SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2026, Vol. 69, Iss. 12, 222101:1–222101:14
https://doi.org/10.1007/s11432-024-5205-7

Expert policy support expansion: alleviating the overfitting issue of adversarial imitation learning for high-dimensional tasks

Fuguang HAN[1,2†], Zhilong ZHANG[1,2,3†], Yichen LI[1,2,3], Feng CHEN[1,2], Yang YU[1,2,3] & Zongzhang ZHANG[1,2*]

[1]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China*
[2]*School of Artificial Intelligence, Nanjing University, Nanjing 210023, China*
[3]*Polixir Technologies, Nanjing 211106, China*

Received 5 January 2026/Revised 22 March 2026/Accepted 18 May 2026/Published online 4 November 2026

Abstract We consider imitation learning (IL), which mimics a policy from finite expert trajectories. A prevalent framework of IL is adversarial imitation learning (AIL) that trains a discriminator to distinguish samples from the expert and the imitator. Although AIL achieves impressive performance on some artificial tasks, we observe that the learned discriminator may overfit to the limited expert data, especially when encountering tasks with high-dimensional state spaces. We find that the imbalance between expert demonstrations and generated samples during the training of the discriminator can affect its training margin, consequently resulting in overfitting of the discriminator. Building upon this, we propose expert policy support expansion (EXPOSE). Apart from the discriminator and the generator, EXPOSE additionally learns two random network distillations (RNDs), namely, an expert RND (e-RND) and a policy RND (p-RND). The p-RND incentivizes the imitator to explore the state-action space, while the e-RND filters out expert-level samples from the imitator and adds them to the expert dataset, explicitly expanding the support of the expert policy. Extensive empirical results on tasks from the DeepMind Control Suite and Meta-World benchmarks demonstrate that EXPOSE achieves an average improvement of 40% in overall performance compared to baselines.

Keywords reinforcement learning, imitation learning, adversarial imitation learning, overfitting, high-dimensional tasks

Citation Han F G, Zhang Z L, Li Y C, et al. Expert policy support expansion: alleviating the overfitting issue of adversarial imitation learning for high-dimensional tasks. Sci China Inf Sci, 2026, 69(12): 222101, https://doi.org/10.1007/s11432-024-5205-7

1 Introduction

In reinforcement learning (RL) [1–7], agents can achieve remarkable performance when trained with well-designed reward functions that incentivize desirable behavior. However, the manual design of reward functions is challenging and time-consuming because it necessitates a substantial foundation of prior knowledge and rigorous testing by human experts. Imitation learning (IL) [8–12] offers a promising approach to alleviate the burden of reward design by leveraging expert demonstrations to learn a policy that mimics expert behavior. Benefiting from that, IL has become increasingly prevalent in a wide range of domains, e.g., automatic driving [13], robot control [14], recommendation systems [15], and real-time strategy games [16].

Adversarial IL (AIL) [17–21] is a widely adopted IL approach. The fundamental concept of AIL involves state-action distribution matching, aiming to minimize the divergence between the expert policy and the training policy through adversarial training of the discriminator and the policy [22, 23]. Recent research has showcased AIL's superior performance across diverse tasks [24, 25]. However, the discriminator, tasked with distinguishing between expert and generated samples, faces the potential overfitting problem when expert samples are limited, presenting a significant challenge for AIL [26–28]. To address this concern, positive-unlabeled generative adversarial imitation learning (PUGAIL) treats the agent's experience as an unlabeled mixture of positive and negative data, framing the training of the discriminator as a positive-unlabeled learning problem [29]. Additionally, Ref. [30] proposed that incorporating regularization techniques during the training of the AIL discriminator could alleviate the overfitting issue. While these efforts have a positive impact on simple tasks, none of them tackle the more severe

* Corresponding author (email: zzzhang@nju.edu.cn)
† These authors contributed equally to this work.

© Science China Press 2026

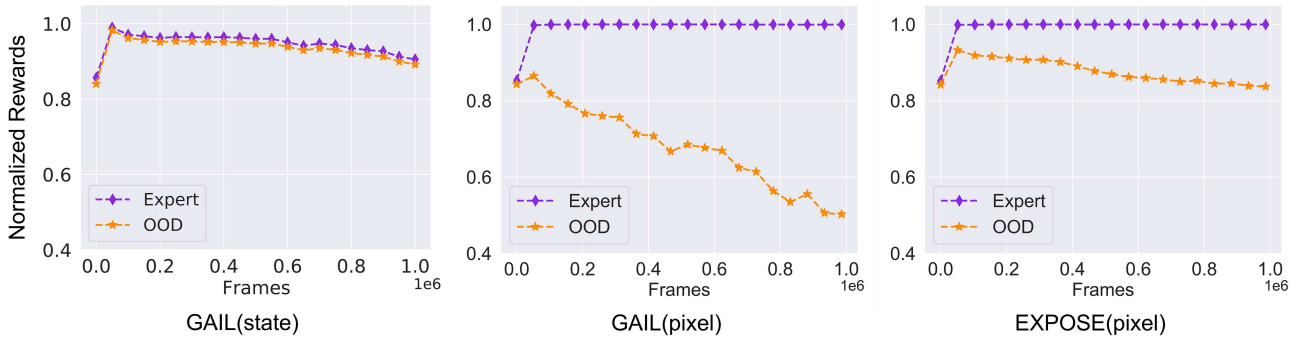


Figure 1 Comparison of discriminator normalized rewards across different representations and algorithms. The purple dashed lines indicate normalized rewards for in-distribution expert data, while the orange dashed lines represent normalized rewards for OOD expert samples.

discriminator overfitting posed in high-dimensional tasks. As a result, their performance tends to be suboptimal in high-dimensional tasks.

