Explorative Imitation Learning: A Path Signature Approach for Continuous Environments

Nathan Gavenski [a,*], **Juarez Monteiro, Felipe Meneguzzi**[b,c], **Michael Luck**[d] **and Odinaldo Rodrigues**[a]

[a]King's College London, London, United Kingdom
[b]University of Aberdeen, Aberdeen, United Kingdom
[c]Pontifícia Universidade Católica do RS, Porto Alegre, Brazil
[d]University of Sussex, Sussex, United Kingdom

Abstract. Some imitation learning methods combine behavioural cloning with self-supervision to infer actions from state pairs. However, most rely on a large number of expert trajectories to increase generalisation and human intervention to capture key aspects of the problem, such as domain constraints. In this paper, we propose Continuous Imitation Learning from Observation (CILO), a new method augmenting imitation learning with two important features: (i) exploration, allowing for more diverse state transitions, requiring less expert trajectories and resulting in fewer training iterations; and (ii) path signatures, allowing for automatic encoding of constraints, through the creation of non-parametric representations of agents and expert trajectories. We compared CILO with a baseline and two leading imitation learning methods in five environments. It had the best overall performance of all methods in all environments, outperforming the expert in two of them.

1 Introduction

One of the most common forms of learning is by watching someone else perform a task and, afterwards, trying it ourselves. As humans, we can observe an action being performed and transfer the acquired knowledge into our reality. In this respect, it is less challenging to achieve a goal in an optimal way by observing how an expert behaves; in the field of computer science, this is Imitation Learning (IL). Unlike conventional reinforcement learning, which depends on a reward function, IL learns from expert guidance, and is concerned with an agent's acquisition of skills or behaviours by observing a 'teacher' perform a given task.

Learning from demonstration is the obvious approach for IL, requiring expert demonstrations, which are 'trajectories' including actions performed along the way to goal completion [11]. Such an approach uses the trajectories to learn an approximate policy that behaves like the expert. Learning from demonstration suffers from two significant drawbacks in practice: poor generalisation in environments with multiple alternative trajectories that achieve a goal, which is bound to occur when the dataset size increases, and the unavailability of data about the expert's actions. *Learning from observation* (LfO) overcomes these limitations by learning a task without direct action information via self-supervision, which increases generalisation [7]. This allows a model to learn from sample executions without action information, which would otherwise be unusable. LfO approaches often rely on techniques from classification to improve sample-efficiency [25] and generalisation [16]. Such agents require fewer expert trajectories, yielding more general approaches that are, hence, adaptable to unseen scenarios. However, these methods still fail to leverage some useful learning features, particularly the use of an exploration mechanism.

Some existing work [4, 6] requires manual intervention in different stages of the process, e.g., the hard-coding of environment goals, which is not feasible in complex environments, such as robotic systems with multifaceted goals. Other work [6, 12, 25] is limited in that learning the environment dynamics depends strongly on previously collected samples that usually do not relate to how the environment dynamics operate under expert behaviour, such as random transitions, or prior knowledge of the dynamics of environments. In addition, maintaining self-supervision [6, 12] for an IL method is important since unlabelled data is more readily available, *e.g.*, from sources that are not necessarily meant for agent learning.

In this paper, we propose a novel LfO approach to IL called Continuous Imitation Learning from Observation (CILO) that addresses the above issues. CILO (i) eliminates the need for manual intervention when using different environments by discriminating between policy and expert; (ii) requires fewer samples for learning by leveraging exploration and exploitation; and (iii) does not require expert-labelled data, thus remaining self-supervised. We evaluated CILO in five widely used continuous environments against a baseline and two leading LfO methods (see Section 4). Our results show that CILO outperformed all of the alternatives, surpassing the expert in two of five environments.

CILO's new mechanisms are model-agnostic and applicable to a wider range of environment dynamics than those of the compared LfO alternatives. We argue that the new mechanisms can be readily incorporated into other IL methods, paving the way for more robust and flexible learning techniques.

2 Problem Formulation

We assume the environment to be an MDP $M = \langle S, A, T, r, \gamma \rangle$, in which S is the state space, A is the action space, T is a transition model, r is the immediate reward function, and γ is the discount factor [21]. Although in general an MDP may carry information regarding the reward and discount factors, we consider that this information is inaccessible to the agent during training, and the learning

* Corresponding Author. Email: nathan.schneider_gavenski@kcl.ac.uk

process does not depend on it. Solving an MDP yields a policy π with a probability distribution over actions, giving the probability of taking an action a in state s. We denote the expert policy by π_ψ.

A common self-supervised approach to solving a task via IL uses an Inverse Dynamic Model $\mathcal{M}$. $\mathcal{M}$ uses a set of state transition samples (s_t, s_{t+1}) to predict the action performed in the transitions. By training $\mathcal{M}$ to infer the actions in the state transitions, these approaches can automatically annotate all expert trajectories $\mathcal{T}^{\pi_\psi}$ with actions without the need for human intervention [23, 16, 6]. The agent policy π_θ then uses these *self-supervised* expert-labelled states $(s^{\pi_\psi}, \hat{a})$ to learn to predict the most likely action given a state $P(a \mid s^{\pi_\psi})$. Torabi et al. [23] show that applying an iterative process in self-supervised IL approaches helps π_θ achieve better performance. Initially, $\mathcal{M}$ uses only single transition samples I^{pre} from π_θ and its randomly initialised weights. At each iteration, these approaches use π_θ to create new samples I^{pos} that are used to fine-tune $\mathcal{M}$. However, using all transitions from π_θ makes this iterative approach susceptible to getting stuck in *local minima* due to class imbalance from the I^{pos} data. Monteiro et al. [16] propose a solution that introduces a goal-aware function to sample from all trajectories at each epoch. This function does not require an aligned goal from the environment, hence it is up to the user to choose a desired goal. If π_θ reaches this goal, the trajectory will be used. Finally, it is sensible to assume that $\mathcal{M}$ is not well-tuned during early iterations and predicts mostly wrong labels. Therefore, Gavenski et al. [6] implement an exploration mechanism that uses the softmax distribution of the output as weights to sample actions proportionally to optimality from the model's prediction. As the confidence in the model increases, it predicts suboptimal actions less than the *maximum a posteriori estimation*. By exploring using the model's confidence, their approach can learn under the exploration and exploitation phases, helping $\mathcal{M}$ to converge faster. Nevertheless, creating a handcrafted goal-aware function and using a softmax distribution as an exploration mechanism requires manual intervention and discrete actions. As a result, these methods become unsuitable for more complex environments where goal achievement is non-trivial to check.

