VALUE FROM OBSERVATIONS: TOWARDS LARGE-SCALE IMITATION LEARNING VIA SELF-IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Imitation Learning from Observation (IfO) offers a powerful way to learn behaviors from large-scale, mixed-quality data. Unlike behavior cloning or offline reinforcement learning, IfO leverages action-free demonstrations and circumvents the need for costly action-labeled demonstrations or carefully crafted reward functions. However, current research focuses on idealized scenarios with tailored, bimodal-quality data distributions. This paper introduces a novel algorithm to learn from datasets with varying quality, moving closer to a paradigm in which imitation learning can be performed iteratively via self-improvement. Our method extends RL-based imitation learning to action-free demonstrations, using a value function to transfer information between expert and non-expert data. Through comprehensive evaluation, we delineate the relation between different data distributions and the applicability of algorithms and highlight the limitations of established methods. Our findings provide valuable insights for developing more robust and practical IfO techniques on a path to scalable behaviour learning.

1 INTRODUCTION

Bringing the cognitive capabilities of large vision and language models to embodied systems is at the forefront of many researchers' attention. Given the nature of the underlying data, robotics has been a popular target domain with a variety of deployed approaches, including prompting based algorithms (Ahn et al., 2022; Jiang et al., 2023; Kwon et al., 2024; Di Palo et al., 2023) and behavior cloning (BC) (Brohan et al., 2022; Bousmalis et al., 2024). We are motivated by a practical future where agents are trained on a large-scale dataset of language and video without requiring explicit action annotations or targeted task prompts. If required, the agent should further autonomously collect data to address knowledge gaps. Enabling learning from such data could overcome two key hurdles to scaling imitation learning: the static limitations of prompt engineering that may hinder generalization and the high cost of large-scale action-labeled demonstrations.

As a first step towards this setting, this paper studies the problem of training agents assuming access to two kinds of data: *expert* demonstrations without action labels, and *background* datasets which include actions but do not necessarily solve any task. So while the agent has access to some action data, it has not directly experienced the task being solved (it may only see solutions within the action-free *expert* data) and does not observe any external rewards. This setup can be directly applied to settings where action labels are difficult to obtain or which would benefit from cross-embodiment transfer, e.g. autonomous driving datasets (CAR) or data collected with the UMI-gripper (Chi et al., 2024). Expanding the *expert* data in our problem setting to include in-the-wild observations of humans solving tasks would allow scaling to cheaply available large-scale data.

Any applicable method has to address two key challenges: a) collect useful data for imitating the expert behaviour and b) learn a robust imitation policy from that data. Collecting data and inferring policies is at the heart of reinforcement learning (RL). Driven by a reward function, RL agents collect data and improve a policy based on the collected, possibly non-expert, experience. The reward serves as key signal for improvement and represents the main judge of *good* and *bad*. However, engineering a reward model can be time intensive (Tirumala et al., 2024; OpenAI et al., 2019; Lee et al., 2020) and cannot capture the breadth of possible contexts and tasks encountered at scale. We are thus interested in algorithms that acquire a notion of *good* and *bad* directly from data. This brings us to imitation learning, in particular Imitation learning from Observations (IfO) which addresses the problem of imitating expert behavior without requiring action annotations (Torabi et al., 2018a;b; Liu et al., 2018). To date, IfO remains an open research problem (Sikchi et al., 2024) and current methods do not match the maturity of related methods in RL, which have been shown to scale to large datasets and models (Springenberg et al., 2024; Chebotar et al., 2021). In contrast, IfO methods have been benchmarked only in small, single-domain settings with ad-hoc choices for data generation (Ma et al., 2022; Zolna et al., 2020; Sikchi et al., 2024). We here suggest a concrete benchmark that moves towards data compositions that better reflect the above large-scale vision.

Concretely, previous work has mainly been focused on offline training on mixed data where a significant amount of expert data was diluted in non-expert data (Ma et al., 2022; Zolna et al., 2020; Sikchi et al., 2024). Instead, we extend previous datasets (Fu et al., 2021) and collect data with a variety of policies of different quality and examine how much performance can be improved across this spectrum. We argue that this setup more accurately captures the scenario where an agent collects its own data for improvement and we underline this with a self-improvement experiment where an agent uses self-collected data to improve.

In addition to forming a new benchmark, we propose a simple, offline IfO method. Our algorithm adapts either SQIL (Reddy et al., 2020) or ORIL (Zolna et al., 2020), two RL-based imitation learning algorithms, to the action free setting. For the SQIL based variant, we simply assign a reward of 1 to *expert* data and a reward of 0 to *background* data before applying a value-function based offline RL method akin to AWR (Wang et al., 2016; Peng et al., 2019). The use of a value function instead of state-action value (aka a Q function) overcomes the lack of expert action annotations. In the second variant, we replace the 0-1 rewards with estimates from a learned discriminator as in Ho & Ermon (2016); Zolna et al. (2020). In both variants, we combine a simple reward function with a learned value-function which transfers expert knowledge from the unlabeled expert data onto the action labeled *background* data. We name our approach Value learning from Observations (VfO).

In summary, we suggest the combination of IfO and iterative self-improvement in order to approach large scale behavior learning from a novel angle and under realistic, scalable data collection assumptions. To this end we introduce a new offline benchmark that is more representative of said setting. We further propose a novel algorithm (VfO), in two variants, that adapts offline RL mechanisms to imitation learning from observations. With a broad set of experiments, we confirm the representative power of our benchmark, underline the competitiveness of our algorithm, and show initial positive results of IfO in conjunction with iterative self-improvement.

2 Related Work

The classical, straightforward approach to imitation learning is behaviour cloning (BC, Osa et al. (2018)), i.e., maximising the likelihood of actions in the dataset. However, this approach requires large numbers of optimal demonstrations. For this reason, a variety of methods have been developed to additionally benefit from suboptimal and other data sources by for instance extrapolating rewards from observations (Brown et al., 2019; Chen et al., 2021), imitation via IRL (Davchev et al., 2021a), or even by using videos from generative models (Bharadhwaj et al., 2024).

In online imitation learning, self-generated agent data represents the best data distribution to learn how to refine agent behaviour (Ross et al., 2011; Swamy et al., 2022; Lavington et al., 2022). Different methods have been proposed to apply the reinforcement learning formalism to address an imitation problem - from classical and deep maximum entropy inverse RL (Ziebart et al., 2008; Wulfmeier et al., 2015; Barnes et al., 2024) to computationally more efficient adversarial imitation learning (Ho & Ermon, 2016; Fu et al., 2017; Wulfmeier et al., 2017; Kostrikov et al., 2019a). When treating imitation as matching of agent visited transitions or divergence minimisation, further divergences have been explored (Ke et al., 2021; Ghasemipour et al., 2020). Transitioning from adversarial learning to more stable value function optimisation, SQIL removes the intermediary classifier from GAIL and instead uses a binary reward (Reddy et al., 2020). IQlearn (Garg et al., 2021) and ValueDICE (Kostrikov et al., 2020) enable transitioning from explicitly defined to implicitly learned rewards. Other non adversarial algorithms include PPIL (Viano et al., 2022), PWIL (Dadashi et al., 2021), and CSIL (Watson et al., 2024). However, online imitation learning can be costly in domains like robotics, is sometimes not even possible and doesn't benefit from existing data sources.

When online data generation is impractical, suboptimal offline datasets can provide an alternative. Practical and scalable algorithms can be derived when using either discriminator (Zolna et al., 2020) or optimal transport (Luo et al., 2023) based rewards together with offline RL. Value (Kim et al., 2022) and model-based approaches (Chang et al., 2021) expand the toolkit. IQLearn can further be shown to be equivalent to BC with dynamics-aware regularisation term (Wulfmeier et al., 2024).

The online and offline settings described above require access to high-quality demonstrations with action annotations, often only attainable via complex tele-operation settings in robotics. The extension towards action-free demonstrations opens considerable scope and has been the target of further methods. A separately trained inverse dynamics model can be applied to label action-free data, enabling behaviour cloning (Radosavovic et al., 2021; Torabi et al., 2018b). Learning rewards provides a path to instead relabel sub-optimal data for RL style optimisation (Eysenbach et al., 2021; Davchev et al., 2021b). Here, adversarial approaches present a common mechanism to learn rewards (Ho & Ermon, 2016). These can be adapted by controlling the discriminator input space, often benefiting from further regularisation (Zhu et al., 2020b; Liu et al., 2020). Value function based imitation methods enabled by inverse Bellman updates and dual formulations of the problem like SMODICE (Ma et al., 2022) and DILO (Sikchi et al., 2024), or variational formulations (Kostrikov et al., 2019b; Garg et al., 2021), bypass the often hard to optimize adversarial objectives and are related to our approach. While mathematically appealing, these methods can still be brittle and harder to scale to the real-world directly from raw observations (Al-Hafez et al., 2023; Watson et al., 2024). Instead, we base our value-based algorithm on a simple RL backbone which draws on decades of experience. We compare performance and show competitiveness against various baselines (including SMODICE and DILO) on a large set of experiments on different domains.