In this paper, we analyze the issue of overfitting as it pertains to discriminators in AIL. Our findings suggest that overfitting to limited expert samples can lead the discriminator to underestimate the support of the expert policy. Interestingly, our comparative analysis reveals that the severity of this overfitting is closely tied to input dimensionality. As illustrated in Figure 1, the discriminator generalizes effectively in low-dimensional state-based tasks (GAIL-state), characterized by the consistent alignment between normalized rewards for out-of-distribution (OOD) expert trajectories and the training data. Conversely, in high-dimensional pixel-based scenarios (GAIL-pixel), a pronounced divergence emerges. While the discriminator assigns near-perfect rewards to the training expert data, the rewards for OOD expert samples—which belong to the same expert support—drop sharply as training progresses. The reward decay on unseen expert data is a clear symptom of the discriminator overfitting to specific visual features rather than the underlying expert policy. The underestimation of the rewards for OOD expert samples has the potential to compromise both the efficacy and robustness of the AIL algorithm.

Drawing insights from our analysis, we improve AIL by mitigating discriminator overfitting through support expansion. As shown in Figure 1 (EXPOSE-pixel), our proposed method effectively maintains high rewards for OOD expert samples even in high-dimensional pixel spaces, preventing the reward coverage from shrinking. The core of this approach lies in augmenting the expert dataset by selecting samples that belong to the support of the expert policy from the generated data. To estimate the expert support in high-dimensional tasks, we employ random network distillation (RND) [31], a method that has been validated for estimating sample distributions in high-dimensional tasks [32]. Specifically, we utilize an expert RND (e-RND) to estimate the expert support and filter out samples belonging to it from the generated samples. These identified samples are treated as positive examples to train the discriminator, effectively augmenting the expert dataset and expanding the expert support. Nevertheless, we observe that relying solely on the e-RND for sample selection may not be sufficiently efficient, as the historical samples might not encompass the required samples during the early stage of training. So, we adopt another RND, namely policy RND (p-RND), to quantify the sample uncertainty of the replay buffer, serving as the intrinsic reward to facilitate policy exploration. This intrinsic reward mechanism enables the collection of a more diverse set of samples, ensuring an ample supply for sample selection by the expert RND. In summary, we refer to this method as expert policy support expansion, abbreviated as EXPOSE.

In the experimental part, we validate the effectiveness of our method by conducting experiments in the pixel-based DeepMind Control Suite (DMControl) [33] and Meta-World tasks [34], finding that EXPOSE shows an improvement of 40% over baselines in terms of sample efficiency and convergence performance. Notably, EXPOSE stands out as the only AIL algorithm that demonstrates convergence potential on the challenging pixel-based Humanoid stand task. In addition, we conduct an ablation study to assess the efficacy of various modules within the algorithm.

2 Background

2.1 Markov decision process

We consider the infinite horizon Markov decision process (MDP) [1,35], which can be denoted by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, p_0, r, \gamma \rangle$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ represents the environment dynamics, $p_0 : \mathcal{S} \to \mathbb{R}$ is the initial state distribution, $r(s, a) \in [-R_{\max}, R_{\max}]$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi : \mathcal{S} \to \mathcal{D}(\mathcal{A})$ is a mapping from states to probability distributions over the action

space. The occupancy measure d_π of the policy π is defined as $d^\pi(s,a) = \pi(a|s)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi, \mathcal{M})$, where $P(s_t = s|\pi, \mathcal{M})$ is the probability that the agent is in state s at time step t when following the policy π [35,36]. The decision-making process unfolds in the following manner: at each time step, the agent observes a state s, chooses an action a from the distribution $\pi(\cdot|s)$, and subsequently receives a reward $r(s,a)$ from the environment. The environment then transitions to a new state s' following the probability distribution $\mathcal{P}(\cdot|s,a)$. In RL, the goal of the agents is to maximize the expected discounted cumulative reward, a.k.a. return $V(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t(s,a)]$. Due to the difficulty of reward design, in IL, agents cannot receive the reward signals but aim to learn an optimal policy by mimicking a set of expert demonstrations $\mathcal{D}^E$. Additionally, we characterize the expert support as $\text{supp}(d^E) = \{(s,a)|d^E(s,a) > 0\}$ and the empirical expert support as $\text{supp}(\hat{d^E}) = \{(s,a)|(s,a) \in \mathcal{D}^E\}$, where d^E is short for d^{π_E}.

2.2 Adversarial imitation learning

IL is a method that enables an agent to make decisions like human experts by learning from expert examples. It aims to learn a good policy from expert demonstrations $\mathcal{D}^E = \{(s_i, a_i)\}_i$ so that the expected return of the agent's policy π matches the expected return of the expert policy, as follows:

$$\min_\pi V(\pi_E) - V(\pi). \tag{1}$$

As a popular IL method, the core objective of AIL is to minimize the divergence between the occupancy measures of the expert policy and the training policy. Notably, in the case of generative AIL (GAIL), the Jensen-Shannon (JS) divergence serves as the distance metric for matching the distribution of the expert policy [17]. According to the dual representation of the JS divergence, we can derive the optimization objective as follows:

$$\max_\pi \min_D \mathbb{E}_\pi[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))], \tag{2}$$

where D is a binary classifier, denoted as the discriminator, distinguishing between samples generated by the policy and those from the expert. The policy π mimics the expert by maximizing the return as determined by the discriminator.

