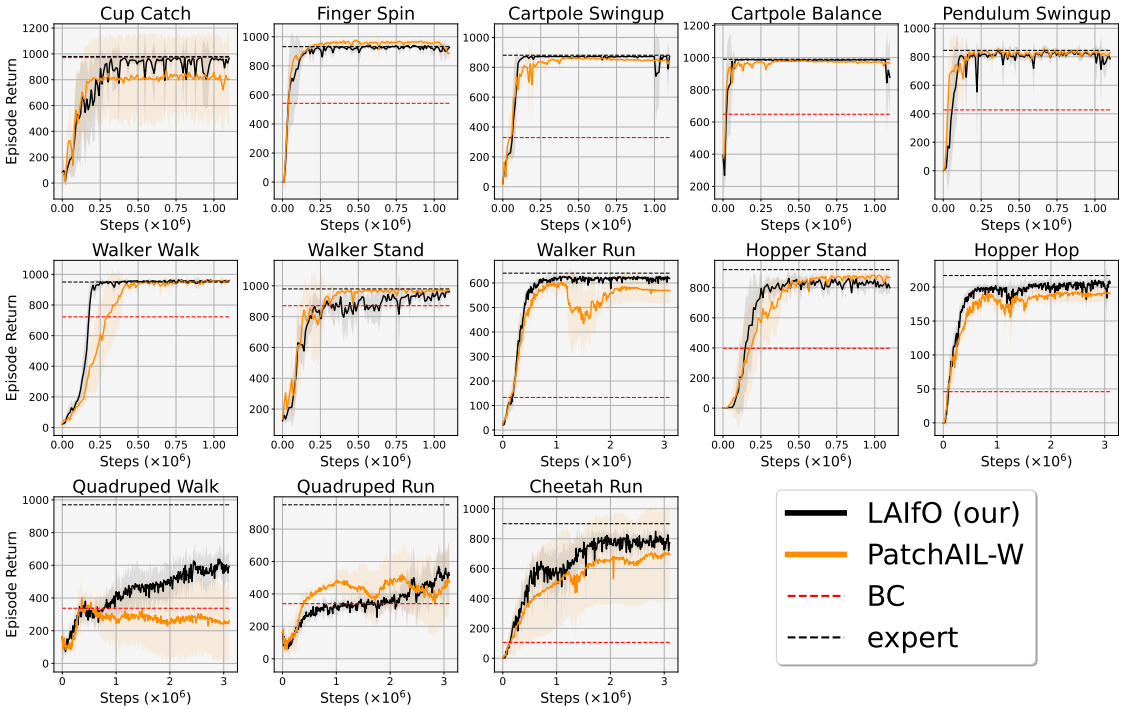

Figure 2: Learning curves for the V-IfO results in Table 2. Plots show the average return per episode as a function of training steps.

Table 3: Experimental results for V-IL (i.e., imitation from experts with partial observability and access to expert actions). We use DDPG to train experts in a fully observable setting and collect 100 episodes of expert data. The experiments are conducted as in Table 2. In the third, fourth and fifth columns, we report mean and standard deviation of final performance over seeds. In the last column, we report the ratio of wall-clock times between the two algorithms to achieve 75% of expert performance. For each task, we **highlight** the highest asymptotic performance between LAIL and VMAIL.

	Expert	BC	LAIL (our)	VMAIL (Rafailov et al., 2021)	Wall-clock time ratio to 75% expert performance (LAIL (our) / VMAIL)
Cup Catch	980	971 ± 9.7	$\mathbf{962 \pm 18.0}$	939 ± 37.4	0.31
Finger Spin	932	542 ± 219	$\mathbf{775 \pm 345}$	476 ± 352	0.21
Cartpole Swingup	881	329 ± 25.0	$\mathbf{873 \pm 3.0}$	512 ± 291	0.13
Cartpole Balance	990	648 ± 36.3	$\mathbf{982 \pm 1.6}$	868 ± 104	0.10
Pendulum Swingup	845	427 ± 142	$\mathbf{825 \pm 45.1}$	723 ± 143	1.43
Walker Walk	960	723 ± 137	$\mathbf{946 \pm 8.5}$	939 ± 9.8	0.40
Walker Stand	980	871 ± 77.7	$\mathbf{893 \pm 106}$	805 ± 309	0.82
Walker Run	640	133 ± 27.8	$\mathbf{625 \pm 5.1}$	516 ± 224	0.58
Hopper Stand	920	398 ± 96.4	$\mathbf{764 \pm 111}$	567 ± 285	0.12
Hopper Hop	217	45.9 ± 22.1	$\mathbf{208 \pm 3.1}$	72.3 ± 73.0	0.23
Quadruped Walk	970	337 ± 50.5	$\mathbf{500 \pm 182}$	223 ± 51.9	NA*
Quadruped Run	950	340 ± 75.2	$\mathbf{697 \pm 102}$	127 ± 66.0	NA**
Cheetah Run	900	106 ± 26.3	$\mathbf{811 \pm 67.9}$	539 ± 367	0.83

* 75% of the expert performance was not achieved by any algorithm.
** 75% of the expert performance was not achieved by VMAIL.

Visual Imitation Learning To answer Question (2), we test LAIL, the V-IL version of LAIfO, and VMAIL (Rafailov et al., 2021) using the same experimental setup that we considered in the V-IfO setting. As for V-IfO, we also compare these algorithms to BC for reference. VMAIL stands for *Variational Model Adversarial Imitation Learning*, and represents a model-based version of generative adversarial IL built upon the variational models presented in Hafner et al. (2019b;a; 2020). LAIL is obtained by simply defining the discriminator in (5) as $D_{\boldsymbol{\chi}} : \mathcal{Z} \times \mathcal{A} \to [0, 1]$ rather than $D_{\boldsymbol{\chi}} : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ as in LAIfO. The results for these experiments are summarized in Table 3, Figure 3, and Figure 4. Compared to VMAIL, we see that *LAIL achieves better asymptotic performance and better computational efficiency.* While both algorithms perform imitation on a latent space $\mathcal{Z}$, LAIL is a *model-free* algorithm that requires a lower number of learnable parameters compared to the model-based VMAIL. VMAIL must learn an accurate world model during training, which can be a challenging and computationally demanding task. The model learning process contributes to higher wall-clock times, and can also lead to instability in the training process for some environments (cf. Figure 4). On the other hand, the model-free approach of LAIL results in stable training that yields faster convergence and better efficiency. Finally, in Table 4 we examine the impact of the amount of expert data on the final results. Throughout these experiments, both LAIL and VMAIL exhibit reliable results and no tangible decrease in performance is observed due to relying on less expert data.

Ablation study In order to answer Question (3), we compare performance for each type of imitation from experts in Table 1. For the partially observable setting, we consider our algorithms LAIL and LAIfO. For the fully observable setting, we consider DAC (Kostrikov et al., 2018) and our implementation of *DAC from Observations (DACfO)*. We provide the full learning curves for DAC and DACfO in Appendix E (cf. Table 6 and Figure 7). The results are summarized in Figure 5, which shows the average normalized return obtained by each algorithm throughout the different tasks in Table 2. These experiments highlight how our algorithms can successfully address the absence of expert actions and partial observability, suffering only marginal performance degradation due to these additional challenges. As explained in our theoretical analysis in Section 4, partial observability is addressed by estimating a latent state representation that successfully approximates a sufficient statistic of the history. On the other hand, marginal degradation due to the absence of expert actions occurs either because we are in the context described by Theorem 2, where the environment reward function does not depend on actions, or because C in Theorem 1 becomes negligible.