3 Method

3.1 Offline Imitation Learning From Observations

We consider learning in a dynamical system modelled as Markov decision process with states $s \in S$, actions $a \in A$, and dynamics $p(s_{t+1}|s_t, a_t)$. In order to learn useful behaviour, the agent has access to two sources of information: a dataset of *expert* state trajectories $\tau_E = (s_1, \dots, s_T) \in D_E$ without actions and a dataset of state-action trajectories from its own embodiment but of unspecified origin and quality $\tau_B = (s_1, a_1, \dots, a_{T-1}, s_T) \in D_B$. We will refer to the latter as *background* dataset. The agent's goal is to obtain a policy $\pi(a|s)$ that imitates the behaviour underlying the expert trajectories. To support scalability, we limit the use of further information (e.g. rewards, domain knowledge) and thus do also not expect the agent to outperform expert performance.

Given the lack of *expert* actions, the agent has to be able to leverage the *background* dataset to understand the dynamics, i.e., the relationship between actions and states. However, similar to prior work on inverse RL (Abbeel & Ng, 2004; Ziebart et al., 2008) as well as for the related problem of offline RL (Schweighofer et al., 2022; Hong et al., 2023), the quality and distribution of the *background* data plays an important role on the achievable performance. In previous work, different sources have been employed, such as agent replay data, a mixture of expert and non-expert data, or data collected with a suboptimal policy. Given that we are interested in an agent that can start from few assumptions and that should be able to leverage the data it collects, we focus on the suboptimal policy case. In order to generate a corresponding benchmark we suggest to train multiple policies using BC but vary the number of demonstrations provided. We then run these policies to collect multiple datasets of varying quality (see Section 4.1).

3.2 VfO: Value from Observation

We introduce a simple IfO method that can effectively learn from observations in the self-improvement setting at hand. For this purpose, we consider two variants of a value-function based approach that learns a state-value function from observations alone. In the first, we assign binary rewards (i.e., 1 for *expert* or 0 for *background*) to the data – thus adapting SQIL (Reddy et al., 2020) to our setting. In the second, we use a learned discriminator that performs a soft *expert* / *background* assignment of each state – thus adapting ORIL (Zolna et al., 2020) to our setting. Other imitation learning-based rewards such as Luo et al. (2023) could also be employed, but note that employing rewards that rely on prior knowledge such as when derived from goal states may impact generality.

Algorithm 1 Value from Observation (VfO)

```
Require: Expert dataset $D_E$, background dataset $D_B$, mixture parameter $\alpha$, temperature $\lambda$, discount $\gamma$, initial policy $\pi_0$, initial value $v_0$. Optional: discriminator $d(s) : S \mapsto [0, 1]$
  for $k \leftarrow 1$ to $K$ iterations do
      $r(s', z) = \begin{cases} d(s') & \text{if discriminator provided} \\ \mathbf{1}_E(z) & \text{otherwise} \end{cases}$      ▷ $\mathbf{1}_E$ is the expert indicator function.
      $L^v_k \leftarrow E_{(s,s',z)\sim(1-\alpha)D_E+\alpha D_B} \left(\gamma v_{k-1}(s') + r(s', z) - v_{k-1}(s)\right)^2$
      $L^\pi_k \leftarrow -E_{(s,s',a,z)\in D_B} \exp\left(\left(\gamma v_{k-1}(s') + r(s', z) - v_{k-1}(s)\right)/\lambda\right) \log(\pi_{k-1}(a|s))$
      $v_k \leftarrow \text{AdamUpdate}(v_{k-1}, L^v_k)$, $\pi_k \leftarrow \text{AdamUpdate}(\pi_{k-1}, L^\pi_k)$
  end for
```

As mentioned above, we resort to learning a state-value function for transferring knowledge from the *expert* data without action labels to the *background* data. We note that a state-action Q-value based offline RL approach cannot be applied in our setting due to a lack of signal on the *background* data: all transitions are labeled with a zero reward in the binary setting or potentially very small rewards in the learned discriminator setting. In contrast, if we apply an approach based on the state-value function v, policy evaluation is possible without knowing the action and can thus leverage a mixture of *expert* and *background* data.

First, we define a virtual policy $\bar{\pi}$ which mixes the *expert* and *background* data-generating processes at each transition:

$$\bar{\pi}(a|s) = p(z = E|s, \alpha)\pi_E(a|s) + p(z = B|s, \alpha)\pi_B(a|s) \tag{1}$$

with mixture coefficient α and where z denotes the latent indicating the origin of the data, either *expert* E or *background* B. This policy is equivalent to deciding at the beginning of an episode whether to follow the implicit expert or background policy. Note that the probability of using π_E or π_B is state-dependent, and will depend on the likelihood of reaching s under each of those policies. With discount factor γ and reward $r(s', z)$, we can define the temporal difference error of a value function for this policy:

$$L_v = E_{(s,s',z)\sim(1-\alpha)D_E+\alpha D_B}(v(s) - (r(s', z) + \gamma v(s')))^2. \tag{2}$$

Leaving aside – for a moment – how the reward can be obtained, we can then utilize this learned value function to find an improved policy by (exponentiated) advantage weighted regression (Peng et al., 2019; Wang et al., 2018) which amounts to weighted supervised learning on the *background* data (which enforces closeness to the policy that generated the data via a temperature λ) and yields the following policy loss:

$$L_\pi = -E_{(s,s',a)\in D_B} \exp((\gamma v(s') + r(s', z) - v(s))/\lambda) \log(\pi(a|s)). \tag{3}$$

An advantage of this presented scheme is its simplicity and use of well established offline RL methods, allowing for an efficient implementation while utilizing insights from many years of RL research such as the use of target networks and how to deal with terminations (see section 4). The full method is described in Algorithm 1 with the key differences to the offline RL setting marked in blue: the reward source and the mixture of *expert* and *background* data.

Binary demonstration-based rewards (VfO-bin) In the simplest setting we avoid any additional learning or estimation bias in the reward function by directly assigning a reward of 1 to *expert* transitions and a reward of 0 to *background* transitions in line with what has been proposed by Reddy et al. (2020). While this might seem trivial, it recovers what a perfect discriminator with infinite capacity would output and removes a layer of complexity. It follows the intuition that we may be able to leverage the value function directly for distinguishing good from bad states in the *background* data, rather than learning an intermediary reward function. We can make this notion more precise by realising that when we only provide a reward of 1 for the *expert* transitions and 0 otherwise, the learned value can be interpreted as:

$$\begin{aligned} v_{\bar{\pi}}(s_t) &= p(z_t = E|s_t) + \gamma E_{a\sim\bar{\pi}(\cdot|s_t), s_{t+1}\sim p(\cdot|s_t,a)} v_{\bar{\pi}}(s_{t+1}), \qquad (4) \\ &= E_{(s_{t+1}, s_{t+2}, \ldots)\sim\bar{\pi}} \sum_{i=0} \gamma^i p(z_{t+i} = E|s_{t+i}) = \sum_{i=0} \gamma^i p(z_{t+i} = E), \qquad (5) \end{aligned}$$

which is the cumulative discounted likelihood of futures states having been visited by the expert within the mixed dataset when starting in s_t. A policy that maximizes this cumulative return thus prioritises visiting expert states.

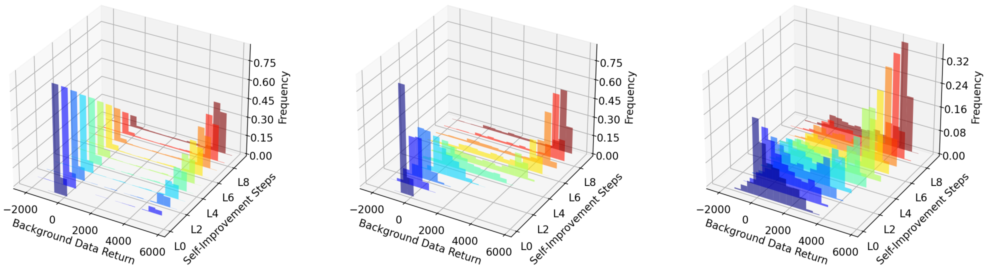
A different way to look at it, is that states that are contained in both the *expert* and *background* data will receive a positive and a negative reward signals. Given that policy improvement relies on weighted regression on the *background* data this is perfectly fine: actions that lead to states which are closer to the *expert* data will automatically receive a higher weight.

Discriminator-based rewards (VfO-disc) In a second setting, we consider learning a discriminator to represent the rewards (Ho & Ermon, 2016). When learning from observations, these usually learn to distinguish expert from non-expert states (Zolna et al., 2020; Ma et al., 2022) in order to derive a reward for learning a policy. We adopt the objective from ORIL (Zolna et al., 2020) and pre-train the discriminator by minimizing

$$L_d = E_{s \sim D_E}[-\log d(s)] + E_{s \sim D_B}[-\log(1 - d(s))], \tag{6}$$

where $d(s) \in [0, 1]$ is a binary classifier and the objective is akin to training a discriminator in generative adversarial learning (Goodfellow et al., 2014; Ho & Ermon, 2016). The discriminator output directly serves as reward similar to was done in Wulfmeier et al. (2017) using the Wasserstein-1 (or Earth-Mover) distance (Arjovsky et al., 2017).

4 EXPERIMENTS



Bimodal

Self-Improvement Proxy

Self-Improvement

Figure 1: Schematic visualisation of key differences between bimodal (left) and self-improvement (right) based data configurations. Even this simple visualisation intuits how different algorithmic properties benefit in each setting. Explicitly learning classifiers or discriminators between demonstrations and other data is intuitively easier with stronger split. The intermediate offline self-improvement benchmark serves as a good proxy and enables quick evaluation bypassing serial dependencies between self-improvement steps. The source of all dataset for individual training steps is described in Section 4 in further detail.