3 Continuous Imitation Learning from Observation

We address the need for manual intervention and for maintaining self-supervision in CILO through two key innovations: an exploration mechanism used when the action predictions are uncertain; and a discriminator to interleave random and current states to improve the prediction of self-supervised actions. CILO achieves this by employing three different models: (i) the inverse dynamic model $\mathcal{M}$ to predict the action responsible for a transition between two states $P(a \mid s_t, s_{t+1})$; (ii) a policy model π_θ that uses the self-supervised labels $\hat{a}$ to imitate the expert π_ψ given a state $P(a \mid s_t)$; and (iii) a discriminator model $\mathcal{D}$ to discriminate between π_ψ and π_θ, creating newer samples for $\mathcal{M}$.

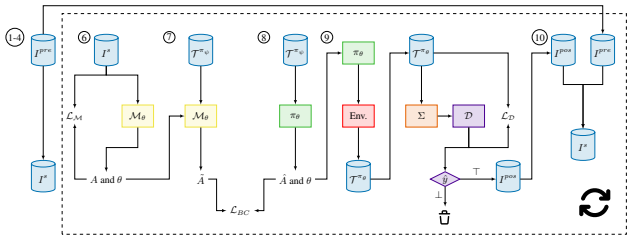


Figure 1. CILO's training cycle.

Algorithm 1 CILO

1: Initialize $\mathcal{M}_\theta$, π_θ, and $\mathcal{D}$ with random weights
2: $I^s \leftarrow I^{pre}$ s.t. $I^{pre} \leftarrow$ samples from π_θ
3: **for** $i \leftarrow 1$ to epochs **do**
4: Improve $\mathcal{M}_\theta$ by TRAINM(I^s)
5: Use $\mathcal{M}_\theta$ with $\mathcal{T}^{\pi_\psi}$ to predict $\hat{A}$
6: Improve π_θ by error$_{\pi_\theta} \leftarrow$ BEHAVIOURCLONING$(\mathcal{T}^{\pi_\psi}, \hat{A})$
7: Use π_θ to solve environments E
8: $\mathcal{T}^{\pi_\theta} \leftarrow \mathcal{T}^{\pi_\theta} \oplus \{(s_0, \hat{a}_0, s_1), \cdots, (s_{t-1}, \hat{a}_{t-1}, s_t)\}$
9: $I^{pos} \leftarrow I^{pos} \oplus \{\forall_{i \in \mathcal{T}^{\pi_\theta}} \mid \mathcal{D}(\beta(\mathcal{T}_i^{\pi_\theta})) \text{ is } \top\}$
10: $I^s \leftarrow I^{pre} \oplus I^{pos}$
11: **if** error$_{\pi_\theta} \leqslant threshold$ **then**
12: Finish training

Algorithm 1 provides an overview of CILO's learning process. First, CILO initialises all models with random weights and uses the random initialised policy to collect random samples I^{pre} from the environment (Lines 1-2). The dynamics model uses these random samples to train in a supervised manner (Function TRAINM, Line 4). These samples are vital since they help $\mathcal{M}$ learn how actions cause environmental transitions without expert behaviour-specific knowledge. TRAINM uses the loss from Eq. 1, where θ are the model's current parameters, S is the vector for state representations, A is the action vector representation, and t is the timestep.

$$\mathcal{L}_\mathcal{M}(I^s, \theta) = \sum_{t=1}^{I^s} |\mathcal{M}_\theta(S_t, S_{t+1}) - A_t| \tag{1}$$

With the updated parameters θ, $\mathcal{M}$ predicts the self-supervised labels $\hat{A}$ to all expert transitions ($\mathcal{T}^{\pi_\psi}$ in Line 5). CILO then uses these expert labelled transitions to train π_θ using behaviour cloning (Function BEHAVIOURALCLONING with Eq. 2) coupled with an exploration mechanism (Line 6).

$$\mathcal{L}_{BC}(I^s) = \sum_{(s_t, s_{t+1}) \in I^s} |\mathcal{M}_\theta(s_t, s_{t+1}) - \pi_\theta(s_t)| \tag{2}$$

The policy then generates new samples ($\mathcal{T}^{\pi_\theta}$) that might help $\mathcal{M}$ approximate the unknown ground-truth actions from the expert (Lines 7-8). Given all new samples, CILO generates path signatures β [2] and uses $\mathcal{D}$ to classify signatures as from the expert or the agent. Line 9 updates the discriminator weights with the classification loss in Eq. 3, where $\mathcal{T}_\beta$ are path signatures for all trajectories from expert and agent, C is for the source of the observation (*expert* and *agent*), y is the ground-truth label, and $\hat{y}$ is the source predicted by $\mathcal{D}$.

$$\mathcal{L}_\mathcal{D}(\mathcal{T}_\beta^{\pi_\psi}, \mathcal{T}_\beta^{\pi_\theta}) = -\sum_{i=1}^{|\mathcal{T}_\beta|} \sum_{j=1}^{|C|} y_{ij} log(\hat{y}_{ij}), \tag{3}$$

Samples classified as expert by $\mathcal{D}$ are added to I^{pos} (Line 9), which is then combined with the original I^{pre} to form I^s (Line 10). CILO uses this updated I^s in each iteration for a specified number of epochs (Line 4) or until it no longer improves (Lines 11-12), with an optional hyperparameter $threshold$.