3 Analysis

In this section, we propose that the limited expert samples in high-dimensional environments can aggravate discriminator overfitting, leading to the underestimation of the expert support.

We argue that training a discriminator entails addressing a binary classification problem with an imbalanced dataset. Specifically, during discriminator training, expert data is designated as positive, while generated samples are labeled as negative. Additionally, the sample size of expert data remains constant, while the agent's experience expands, leading to an inherent imbalance between positive and negative samples. In this context, we elaborate on how the overfitting of the discriminator increases its classification error. The following presents relevant theoretical explanations for this problem.

Definition 1 (Training margin [37])**.** In a binary classification task, we assume the input space is $\mathbb{R}^m$, and the label space is $\mathcal{Y} = \{0, 1\}$. Let x denote the input and $y \in \mathcal{Y}$ denote its true label. For a model $f : \mathbb{R}^m \to \mathbb{R}^2$, the margin of an example (x, y) is defined as

$$\gamma(x,y) = f(x)_y - \max_{k \neq y} f(x)_k, \tag{3}$$

where $f(x)_k$ represents the output score for the k-th class. Furthermore, the training margin for a specific class $j \in \{0, 1\}$ is defined as the minimum margin across all training examples belonging to that class

$$\gamma_j = \min_{(x,y):y=j} \gamma(x,y). \tag{4}$$

Note that Figure 2 gives an illustration of the training margin.

Theorem 1 (Generalization error with pseudo-experts [37])**.** Let $\mathcal{F}$ be a class of functions with complexity $C(\mathcal{F})$. At iteration k, let $D_1^{(k)}$ denote the positive dataset with size $n_1^{(k)}$, and margin $\gamma_1^{(k)}$. Assume that a fraction η_k

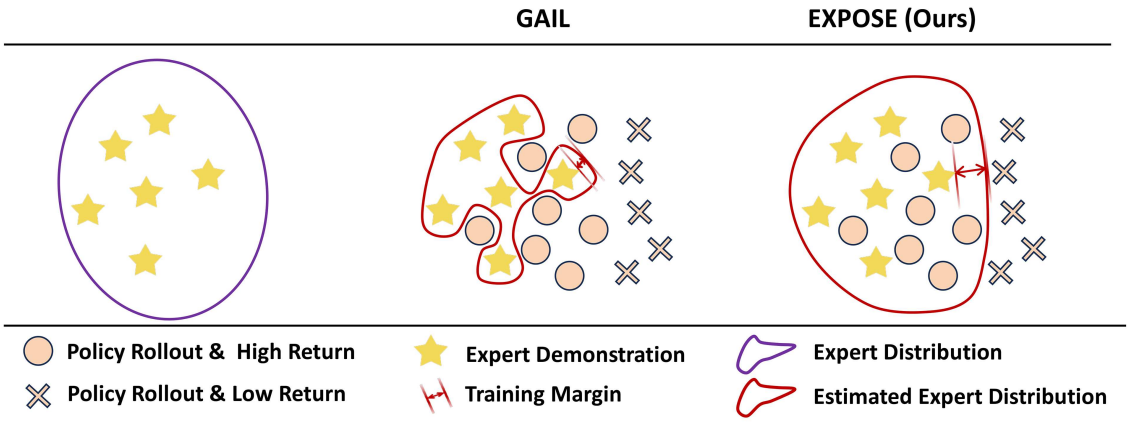


Figure 2 Illustration of the discriminator's overfitting issue. The typical GAIL algorithm strictly distinguishes between expert data and policy rollout data, forming small training margins, while EXPOSE alleviates the overfitting issue by expert support expansion.

of samples in $D_1^{(k)}$ are false positives, then for any $\delta > 0$, with probability at least $1 - \delta$, the generalization error satisfies

$$L_1^{(k)}[f] \leqslant \frac{1}{\gamma_1^{(k)}} \sqrt{\frac{C(\mathcal{F})}{n_1^{(k)}}} + \frac{\log(1/\delta)}{\sqrt{n_1^{(k)}}} + \mathcal{O}(\eta_k).$$

Theorem 1 provides an upper bound for the classification error in the presence of pseudo-expert samples. The bound shows that the generalization error is inversely related to the training sample size $n_1^{(k)}$ and the margin $\gamma_1^{(k)}$, and increases with the complexity of the hypothesis class $\mathcal{F}$. In addition, the term $\mathcal{O}(\eta_k)$ explicitly captures the effect of false positives introduced by the pseudo-expert set. When the filtering error is controlled, the influence of noisy pseudo-experts is transient and does not dominate the generalization behavior in later stages.

We consider the training of the discriminator as a binary classification task and assume the label of the expert sample is 1. In the context of high-dimensional environments with limited expert demonstrations (where n_1 is small), the discriminators often exhibit strong fitting capabilities to capture the informative features from high-dimensional inputs, resulting in a large $C(\mathcal{F})$. Additionally, overfitting tends to reduce the training margin of the positive class γ_1. Consequently, the expert 0-1 loss $L_1[f]$ becomes large due to the combined impact of these factors. This increase in $L_1[f]$ signifies that numerous expert samples are incorrectly classified as negative examples, ultimately causing the underestimation of the expert support.