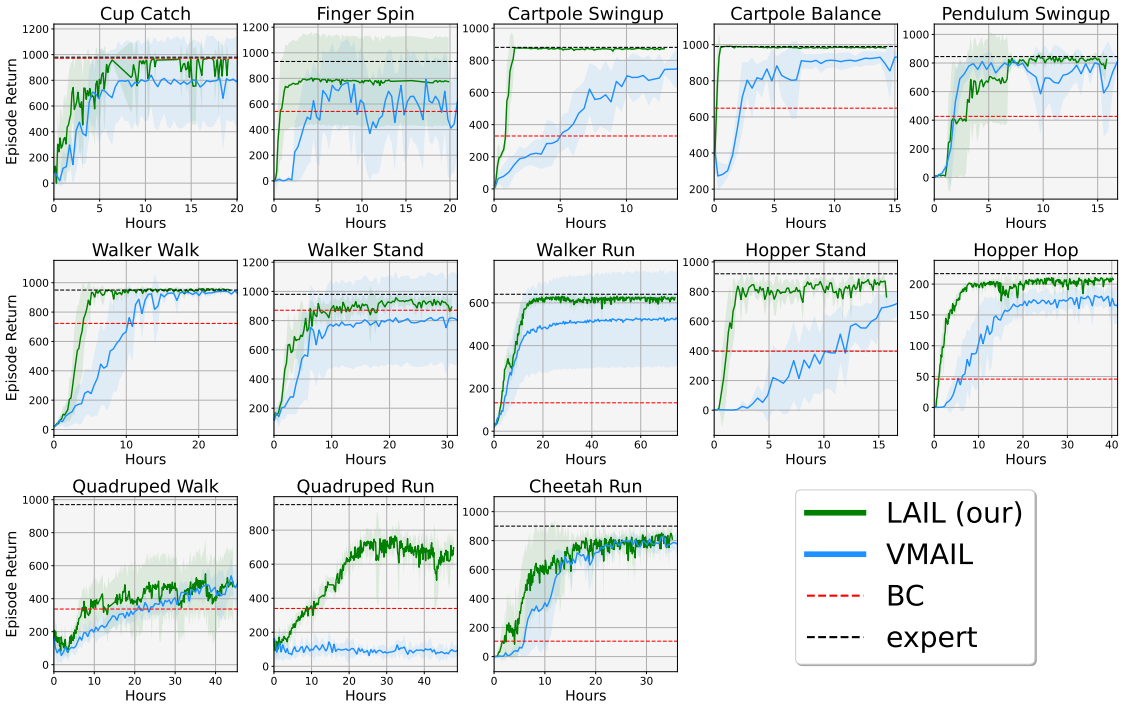

Figure 3: Learning curves for the V-IL results in Table 3. Plots show the average return per episode as a function of wall-clock time. LAIL outperforms VMAIL in terms of both asymptotic performance and computational efficiency.

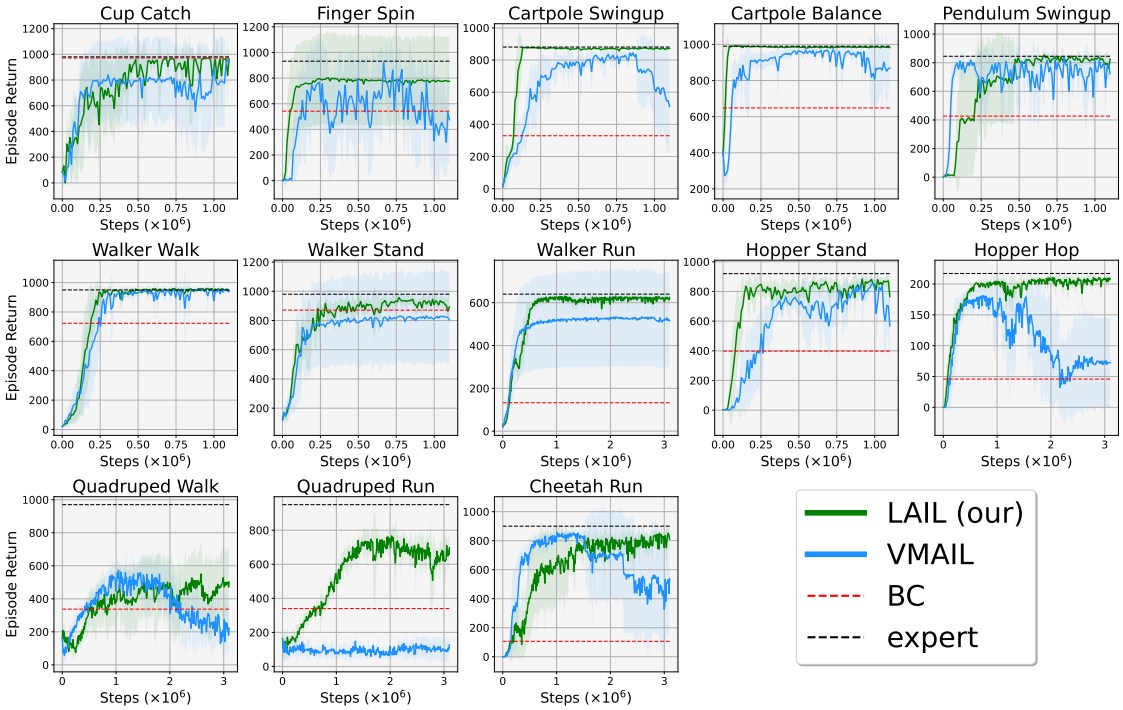

Figure 4: Learning curves for the V-IL results in Table 3. Plots show the average return per episode as a function of training steps.

Table 4: Performance for different numbers of expert episodes on Walker Run. Except for the number of expert episodes, the experiments are conducted as in Table 2 and Table 3. We report mean and standard deviation after training for 10^6 frames. For both the V-IL and the V-IfO settings, we **highlight** the highest performance.

Walker Run	n Expert episodes				
	1	10	20	50	100
BC	28.6 ± 1.9	70.7 ± 1.8	87.0 ± 1.8	113 ± 1.8	133 ± 1.8
VMAIL (Rafailov et al., 2021)	520 ± 224	$\mathbf{630 \pm 21.9}$	517 ± 221	$\mathbf{619 \pm 11.8}$	524 ± 222
LAIL (our)	$\mathbf{614 \pm 7.6}$	626 ± 7.3	$\mathbf{627 \pm 6.8}$	611 ± 22.5	$\mathbf{613 \pm 20.3}$
PatchAIL-W (Liu et al., 2023)	494 ± 199	561 ± 36.0	586 ± 39.7	604 ± 22.5	598 ± 34.7
LAIfO (our)	$\mathbf{612 \pm 22.8}$	$\mathbf{620 \pm 6.4}$	$\mathbf{623 \pm 7.6}$	$\mathbf{618 \pm 10.1}$	$\mathbf{621 \pm 4.8}$

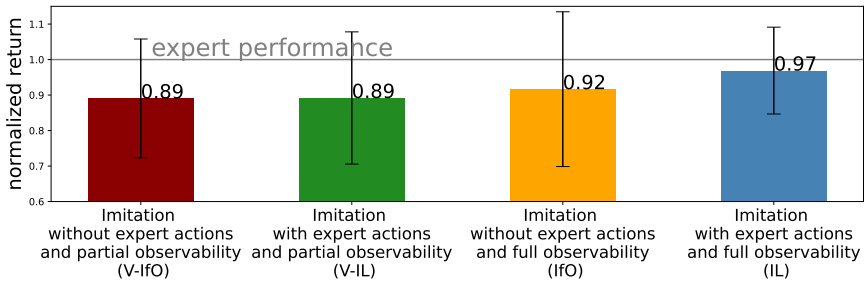

Figure 5: Normalized returns obtained by each type of imitation from experts over the tasks in Table 2. For each task and random seed, normalized returns represent performance divided by the expert performance for the considered task (Expert column in Table 2 and Table 3). For each type of imitation from experts, we plot mean and standard deviation over the full set of runs. The performance of our algorithms in the partially observable setting are comparable to the performance in the fully observable setting, and the absence of expert actions and partial observability leads only to marginal performance degradation.

Improving RL using expert videos We answer Question (4) by applying LAIfO to the problem in (9) for the humanoid from pixels environment. We consider the state-of-the-art model-free RL from pixels algorithms, DrQv2 (Yarats et al., 2021) as a baseline. The results are illustrated in Figure 6. By leveraging expert videos, we see that our algorithm significantly outperforms the baseline in terms of sample efficiency. This result highlights the value of leveraging expert videos to improve the sample efficiency of RL algorithms, in particular when dealing with challenging tasks.