The exploration mechanism allows CILO to deviate from its original action distribution according to the model certainty. This behaviour is helpful during early iterations when $\mathcal{M}$ is unsure about which action might be responsible for a specific transition. Since random samples can be very different from the expert's transitions, we can assume that the model does not learn to recognise these transitions and generalises poorly. Here, we assume that the environment is stochastic, in that multiple actions might occur with a non-zero probability of transition between any pair of states. $\mathcal{D}$'s key objective is to discard trajectories that could result in $\mathcal{M}$ getting stuck in

bad local minima, and for instance, stop predicting specific actions (underfitting). Without a discriminator, it would be difficult to ignore signatures that differ considerably from the ground-truth without an environment-specific hyperparameter, hence reducing the method's generalisability. Finally, combining these mechanisms makes CILO more sample efficient, allowing for oversampling without misrepresenting the action distributions and overfitting. Figure 1 shows the CILO learning cycle in more detail with the different loss functions.

3.1 Exploration

Exploration is vital for IL methods that use dynamics models to learn how the expert behaves. It enables policy divergence when the dynamics model is uncertain and increases state diversity, which helps the model approximate labelled transitions from unlabelled ones (expert). CILO borrows an exploration mechanism from reinforcement learning in continuous domains, in which each action in a policy consists of two outputs: the mean and standard deviation to sample from a Gaussian distribution. However, unlike traditional reinforcement learning, where a policy receives feedback in the form of the reward function, IL lacks this information. Thus, for a model $\mathbb{M}$ and parameters θ ($\mathbb{M}_\theta$), we employ the sampling mechanism in Eq. 4, where $\boldsymbol{\pi}$ is the usual mathematical constant $3.14\ldots$ and ε, as defined in Eq. 5, is used as standard deviation, where a is the ground-truth action (or pseudo-labels from $\mathcal{M}$) and $\hat{a}$ is the action predicted by the model:

$$\tilde{a}_{\mathbb{M}_\theta} = \frac{1}{\varepsilon\sqrt{2\boldsymbol{\pi}}} e^{-\frac{(s_t^e - \mathbb{M}_\theta(S))}{2\varepsilon^2}} \tag{4}$$

$$\varepsilon = \|a - \hat{a}\|^p \tag{5}$$

In Eq. 4, $\mathbb{M}$ is either $\mathcal{M}$ or π, and θ are the parameters of the model updated for the epoch. Notice that when $p = 1$, the model $\mathbb{M}$ uses the absolute value between the predicted and ground-truth labels $\|a - \hat{a}\|$ and this allows for higher exploration.

Observation 1. *If $\mathcal{L}$ is a loss function that monotonically decreases a model's $\mathbb{M}$ error as it approximates the ground-truth function, eventually $\|a - \hat{a}\| < 1$. If we then use $p > 1$ in Eq. 5, ε will exponentially decrease.*

Given all of the above, Eq. 5 offers a trade-off between exploration and exploitation. Since ε is the standard deviation for the exploration function, as the model's predictions get closer to the ground-truth and pseudo-labels, the clusters will have lower variance because the exploration ratio is directly correlated to the model's error.

In Alg. 1, functions TRAINM (ln. 4) and BEHAVIOURCLONING (ln. 6) use this adaptation to adjust the exploration ratio depending on how close the model's predictions are to the ground-truth (or pseudo-labels), in accordance with the standard deviation of the Gaussian distribution. This mechanism also has the benefit of not having to predict information beyond the agent's actions, such as standard deviation, instead obtaining this directly from the model's error. For deterministic behaviour, we can assume that the standard deviation for the model is 0 and use the model's output since sampling from a Gaussian distribution with average x and deviation 0 equals x.

3.2 Goal-aware function

Developing a goal-aware function may not be a trivial task. For environments with a well-defined goal, such as CartPole [1], which defines the goal to be balancing the pole for 195 steps, a goal-identification function could simply classify all trajectories that reach 195 steps as optimal. In this work, we formally define trajectories as:

Definition 1. *A trajectory τ is a finite sequence of states $(s_1, \ldots, s_n)$ where for each $1 \leq i < n$, s_{i+1} is obtained from s_i via the execution of some action. We use the term $(\tau_t^1, \tau_t^2, \ldots, \tau_t^d) \in \mathbb{R}^d$ to denote the particular state s_t ($1 \leqslant t \leqslant n$) within the trajectory τ.*

However, recall that in the context of IL, the agent has no access to the reward signal, and as environments grow in complexity, such a function becomes even harder to encode. By contrast, some environments have no prescribed goal. For example, the Ant environment requires the agent to walk as far as possible without falling, but with no defined cap on the number of time steps [19]. Thus, existing IL approaches [16, 6, 5] often rely on manually defined goal-aware functions, which have the benefit of dispensing with the alignment of the environment's goal. For example, we might define a specific trajectory as required in the Ant environment. Unless the agent reaches all points in this trajectory, our goal-aware function does not classify the episode as successful. However, this creates a degree of unwanted complexity in a learning algorithm and a cumbersome process as the number of environments grows. Yet, trajectories may carry relevant information for CILO since they approximate I^s's samples from $\mathcal{T}^{\pi_\psi}$ [6]. Therefore, CILO tries to classify trajectories that are close to $\mathcal{T}^{\pi_\psi}$ instead of successful ones.

Nevertheless, identifying whether samples are near $\mathcal{T}^{\pi_\psi}$ is also difficult. If we consider a stationary agent, we might discard samples that allow $\mathcal{M}$ to better predict transitions due to their distance to the π_ψ states alone. But, if we consider whole trajectories, it might be difficult to identify middling trajectories needed to close the gap between $\mathcal{T}^{\pi_\theta}$ and $\mathcal{T}^{\pi_\psi}$, and better generalise [7]. Therefore, CILO needs a function that (i) simplifies comparisons between trajectories and (ii) allows $\mathcal{M}$ to receive suboptimal samples.