We aim to evaluate the two settings schematically shown in Figure 1: bimodal, where the *background* data exhibits a bimodal distribution with a clear gap; and self-improvement, with a more nuanced distribution without a clear expert mode. While the bimodal setting is commonly employed for evaluation in previous work we argue that the self-improvement setting can better serve as an offline proxy for the online iterative self-improvement that motivates our paper. We thus also investigate how representative this proxy is when compared to actual self-improvement style data collection on a subset of the simulation domains. The following questions guide our experiments:

- How do different data distributions affect the performance across algorithms and how effective is VfO? (Sections 4.3 and 4.4)?
- Does the offline self-improvement proxy correlate with full iterative self-improvement (Section 4.6)?
- Can VfO deal with complex inputs such as images (Section 4.5)?

4.1 Datasets and Benchmarks

Our experiments are based on two established benchmarks: D4RL (Fu et al., 2021) with many baseline results in the literature and Robomimic (Mandlekar et al., 2021), from which the robosuite (Zhu et al., 2020a) tasks in particular provide a more realistic challenge for our algorithms – and include tasks that require the ability to process image data. D4RL contains data from various OpenAI gym (Brockman et al., 2016) MuJoCo (Todorov et al., 2012) environments, from which we use the Ant, HalfCheetah, Hopper and Walker2D domains. The expert data for these environments comes from policies trained via RL. For each domain, there are 1000 expert demonstrations available; and we drop the action information to obtain the IfO setting. The Robomimic benchmark contains a variety of simulated and real robotics datasets of human demonstrations. From the simulated domains, we use Lift, PickPlaceCan and NutAssembly. For each of these tasks, there are 200 demonstrations.

The most realistic, scalable source of *background* robot data is agent-generated. Therefore in contrast to existing datasets – which are mostly bimodal (very high and low performance) – we target self-improvement as a data source. To emulate the sequential nature of different quality levels that we would expect during self-improvement, we introduce a proxy Self-Improvement Benchmark (SIBench) which uses data generated by a set of varying policies.

To produce data for this benchmark we use a set of policies with various levels of performance; which we train via BC with varying numbers of demonstrations. For each task τ we train a set of BC policies $\{\pi_\tau^d | d \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}, d \leq N_\tau\}$, where d indicates the number of demonstrations that the policy was trained on and N_τ is the number of demonstrations available. For each of these policies, we collect 1000 episodes, which we use as our *background* data[1]. This setting leads to considerably faster experimental iteration compared to running a full self-improvement experiment (since the data is pre-generated and fixed) and creates a consistent benchmark for fair comparisons; but comes at the cost of removing the data generation or exploration process from the analysis. Importantly, looking at the return distributions for all settings in Figure 1 we can observe that this proposed *background* data is qualitatively close to the data encountered in self-improvement. Further, the different *background* datasets will exhibit different levels of overlap with the *expert* data, ranging from a scenario where the *expert* data is mostly out of distribution to a regime where the *expert* data is contained in the *background* data.

For comparability with prior work and to investigate the added value of our evaluation scheme, we also constructed datasets of what we refer to as bimodal data composed of expert demonstrations and trajectories generated with a random policy (i.e., actions sampled from a uniform distribution). To obtain a more complete picture we sweep over the data mixture: we interpolate linearly from 1000 random demonstrations to 1000 expert demonstration.

Using simulated rollouts of the stochastic policy, we report returns and success rates[2] averaged over 5 seeds and over the last 1e5 training steps to reduce noise (the last 5e5 for Robomimic). We display most results as the difference between the average return observed in the data and the average return obtained from the policy to be evaluated. In contrast to simply reporting policy returns, we argue that this clearly visualises improvement of imitation learning algorithms and allows for better comparison against baselines via improved resolution.

4.2 Algorithms and Baselines

We compare our algorithm to a broad set of baselines including behaviour cloning (BC) on the *background* data, BCO (Torabi et al., 2018a), SMODICE (Ma et al., 2022), DILO (Sikchi et al., 2024)[3]. Finally, to provide a performance upper bound and thereby support usefulness of the *background* data, we also report results for Advantage-Weighted Regression (AWR; Peng et al., 2019) trained with ground-truth rewards available for the *background* data. This can be interpreted as an oracle algorithm that does not perform IfO (and is not directly comparable) but serves as indicator of what could be learned from the data. Please refer to Appendix A for further implementation details.

[1] The data will be publicly released following the manuscript's review decision.

[2] Note, that this may not be the best metrics for imitation learning performance, as it may provide a distorted view of imitation. E.g. success rate may be blind to any improvement as long as the task is not solved.

[3] Re-implemented in a shared code base for improved comparability. See Appendix B for details. Where possible we verify performance against reported results in prior work.

4.3 SIBENCH RESULTS

The results for the D4RL tasks of our SIBench data are shown in Figure 2 (see plots with absolute returns in Appendix E and SQIL with privileged actions in Appendix G). Both VfO-bin and VfO-disc perform well on the Ant and HalfCheetah tasks, getting close to the oracle AWR performance across the full spectrum of *background* data. All methods underperform on the Hopper task. On Walker2D, VfO-disc performs on par with AWR and better than VfO-bin. A possible explanation for VfO-bin's decreased performance could be its lack of immediate reward on the *background* data which could impact its performance on cyclic tasks. SMODICE and DILO perform poorly in comparison on all tasks, only improving on the data on scattered occasions. Generally in this more realistic settings VfO performs remarkably strong; close to 'oracle' performance in many settings despite it having to deal with lack of reward information and lack of action data on the expert demonstrations. The baselines SMODICE and DILO clearly underperform; we hypothesize that similar to residual gradient algorithms in RL (Baird, 1995), which do not make use of stop-gradients or target networks, the signal from the bellman residual may be very weak when there is significant overlap between good and bad trajectories, such as is the case for the *background* data here.

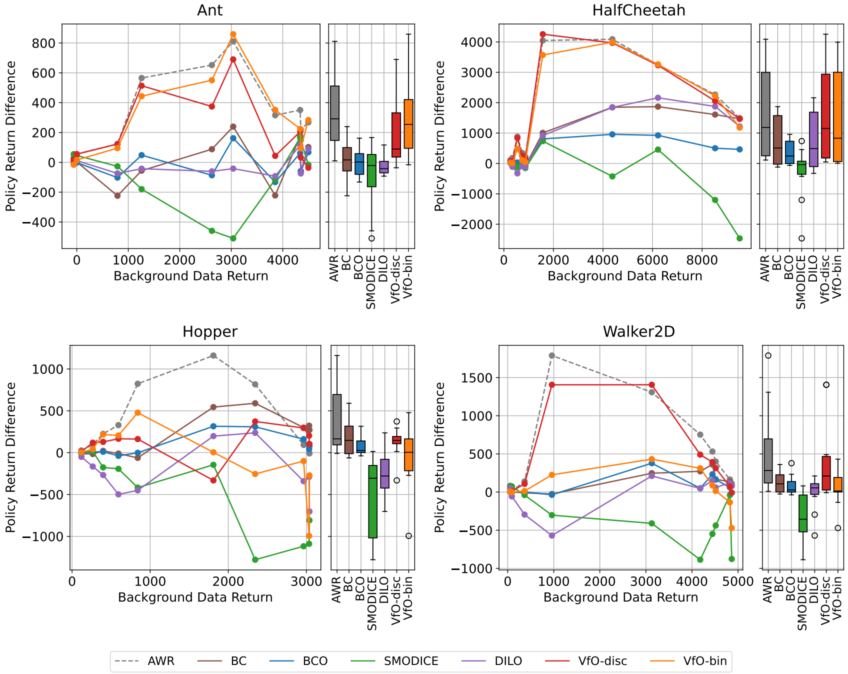


Figure 2: Difference in cumulative return of various algorithms on D4RL tasks using the SIBench data. We plot the average return in the background data against the difference in return (relative to the background data return) achieved by an algorithm. I.e. positive differences mean the policy produced by the algorithm is better than the policy that generated the background data. The boxplots average across the spectrum of background data. AWR, VfO-disc, VfO-bin all show good improvement across the spectrum of *background* data with the oracle AWR performing best. SMODICE and DILO only rarely improve on the data.

Figure 3 depicts the results for the Robomimic tasks of the SIBench data. Overall the results are less conclusive here and this could be related to how improvement is measured: Given that there is no dense reward for these tasks, we resort to success, which is much less indicative of learning progress, i.e. behavior could become more similar to the expert demonstration without higher success rate. Nevertheless, VfO-bin is able to yield positive improvement across all tasks and performs on par or better than AWR, while VfO-disc, SMODICE, and DILO perform worse. Comparing AWR against VfO effectively also compares the underlying driving sources of information, i.e. reward annotations against demonstration. In a scenario with sparse rewards, such as for Robomimic, it is entirely possible that relying on demonstrations allows for better performance.