The overall success of the algorithm hinges on increasing the training data size n_1 while keeping the noise penalty $\mathcal{O}(\eta_k)$ strictly bounded. In standard AIL, the reward margin between expert and policy samples inherently collapses as the policy improves, which would normally lead to an explosion of η_k if the discriminator were used as the filter.

In contrast, our e-RND filtering operates in a decoupled representation space defined by the prediction error $l_\psi(s, a)$. The expert-trained RND model defines a stationary reference manifold. Because we enforce a conservative threshold ($\alpha = 0.98$), the representation-space separation between accepted and rejected samples remains strictly preserved. Policy-generated samples, even when achieving high rewards, are rejected if they exhibit measurable deviation from the expert manifold in this structural space. This mechanism ensures η_k remains bounded throughout training, allowing the significant benefits of dataset expansion to vastly outweigh the minor costs of noisy approximation. As a result, the benefit of enlarging the support outweighs the impact of noisy samples, leading to improved generalization and a more accurate estimation of the expert distribution.

4 AIL via expert policy support expansion

In this section, we introduce an AIL algorithm via EXPOSE. Subsection 4.1 describes how to leverage e-RND to filter samples from the replay buffer and expand the support of the expert policy. Subsection 4.2 introduces the p-RND to encourage policy exploration and make unbiased corrections to the expert distribution. The overall framework of EXPOSE is depicted in Figure 3.

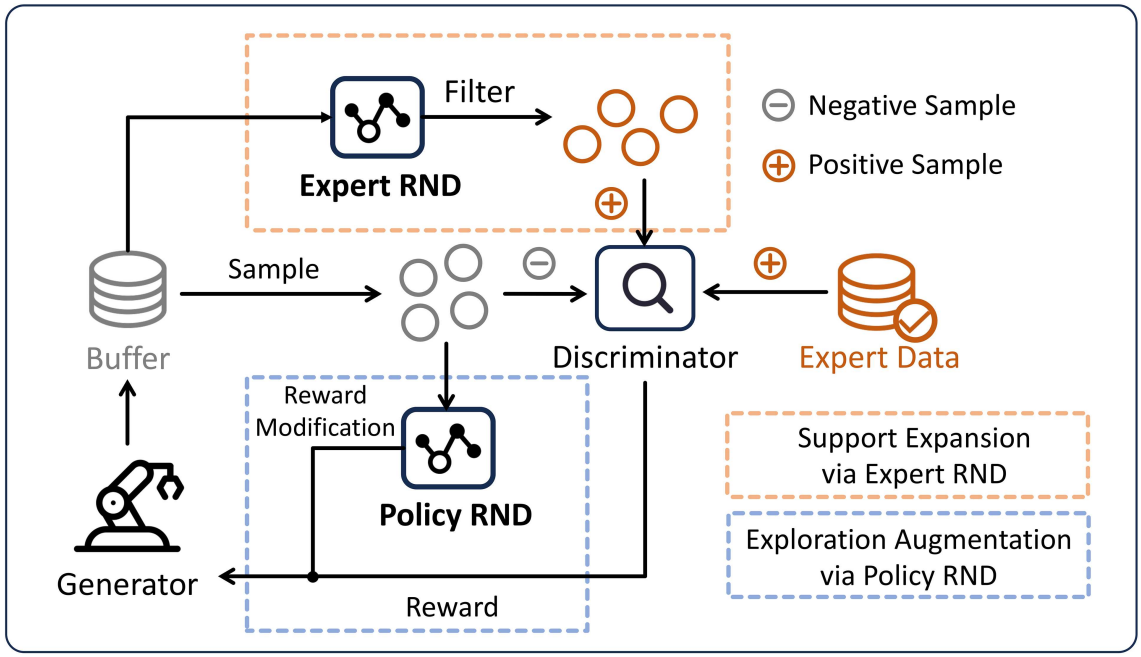


Figure 3 Overall framework of EXPOSE. We utilize an e-RND to filter additional positive samples from the replay buffer to expand the expert support, while one p-RND is applied to encourage the policy to explore unfamiliar state-action pairs.

4.1 Support expansion by expert RND

Drawing insights from our analysis above, we improve AIL by mitigating discriminator overfitting through support expansion. Specifically, during the training process, we store all historical samples obtained by executing the policy in the environment in the replay buffer. When updating the discriminator in each iteration, we select samples belonging to the expert support from the replay buffer and use them as positive examples to train the discriminator.

To estimate the expert support in high-dimensional environments, we use RND [31], a commonly used method for distribution estimation for high-dimensional tasks. RND involves two randomly initialized networks: a target network $g_{\hat{\theta}}: \mathcal{X} \to \mathbb{R}^m$ and a predictor network $g_\theta: \mathcal{X} \to \mathbb{R}^m$, where θ and $\hat{\theta}$ are respectively the parameters of the predictor and target networks. During the training process, $g_{\hat{\theta}}$ is fixed, and g_θ is optimized to minimize the expected squared prediction error

$$L(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} l_\theta(x) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left(g_\theta(x) - g_{\hat{\theta}}(x)\right)^2. \tag{5}$$

Assume that $\mathcal{D}$ is sampled i.i.d. from a distribution d. Larger prediction errors indicate fewer similar examples are seen by the predictor. So we infer that the predicted error can be considered an approximation of the underlying distribution d.