For the first problem, previous work [17] dealt with the issue of trajectory length by using the average of all states up to a point in time to account for the trajectory changes. Conversely, we use path signatures [2], which are fixed-length feature vectors that are used to represent multi-dimensional time series (i.e., trajectories). A path signature is computed by the function β comprehensively defined in Section 3 of Yang et al. [24], succinctly summarised in the definition below (see Supplementary Material [9] for more detail).[1]

Definition 2. *Let a trajectory τ of a countable length between $[1, n]$ ($n \in \mathbb{N}$), where each state is a vector in $\mathbb{R}^d$ with dimensions indexed by a collection of indices $i_1, \cdots, i_k \in \{1, \cdots, d\}$. Let the recursively computed path signature β for a trajectory τ for any $k \geqslant 1$ and time t ($1 \leqslant t \leqslant n$) be:*

$$\beta(\tau)_{1,t}^{i_1,\cdots,i_k} = \int_{1<s\leqslant t} \beta(\tau)_{1,s}^{i_1,\cdots,i_{k-1}} d\tau_s^{i_k}. \tag{6}$$

Then, the signature of a trajectory $\tau : [1, n] \to \mathbb{R}^d$ is the collection of all the iterated integrals of τ:

$$\begin{aligned}\beta(\tau)_{1,n}^{1,\ldots,i_k} = \Big(1, \beta(\tau)_{1,n}^{1}, \cdots, \beta(\tau)_{1,n}^{d}, \beta(\tau)_{1,n}^{1,1}, \cdots, \\ \beta(\tau)_{1,n}^{1,d}, \beta(\tau)_{1,n}^{2,1}, \ldots, \beta(\tau)_{1,n}^{i_1,i_2,\cdots,i_k}\Big),\end{aligned} \tag{7}$$

where the zero-th term is conventionally equal to 1, and k is defined as the k-th level of the signature, which defines the finite collection of all terms $\beta(\tau)_{1,n}^{i_1,\cdots,i_k}$ for the multi-index of length k. For example, when $k = d$, the last term would be $\beta(\tau)_{1,n}^{d,d,\cdots,d}$.

[1] In our experiments, β (Line 9, Algorithm 1) was computed using the implementation provided by [13].

Path signatures allow CILO to solve the issue of comparing two trajectories and encoding different characteristics that may be relevant when classifying how close a new trajectory is from $\mathcal{T}^{\pi_\psi}$. By using path signatures generated from trajectories in $\mathcal{T}^{\pi_\theta}$ and $\mathcal{T}^{\pi_\psi}$, CILO benefits from: (i) a common signature size, regardless of the original length of trajectories, helping the discriminator not to discriminate against longer trajectories; (ii) independence of environment characteristics embedded in the data (avoiding the need for re–parametrisation for each environment); and (iii) the preservation of the uniqueness of trajectories via the non-linearity of the signatures.

The use of signatures still requires some manual intervention in CILO to define how close a trajectory needs to be before adding it to I^s (i.e., an appropriate similarity threshold). To prevent the need for manually defining this threshold, CILO uses a discriminator model $\mathcal{D}$ to discriminate between π_θ and π_ψ trajectories, which optimises Eq. 3. This yields a non-greedy sampling mechanism by using a model to classify expert and non-expert trajectories.

In summary, CILO's goal-aware function works by computing a signature $\beta(\tau)$ of a trajectory τ and feeds it into the discriminator model $\mathcal{D}$, which classifies whether the source of the trajectory is π_θ or π_ψ. If $\mathcal{D}$ classifies the source of an agent's trajectory as the expert, then CILO appends the trajectory into I^s, helping $\mathcal{M}$ better understand how the transition function T works in the environment.

3.3 Sample efficiency

Besides approximating the expert policy, IL methods focus on efficiently using expert samples. This focus happens since expert samples are hard to obtain. Thus, creating more efficient methods, *i.e.*, that require fewer samples, allows for more useable approaches. Some recent strategies [12, 25] minimise the number of required samples but depend on strong assumptions (see Section 4.2) or manual intervention for each new environment. For comparison, CILO uses 10 expert episodes – a number similar to Zhu et al. [25] and Kidambi et al. [12], but without requiring manual intervention for each environment. CILO relies on up-scaling $\mathcal{T}^{\pi_\psi}$ to increase the number of observations π_θ sees before interacting with the environment. Although trivial, this strategy works because CILO is self-supervised and has an exploration mechanism. This strategy helps in two ways: (i) for each epoch all pseudo-labels differ in all transitions due to the exploration mechanism (Line 4, Algorithm 1); and (ii) increasing the number of samples π_θ receives allows for more updates before sampling new experiences from the environment. By applying its exploration mechanism to each observation individually and sampling exploration values from a distribution, CILO ensures that each observation has unique action values, reducing the risk of misrepresenting the ground-truth action distribution.

4 Experimental Results

We compared CILO's results against three key related methods. Behavioral Cloning from Observations (BCO) [23], which is usually used as a baseline, and two of the most efficient LfO methods: Off-Policy Imitation Learning from Observations (OPOLO) [25], and Model-Based Imitation Learning From Observation Alone (MobILE) [12]. We experimented with five commonly used environments: Ant, Half Cheetah, Hopper, Swimmer, and Pendulum.[2] Each method was run for 50 episodes, with the environment reset when the agent falls or after $1,000$ steps. Each episode was run using random seeds to test the agent's ability to generalise.

4.1 Implementation and Metrics

We used PyTorch to implement our agent and optimise the loss functions in Eq. 1-3 via Adam [14] and Imitation Datasets [8] to collect the expert data. As for the exploration mechanism in Eq. 5, we use $p = 1$ for ε due to all environments actions being in the interval $[-1, 1]$, and using $p > 1$ would significantly diminish the gap between predicted and ground-truth actions (as defined in Definition 1). In the supplementary material [9], we provide all learning rates, a link for the official implementation and discuss hyperparameter sensitivity in more detail, but we note that CILO is not very sensitive to precise hyperparameters.

We evaluated all approaches using the *Average Episodic Reward* (AER) metric (Eq. 8) and use *Performance* ($\mathcal{P}$) (Eq. 9). AER is the average accumulated reward for a policy π over n number of episodes in t number of steps:

$$AER(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{t}\gamma^t r(s_{ij}, \pi(s_{ij})). \tag{8}$$

On the other hand, $\mathcal{P}$ normalises between random and expert policies rewards, where performance 0 corresponds to random policy π_ξ performance, and 1 is for expert policy π_ψ performance.

$$\mathcal{P}_\tau(\pi) = \frac{AER(\pi) - AER(\pi_\xi)}{AER(\pi_\psi) - AER(\pi_\xi)} \tag{9}$$

Note that a negative value for $\mathcal{P}$ indicates a reward for the agent lower than a random agent's and a value higher than 1 indicates that the agent's reward is higher than the expert's. All results in Table 1 are the average and standard deviation in five different experiments. We do not report accuracy since achieving high accuracy does not necessarily translate into a high reward for the agent.