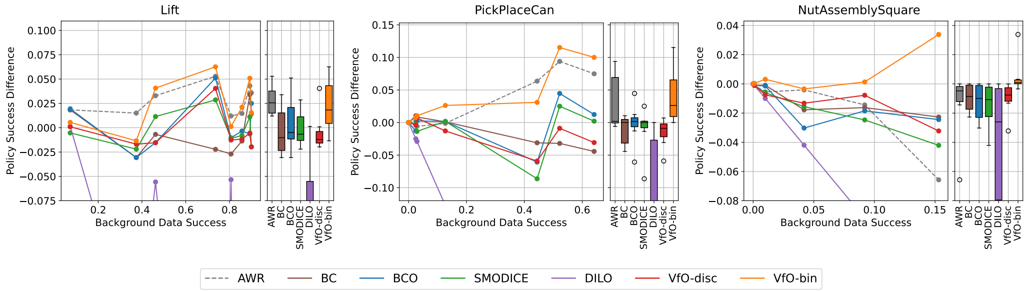


Figure 3: Difference in success of various algorithms on Robomimic tasks using the SIBench data. AWR and VfO-bin mostly yield good improvement. DILO, SMODICE, and VfO-disc have more troubles generating improvement.

4.4 Bimodal Results

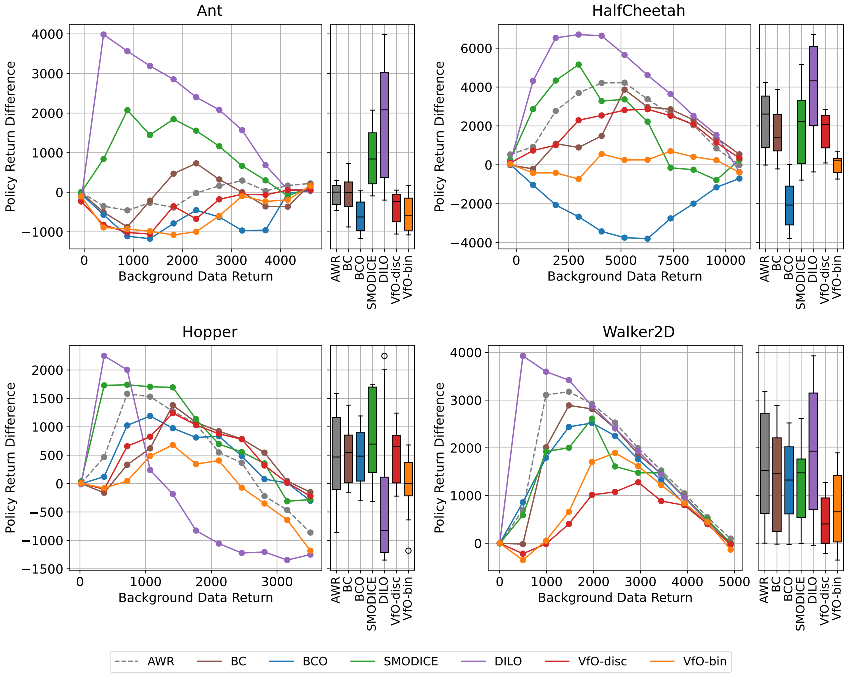


Figure 4: Difference in cumulative return of various algorithms on D4RL tasks using the bimodal data. As reported in previous work, SMODICE and DILO exhibit strong improvement when the data is composed of a little amount of expert demonstrations. VfO-bin and VfO-disc both underperform in that case.

To investigate how SIBench data differs from existing data mixtures and to enable comparability with previous work, we report results for the bimodal D4RL data in Figure 4. DILO and SMODICE exhibit strong improvement, particularly when the data is composed of relatively few expert demonstrations. In most cases, DILO reaches expert performance for the third data point (200 expert demonstrations) which is in line with results from Sikchi et al. (2024). Improvement then degrades smoothly with increasing data quality as there is less room for improvement.

VfO-bin and VfO-disc underperform when compared to BC. The improvement of BC itself can be attributed to the bimodal state distribution and to dynamic effects that can lead BC to pick the more consistent underlying policy (Zhang et al., 2024). These results are also in line with our hypothesis that bimodal data of this type turns imitation learning into a filtering problem of separating the good from bad trajectories and may not effectively measure the properties of imitation learning algorithms that matter in practical self-improvement scenarios. Instead, the mostly distinct *expert* and *background* state distributions render it more important to pick the right action when the distributions bifurcate. VfO likely struggles to do so because the learned values are not sufficiently discriminative. However, lowering temperatures to increase the effect incurs instabilities.

4.5 Vision-based Results

We also investigate the ability to learn from image observations (see Figure 5), as is often required in real-world robotics applications. For this we report results on the Robomimic tasks using the SIBench image data. While improvement is more difficult to measure here, we can observe some improvement in the Lift domain for AWR and VfO-bin, again highlighting the fact that our simple VfO scheme is a strong algorithm even in high-dimensional, difficult settings.

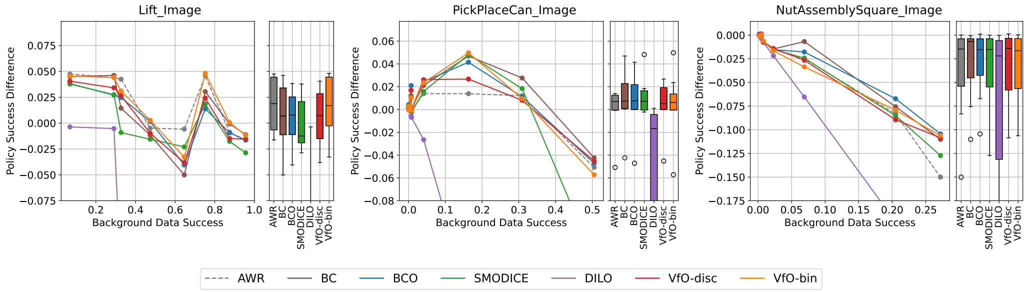


Figure 5: Difference in success of various algorithms on Robomimic tasks using the SIBench image data. In this difficult setup both VfO-bin and AWR manage to achieve some improvement on Lift.

4.6 Iterative Self-Improvement Results

To confirm the validity of our SIBench proxy, we run self-improvement experiments akin to what we envisioned in the introduction. In these experiments after learning an initial policy from a seed dataset, we collect 1,000 episodes to form a new dataset for the next learning iteration (and then repeat this process in an improvement loop). We perform 20 iterations in total and seed with data with bad but non-zero performance in order to avoid regions with low signal-to-noise ratios. In order to observe correlation with SIBench we pick VfO-bin as our method to benchmark (for which we expect good performance) and use SMODICE as a baseline. We again run AWR (assuming rewards on all data) to compare to a form of oracle performance. Further baselines are shown in Appendix F.

Figures 6 and 7 plot policy performance against the performance of the input data for each self-improvement iteration. A saw-tooth pattern is observed when consistent improvement is achieved during iterative self-improvement: The achieved performance of one iteration (y-axis) is used as base performance for the next iteration (x-axis), thus the projections onto the diagonal. A box-tooth pattern is observed when performance increases and decreases alternate. This might be caused by oscillating effects, such as the temporary emergence of stationary regions: During one iteration a

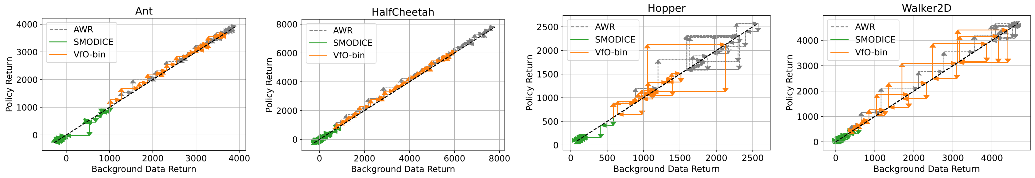


Figure 6: Self-improvement experiments for D4RL tasks. We evaluate VfO-bin, SMODICE, and AWR with ground-truth rewards. Starting with low-performance initial policies, we generate data to train the next iteration of policies for each algorithm and iterate. The policy return (averaged over 1000 episodes) at each iteration is projected on the diagonal and used as *background* data for the next step. Both AWR with ground-truth rewards and VfO-bin lead to strong results.

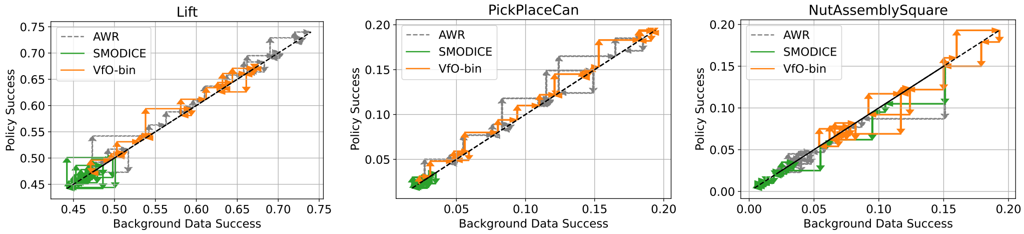


Figure 7: Self-improvement experiments for Robomimic tasks. Starting from different initial performances we can observe whether performance increases or not. Results are mixed, VfO-bin is good on two out of three tasks. Given AWR's dependency on informative rewards (here: sparse), VfO-bin can slightly outperform offline RL in this setting.

policy might decide to remain stationary within a region of high expert density. This would however likely be resolved during the next self-improvement iteration due to an over-proportional visitation of these states.