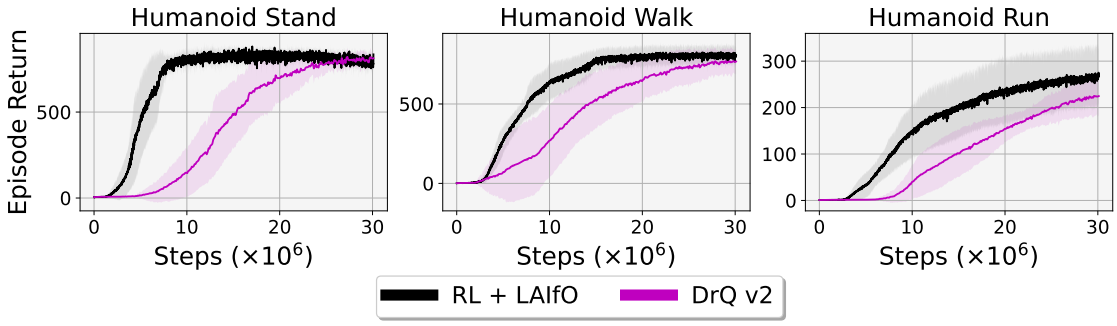

Figure 6: Performance using the multi-objective RL framework in (9) on the humanoid environment. The experiments are designed as in Table 2. We report mean and standard error over seeds. For DrQv2, note that we use as comparison the curves provided in the official Github repository of the paper.

7 Conclusion

In this work, we formally analyzed the V-IfO problem and introduced our algorithm LAIfO as an effective solution. We experimentally showed that our approach matches the performance of state-of-the-art V-IL and V-IfO methods, while requiring significantly less computational effort due to our model-free approach in latent space. Furthermore, we showed how LAIfO can be used to improve the efficiency and asymptotic performance of RL methods by leveraging expert videos.

Limitations and future work Despite the advancement in addressing the V-IfO problem, it is important to understand the limitations of our approach. The primary limitation arises from the assumption that the expert and the agent act within the same POMDP. In realistic scenarios, such alignment rarely occurs, emphasizing the need for methods that can handle dynamics mismatch and visual domain adaptation. This is a crucial next step towards enabling successful learning from expert videos. Furthermore, throughout this work we have used adversarial learning for divergence minimization between distributions. Adversarial learning can introduce optimization challenges and stability issues. While we propose practical solutions to mitigate these problems, exploring alternatives to this framework offers another interesting avenue for future research. Additionally, from an experimental standpoint, our emphasis has been on robotics control tasks. In the future, we plan to address navigation tasks, considering not only third-view perspectives but also egocentric camera viewpoints. In this context, a challenging and relevant consideration is the correspondence problem, i.e., the problem of enabling egocentric policy learning directly from third-view videos of experts. Finally, we are interested in testing our algorithms on more realistic scenarios that go beyond simulated environments. Our long-term goal is to provide solutions for real-world problems, such as vehicle navigation and robotic manipulation, and to enable imitation directly from videos of biological systems such as humans and animals.

8 Acknowledgements

Vittorio Giammarino and Ioannis Ch. Paschalidis were partially supported by the NSF under grants IIS-1914792, CCF-2200052, and ECCS-2317079, by the ONR under grant N00014-19-1-2571, by the DOE under grant DE-AC02-05CH11231, by the NIH under grant UL54 TR004130, and by Boston University. James Queeney was exclusively supported by Mitsubishi Electric Research Laboratories.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first International Conference on Machine Learning*, pp. 1, 2004.

Karl J. Astrom. Optimal control of markov decision processes with incomplete state estimation. *J. Math. Anal. Applic.*, 10:174–205, 1965.

Christopher G. Atkeson and Stefan Schaal. Robot learning from demonstration. In *Proceedings of the fourteenth International Conference on Machine Learning*, pp. 12–20, 1997.

Feryal Behbahani, Kyriacos Shiarlis, Xi Chen, Vitaly Kurin, Sudhanshu Kasewa, Ciprian Stirbu, Joao Gomes, Supratik Paul, Frans A. Oliehoek, Joao Messias, et al. Learning from demonstration in the wild. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 775–781. IEEE, 2019.

Lionel Blondé and Alexandros Kalousis. Sample-efficient imitation learning via generative adversarial nets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3138–3148. PMLR, 2019.

Lionel Blondé, Pablo Strasser, and Alexandros Kalousis. Lipschitzness is all you need to tame off-policy generative adversarial imitation learning. *Machine Learning*, 111(4):1431–1521, 2022.

Edoardo Cetin and Oya Celiktutan. Domain-robust visual imitation learning with mutual information constraints. *arXiv preprint arXiv:2103.05079*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Zhihao Cheng, Liu Liu, Aishan Liu, Hao Sun, Meng Fang, and Dacheng Tao. On the guaranteed almost equivalence between imitation learning from observation and demonstration. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. *arXiv preprint arXiv:2006.04678*, 2020.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng. Learning belief representations for imitation learning in POMDPs. In *Uncertainty in Artificial Intelligence*, pp. 1061–1071. PMLR, 2020.

Tanmay Gangwani, Yuan Zhou, and Jian Peng. Imitation learning from observations under transition model disparity. *arXiv preprint arXiv:2204.11446*, 2022.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Proceedings of the Conference on Robot Learning*, pp. 1259–1277. PMLR, 2020.

Vittorio Giammarino, Matthew F Dunne, Kylie N Moore, Michael E Hasselmo, Chantal E Stern, and Ioannis Ch Paschalidis. Combining imitation and deep reinforcement learning to human-level performance on a virtual foraging task. *Adaptive Behavior*, 2023a.

Vittorio Giammarino, James Queeney, Lucas C. Carstensen, Michael E. Hasselmo, and Ioannis Ch. Paschalidis. Opportunities and challenges from using animal videos in reinforcement learning for navigation. *IFAC-PapersOnLine*, 56(2):9056–9061, 2023b. 22nd IFAC World Congress.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30, 2017.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the thirty-sixth International Conference on Machine Learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zak Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. *arXiv preprint arXiv:2210.07729*, 2022.

Haresh Karnan, Faraz Torabi, Garrett Warnell, and Peter Stone. Adversarial imitation learning from video using a state observer. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2452–2458. IEEE, 2022.

Rahul Kidambi, Jonathan Chang, and Wen Sun. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34:28598–28611, 2021.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *Proceedings of the thirty-seventh International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020a.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems*, 33:19884–19895, 2020b.

Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems*, 33:741–752, 2020.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Minghuan Liu, Tairan He, Weinan Zhang, Shuicheng Yan, and Zhongwen Xu. Visual imitation learning with patch rewards. *arXiv preprint arXiv:2302.00965*, 2023.

YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1118–1125. IEEE, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Andrew Y. Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth International Conference on Machine Learning*, volume 1, pp. 2, 2000.

Ryo Okumura, Masashi Okada, and Tadahiro Taniguchi. Domain-adversarial and-conditional state space model for imitation learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5179–5186. IEEE, 2020.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1, 1988.

James Queeney, Ioannis Ch. Paschalidis, and Christos G. Cassandras. Generalized proximal policy optimization with sample reuse. *Advances in Neural Information Processing Systems*, 34:11909–11919, 2021.

James Queeney, Ioannis Ch. Paschalidis, and Christos G. Cassandras. Generalized policy improvement algorithms with theoretically supported sample reuse. *arXiv preprint arXiv:2206.13714*, 2022.

Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Visual adversarial imitation learning using variational models. *Advances in Neural Information Processing Systems*, 34:3016–3028, 2021.

Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:5402–5415, 2021.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*, 2019.

Alexander Reske, Jan Carius, Yuntao Ma, Farbod Farshidian, and Marco Hutter. Imitation learning from mpc for quadrupedal multi-gait control. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5014–5020. IEEE, 2021.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh Annual Conference on Computational Learning Theory*, pp. 101–103, 1998.