4.2 Results

We trained all methods using 10 expert trajectories. Table 1 shows how each method performed in the five environments. CILO had the best overall results in all environments. It consistently achieved results similar to the expert, surpassing it on Ant and Swimmer and achieving the maximum reward for the Pendulum environment. CILO's performance was close to the expert's in Hopper but a little lower in HalfCheetah – likely due to the higher standard deviation from the ground-truth actions in both environments. In the Swimmer environment, BCO and OPOLO achieved AER and performance similar to the expert, while CILO outperformed it by 0.29 points. The same happened in the Ant environment, where CILO surpassed the expert by ≈ 484 reward points. We hypothesise this is due to CILO's explorative nature and its ability to acquire new samples that the discriminator judges to come from the expert.

Comparing CILO to other methods, we see that OPOLO had the closest performance to CILO's in almost all environments. We attribute CILO's better performance than OPOLO's due to the fact that OPOLO's problem formulation assumes that the environment follows an injective MDP, which cannot be guaranteed with random seeds. For this work, we believe that it is more important for an agent to be able to correct its initial states into a successful trajectory than to be optimal in a single setting. Moreover, we notice that for the Pendulum environment, OPOLO only achieved the optimal reward when

[2] The Supplementary Material [9] briefly describes these environments and the neural networks topology.

Table 1. CILO and baselines AER and $\mathcal{P}$ results for all environments. All results are the average of 50 trajectories.

Algorithm	Metric	Ant	Pendulum	Swimmer	Hopper	HalfCheetah
Random	AER	-65.11 ± 106.16	5.70 ± 3.26	0.73 ± 11.44	17.92 ± 16.02	-293.13 ± 82.12
	$\mathcal{P}$	0	0	0	0	0
Expert	AER	5544.65 ± 76.11	1000 ± 0	259.52 ± 1.92	3589.88 ± 2.43	7561.78 ± 181.41
	$\mathcal{P}$	1	1	1	1	1
CILO	AER	$\mathbf{6091 \pm 801.2}$	$\mathbf{1000 \pm 0}$	$\mathbf{334.6 \pm 3.45}$	$\mathbf{3589 \pm 178.2}$	$\mathbf{7100.6434 \pm 90.1775}$
	$\mathcal{P}$	$\mathbf{1.0974 \pm 0.1372}$	$\mathbf{1 \pm 0}$	$\mathbf{1.2901 \pm 0.0128}$	$\mathbf{0.9998 \pm 0.0487}$	$\mathbf{0.9413 \pm 0.0115}$
OPOLO	AER	5508.6807 ± 930.7590	$\mathbf{1000 \pm 0}$	253.3297 ± 3.4771	3428.6405 ± 420.3285	7004.65 ± 568.66
	$\mathcal{P}$	0.9935 ± 0.1659	$\mathbf{1 \pm 0}$	0.9761 ± 0.0134	0.9549 ± 0.1177	0.9291 ± 0.0724
MobILE	AER	995.5 ± 25.65	111.7 ± 31.25	130.7 ± 24.36	2035 ± 192.95	4721.5 ± 364.5
	$\mathcal{P}$	0.1891 ± 0.0047	0.1066 ± 0.0313	0.5022 ± 0.0968	0.5647 ± 0.0531	0.5647 ± 0.0454
BCO	AER	1529 ± 980.86	521 ± 178.9	257.38 ± 4.28	1845.66 ± 628.41	3881.10 ± 938.81
	$\mathcal{P}$	0.2842 ± 0.1724	0.5675 ± 0.1785	0.9917 ± 0.0166	0.5177 ± 0.1765	0.5117 ± 0.1217

clipping the actions between $[-1, 1]$, which CILO does not require. When the actions are not clipped, OPOLO accumulates ≈ 9.55 reward points, a performance similar to the random policy. We opted not to clip CILO's actions, so that the method would not require any previous environment knowledge.

BCO requires more expert trajectories to achieve better results. In its original work, BCO used 5×10^5 samples for $\mathcal{M}$ and more than $1,000$ expert trajectories for its policy, which may be unrealistic for many domains. Nevertheless, BCO achieved almost expert results in the Swimmer environment and higher rewards than MobILE in almost all other environments, with the exception of HalfCheetah. We believe BCO outperforms MobILE because the latter assumes that each environment has a fixed initial state, which does not happen since the gym suite alters each initial state according to some parametrised intervals and its current seed.

As for MobILE, we used the same number of trajectories as in its original work. We observe that MobILE suffers from three different issues: (i) it is ensemble; (ii) has domain knowledge embedded into the algorithm (not publicly available); and (iii) its results are difficult to reproduce, because of the large number of hyperparameters on which they depend. During our experimentation, we observed that some approaches underperform when using an expert with strict movement constraints. To some extent, when obtaining $\mathcal{T}^{\pi_\psi}$, all environments are susceptible to this, but MoBILE was especially impacted. We believe that this strict movement pattern is difficult for all methods to learn since the impact of the variations cannot be immediately perceived. The lack of reproducibility is a major drawback of MobILE, from which CILO does not suffer. By using path signatures, which is a non-parametric encoding technique, CILO is left with only two different parameters: the network size and the learning rate. We used Smith's work (2017) as a guide for finding optimal values for these parameters.[3]

Finally, in environments with broadly distributed expert actions like Ant, Pendulum, and Hopper, CILO matches expert performance in fewer iterations than the other methods. However, in environments where actions are more concentrated (Swimmer and HalfCheetah), CILO takes longer to match the expert.