The plots confirm that whenever we see performance improvement in SIBench we also attain self-improvement when collecting data online, confirming the representative power of our offline proxy. For Ant, HalfCheetah, and Walker2D the self-improvement has not converged within the alotted iterations. Interestingly, for Hopper VfO-bin converges roughly where the zero-crossing of the performance lies in SIBench between 1'500 and 2'000. Additionally, self-improvement is also obtained for Robomimic when starting from data that shows positive SIBench improvement.

Additionally, VfO-bin clearly outperforms SMODICE in this setting and, remarkably, obtains performance similar to the AWR oracle in all settings. We want to highlight that this is a highly non trivial result, bootstrapping imitation learning to mastery via self-collection starting from low signal (near random data) is an open problem in imitation learning (see e.g. Sun et al. (2017)).

5 Limitations and Future Opportunities

While a key application of IfO targets transfer from direct human provided (third person) demonstrations of a task rather than trained first person control, all presented experiments are limited to consistent embodiment between demonstrations and additional data source. Intermediate steps in this direction might utilize state estimation techniques to map correspondences between robot and human states (Luo et al., 2024), but the final goal should remain to exploit existing semantic understanding in pre-trained vision-language and other foundation models (Stone et al., 2023; Yuan et al., 2024; Zitkovich et al., 2023; Wulfmeier et al., 2023; Majumdar et al., 2023). The considerable computational requirements of such models renders the iterative offline learning setting we describe in Section 4.1 more tractable than the pure online learning setting. Self-generated data remains however the most targeted path to obtain the most relevant *background* data for agent training.

6 Conclusions

Imitation learning from observation has the potential to become a principal component of large-scale behaviour learning. We advance this paradigm by suggesting the use of IfO in conjunction with self-improvement. We provide a novel offline benchmark, which we find to be much more representative of this self-improvement setting when compared to existing benchmarks. We also present a simple algorithm (VfO) that builds on ideas from SQIL, ORIL, and AWR to effectively train agents by relying on offline reinforcement learning as the mechanism to learn to imitate. Remarkably, across nearly all experiments in our analysis, VfO is competitive with RL from ground-truth rewards when using just a few action-free trajectories as the defining good behaviour. VfO, and IfO in general, provides an efficient path to real-world RL, where reward function design often becomes the restricting factor.

REFERENCES

Carla leaderboard. http://leaderboard.carla.org.

P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on Machine learning*, 2004.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022. URL https://arxiv.org/abs/2204.01691.

Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, and Jan Peters. Ls-iq: Implicit reward regularization for inverse reinforcement learning. In *ICLR*, 2023. URL https://openreview.net/pdf?id=o3Q4m8jg4BR.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine learning proceedings 1995*, pp. 30–37. Elsevier, 1995.

Matt Barnes, Matthew Abueg, Oliver F. Lange, Matt Deeds, Jason Trader, Denali Molitor, Markus Wulfmeier, and Shawn O'Banion. Massively scalable inverse reinforcement learning in google maps. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=z3L59iGALM.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation, 2024. URL https://arxiv.org/abs/2409.16283.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Fernandes Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott Reed, Sergio Gómez Colmenarejo, Jonathan Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Thomas Rothörl, Jose Enrique Chen, Yusuf Aytar, David Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=vsCpILiWHu.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *CoRR*, abs/1606.01540, 2016.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Jonathan Daniel Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=7PkfLkyLMRM`.

Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conference on robot learning*, pp. 1262–1277. PMLR, 2021.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Oilvier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. *ArXiv*, abs/2406.01660, 2024.

Robert Dadashi, Leonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=TtYSU29zgR`.

Todor Davchev, Sarah Bechtle, Subramanian Ramamoorthy, and Franziska Meier. Learning time-invariant reward functions through model-based inverse reinforcement learning. *arXiv preprint arXiv:2107.03186*, 2021a.

Todor Davchev, Oleg O. Sushkov, Jean-Baptiste Regli, Stefan Schaal, Yusuf Aytar, Markus Wulfmeier, and Jonathan Scholz. Wish you were here: Hindsight goal selection for long-horizon dexterous manipulation. *ArXiv*, abs/2112.00597, 2021b.

Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards a unified agent with foundation models. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=VXeoK3fJZhW`.

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

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL `https://arxiv.org/abs/2004.07219`.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-learn: Inverse soft-q learning for imitation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=Aeo-xqtb5p`.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36:4985–5009, 2023.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *ArXiv*, abs/2210.11610, 2022.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *Fortieth International Conference on Machine Learning*, 2023.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pp. 313–329. Springer, 2021.

Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BrPdX1bDZkQ.

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. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=Hk4fpoA5Km.

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

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyg-JC4FDr.

Teyun Kwon, Norman Di Palo, and Edward Johns. Language models as zero-shot trajectory generators. *IEEE Robotics and Automation Letters*, 9(7):6728–6735, 2024. doi: 10.1109/LRA.2024.3410155.

Thomas Lampe, Abbas Abdolmaleki, Sarah Bechtle, Sandy Huang, Jost Tobias Springenberg, Michael Bloesch, Oliver Groth, Roland Hafner, Tim Hertweck, Michael Neunert, Markus Wulfmeier, Jingwei Zhang, Francesco Nori, Nicolas Manfred Otto Heess, and Martin A. Riedmiller. Mastering stacking of diverse shapes with large-scale iterative reinforcement learning on real robots. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7772–7779, 2023.

Jonathan Wilder Lavington, Sharan Vaswani, and Mark Schmidt. Improved policy optimization for online imitation learning. In *Conference on Lifelong Learning Agents*, pp. 1146–1173. PMLR, 2022.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

Fangchen Liu, Zhan Ling, Tongzhou Mu, and Hao Su. State alignment-based imitation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rylrdxHFDr.

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, 2018. doi: 10.1109/ICRA.2018.8462901.

Jianlan Luo, Perry Dong, Yuexiang Zhai, Yi Ma, and Sergey Levine. RLIF: Interactive imitation learning as reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oLLZhbBSOU.

Yicheng Luo, Samuel Cohen, Edward Grefenstette, Marc Peter Deisenroth, et al. Optimal transport for offline imitation learning. In *11th International Conference on Learning Representations (ICLR)*, 2023.

Yecheng Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14639–14663. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ma22a.html.

Yecheng Jason Ma, Joey Hejna, Ayzaan Wahid, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, Jonathan Tompson, Osbert Bastani, Dinesh Jayaraman, Wenhao Yu, Tingnan Zhang, Dorsa Sadigh, and Fei Xia. Vision language models are in-context value learners. 2024. URL https://api.semanticscholar.org/CorpusID:273877849.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.

Liyuan Mao, Haoran Xu, Weinan Zhang, and Xianyuan Zhan. Odice: Revealing the mystery of distribution correction estimation via orthogonal-gradient update. *arXiv preprint arXiv:2402.00348*, 2024.

Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Shane Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *ArXiv*, abs/2006.03647, 2020.

OpenAI, :, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Jozefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Ponde de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning, 2019.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL https://arxiv.org/abs/1910.00177.

Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto, and Jitendra Malik. State-only imitation learning for dexterous manipulation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7865–7871, 2021. doi: 10.1109/IROS51168.2021.9636557.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. {SQIL}: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xKd24twB`.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Martin Riedmiller, Jost Tobias Springenberg, Roland Hafner, and Nicolas Heess. Collect & infer-a fresh look at data-efficient reinforcement learning. In *Conference on Robot Learning*, pp. 1736–1744. PMLR, 2022.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Kajetan Schweighofer, Marius-constantin Dinu, Andreas Radler, Markus Hofmarcher, Vihang Prakash Patil, Angela Bitto-nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. A dataset perspective on offline reinforcement learning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 470–517. PMLR, 22–24 Aug 2022. URL `https://proceedings.mlr.press/v199/schweighofer22a.html`.

Harshit Sikchi, Caleb Chuck, Amy Zhang, and Scott Niekum. A dual approach to imitation learning from observations with offline datasets, 2024. URL `https://arxiv.org/abs/2406.08805`.

Jost Tobias Springenberg, Abbas Abdolmaleki, Jingwei Zhang, Oliver Groth, Michael Bloesch, Thomas Lampe, Philemon Brakel, Sarah Maria Elisabeth Bechtle, Steven Kapturowski, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Offline actor-critic reinforcement learning scales to large models. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=tl2qmO5kpD`.

Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3309–3318. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/sun17d.html`.

Gokul Swamy, Nived Rajaraman, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, Jiantao Jiao, and Kannan Ramchandran. Minimax optimal online imitation learning via replay estimation, 2022.

Dhruva Tirumala, Markus Wulfmeier, Ben Moran, Sandy Huang, Jan Humplik, Guy Lever, Tuomas Haarnoja, Leonard Hasenclever, Arunkumar Byravan, Nathan Batchelor, Neil Sreendra, Kushal Patel, Marlon Gwira, Francesco Nori, Martin A. Riedmiller, and Nicolas Manfred Otto Heess. Learning robot soccer from egocentric vision with deep reinforcement learning. *ArXiv*, abs/2405.02425, 2024. URL `https://api.semanticscholar.org/CorpusID:269605913`.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 4950–4957. AAAI Press, 2018a. ISBN 9780999241127.

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

Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4iEoOIQ7nL.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn, 2016.

Qing Wang, Jiechao Xiong, Lei Han, peng sun, Han Liu, and Tong Zhang. Exponentially weighted imitation learning for batched historical data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/4aec1b3435c52abbdf8334ea0e7141e0-Paper.pdf.