RND leverages the neural network's strong generalization capabilities to estimate the distribution with limited samples, which has been validated in several studies [31,32]. In our work, we train an e-RND g_ψ on a limited set of expert data to estimate the expert distribution. Similar to previous studies [32], we assume that the distribution of the expert policy is negatively correlated with the predicted error

$$d^E(s,a) \propto \frac{1}{l_\psi(s,a)} = \frac{1}{\left(g_\psi(s,a) - g_{\hat{\psi}}(s,a)\right)^2}. \tag{6}$$

After estimating the expert distribution by training the e-RND on the expert dataset, we naturally extend it to obtain the support of the expert policy. We assert that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, if the output of e-RND $l_\psi(s,a)$ is less than a predefined constant multiplying the maximum predicted error of e-RND over the expert dataset, then (s,a) belongs to the expert support, as depicted below

$$\begin{aligned} &\mathrm{supp}_\psi(d^E) = \{(s,a) | l_\psi(s,a) \leqslant l_\psi(\alpha, \mathcal{D}^E)\}, \\ &\text{where } l_\psi(\alpha, \mathcal{D}^E) = \alpha \max_{(s^e,a^e) \in \mathcal{D}^E} l_\psi(s^e, a^e), \end{aligned} \tag{7}$$

where α serves as a hyperparameter controlling the extent of the estimated expert support. Increasing α will expand the filtering scope, but it may lead to the inclusion of more samples that do not belong to the expert support. To prevent suboptimal samples from being misclassified as expert-like, we employ a stringent novelty threshold α, ensuring that only samples with high structural similarity to the original expert data are integrated into the support expansion. Furthermore, the adversarial dynamics of AIL function as an implicit error-correction mechanism: as the discriminator is continuously optimized against the agent's evolving trajectories, any reward signal derived from potential false positives is ephemeral. This self-correcting balance suppresses spurious rewards as training progresses, effectively anchoring the policy within the true expert distribution.

In practical implementation, before initiating the training of the AIL algorithm, we first train an e-RND to estimate expert support. In each iteration, we leverage the e-RND to select a batch of samples from the replay buffer that falls within the expert support. These samples are then used to augment the positive examples for updating the discriminator. Revisiting the discriminator training paradigm, negative examples ($\mathcal{D}^-$) are sampled from the replay buffer, while positive examples ($\mathcal{D}^+$) are sampled from the expert dataset. Introducing $\mathcal{D}_+^-$ as the dataset obtained through screening the replay buffer by e-RND, the optimization objective for the discriminator becomes as follows:

$$\min_D \mathbb{E}_{\mathcal{D}^+\cup\mathcal{D}_+^-}[\log(1-D(s,a))]+\mathbb{E}_{\mathcal{D}^-}[\log D(s,a)]. \tag{8}$$

By such an augmentation, the discriminator's estimation range of expert support is expanded, bringing the estimated expert support closer to the true expert support, effectively mitigating the discriminator's overfitting issue in high-dimensional tasks.

4.2 Exploration augmentation by policy RND

We notice that depending solely on e-RND for sample selection may not be efficient enough, as the historical samples might not cover the necessary examples, especially in the early stage of training. To illustrate, imagine an extreme scenario where the replay buffer lacks any sample from the expert support. In such a case, the e-RND is unable to filter out any sample to augment the expert dataset, thus not contributing any positive impact to the AIL algorithm.

To enhance the efficiency of our method, we introduce an additional exploration mechanism during policy training. An effective exploration strategy aids the policy in acquiring a diverse range of experiences, thereby increasing the probability that the replay buffer contains more samples belonging to the expert support but not present in the expert dataset. Inspired by RND [31], we employ another RND network g_ϕ, termed p-RND, to quantify the uncertainty of experiences. This uncertainty estimation serves as an intrinsic bonus, combined with the output of the discriminator, acting as the reward for policy training, as illustrated below:

$$\begin{aligned} r(s,a) =& r_D(s,a)+\beta l_\phi(s,a)\\ =& r_D(s,a)+\beta\big(g_\phi(s,a)-g_{\hat{\phi}}(s,a)\big)^2, \end{aligned} \tag{9}$$

where β is a hyperparameter that controls the proportion of the intrinsic reward. As β increases, the policy leans towards exploration, favoring visits to regions with higher uncertainty.

Furthermore, after obtaining e-RND l_ψ and p-RND l_ϕ, we can refine the target distribution of AIL to ensure it is an unbiased estimate of the expert distribution by multiplying the ratio of l_ϕ and l_ψ for the positive examples of the discriminator

$$d^+(s,a)=\lambda d^E(s,a)+(1-\lambda)\frac{1}{Z}\frac{l_\phi(s,a)}{l_\psi(s,a)}d_+^-(s,a), \tag{10}$$

where λ serves as a hyperparameter controlling the proportion of augmented samples, and $Z=\sum_{s,a}\frac{l_\phi(s,a)}{l_\psi(s,a)}d_+^-(s,a)$ is the normalization factor. By doing this, the optimization objective for the discriminator becomes as follows:

$$\min_D \lambda\mathbb{E}_{\mathcal{D}^+}[\log(1-D(s,a))]+\frac{1-\lambda}{Z}\mathbb{E}_{\mathcal{D}_+^-}\Big[\frac{l_\phi(s,a)}{l_\psi(s,a)}\log(1-D(s,a))\Big]+\mathbb{E}_{\mathcal{D}^-}[\log D(s,a)]. \tag{11}$$

Lemma 1. Assume that $\hat{d}$ is an unbiased estimate of distribution d over $\mathcal{X}$, for any mutually exclusive partition over $X=X_1\cup X_2$, satisfying $X_1\cap X_2=\emptyset$. Let $d_1(x)=\frac{1}{Z}d(x)$ if $x\in X_1$ is the corresponding distribution over X_1, where Z is the normalization factor. We can deduce that $\hat{d}(x)\propto d_1(x)$.