5 Discussion

In this section, we consider some key aspects of CILO's behaviour: (i) how CILO learns with different sample amounts; (ii) how it approximates predictions to the ground-truth actions of the expert; (iii) how similar each signature becomes to all trajectories over time; (iv) how different action distributions affect CILO; and (v) how I^s behaves over time.

[3] We followed the original network topology for a fair comparison.

5.1 Sample Efficiency

In order to understand CILO's sample efficiency, we experimented with three different amounts of expert episodes in the Ant environment. Ant provides an ideal setting due to its balanced learning complexity and shorter training times. Table 2 shows the AER and $\mathcal{P}$ results using 1, 10, and 100 trajectories. As expected, CILO does not achieve good results when using a single trajectory. This is because π_θ has no information regarding different initialisation and trajectory deviations. This behaviour is intrinsic to behavioural cloning where, without sufficient information, the policy tends not to generalise [15].

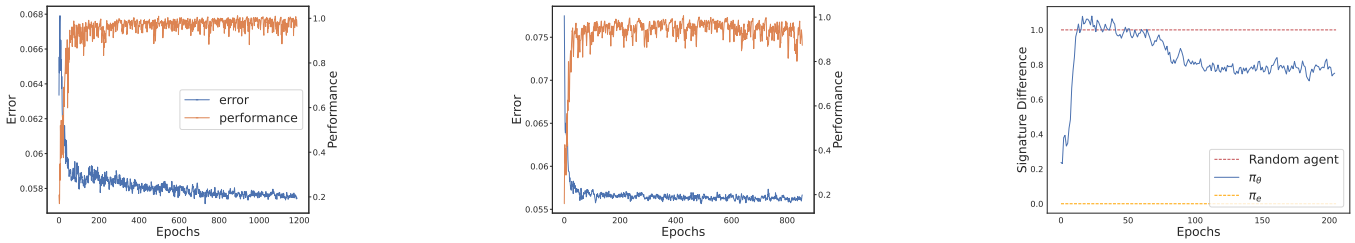
Interestingly, CILO achieves 65 *fewer* reward points when using 100 trajectories than when it uses 10. We attribute this to the following: (i) when used in a LfO scenario, BC methods usually fail to scale according to the number of samples due to *compounding error* [22]; and (ii) increasing the number of expert samples decreases the deviation from π_ψ trajectories, resulting in overfitting and a worse π_θ. Since it achieves expert results for almost all environments, we do not consider this behaviour a limitation of CILO. Nevertheless, we hypothesise that using different strategies might result in an increase in performance when its data pool is increased. We also hypothesise that using incomplete or faulty trajectories might help CILO since it would not have so much data for all points in a trajectory, reducing overfit. Fine-tuning the exact number of expert trajectories requires some experimentation for each environment.

5.2 Ground-truth error over time

A concern for self-supervised IL methods is how to approximate pseudo-labels to ground-truth actions from the expert. However, approximating I^s samples to those from the experts is not always best. There might be samples that are between the experts' and I^{pre} that can help $\mathcal{M}$ smoothly close the gap between equally distributed and the ground-truth distribution [5]. Since CILO uses exploration to learn and this exploration mechanism relies on $\mathcal{M}$'s error, achieving lower error margins early might lead to less exploration and poorer results. It would therefore be better to have a consistent stream of new samples, to maintain the error marginally high but not a significant number of new samples since this could keep $\mathcal{M}$'s error too high or even collapse the network, *i.e.*, updating all weights drastically and requiring a higher learning rate.

Table 2. CILO's AER and $\mathcal{P}$ values for different Ant dataset sizes.

Trajectories	AER	$\mathcal{P}$
1	1003 ± 1999	0.18
10	$\mathbf{6091 \pm 801.2}$	**1.1**
100	6026 ± 725.86	1.09



(a) π_θ trained with a scheduler. (b) π^*_θ trained with no scheduler. (c) Signature difference in Ant-v3.

Figure 2. (a) and (b) show ground-truth error for $\mathcal{M}$ and $\mathcal{P}$. (c) shows the normalised difference between π_e, π_θ and random signatures: 0 is equivalent to expert, and $\geqslant 1$ means equal or worse than random policy signature.

Table 3 shows that using two different procedures and achieving two different policies with different error margins and weights yields similar results for the error margins in both methods but not performance and AER. Figure 2a shows the learning results (error and performance) for π_θ, which uses weight decay and a scheduler during training, and Figure 2b shows π^*_θ, with no weight decay and a learning rate scheduler. We observe that both policies achieve a similarly consistent error margin. However, when comparing the average performances in a single episode, π_θ achieves 29 more reward points with a lower variation. While using different strategies for classifying the action might help CILO with this behaviour, we use a similar topology to the one used by the models compared.

Table 3. $\mathcal{M}$'s ground-truth error and policy's AER and $\mathcal{P}$ for Ant.

Method	Error	AER	$\mathcal{P}$
π_θ	**0.0571**	**5610**	**1.0116**
π^*_θ	0.0556	5581	1.0065

5.3 *Signature approximation over time*

Given Definition 2, trajectories that are similar should be closer in the feature space, while those that do not share any states should be farther apart. Figure 2c shows the Manhattan distance between π_θ and π_ψ trajectories during the first 200 iterations. The difference is normalised between trajectories from random and expert agents. Hence, a difference greater than 1 means that the agent's signature path is farther from π_ψ than a random agent's. As expected, during early iterations, CILO produces episodes that are farther than the random agent since $\mathcal{M}$ has to learn state transitions before π_θ can learn how to behave in the environment. We see similar behaviour from the discriminator $\mathcal{D}$. In the initial iterations, it allows multiple trajectories to be appended to I^s due to its poor performance in discriminating between generated and expert trajectories. Once $\mathcal{D}$ learns to classify correctly, it is only 'fooled' by $\approx 18\%$ of trajectories.