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

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Proceedings of The sixth Conference on Robot Learning*, pp. 654–665. PMLR, 2023.

Bradly C. Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017.

Umar Syed and Robert E. Schapire. A game-theoretic approach to apprenticeship learning. *Advances in Neural Information Processing Systems*, 20, 2007.

Umar Syed, Michael Bowling, and Robert E. Schapire. Apprenticeship learning using linear programming. In *Proceedings of the twenty-fifth International Conference on Machine Learning*, pp. 1032–1039, 2008.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018a.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018b.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *Advances in Neural Information Processing Systems*, 32, 2019.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

Jimuyang Zhang and Eshed Ohn-Bar. Learning by watching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12711–12721, 2021.

Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. Selfd: self-learning large-scale driving policies from the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17316–17326, 2022.

Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in Neural Information Processing Systems*, 33:12402–12413, 2020.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, Anind K. Dey, et al. Maximum entropy inverse reinforcement learning. In *Twenty-Third AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Broader impacts

Learning from expert videos will have a transformative impact on robotics, revolutionizing the way robots acquire skills, collaborate with humans, and operate in various domains. This technology holds the potential to enhance efficiency and accessibility across industries, while also fostering human-robot collaboration.

Despite the numerous positive implications, it is important to consider potential negative aspects that may arise from this technology. As in all data-driven methods, biases in the data should always be a cause for concern. Expert observations may inadvertently contain biases, reflecting the preferences or limitations of the demonstrators. If these biases are not properly addressed, robots could potentially perpetuate and amplify those biases in their actions, leading to unsafe or unfair behaviors.

We believe it is crucial to address these potential negative implications through careful design, robust validation, and ongoing research. Responsible development and deployment of machine learning in robotics should consider ethical considerations, promote fairness and transparency, and ensure the technology benefits society as a whole.

B Auxiliary results

In the following, we introduce a series of auxiliary definitions and results which are then used to prove Theorems 1 and 2.

Definition 1 (f-divergence)**.** *Let P and Q be two probability distributions over a measurable space $(\Omega, \mathcal{F})$, such that P is absolutely continuous with respect to Q. Then, for a convex function $f : [0, \infty) \to (-\infty, \infty]$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t\to 0^+} f(t)$, the f-divergence is defined as*

$$\mathbb{D}_f(P||Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right], \tag{11}$$

where $P = \frac{dP}{d\mu}$ and $Q = \frac{dQ}{d\mu}$ with μ any dominating probability measure (Polyanskiy & Wu, 2022).

From Definition 1, the following f-divergences can be derived:

- Total variation distance: $f(x) = \frac{1}{2}|x - 1|$,

$$\mathbb{D}_{\mathrm{TV}}(P, Q) \triangleq \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{P}{Q} - 1\right|\right] = \frac{1}{2}\int |P - Q| d\mu. \tag{12}$$

- Jensen-Shannon divergence: $f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$,

$$\mathbb{D}_{\mathrm{JS}}(P||Q) \triangleq \mathbb{D}_{\mathrm{KL}}\Big(P\Big|\Big|\frac{P+Q}{2}\Big) + \mathbb{D}_{\mathrm{KL}}\Big(Q\Big|\Big|\frac{P+Q}{2}\Big). \tag{13}$$

Lemma 1 (Relation between f-divergences)**.** *The following inequality holds:*

$$\mathbb{D}_{\mathrm{TV}}(P, Q) \leq \sqrt{\mathbb{D}_{\mathrm{JS}}(P||Q)}. \tag{14}$$

Proof. For (14), let $M = \frac{P+Q}{2}$. Note that

$$\begin{aligned} 2\mathbb{D}_{\mathrm{TV}}(P, M) &= \int |P - M| d\mu \\ &= \int \Big|P - \Big(\frac{P+Q}{2}\Big)\Big| d\mu \\ &= \frac{1}{2}\int |P - Q| d\mu \\ &= \mathbb{D}_{\mathrm{TV}}(P, Q). \end{aligned}$$

Similarly, $2\mathbb{D}_{\mathrm{TV}}(Q, M) = \mathbb{D}_{\mathrm{TV}}(P, Q)$. Therefore,

$$\begin{aligned}(\mathbb{D}_{\mathrm{TV}}(P,Q))^2 =& \frac{1}{2}(\mathbb{D}_{\mathrm{TV}}(P,Q))^2 + \frac{1}{2}(\mathbb{D}_{\mathrm{TV}}(P,Q))^2 \\ =& 2(\mathbb{D}_{\mathrm{TV}}(P,M))^2 + 2(\mathbb{D}_{\mathrm{TV}}(Q,M))^2 \\ \leq& \mathbb{D}_{\mathrm{KL}}(P||M) + \mathbb{D}_{\mathrm{KL}}(Q||M) \qquad (15)\\ =& \mathbb{D}_{\mathrm{JS}}(P||Q), \qquad (16)\end{aligned}$$

where in (15) we used Pinsker's inequality and in (16) the definition of Jensen-Shannon divergence from (13). The result follows by taking the square root of both sides. □

Lemma 2. *Consider a POMDP, under the assumption that the agent $\pi_{\boldsymbol{\theta}}$ and the expert π_E act on the same POMDP and with $\mathbb{P}(z'|z,a)$ defined in (2). Then, the following holds:*

$$\mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z'), \rho_{\pi_E}(z,a,z')) = \mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,a), \rho_{\pi_E}(z,a)). \tag{17}$$

Proof. We have that

$$\begin{aligned}\mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z'), \rho_{\pi_E}(z,a,z')) =& \mathbb{E}_{(z,a,z')\sim\rho_{\pi_E}(z,a,z')}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z')}{\rho_{\pi_E}(z,a,z')}\right)\right] \\ =& \mathbb{E}_{(z,a,z')\sim\rho_{\pi_E}(z,a)\mathbb{P}(z'|z,a)}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(z,a)\mathbb{P}(z'|z,a)}{\rho_{\pi_E}(z,a)\mathbb{P}(z'|z,a)}\right)\right] \qquad (18)\\ =& \mathbb{E}_{(z,a)\sim\rho_{\pi_E}(z,a)}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(z,a)}{\rho_{\pi_E}(z,a)}\right)\right] \\ =& \mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,a), \rho_{\pi_E}(z,a)).\end{aligned}$$

This result was similarly proved for fully observable MDPs in Yang et al. (2019). We generalize it for the POMDP case using the definition of $\mathbb{P}(z'|z,a)$ in (2), which does not depend on any policy but only on the environment. By assumption, expert and agent act on the same POMDP and this yields the step in (18) from which the proof follows. □

Lemma 3. *Consider a POMDP, under the assumption that the agent $\pi_{\boldsymbol{\theta}}$ and the expert π_E act on the same POMDP. Then, the following holds:*

$$\begin{aligned}\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,a), \rho_{\pi_E}(z,a)) \leq\ & \mathbb{E}_{(z,z')\sim\rho_{\pi_{\boldsymbol{\theta}}}(z,z')}\Big[\mathbb{D}_{\mathrm{TV}}\big(\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z'), \mathbb{P}_{\pi_E}(a|z,z')\big)\Big] \\ &+ \mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')).\end{aligned} \tag{19}$$