Joe Watson, Sandy Huang, and Nicolas Heess. Coherent soft imitation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

Markus Wulfmeier, Ingmar Posner, and P. Abbeel. Mutual alignment transfer learning. *ArXiv*, abs/1707.07907, 2017.

Markus Wulfmeier, Arunkumar Byravan, Sarah Bechtle, Karol Hausman, and Nicolas Manfred Otto Heess. Foundations for transfer in reinforcement learning: A taxonomy of knowledge modalities. *ArXiv*, abs/2312.01939, 2023. URL https://api.semanticscholar.org/CorpusID:265609417.

Markus Wulfmeier, Michael Bloesch, Nino Vieillard, Arun Ahuja, Jorg Bornschein, Sandy Huang, Artem Sokolov, Matt Barnes, Guillaume Desjardins, Alex Bewley, Sarah Maria Elisabeth Bechtle, Jost Tobias Springenberg, Nikola Momchev, Olivier Bachem, Matthieu Geist, and Martin Riedmiller. Imitating language via scalable inverse reinforcement learning, 2024. URL https://arxiv.org/abs/2409.01369.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Edwin Zhang, Vincent Zhu, Naomi Saphra, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham M. Kakade, and Eran Malach. Transcendence: Generative models can outperform the experts that train them, 2024. URL https://arxiv.org/abs/2406.11741.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020a.

Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in neural information processing systems*, 33:12402–12413, 2020b.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

Konrad Zolna, Alexander Novikov, Ksenia Konyushkova, Caglar Gulcehre, Ziyu Wang, Yusuf Aytar, Misha Denil, Nando de Freitas, and Scott Reed. Offline learning from demonstrations and unlabeled experience, 2020. URL https://arxiv.org/abs/2011.13885.

A Implementation Details

The policies are simple MLPs with two hidden layers of either 512 units for D4RL or 1024 units for Robomimic. For the VfO algorithms we set the temperature parameter λ to 1 on the D4RL tasks and to 0.1 on the Robomimic tasks (see Appendix D for an overview of parameters and a sensitivity analysis). For the methods that learn value functions, we use a target network that gets updated every 200 steps. Some additional implementation details are worth pointing out: We use a multi-scale encoder similar to the one used in the Perceiver Actor-Critic model (Springenberg et al., 2024) that circumvents issues with saturation of non-linearities or insensitivity to lower amplitude signals. This setting reduces dependence on exact input normalisation and can simplify later extensions to multi-task scenarios. Instead of continuous action predictions, we discretize the actions in 101 uniformly spaced bins for which we learn a categorical distribution as is common for recent transformer architectures in control domains (Reed et al., 2022). However, we apply a Gaussian kernel on the last layer to provide sufficient inductive bias in the low data regime. For tasks with terminations, we further bootstrap the values by assuming the agent to continue receiving the same reward, i.e. $v(s_t) = \frac{r_t}{1-\gamma}$. This increases the effect of terminating states and improved performance on terminating tasks. All algorithms are trained for 1e6 learning steps.

B Baselines Details

BCO For the BCO (Torabi et al., 2018a) baseline, we train an inverse dynamics model to predict $p(a_t|s_t, s_{t+1})$ on the background data and subsequently use this to label the demonstrations. After this, regular BC learning is done on a $50/50$ mixture of the background data and the now action annotated demonstrations. To implement this, we used the same architecture for the dynamics model as we used for the policies in our experiments.

SMODICE Like in the paper that proposes SMODICE (Ma et al., 2022), we first train a discriminator network to distinguish between the expert and background data based on the observations/states. We apply early stopping at 10000 steps as we found this to be beneficial in preliminary experiments. Subsequently, we learn the value function as suggested and derive weights for weighted BC. Unlike the original paper, we don't apply entropy regularization to policy and don't apply gradient penalization to the discriminator. The SMODICE discriminator has the same architecture as the discriminator for VfO-disc.

DILO This algorithm learns a state-state value function that takes as input *two* adjacent states and does policy improvement via AWR. We implemented the same loss as in Sikchi et al. (2024). Unlike in that work, we found the orthogonal gradient method from Mao et al. (2024) to lead to worse learning stability than simply using the *true-gradient* update in which no target networks or stop-gradients are used. Despite not using the orthogonal gradient update, DILO was still the strongest baseline for the bimodal data. We hypothesize that this difference in results could be due to architectural differences like the discretized actions and multi-scale encoder that we used. For the AWR part of the algorithm we use a temperature of 10 – note that this would be a setting of 0.1 in the notation of Sikchi et al. (2024) where the parameter τ is the inverse of our temperature parameter λ. We found that increasing the temperature further led to more stable results on the Robomimic but at the cost of essentially turning the algorithm into BC. Otherwise, the settings for this baseline are the same as for the other methods. At the time of writing, the code for DILO is not available yet, so there could be more subtle implementation differences.

C Extended Related Work

In addition to the areas described in the main paper, our work strongly relates to and directly builds on research on self-improvement including fundamental reinforcement learning research. The classical idea to have a policy generate its own data to learn and adapt has a long-standing history in RL (Wulfmeier et al., 2023). More recently various works explicitly split the data generation and learning processes (Riedmiller et al., 2022) via model-based (Matsushima et al., 2020) and model-free RL approaches (Lampe et al., 2023; Bousmalis et al., 2024; Springenberg et al., 2024). While many of these works rely on externally defined reward functions, related signals, or vision-language models as reward sources (Ma et al., 2024), ours directly uses demonstration data to define optimal behaviour (Abbeel & Ng, 2004). Self-improvement research has further gained strong relevance for other foundation model applications such as language modelling (Huang et al., 2022; Choi et al., 2024).

D Hyperparameters

Table 1 provides an overview of the employed parameters. Further Figures 8 and 9 provide hyperparameter ablation for the temperature λ and the mixing parameter α. They confirm that for both there is a wide range of parameters that enable improvement.

Hyperparameter	Value
learning rate	3e-4
batch size	256
MLP layers D4RL	(512, 512)
MLP layers Robomimic	(1024, 1024)
target network update period	200
weight decay D4RL	0
weight decay Robomimic	0.1
λ (temperature) D4RL	1.0
λ (temperature) Robomimic	0.1
α (mixture ratio)	0.5
γ (discount ratio)	0.99

Table 1: Shared Algorithmic Hyperparameters

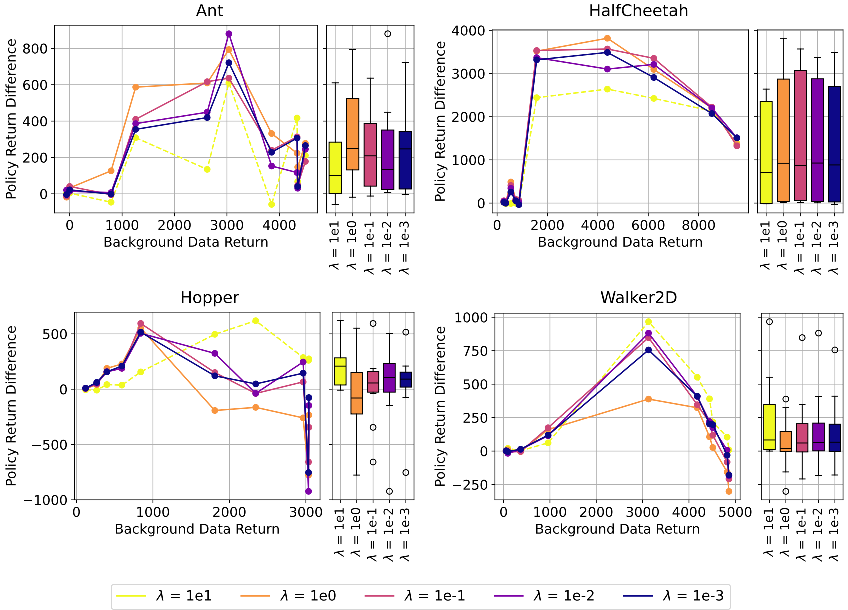


Figure 8: Difference in cumulative return of VfO-bin on D4RL tasks using the SIBench data for different temperature parameters and mixing parameter 0.5. We plot the average return in the background data against the return of the trained policy. We can observe a fairly wide range of hyperparameter settings leading to improvement.

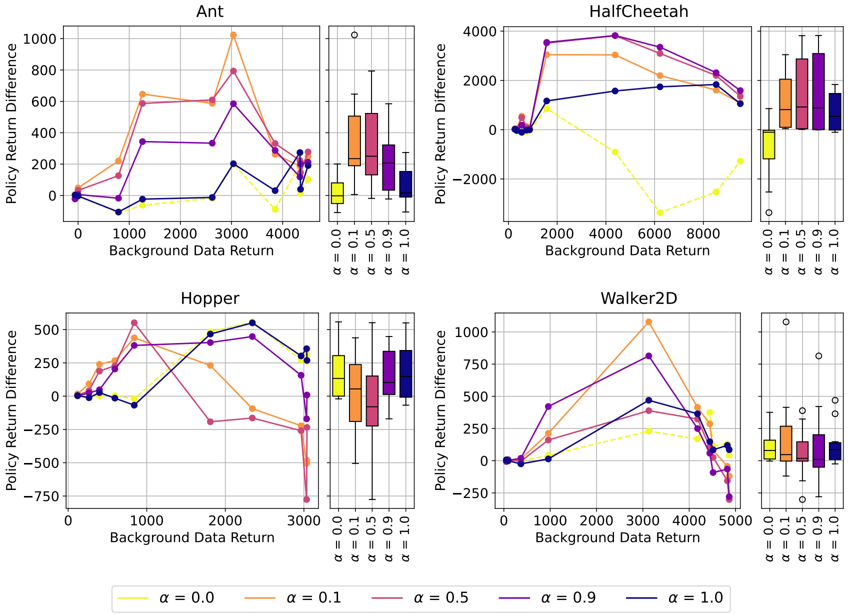


Figure 9: Difference in cumulative return of VfO-bin on D4RL tasks using the SIBench data for different mixing parameters and temperature 1.0. We plot the average return in the background data against the return of the trained policy. Except for Hopper, we can observe that picking a parameter between 0.1 and 0.9 yields consistent improvement.