Proof. For any $x,x'\in X_1$, assuming $d_1(x)>0$ and $d_1(x')>0$, it follows that $\frac{d_1(x')}{d_1(x)}=\frac{d(x')}{d(x)}=\frac{\hat{d}(x')}{\hat{d}(x)}$. The second equality stems from the fact that $\hat{d}$ is an unbiased estimate of the distribution d. Therefore, we can assert that $\frac{d_1(x)}{\hat{d}(x)}=\frac{d_1(x')}{\hat{d}(x')}$, which implies that $\hat{d}(x)\propto d_1(x)$.

Proposition 1 (Unbiased estimate of expert distribution)**.** If $\frac{1}{l_\psi}$ and $\frac{1}{l_\phi}$ are unbiased estimates of the expert distribution and the replay buffer distribution, respectively, then Eq. (10) is also an unbiased estimate of the expert distribution.

Proof. According to Lemma 1 and the fact that $\frac{1}{l_\phi}$ is an unbiased estimate of the replay buffer distribution, we observe that $\frac{1}{l_\phi} \propto d_+^-$. Consequently, we can establish that $\frac{l_\phi}{l_\psi} \propto \frac{d^E}{d_+^-}$ due to the relationship $\frac{1}{l_\psi} \propto d^E$. Therefore, assuming the normalization factor $Z = \sum_{s,a} \frac{l_\phi(s,a)}{l_\psi(s,a)} d_+^-(s,a)$, the term $\frac{1}{Z}\frac{l_\phi(s,a)}{l_\psi(s,a)} d_+^-$ serves as an unbiased estimate of expert distribution. Additionally, since d^+ is a linear combination of two unbiased estimates of the expert distribution, we conclude that d^+ is also an unbiased estimate of the expert distribution.

We provide the detailed pseudo-code in Algorithm 1.

Algorithm 1 EXPOSE.

Require: Expert demonstrations $\mathcal{D}^E$.
  Initialize policy π_θ, discriminator d_ω;
  Initialize e-RND g_ψ, p-RND g_ϕ;
  Initialize replay buffer $\mathcal{R} \leftarrow \emptyset$;
  for $n = 1, 2, \ldots, N$ **do**
    Update e-RND g_ψ with $\mathcal{D}^E$ via (5);
  end for
  for $k = 1, 2, \ldots, K$ **do**
    Collect data $\mathcal{D}^k = \{(s, a, s')\}$ by π_θ^{k-1};
    Merge $\mathcal{D}^k$ into $\mathcal{R}$;
    Sample $\mathcal{D}^+$ from $\mathcal{D}^E$ and $\mathcal{D}^-$ from $\mathcal{R}$;
    Select $\mathcal{D}_+^-$ from $\mathcal{D}^-$ by g_ψ via (7);
    Update discriminator d_ω via (11);
    Update p-RND g_ϕ with $\mathcal{D}^-$ via (5);
    Compute reward r via (9);
    Update policy π_θ^k to maximize $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$ in an RL approach;
  end for
Ensure: Policy π_θ^K.

5 Experiments

In experiments, we carefully assess EXPOSE across various high-dimensional tasks. We first conduct a fair comparison against baseline methods on a wide range of high-dimensional tasks from DMControl [33] and Meta-World [34], demonstrating the effectiveness of EXPOSE in handling challenging high-dimensional tasks. Subsequently, we conduct ablation studies to investigate the impact of each module within EXPOSE. Finally, to validate the precision of e-RND in identifying expert samples, we demonstrate the performance of e-RND in terms of the ratio of expert samples filtered out across various datasets.

5.1 Experimental setup

Environment. We experiment with 8 tasks from the DMControl and 4 tasks from the Meta-World. DMControl stands as a prominent benchmark in RL, providing a comprehensive set of continuous control tasks and serving as a versatile platform for assessing the adaptability and generalization capabilities of RL models. Meta-World also comprises a vast array of benchmark tasks, incorporating diverse environments that span manipulation, locomotion, and perception challenges. For DMControl tasks, we collect 10 expert trajectories using an expert policy trained with DrQ-v2 [38]. As for Meta-World tasks, we experiment with one expert trajectory per task, obtained from their open-source implementation [34]. We conduct 5 trials for each task, each with different random seeds.

Baselines. Here is a detailed description of the baselines we have chosen for comparison.

• **Discriminator-actor-critic (DAC)** [24]. DAC is a sample-efficient off-policy algorithm designed for AIL. It addresses the challenges of sample inefficiency and reward bias, which are common in traditional AIL methods. The algorithm integrates an off-policy reinforcement learning approach to significantly reduce the number of interactions required between the policy and the environment for effective learning. This makes DAC particularly suitable for scenarios where data collection is costly or limited.