As π_θ increases its performance, the distance between π_θ and π_ψ signatures decreases. Similarly, $\mathcal{D}$ has a harder time distinguishing from expert and π_θ. We observe that $\mathcal{D}$'s results are as expected. By allowing these early trajectories to append into I^s, which had not achieved any goals, it allows $\mathcal{M}$ to learn from samples outside its randomly distributed ones. Since it only allows a few samples, $\mathcal{M}$ does not stop to predict actions due to skewed samples. But as $\mathcal{D}$ improves its classification performance, it forces π_θ indirectly to be closer to the expert behaviour, therefore, achieving higher rewards. Using the gradient signal from $\mathcal{D}$ is likely to improve π_θ's performance further, but this adaptation would require the policy also to predict the next state, *e.g.*, in a mechanism similar to the one used in Edwards et al.'s work [4].

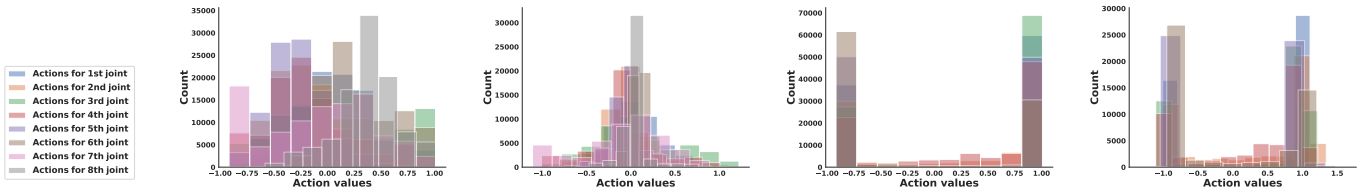
5.4 *Effects of Gaussian exploration*

Since we observed that CILO has a different behaviour for environments with different action distributions, we analyse Ant and HalfCheetah to understand the disparities between π_θ and π_ψ action predictions. Figure 3 displays all distributions for 50 trajectories from the expert and trained policies. Note that for these actions, π_θ is not using its exploration mechanism, that is, the policy is greedy. In all environments, the distribution from π_θ actions differs from the expert ones. However, we observe that π_θ actions have a higher intra-cluster variance than the expert ones. We believe this behaviour is due to CILO's exploration mechanism sampling from a Gaussian distribution, making it learn to have a higher variance around the average of an action (considering the error rate from Table 3). Therefore, the exploration mechanism makes it difficult to approximate distributions that do not follow this pattern, such as HalfCheetah.

We also note that CILO has more difficulty achieving better results in environments with sparse action distributions. If we compare Figures 3c and 3d, it is evident that CILO achieves actions near both limits, *i.e.*, -1 and 1; however, it has a harder time predicting actions near the limit. In contrast, although both distributions from Figures 3a and 3b are unequal, we observe a more concentrated action cluster around 0, which helps π_θ achieve better results. We see this behaviour as a limitation of CILO since selecting a new sampling distribution would require knowing beforehand how an expert behaves. However, we also hypothesise that training for a period without exploration and fine-tuning π_θ with $\mathcal{M}$'s pseudo-labels would minimise this impact. Further training π_θ with no exploration, we observe an increase in all environments, although not significantly.

5.5 *I^s size over time*

The use of the discriminator $\mathcal{D}$ allows CILO to start with fewer random samples since it appends samples on almost every iteration. However, increasing I^s on each epoch can create issues if the number of samples grows exponentially. Therefore, we plot in Figure 4 the size of I^s for each environment and epoch for the first 450 epochs. It is important to note that CILO usually reaches expert performance before its first 100 epochs. We observe that for most environments, CILO has a lower slope for appending I^{pos} into its dataset. This behaviour is excellent since it means that CILO is less likely to create data pool sizes that would transform it to be inefficient. Furthermore, when we consider that in Torabi et al.'s work [23], 5×10^5 transitions are needed to learn the inverse dynamic model (≈ 50 epochs),



(a) Ant (π_ψ) (b) Ant (π_θ) (c) HalfCheetah (π_ψ) (d) HalfCheetah (π_θ)

Figure 3. Distribution of expert actions for Ant and HalfCheetah environments.

this behaviour allows for less preparation when learning an agent. Nevertheless, Figure 4 also present two other behaviours.

For the InversePendulum environment, CILO gets almost no sample variation when compared to the other methods in the early stages. However, after approximately 150 epochs, π_θ yields trajectories more similar to the expert, which deteriorates $\mathcal{D}$ accuracy and increases I^s quite substantially. In this environment, we use this behaviour as a form of signal to stop since the reward does not improve. Figure 4 has an inset graph showing the first 150 epochs for the InvertedPendulum environment. In it, we observe during its first epochs, CILO appends samples in a lower rhythm.

For both HalfCheetah and Ant environments, we observe a linear pattern from the samples added into I^s. This behaviour is not desired, resulting in a training procedure that takes around 3 and 1.5 times longer to finish than all the other environments for HalfCheetah and Ant. To mitigate this problem, CILO could implement a forgetting mechanism to get rid of some samples in each epoch either by random selection or using the chronological order of insertion. However, it should not keep its initial sample pool size, considering it has a smaller dataset and changing it could make $\mathcal{M}$ susceptible to covariate shift. We hypothesise that adding samples up until an upper limit would be a better approach, eliminating samples from I^s in each epoch as needed to keep the pool size within the limit.

6 Related Work

The simplest form of imitation learning from observation is Behavioral Cloning (BC) [18], which treats imitation learning as a supervised problem. It uses samples (s_t, a, s_{t+1}) from an expert consisting of a state, action and subsequent state to learn how to approximate the agent's trajectory to the expert's. However, such an approach becomes costly for more complex scenarios, requiring more samples and information about the action effects on the environment. For example, solving Atari requires approximately 100 times more samples than CartPole. Generative Adversarial Imitation Learning (GAIL) [10] solves this issue by matching the state-action frequencies from the agent to those seen in the demonstrations, creating a policy with action distributions that are closer to the expert. GAIL uses adversarial training to discriminate state-actions either from the agent or the expert while minimising the difference between both.

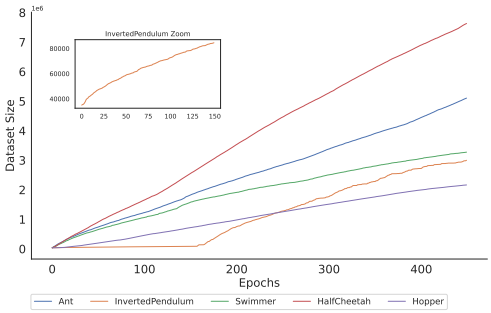


Figure 4. Size of $I^s \times$ epochs for all environments.