E Absolute Return Plots

The relative plots in Figures 2 to 5 are not very common. We thus provide the absolute counterparts in Figures 10 to 13.

F Further Self-Improvement Baselines

We ran further self-improvement baselines including DILO, BC, and BCO. Figure 14 shows the average returns against the self-improvement iteration. It further confirms that methods bad performance on SIBench are less suited to attain positive self-improvement.

G Offline SQIL with Privileged Expert Actions

Figures 15 and 16 show results for our offline implementation of SQIL with privileged access to expert actions. We provide hyperparameter ablation for the temperature λ and the mixing parameter α and can observe that a higher mixing parameter is required to avoid overfitting on the scarce expert actions.

H Improvement Plots with Standard deviations

Figure 17 shows the mean and standard deviation of differences in cumulative return of AWR and VfO-disc on D4RL tasks using the SIBench data. Except for Hopper, VfO-disc achieves consistent improvement and exhibits a variance similar to that of AWR.

We further also provide results with an updated evaluation procedure. Here, only the final model weights from each training run are evaluated, using 1000 simulated rollouts each. We keep training with 5 different seeds to capture the distribution of average returns and report mean and standard deviation across training seeds. These are plotted in Figure 18. Adding more rollouts reduces the

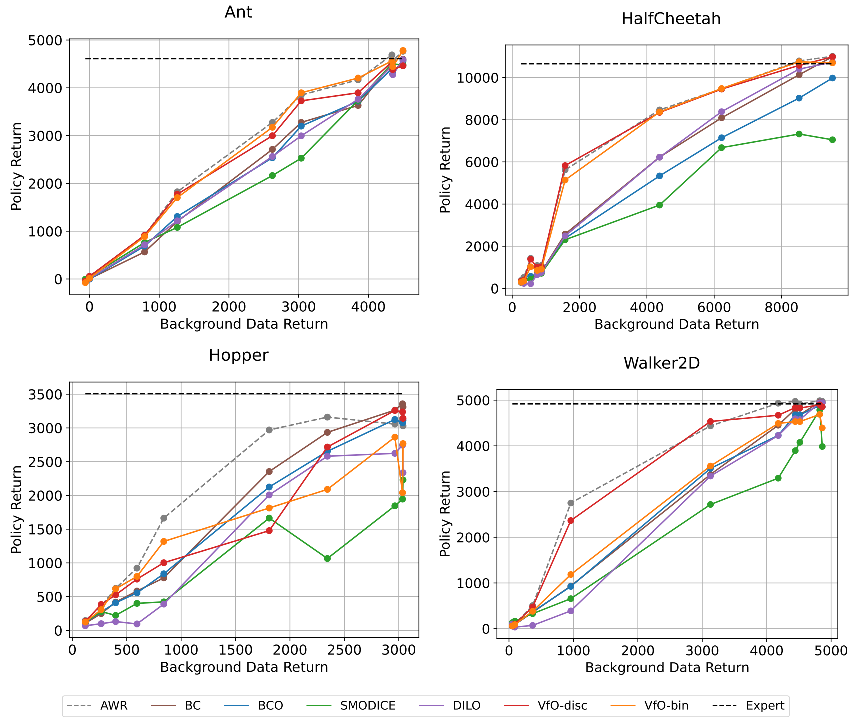


Figure 10: Cumulative return of various algorithms on D4RL tasks using the SIBench data. We plot the average return in the background data against the return of the trained policy. AWR, VfO-disc, VfO-bin all show good improvement across the spectrum of *background* data with the oracle AWR performing best. The corresponding relative plots can be seen in Figure 2.

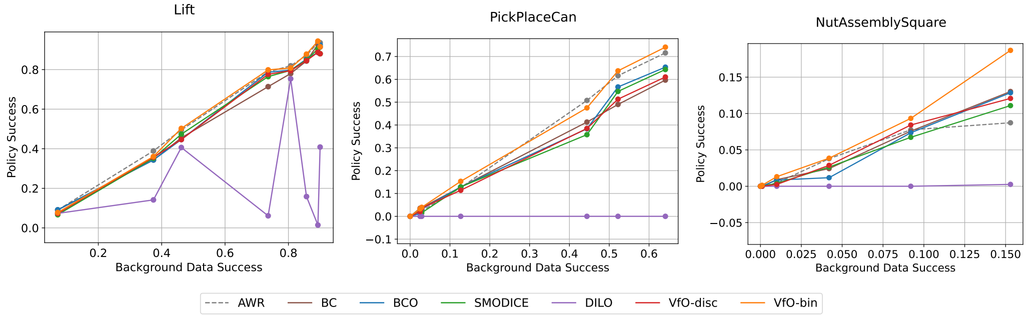


Figure 11: Success of various algorithms on Robomimic tasks using the SIBench data. As the absolute performance range is considerably larger than the differences due to highly different initial data quality, relative rankings require a closer look. AWR and VfO-bin mostly yield good improvement. The corresponding relative plots can be seen in Figure 3.

variance of the results and we can now claim that most VfO model weights attain good performance improvement.

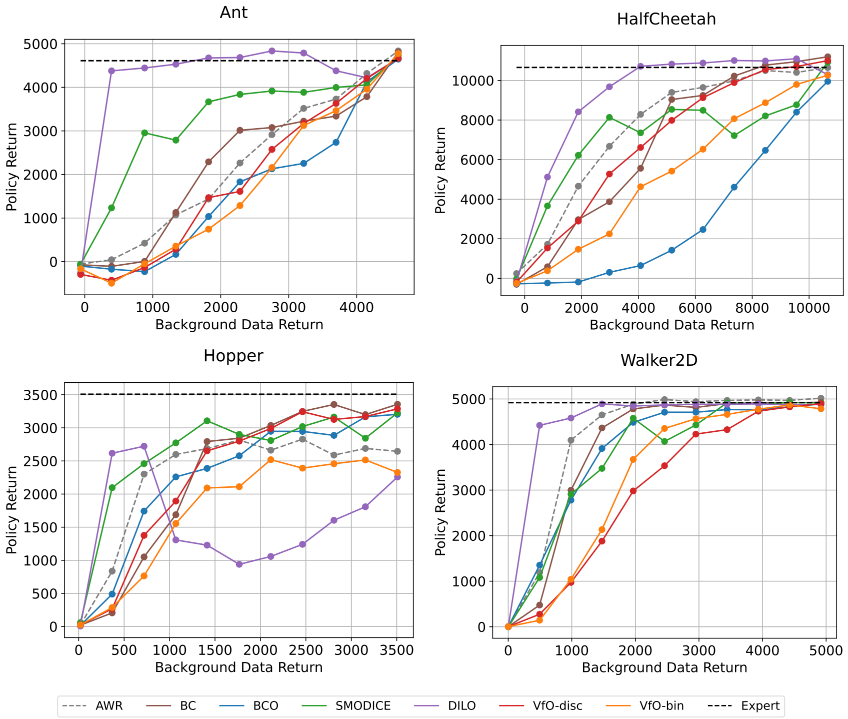


Figure 12: Cumulative return of various algorithms on D4RL tasks using the bimodal data. As reported in previous work, SMODICE and DILO exhibit strong improvement when the data is composed of a little amount of expert demonstrations. The corresponding relative plots can be seen in Figure 4.

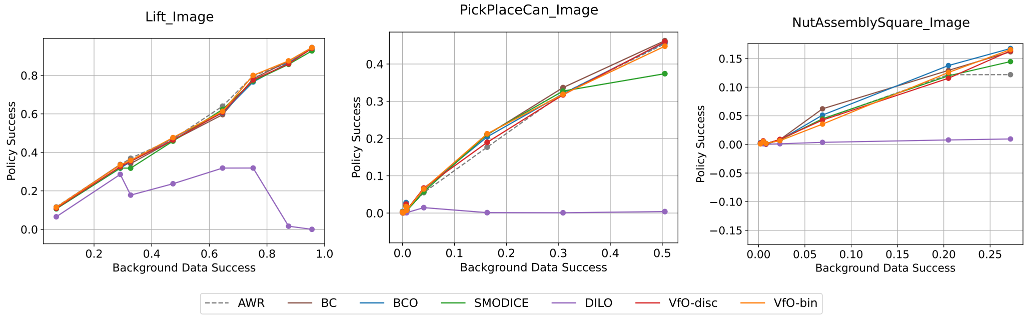


Figure 13: Success of various algorithms on Robomimic tasks using the SIBench image data. Improvement is difficult to discern in these plots. The corresponding relative plots can be seen in Figure 5.