• **PUGAIL** [29]. Originally designed to tackle the issue of discriminator overfitting in AIL, PUGAIL employs a positive-unlabeled learning framework and learns from both positive (expert demonstrations) and unlabeled (agent experiences) data. This method is adapted to an off-policy setting in our experiments to ensure a fair comparison.

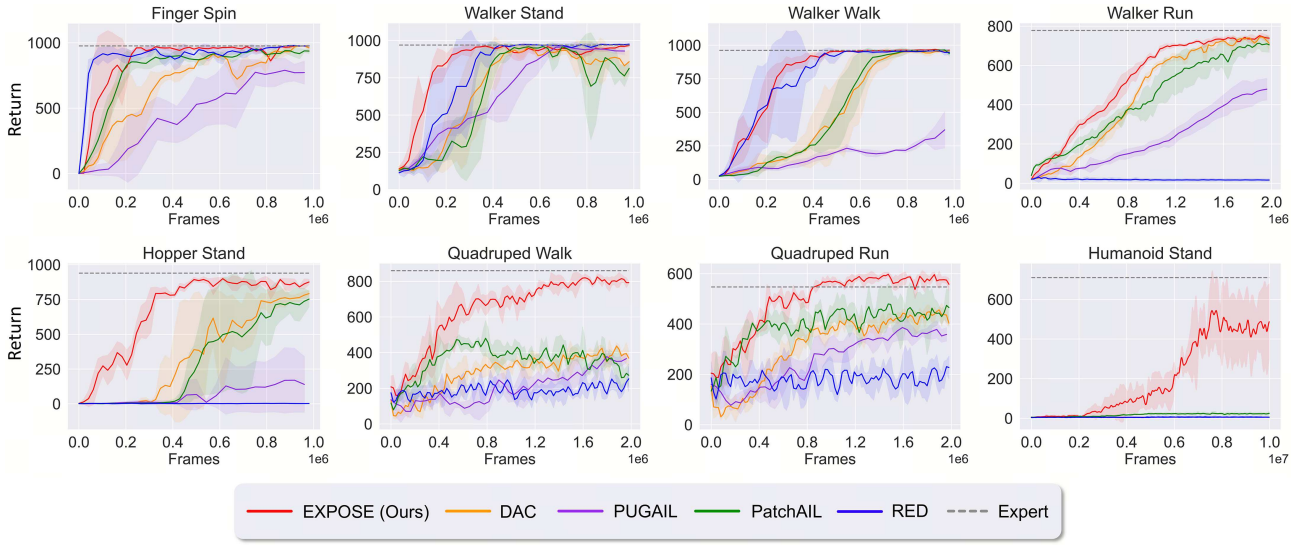


Figure 4 Learning curves on 8 pixel-based DMControl tasks with 10 expert trajectories. The shaded regions represent ±1 standard deviation across 5 seeds.

Table 1 Average returns on 8 DMControl tasks at the end of training over 5 random seeds using 10 expert trajectories. The best results on each task are bold.

	EXPOSE	DAC	PUGAIL	PatchAIL	RED	Expert
Finger spin	**977±1**	969±3	**977±1**	922±43	**977±2**	976±9
Walker stand	966±10	929±47	867±5	828±42	**979±3**	970±20
Walker walk	**963±2**	956±9	950±2	911±42	958±6	961±17
Walker run	742±5	741±3	**743±27**	708±20	16±5	778±10
Hopper stand	**858±24**	772±57	793±34	769±35	3±0	939±9
Quadruped walk	**799±47**	409±14	388±13	293±64	251±63	859±41
Quadruped run	**551±26**	424±37	373±45	432±31	204±81	547±136
Humanoid stand	**441±117**	4±0	5±0	22±3	5±1	710±92

• **PatchAIL** [39]. PatchAIL represents an innovative approach in AIL by deriving rewards at the patch level from pixel observations. This method provides a more nuanced and precise characterization of rewarding signals, which is particularly beneficial in tasks where visual input is high-dimensional and complex. By focusing on local image patches, PatchAIL can better capture the intricacies of expert demonstrations, leading to more effective imitation.

• **RED** [32]. RED constructs a reward function using RND and avoids the adversarial training typically associated with discriminator-based AIL methods. By constructing a static reward function that captures the novelty of states, RED provides a consistent and stable reward signal for the agent to learn from, thus simplifying the overall learning process.

In our experiments, EXPOSE employs the same set of hyperparameters across most tasks. Specifically, we set the parameter α in (7) to 0.98, while initializing β in (9) at 1 and progressively decaying it to 0. Detailed discussions on additional hyperparameters and a comprehensive parameter sensitivity study can be found in the Appendixes A and B.

5.2 Main results

We compare our method with all baseline methods on 8 DMControl tasks and 4 Meta-World tasks. In Table 1, we present the average returns on DMControl tasks at the end of training, and we show the learning curve of each method in Figure 4. Table 1 and Figure 4 illustrate that the EXPOSE method exhibits superior efficiency and convergence across almost all DMControl tasks.