Proof. From Lemma 2, we can write $\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,a), \rho_{\pi_E}(z,a)) = \mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z'), \rho_{\pi_E}(z,a,z'))$. Then,

$$\begin{aligned}&\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z'), \rho_{\pi_E}(z,a,z')) \\ &= \frac{1}{2}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{A}} \big|\rho_{\pi_{\boldsymbol{\theta}}}(z,a,z') - \rho_{\pi_E}(z,a,z')\big| da dz' dz \\ &= \frac{1}{2}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{A}} \big|\rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') - \rho_{\pi_E}(z,z')\mathbb{P}_{\pi_E}(a|z,z')\big| da dz' dz \\ &= \frac{1}{2}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{A}} \big|\rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') - \rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_E}(a|z,z') \\ &\qquad + \rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_E}(a|z,z') - \rho_{\pi_E}(z,z')\mathbb{P}_{\pi_E}(a|z,z')\big| da dz' dz \\ &\leq \frac{1}{2}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{A}} \big|\rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') - \rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_E}(a|z,z')\big| da dz' dz \\ &\quad + \frac{1}{2}\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{A}} \big|\rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}_{\pi_E}(a|z,z') - \rho_{\pi_E}(z,z')\mathbb{P}_{\pi_E}(a|z,z')\big| da dz' dz \\ &= \mathbb{E}_{(z,z')\sim\rho_{\pi_{\boldsymbol{\theta}}}(z,z')}\Big[\mathbb{D}_{\mathrm{TV}}\big(\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z'), \mathbb{P}_{\pi_E}(a|z,z')\big)\Big] + \mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')),\end{aligned} \tag{20}$$

where (20) follows from the triangle inequality. □

Lemma 4. *Consider a POMDP and let z_t be a latent state representation such that $\mathbb{P}(s_t|z_t, a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Then, the filtering posterior distributions $\mathbb{P}(s_t|z_t)$ and $\mathbb{P}(s_{t+1}, s_t|z_{t+1}, z_t)$ do not depend on the policy π but only on the environment.*

Proof. The proof follows by applying Bayes rule and considering the definition of the latent variable z_t. We start from $\mathbb{P}(s_t|z_t)$:

$$\begin{aligned}\mathbb{P}(s_t|z_t) =& \mathbb{P}(s_t|x_t, a_{t-1}, z_{t-1}) \\ =& \frac{\mathbb{P}(x_t|s_t, a_{t-1}, z_{t-1})\mathbb{P}(s_t|a_{t-1}, z_{t-1})}{\mathbb{P}(x_t|a_{t-1}, z_{t-1})} \\ =& \frac{\mathcal{U}(x_t|s_t)\int_{\mathcal{S}}\mathcal{T}(s_t|s_{t-1}, a_{t-1})\mathbb{P}(s_{t-1}|z_{t-1})ds_{t-1}}{\int_{\mathcal{S}}\int_{\mathcal{S}}\mathcal{U}(x_t|s_t)\mathcal{T}(s_t|s_{t-1}, a_{t-1})\mathbb{P}(s_{t-1}|z_{t-1})ds_t ds_{t-1}},\end{aligned}$$

where the denominator can be seen as a normalizing factor. Note that there is no dependence on the policy π. Similarly, for $\mathbb{P}(s_{t+1}, s_t|z_t, z_{t+1})$ we have that

$$\begin{aligned}\mathbb{P}(s_{t+1}, s_t|z_t, z_{t+1}) =& \mathbb{P}(s_t, s_{t+1}|x_{t+1}, a_t, z_t) \\ =& \frac{\mathbb{P}(x_{t+1}|s_t, s_{t+1}, a_t, z_t)\mathbb{P}(s_t, s_{t+1}|a_t, z_t)}{\mathbb{P}(x_{t+1}|a_t, z_t)} \\ =& \frac{\mathcal{U}(x_{t+1}|s_{t+1})\mathcal{T}(s_{t+1}|s_t, a_t)\mathbb{P}(s_t|z_t)}{\int_{\mathcal{S}}\int_{\mathcal{S}}\mathcal{U}(x_{t+1}|s_{t+1})\mathcal{T}(s_{t+1}|s_t, a_t)\mathbb{P}(s_t|z_t)ds_{t+1}ds_t},\end{aligned}$$

which does not depend on π since $\mathbb{P}(s_t|z_t)$ does not depend on π. □

Remark 1 (Remark on the assumptions in Lemma 4)**.** The assumption $\mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$ in Lemma 4 is crucial to prove the independence of $\mathbb{P}(s_t|z_t)$ on the policy π. Note that by only considering $\mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t})$, we have

$$\begin{aligned}\mathbb{P}(s_t|z_t) =& \mathbb{P}(s_t|x_t, z_{t-1}) \\ =& \frac{\mathbb{P}(x_t|s_t, z_{t-1})\mathbb{P}(s_t|z_{t-1})}{\mathbb{P}(x_t|z_{t-1})} \\ =& \frac{\mathcal{U}(x_t|s_t)\int_{\mathcal{A}}\int_{\mathcal{S}}\mathcal{T}(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|z_{t-1})\mathbb{P}(s_{t-1}|z_{t-1})ds_{t-1}da_{t-1}}{\int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{S}}\mathcal{U}(x_t|s_t)\mathcal{T}(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|z_{t-1})\mathbb{P}(s_{t-1}|z_{t-1})ds_{t-1}da_{t-1}ds_t},\end{aligned}$$

so $\mathbb{P}(s_t|z_t)$ depends on the policy π.

The posteriors $\mathbb{P}_\pi(z_t|s_t, a_t)$ and $\mathbb{P}_\pi(z_{t+1}, z_t|s_{t+1}, s_t)$ can be expressed using Bayes rule as

$$\begin{aligned}\mathbb{P}_\pi(z|s, a) &= \frac{\pi(a|z)\mathbb{P}(s|z)d_\pi(z)}{\rho_\pi(s, a)}, \\ \mathbb{P}_\pi(z, z'|s, s') &= \frac{\mathbb{P}(s, s'|z, z')\rho_\pi(z, z')}{\rho_\pi(s, s')}.\end{aligned}$$

Without any additional assumptions, we cannot express these posteriors without an explicit dependence on the policy π and, therefore, we cannot claim that they are policy independent.

Theorem 3 (From Rafailov et al. 2021)**.** *Consider a POMDP and let z_t be a latent state representation such that $\mathbb{P}(s_t|z_t, a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Given $\mathbb{D}_f$ a generic f-divergence and under the assumption that the agent $\pi_{\boldsymbol{\theta}}$ and the expert π_E share the same POMDP, the following inequality holds:*

$$\mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(s, a), \rho_{\pi_E}(s, a)) \leq \mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z, a), \rho_{\pi_E}(z, a)). \tag{21}$$

Proof. We have that

$$
\begin{aligned}
\mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(s,a),\rho_{\pi_E}(s,a)) &= \mathbb{E}_{(s,a)\sim\rho_{\pi_E}(s,a)}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,a)}{\rho_{\pi_E}(s,a)}\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\rho_{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{z\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(z|s,a)}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,a)}{\rho_{\pi_E}(s,a)}\right]\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\rho_{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{z\sim\mathbb{P}_{\pi_E}(z|s,a)}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,a)}{\rho_{\pi_E}(s,a)}\frac{\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(z|s,a)}{\mathbb{P}_{\pi_E}(z|s,a)}\right]\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\rho_{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{z\sim\mathbb{P}_{\pi_E}(z|s,a)}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,a,z)}{\rho_{\pi_E}(s,a,z)}\right]\right)\right] \\
&\leq \mathbb{E}_{(s,a)\sim\rho_{\pi_E}(s,a)}\mathbb{E}_{z\sim\mathbb{P}_{\pi_E}(z|s,a)}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,a,z)}{\rho_{\pi_E}(s,a,z)}\right)\right] \qquad (22) \\
&= \mathbb{E}_{(z,a)\sim\rho_{\pi_E}(z,a)}\mathbb{E}_{s\sim\mathbb{P}(s|z)}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(z,a)\mathbb{P}(s|z)}{\rho_{\pi_E}(z,a)\mathbb{P}(s|z)}\right)\right] \qquad (23) \\
&= \mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,a),\rho_{\pi_E}(z,a)).
\end{aligned}
$$

This proof follows by applying Jensen's inequality in (22) and by noticing that $\mathbb{P}(s|z)$ in (23) does not depend on the policy but exclusively on the environment (Lemma 4). □

C Theoretical analysis proofs

Theorem 1 restated: Consider a POMDP, and let $\mathcal{R}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and $z_t=\phi(x_{\leq t})$ such that $\mathbb{P}(s_t|z_t,a_t)=\mathbb{P}(s_t|z_t)=\mathbb{P}(s_t|x_{\leq t},a_{<t})$. Then, the following inequality holds:

$$
\big|J(\pi_E)-J(\pi_{\boldsymbol{\theta}})\big| \leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'),\rho_{\pi_E}(z,z')\big)+C,
$$

where $R_{\max}=\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}|\mathcal{R}(s,a)|$ and

$$
C=\frac{2R_{\max}}{1-\gamma}\mathbb{E}_{\rho_{\pi_{\boldsymbol{\theta}}}(z,z')}\big[\mathbb{D}_{\mathrm{TV}}\big(\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z'),\mathbb{P}_{\pi_E}(a|z,z')\big)\big]. \qquad (24)
$$