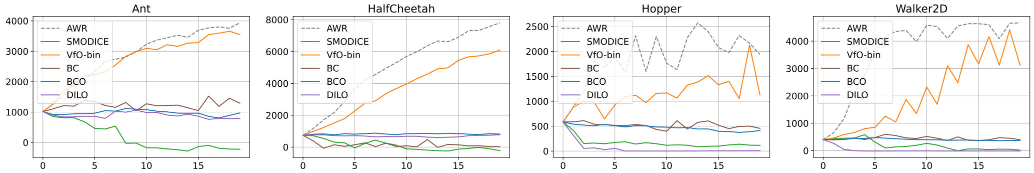


Figure 14: Self-improvement experiments for D4RL tasks. We evaluate VfO-bin, SMODICE, DILO, BC, BCO, and AWR with ground-truth rewards. Starting with low-performance initial policies, we generate data to train the next iteration of policies for each algorithm and iterate. Both AWR with ground-truth rewards and VfO-bin lead to strong results.

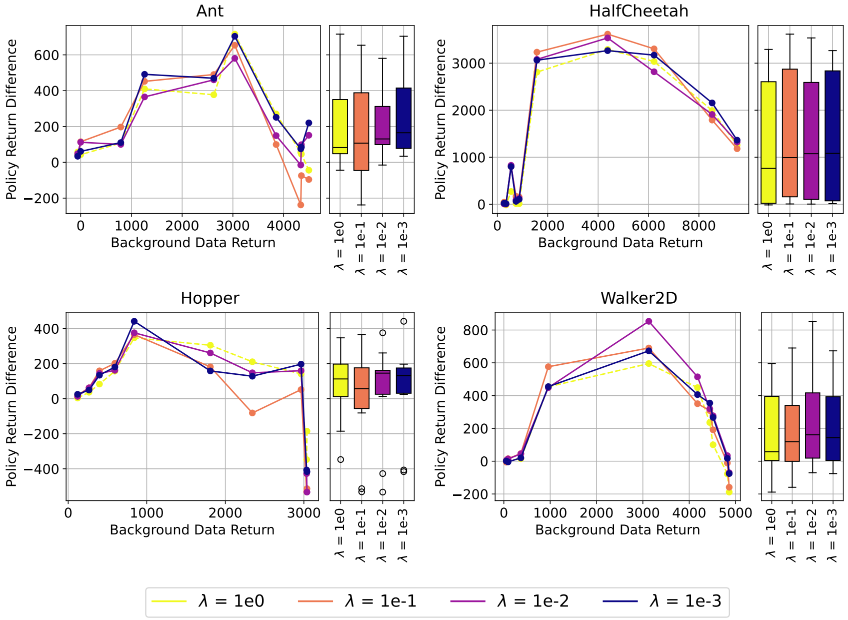


Figure 15: Difference in cumulative return of offline SQIL (with privileged expert actions) on D4RL tasks using the SIBench data for different temperature parameters and mixing parameter 0.9. We plot the average return in the background data against the return of the trained policy. We can observe a fairly wide range of hyperparameter settings leading to improvement.

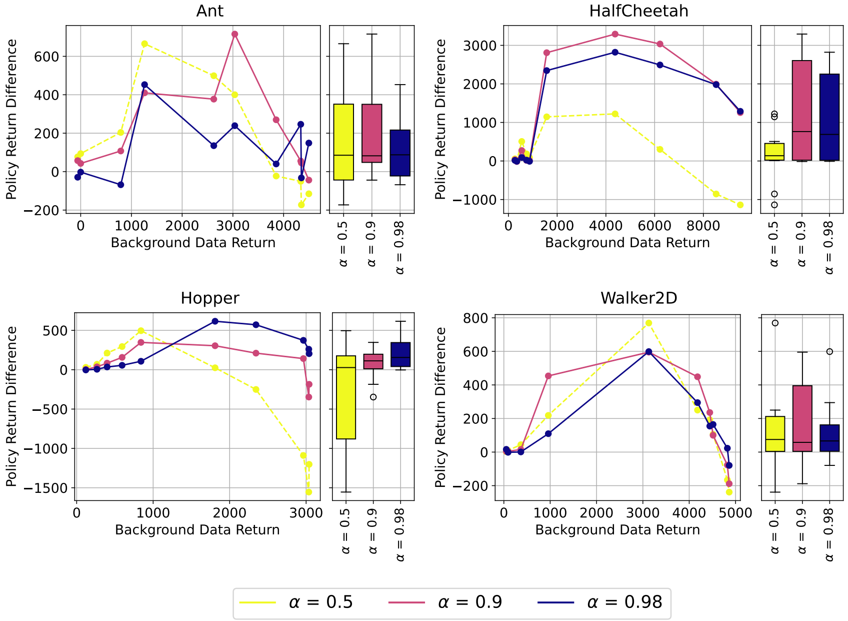


Figure 16: Difference in cumulative return of offline SQIL (with privileged expert actions) on D4RL tasks using the SIBench data for different mixing parameters and temperature 1.0. We plot the average return in the background data against the return of the trained policy. In comparison to the action-free VfO, we need a higher mixing parameter to avoid overfitting on the scarce expert data.

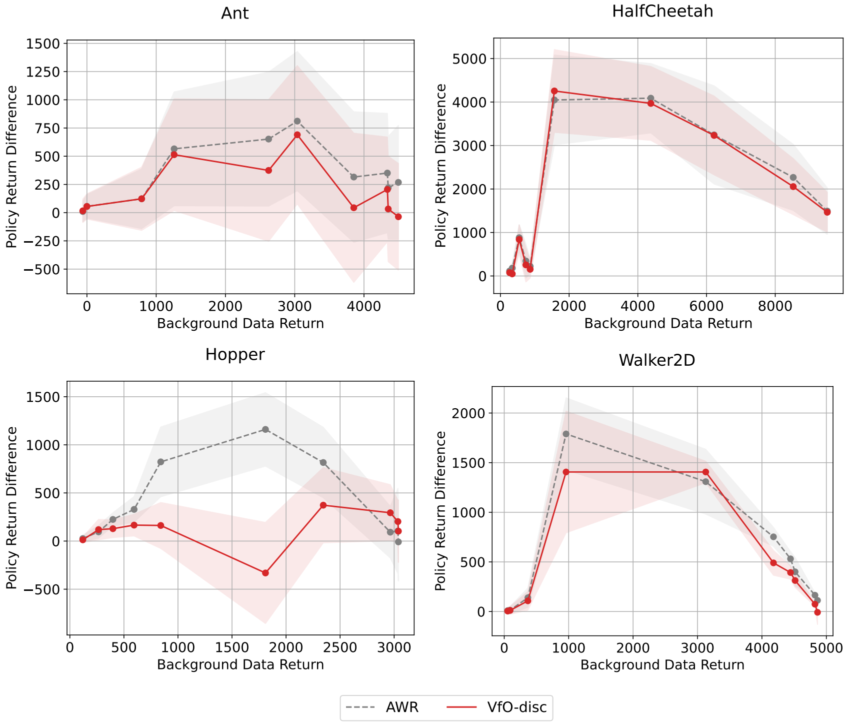


Figure 17: Mean and standard deviation of differences in cumulative return of AWR and VfO-disc on D4RL tasks using the SIBench data. We plot the average return across 5 seeds and across the last 1e5 steps of training.

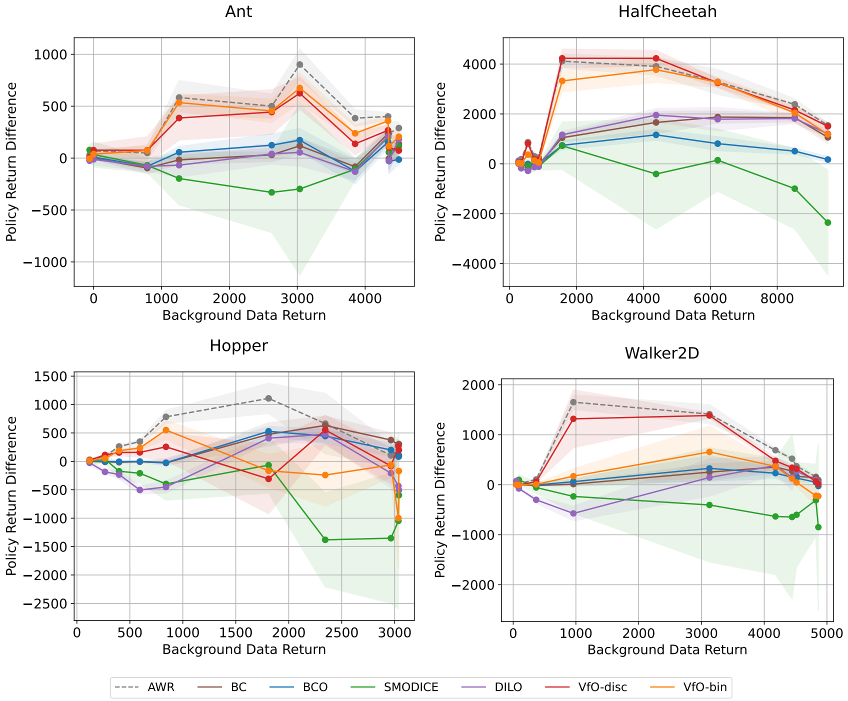


Figure 18: Mean and standard deviation of differences in cumulative return of for different algorithm on D4RL tasks using the SIBench data.