Recent self-supervised approaches [23, 6] that learn from observations use the expert's transitions $(s_t^{\pi_\psi}, s_{t+1}^{\pi_\psi})$ and leverage random transitions (s_t, a, s_{t+1}) to learn the inverse dynamics of the environment, and afterwards generate pseudo-labels for the expert's trajectories. Imitating Latent Policies from Observation (ILPO) [4] differs from such work by trying to estimate the probability of a latent action given a state. Within a limited number of environment steps, it remaps latent actions to corresponding ones. More recently, Off-Policy Learning from Observations (OPOLO) [25] uses a dual-form of the expectation function and an adversarial structure to achieve off-policy LfO. Model-Based Imitation Learning from Observation Alone (MobILE) [12] uses the same adversarial techniques, which rely on an objective discriminator coupled with exploration to diverge from its actions when far from the expert.

7 Conclusions and Future Work

In this paper, we proposed Continuous Imitation Learning from Observation (CILO), a new LfO method combining an exploration mechanism and path signatures. CILO (i) does not require prior domain knowledge or information about the expert's actions; (ii) has sample efficiency superior or equal to the state-of-the-art LfO alternatives; and (iii) approximates (sometimes surpassing) expert performance. CILO achieves these results due to two key contributions. Firstly, the use of a discriminator paired with path signatures, allows CILO to acquire more diverse state transition samples while increasing sample quality. Secondly, the exploration mechanism, which uses the model's error rate to sample from a normal distribution, allows for a more dynamic exploration of the environment. As a result, the exploration ratio decreases as the model learns to approximate from the ground-truth labels. More importantly, these two innovations are completely model-agnostic, allowing them to be used in other IL methods without requiring major changes. We would argue that the innovations we proposed pave the way for IL models that generalise better and require less expert training data.

Our next step is to investigate different exploration mechanisms to better fit the policy needs of specific environments. We would also like to experiment with different forms of adversarial learning to embed CILO's current discriminator into the policy loss function. Considering the path signatures are differentiable, it would be possible to backpropagate the gradients from the discriminator into the policy. This change would allow us to see if a direct signal from the enhanced loss function could improve the action prediction of the inverse dynamic model.

Acknowledgements

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org) and made possible via King's Computational Research, Engineering and Technology Environment (CREATE) [3].

References

[1] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, 1(5):834–846, Sep 1983.

[2] I. Chevyrev and A. Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

[3] K. C. L. e Research team. King's computational research, engineering and technology environment (create), 2023. URL https://doi.org/10.18742/rnvf-m076.

[4] A. D. Edwards, H. Sahni, Y. Schroecker, and C. L. Isbell. Imitating latent policies from observation. In *Proceedings of the 36th International Conference on Machine Learning*, ICML 2019, pages 1755–1763, 2019.

[5] N. Gavenski. Self-supervised imitation learning from observation. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2021.

[6] N. Gavenski, J. Monteiro, R. Granada, F. Meneguzzi, and R. C. Barros. Imitating unknown policies via exploration. In *International British Machine Vision Virtual Conference*, pages 1–8, 2020. URL https://www.bmvc2020-conference.com/conference/papers/paper_0774.html.

[7] N. Gavenski, J. Monteiro, A. Medronha, and R. C. Barros. How resilient are imitation learning methods to sub-optimal experts? In J. C. Xavier-Junior and R. A. Rios, editors, *Intelligent Systems*, pages 449–463, Cham, 2022. Springer International Publishing. ISBN 978-3-031-21689-3.

[8] N. Gavenski, M. Luck, and O. Rodrigues. Imitation learning datasets: A toolkit for creating datasets, training agents and benchmarking. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 2800–2802, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

[9] N. Gavenski, J. Monteiro, F. Meneguzzi, M. Luck, and O. Rodrigues. Explorative imitation learning: A path signature approach for continuous environments. *King's College Pure*, 2024. URL https://kclpure.kcl.ac.uk/portal/en/publications/explorative-imitation-learning-a-path-signature-approach-for-cont. Full version of this paper.

[10] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

[11] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys*, 50(2):21:1–21:35, 2017.

[12] R. Kidambi, J. Chang, and W. Sun. Mobile: Model-based imitation learning from observation alone. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28598–28611. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/f06048518ff8de2035363e00710c6a1d-Paper.pdf.

[13] P. Kidger and T. Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In *International Conference on Learning Representations*, 2021. https://github.com/patrick-kidger/signatory.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] H. M. Le and Y. Yue. Imitation learning tutorial, 2018. URL https://sites.google.com/view/icml2018-imitation-learning. ICML Presentation.

[16] J. Monteiro, N. Gavenski, R. Granada, F. Meneguzzi, and R. Barros. Augmented behavioral cloning from observation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[17] B. S. Pavse, F. Torabi, J. Hanna, G. Warnell, and P. Stone. RIDM: Reinforced inverse dynamics modeling for learning from a single observed demonstration. *IEEE Robotics and Automation Letters*, 5(4): 6262–6269, oct 2020.

[18] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Proceedings of the 1st Conference on Neural Information Processing Systems*, NIPS 1988, pages 305–313, 1988.

[19] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[20] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[21] R. S. Sutton and A. G. Barto. *Reinforcement learning: An Introduction*, volume 2. MIT press Cambridge, 2018.

[22] G. Swamy, S. Choudhury, J. A. Bagnell, and S. Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

[23] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 4950–4957, 2018.

[24] W. Yang, T. Lyons, H. Ni, C. Schmid, and L. Jin. Developing the path signature methodology and its application to landmark-based human action recognition. In *Stochastic Analysis, Filtering, and Stochastic Optimization*, pages 431–464. Springer, 2022.

[25] Z. Zhu, K. Lin, B. Dai, and J. Zhou. Off-policy imitation learning from observations. *Advances in Neural Information Processing Systems*, 33: 12402–12413, 2020.