Proof. Note that $(1-\gamma)\cdot J(\pi)=\mathbb{E}_{(s,a)\sim\rho_\pi(s,a)}[\mathcal{R}(s,a)]$. Then,

$$
\begin{aligned}
\big|J(\pi_E)-J(\pi_{\boldsymbol{\theta}})\big| &= \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,a)\sim\rho_{\pi_E}}[\mathcal{R}(s,a)]-\mathbb{E}_{(s,a)\sim\rho_{\pi_{\boldsymbol{\theta}}}}[\mathcal{R}(s,a)]\Big| \\
&\leq \frac{1}{1-\gamma}\int_{\mathcal{S}}\int_{\mathcal{A}}\Big|\mathcal{R}(s,a)\Big|\Big|\rho_{\pi_E}(s,a)-\rho_{\pi_{\boldsymbol{\theta}}}(s,a)\Big|dads \\
&\leq \frac{R_{\max}}{1-\gamma}\int_{\mathcal{S}}\int_{\mathcal{A}}\Big|\rho_{\pi_E}(s,a)-\rho_{\pi_{\boldsymbol{\theta}}}(s,a)\Big|dads \\
&= \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(s,a),\rho_{\pi_E}(s,a)).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\big|J(\pi_E)-J(\pi_{\boldsymbol{\theta}})\big| &\leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(s,a),\rho_{\pi_E}(s,a)) \\
&\leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(z,a),\rho_{\pi_E}(z,a)) \qquad (25) \\
&\leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'),\rho_{\pi_E}(z,z')\big)+C, \qquad (26)
\end{aligned}
$$

where C is in (24). Note that (25) follows from Theorem 3 and (26) from Lemma 3. □

Corollary 1. *Consider the same conditions as in Theorem 1. Provided the injectivity of $\mathbb{P}(z'|z,a)$ with respect to a, then*

$$\big|J(\pi_E) - J(\pi_{\boldsymbol{\theta}})\big| \leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big).$$

Proof. The proof follows similarly to Yang et al. (2019) in the MDP case. Recall

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') = \frac{\mathbb{P}(z'|z,a)\pi_{\boldsymbol{\theta}}(a|z)}{\int_{\mathcal{A}}\mathbb{P}(z'|z,\bar{a})\pi_{\boldsymbol{\theta}}(\bar{a}|z)d\bar{a}},$$

and

$$\mathbb{P}(z'|z,a) = \int_{\mathcal{S}}\int_{\mathcal{S}}\int_{\mathcal{X}}\mathbb{P}(z'|x',a,z)\mathcal{U}(x'|s')\mathcal{T}(s'|s,a)\mathbb{P}(s|z)dx'ds'ds.$$

We can write $\mathbb{P}(z'|z,a) = \delta(z' - f(z,a))$, where $f : \mathcal{Z}\times\mathcal{A}\to\mathcal{Z}$ and δ is a Dirac delta function. It follows that

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z') = \mathbb{P}_{\pi_E}(a|z,z') = \delta(z' - f(z,a)),$$

which yields

$$\mathbb{E}_{(z,z')\sim\rho_{\pi_{\boldsymbol{\theta}}}(z,z')}\Big[\mathbb{D}_{\mathrm{TV}}\big(\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(a|z,z'), \mathbb{P}_{\pi_E}(a|z,z')\big)\Big] = 0.$$

Using this result in Theorem 1 proves the corollary. □

Theorem 2 restated: Consider a POMDP, and let $\mathcal{R}:\mathcal{S}\times\mathcal{S}\to\mathbb{R}$ and $z_t = \phi(x_{\leq t})$ such that $\mathbb{P}(s_t|z_t,a_t) = \mathbb{P}(s_t|z_t) = \mathbb{P}(s_t|x_{\leq t}, a_{<t})$. Then, the following inequality holds:

$$\big|J(\pi_E) - J(\pi_{\boldsymbol{\theta}})\big| \leq \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big),$$

where $R_{\max} = \max_{(s,s')\in\mathcal{S}\times\mathcal{S}}|\mathcal{R}(s,s')|$.

Proof. Note that $(1-\gamma)\cdot J(\pi) = \mathbb{E}_{(s,s')\sim\rho_\pi(s,s')}[\mathcal{R}(s,s')]$. Then,

$$\begin{aligned}
\big|J(\pi_E) - J(\pi_{\boldsymbol{\theta}})\big| &= \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,s')\sim\rho_{\pi_E}}[\mathcal{R}(s,s')] - \mathbb{E}_{(s,s')\sim\rho_{\pi_{\boldsymbol{\theta}}}}[\mathcal{R}(s,s')]\Big| \\
&\leq \frac{1}{1-\gamma}\int_{\mathcal{S}}\int_{\mathcal{S}}\Big|\mathcal{R}(s,s')\Big|\Big|\rho_{\pi_E}(s,s') - \rho_{\pi_{\boldsymbol{\theta}}}(s,s')\Big|ds'ds \\
&\leq \frac{R_{\max}}{1-\gamma}\int_{\mathcal{S}}\int_{\mathcal{S}}\Big|\rho_{\pi_E}(s,s') - \rho_{\pi_{\boldsymbol{\theta}}}(s,s')\Big|ds'ds \\
&= \frac{2R_{\max}}{1-\gamma}\mathbb{D}_{\mathrm{TV}}(\rho_{\pi_{\boldsymbol{\theta}}}(s,s'), \rho_{\pi_E}(s,s')).
\end{aligned}$$

In order to conclude the proof, we have to show that

$$\mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(s,s'), \rho_{\pi_E}(s,s')\big) \leq \mathbb{D}_{\mathrm{TV}}\big(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'), \rho_{\pi_E}(z,z')\big). \tag{27}$$

Consider a generic f-divergence as in Definition 1. Then,

$$
\begin{aligned}
&\mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(s,s'),\rho_{\pi_E}(s,s')) \\
&\quad= \mathbb{E}_{(s,s')\sim\rho_{\pi_E}(s,s')}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,s')}{\rho_{\pi_E}(s,s')}\right)\right] \\
&\quad= \mathbb{E}_{(s,s')\sim\rho_{\pi_E}(s,s')}\left[f\left(\mathbb{E}_{(z,z')\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(z,z'|s,s')}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,s')}{\rho_{\pi_E}(s,s')}\right]\right)\right] \\
&\quad= \mathbb{E}_{(s,s')\sim\rho_{\pi_E}(s,s')}\left[f\left(\mathbb{E}_{(z,z')\sim\mathbb{P}_{\pi_E}(z,z'|s,s')}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,s')}{\rho_{\pi_E}(s,s')}\frac{\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(z,z'|s,s')}{\mathbb{P}_{\pi_E}(z,z'|s,s')}\right]\right)\right] \\
&\quad= \mathbb{E}_{(s,s')\sim\rho_{\pi_E}(s,s')}\left[f\left(\mathbb{E}_{(z,z')\sim\mathbb{P}_{\pi_E}(z,z'|s,s')}\left[\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,s',z,z')}{\rho_{\pi_E}(s,s',z,z')}\right]\right)\right] \\
&\quad\leq \mathbb{E}_{(s,s')\sim\rho_{\pi_E}(s,s')}\mathbb{E}_{(z,z')\sim\mathbb{P}_{\pi_E}(z,z'|s,s')}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(s,s',z,z')}{\rho_{\pi_E}(s,s',z,z')}\right)\right] \qquad (28) \\
&\quad= \mathbb{E}_{(z,z')\sim\rho_{\pi_E}(z,z')}\mathbb{E}_{(s,s')\sim\mathbb{P}(s,s'|z,z')}\left[f\left(\frac{\rho_{\pi_{\boldsymbol{\theta}}}(z,z')\mathbb{P}(s,s'|z,z')}{\rho_{\pi_E}(z,z')\mathbb{P}(s,s'|z,z')}\right)\right] \\
&\quad= \mathbb{D}_f(\rho_{\pi_{\boldsymbol{\theta}}}(z,z'),\rho_{\pi_E}(z,z')), \qquad (29)
\end{aligned}
$$

from which (27) follows by considering $f(x) = \frac{1}{2}|x-1|$. Note that (28) uses Jensen's inequality and (29) follows considering that $\mathbb{P}(s,s'|z,z')$ is policy independent (Lemma 4). Hence, $\mathbb{P}(s,s'|z,z')$ depends only on the environment which is, by assumption, the same for both the expert and the agent. □

D LAIfO algorithm and hyperparameters

Algorithm 1: LAIfO

Inputs:
Expert observations: $(x_n)_{0:N} \in \mathcal{B}_E$.
$\pi_{\boldsymbol{\theta}}$, $D_{\boldsymbol{\chi}}$, $Q_{\boldsymbol{\psi}_1}$, $Q_{\boldsymbol{\psi}_2}$, $\phi_{\boldsymbol{\delta}}$: networks for policy, discriminator, Q functions, and feature extractor.
T_{train}, $\sigma(t)$, d, aug, c, τ, B, α, α_D, λ, γ: training steps, scheduled standard deviation, frames stack dimension, data augmentation, clip value, target update rate, batch size, learning rate, discriminator learning rate, discriminator gradient penalty weight, and discount factor.

for $t = 1, \ldots, T_{\text{train}}$ **do**
- $\sigma_t \leftarrow \sigma(t)$
- **if** $t \geq d - 1$ **then**
  - $z_t \leftarrow \phi_{\boldsymbol{\delta}}(x_{t-d+1:t})$
- **else**
  - $z_t \leftarrow \phi_{\boldsymbol{\delta}}(x_{0:t})$
- $a_t \leftarrow \pi_{\boldsymbol{\theta}}(z_t) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_t^2)$
- $s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)$ and $x_{t+1} \sim \mathcal{U}(\cdot|s_{t+1})$
- $\mathcal{B} \leftarrow \mathcal{B} \cup (x_t, a_t, x_{t+1})$
- UpdateDiscriminator($\mathcal{B}$, $\mathcal{B}_E$)
- UpdateCritic($\mathcal{B}$)
- UpdateActor($\mathcal{B}$)

begin UpdateDiscriminator
- $\{(x_{t-d+1:t}, x_{t-d+2:t+1})\} \sim \mathcal{B}_E$ and $\{(x_{t-d+1:t}, x_{t-d+2:t+1})\} \sim \mathcal{B}$ (sample B transitions)
- $z_t \leftarrow \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t-d+1:t}))$ and $z_{t+1} \leftarrow \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t-d+2:t+1}))$ for both agent and expert
- Update $D_{\boldsymbol{\chi}}$ to minimize (5) with learning rate α_D and gradient penalty weight λ

begin UpdateCritic
- $\{(x_{t-d+1:t}, a_t, x_{t-d+2:t+1})\} \sim \mathcal{B}$ (sample B transitions)
- $z_t \leftarrow \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t-d+1:t}))$ and $z_{t+1} \leftarrow \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t-d+2:t+1}))$
- $a_{t+1} \leftarrow \pi_{\boldsymbol{\theta}}(z_{t+1}) + \epsilon$ and $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma_t^2), -c, c)$
- Update $Q_{\boldsymbol{\psi}_1}$, $Q_{\boldsymbol{\psi}_2}$ and $\phi_{\boldsymbol{\delta}}$ to minimize (7) with $r_{\boldsymbol{\chi}}(z_t, z_{t+1})$ in (8) and learning rate α
- $\bar{\boldsymbol{\psi}}_k \leftarrow (1-\tau)\bar{\boldsymbol{\psi}}_k + \tau\boldsymbol{\psi}_k \quad \forall k \in \{1, 2\}$

begin UpdateActor
- $\{x_{t-d+1:t}\} \sim \mathcal{B}$ (sample B observations)
- $z_t \leftarrow \phi_{\boldsymbol{\delta}}(\text{aug}(x_{t-d+1:t}))$
- $a_t \leftarrow \pi_{\boldsymbol{\theta}}(z_t) + \epsilon$ and $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma_t^2), -c, c)$
- Update $\pi_{\boldsymbol{\theta}}$ using DDPG (Lillicrap et al., 2015) with learning rate α

Table 5: Hyperparameter values for LAIfO experiments.

Hyperparameter Name	Value
Frames stack (d)	3
Discount factor (γ)	0.99
Image size	84×84
Batch size (B)	256
Optimizer	Adam
Augmentation	Crop
Learning rate (α)	10^{-4}
Discriminator learning rate (α_D)	4×10^{-4}
Discriminator gradient penalty weight (λ)	10
Target update rate (τ)	0.01
Clip value (c)	0.3

Table 6: Experimental results for IL and IfO (i.e., imitation in fully observable environments, with and without access to expert actions, respectively). We use DDPG to train experts in a fully observable setting and collect 100 episodes of expert data. The experiments are conducted as in Table 2. We report mean and standard deviation of final performance over seeds.

	Expert	DAC (Kostrikov et al., 2018)	DACfO	DAC w/o regularizer in (6)	DACfO w/o regularizer in (6)
Cup Catch	980	$\mathbf{974 \pm 2.2}$	971 ± 3.9	746 ± 302	859 ± 143
Finger Spin	932	$\mathbf{939 \pm 14.2}$	913 ± 11.8	918 ± 3.1	925 ± 4.9
Cartpole Swingup	881	$\mathbf{717 \pm 321}$	715 ± 320	477 ± 259	284 ± 312
Cartpole Balance	990	$\mathbf{989 \pm 2.3}$	987 ± 1.1	278 ± 265	837 ± 303
Pendulum Swingup	845	849 ± 25.1	621 ± 320	832 ± 38.5	$\mathbf{851 \pm 26.0}$
Walker Walk	960	$\mathbf{957 \pm 5.1}$	$\mathbf{957 \pm 4.3}$	920 ± 40.1	919 ± 67.8
Walker Stand	980	985 ± 3.0	982 ± 4.6	857 ± 192	$\mathbf{986 \pm 3.0}$
Walker Run	640	$\mathbf{636 \pm 4.4}$	$\mathbf{636 \pm 4.4}$	51.5 ± 15.5	42.2 ± 9.8
Hopper Stand	920	$\mathbf{812 \pm 126}$	796 ± 176	335 ± 152	361 ± 138
Hopper Hop	217	$\mathbf{213 \pm 2.0}$	212 ± 1.9	208 ± 6.7	163 ± 81.8
Quadruped Walk	970	$\mathbf{943 \pm 29.0}$	826 ± 175	126 ± 50.1	207 ± 73.8
Quadruped Run	950	$\mathbf{919 \pm 13.4}$	870 ± 91.7	109 ± 62.3	133 ± 61.6
Cheetah Run	900	$\mathbf{871 \pm 33.9}$	710 ± 355	152 ± 288	31.4 ± 20.4

Table 7: Ablation study on the importance of data augmentation in LAIfO and LAIL. We use DDPG to train experts in a fully observable setting and collect 100 episodes of expert data. The experiments are conducted as in Table 2. We report mean and standard deviation of final performance over seeds after training for 10^6 frames.

	Expert	BC	LAIfO	LAIL	LAIfO w/o data augmentation	LAIL w/o data augmentation
Cup Catch	980	971 ± 36.0	$\mathbf{967 \pm 7.6}$	962 ± 18.0	929 ± 35.8	918 ± 36.0
Finger Spin	932	542 ± 219	$\mathbf{926 \pm 10.7}$	775 ± 345	758 ± 339	889 ± 12.8
Walker Walk	960	723 ± 137	$\mathbf{960 \pm 2.2}$	946 ± 8.6	541 ± 247	577 ± 259
Walker Stand	960	871 ± 77.7	$\mathbf{961 \pm 20.0}$	893 ± 106	729 ± 144	765 ± 118
Walker Run	640	133 ± 27.8	618 ± 4.6	$\mathbf{625 \pm 5.1}$	198 ± 82.4	327 ± 100
Hopper Stand	920	398 ± 96.4	$\mathbf{800 \pm 46.7}$	764 ± 111	658 ± 60.8	695 ± 81.0
Hopper Hop	217	45.9 ± 22.1	193 ± 12.9	$\mathbf{202 \pm 4.8}$	149 ± 66.9	170 ± 38.5
Quadruped Walk	970	337 ± 50.5	397 ± 67.0	$\mathbf{404 \pm 192}$	218 ± 77.2	201 ± 73.0

E Experiments curves and additional ablations

All the experiments are run using Nvidia-A40 GPUs on an internal cluster. For each algorithm, we run two experiments in parallel on the same GPU and each experiment takes 1 to 10 days depending on the simulated environment and the considered algorithm. For all the implementation details refer to our code[1].

Table 6 summarizes the asymptotic performance of DAC and DACfO in the fully observable setting. We highlight in Table 6 the importance of the discriminator penalty term in (6) by comparing DAC and DACfO to results that do not include this regularizer in the adversarial loss in (5). Figure 7 shows the full learning curves as a function of training steps for Table 6. The results of DAC and DACfO in Table 6 are used in Figure 5. Moreover, Table 7 and Figure 8 show the results for the ablation study on data augmentation in LAIL and LAIfO. These results highlight the importance of data augmentation on the latent variable $z \in \mathcal{Z}$ inference problem, as we observe a decrease in performance when data augmentation is not used. Finally, Figure 9 shows the full learning curves for the results in Table 4.

[1]https://github.com/VittorioGiammarino/AIL_from_visual_obs/tree/LAIfO

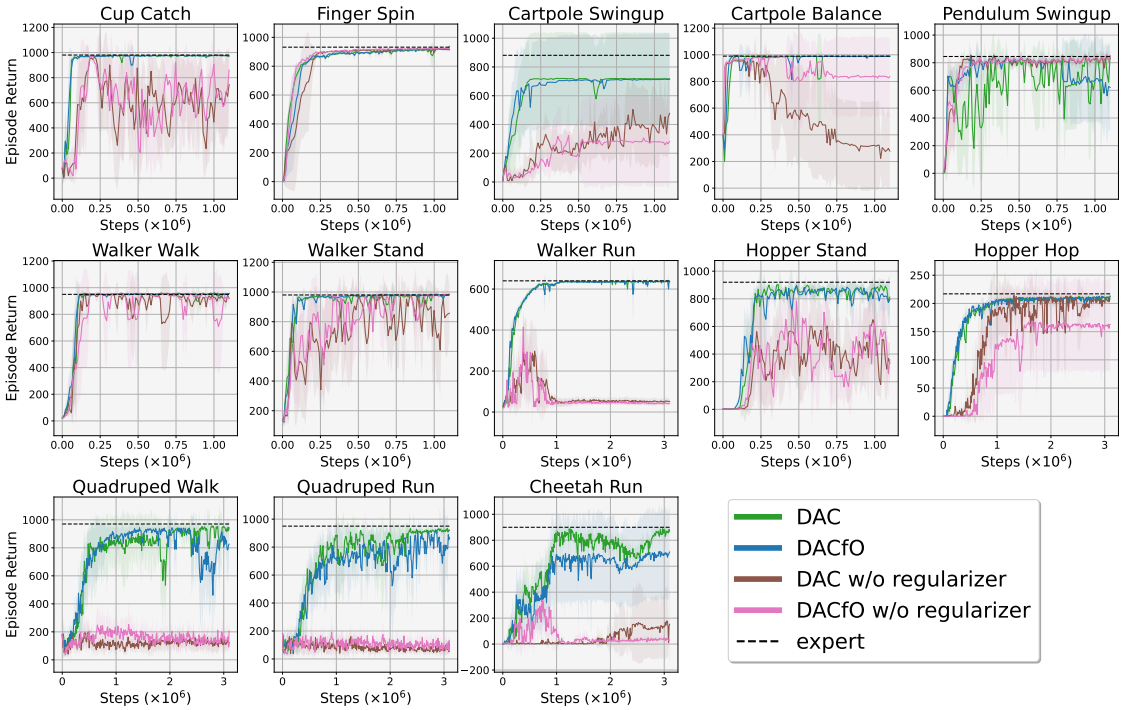

Figure 7: Learning curves for the fully observable IL and IfO settings in Table 6. Plots show the average return per episode as a function of training steps.

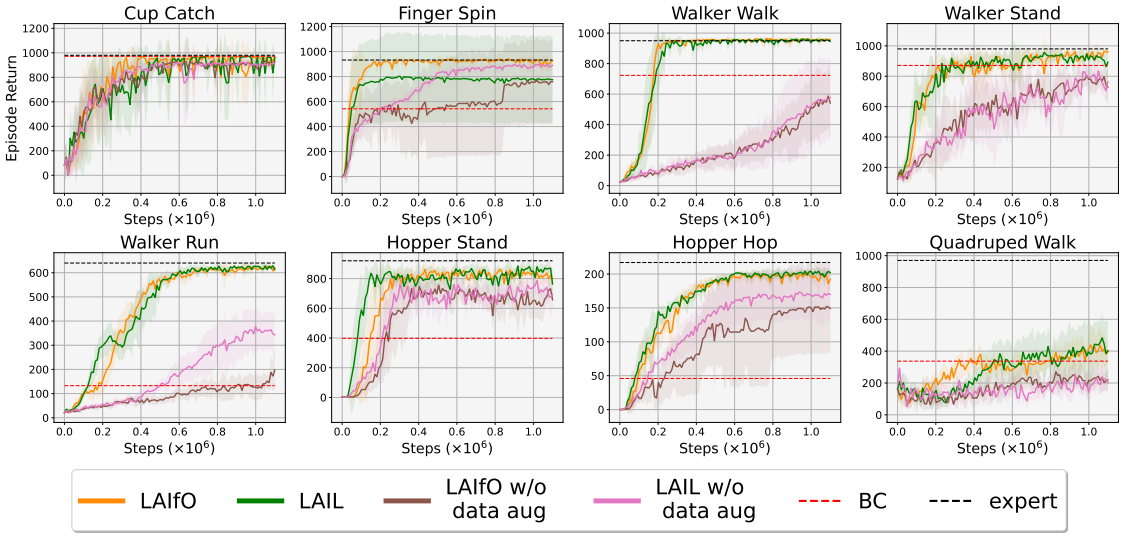

Figure 8: Learning curves for the results in Table 7. Plots show the average return per episode as a function of training steps.

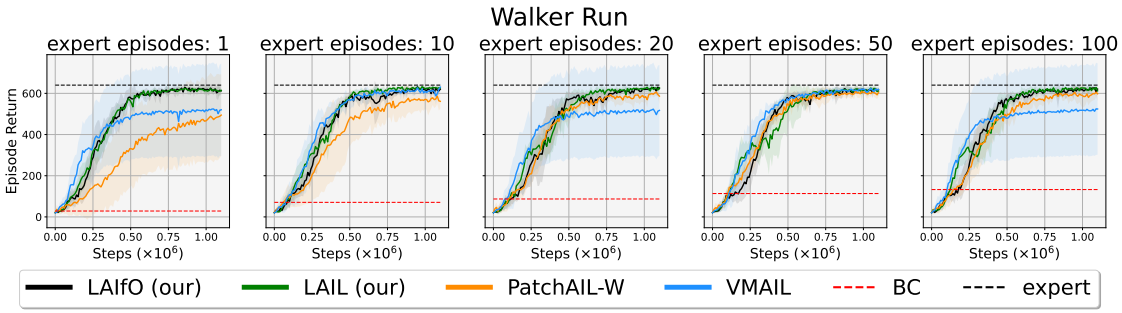

Figure 9: Learning curves for the results in Table 4. Plots show the average return per episode as a function of training steps.