Specifically, EXPOSE achieves better convergence performance than other baseline methods. We find that DAC performs well on several tasks, but fails to converge to a performance close to that of the expert policy on some difficult tasks, such as Quadruped walk, Quadruped run, and Humanoid stand. However, EXPOSE can converge to performance close to the expert policy on most tasks except Humanoid stand. Furthermore, in terms of convergence

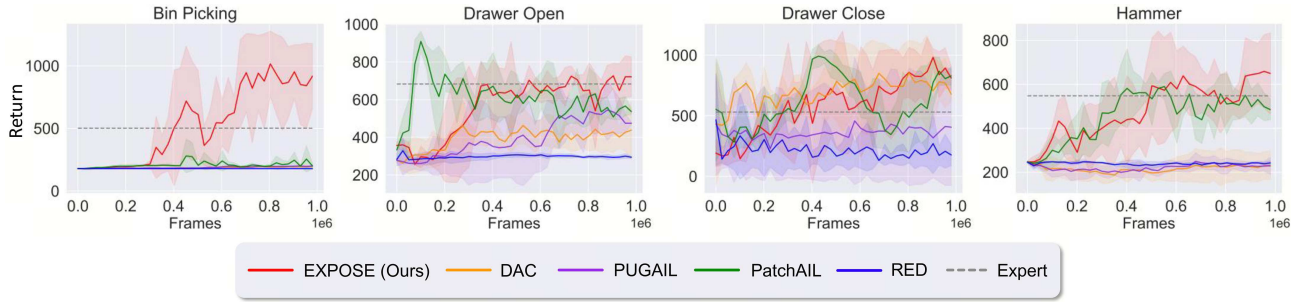


Figure 5 Learning curves on 4 pixel-based Meta-World tasks with 1 expert trajectory.

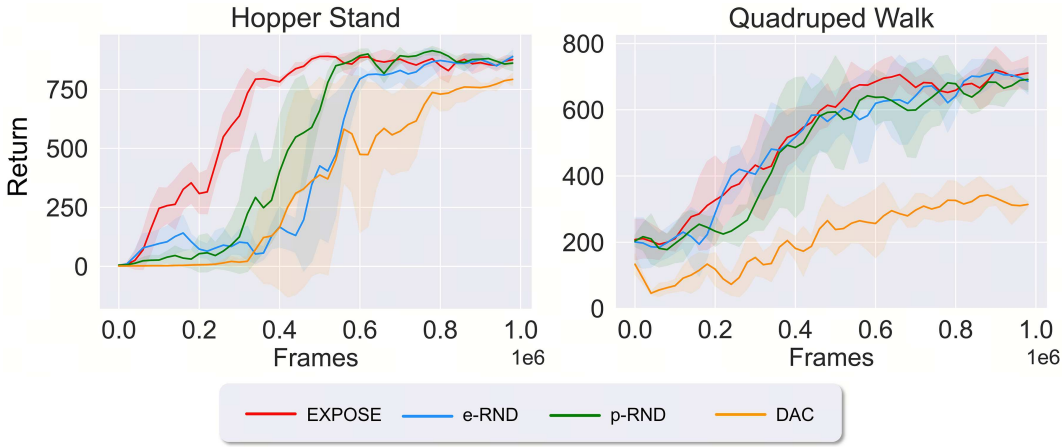


Figure 6 Ablation study on the effect of e-RND and p-RND on Hopper stand and Quadruped walk.

speed, EXPOSE is also ahead of other methods on almost all tasks. It is noteworthy that RED demonstrates markedly enhanced sample efficiency in the Finger spin, Walker stand, and Walker walk tasks when compared to DAC. This improvement is achieved by estimating expert support to construct the reward function, thus avoiding adversarial training. This highlights the feasibility of using RND for expert support estimation. However, the effectiveness of RED is hampered by the lack of robust convergence properties of the constructed reward function compared to the AIL algorithm, leading to failure in all other tasks.

In Figure 5, we present the learning curves for each method on the Meta-World tasks with a single expert trajectory. Across these four tasks, where only a single expert trajectory is available and the difficulty is relatively high, the overfitting of the AIL algorithm is exacerbated. PatchAIL successfully trains a meaningful policy by virtue of its capability to extract patch rewards, providing a more precise reward characterization in the Hammer task, but still fails to converge in the Bin picking task. EXPOSE augments the expert data by acquiring a broader range of diverse expert data through the filtering of e-RND. Remarkably, it achieves superior performance compared to the expert policy even when only one trajectory is provided.

5.3 Ablation study

In this part, we meticulously performed a series of ablation experiments to evaluate the individual impacts of e-RND in selecting positive samples for discriminator training and p-RND in fostering exploration within the EXPOSE algorithm. Figure 6 succinctly illustrates the comparative outcomes for the tasks of Hopper stand and Quadruped walk, with an exhaustive compilation of results for all other tasks reserved for the Appendix B.

In these challenging tasks, the DAC method struggles to reach expert policy performance without the stabilizing effects of e-RND and p-RND, particularly failing in the Quadruped walk task. As shown in Figure 6, p-RND exhibits faster convergence than DAC, mainly due to its use of intrinsic rewards to enhance exploration. Meanwhile, e-RND improves convergence by relabeling expert samples in the replay buffer as positive, reducing the discriminator's overfitting.

The EXPOSE algorithm effectively combines e-RND and p-RND. p-RND increases sample diversity in the replay buffer, enabling e-RND to identify a wider range of positive samples and counteract overfitting. The combined effect of e-RND and p-RND surpasses their individual performances, demonstrating the advantage of their synergy.

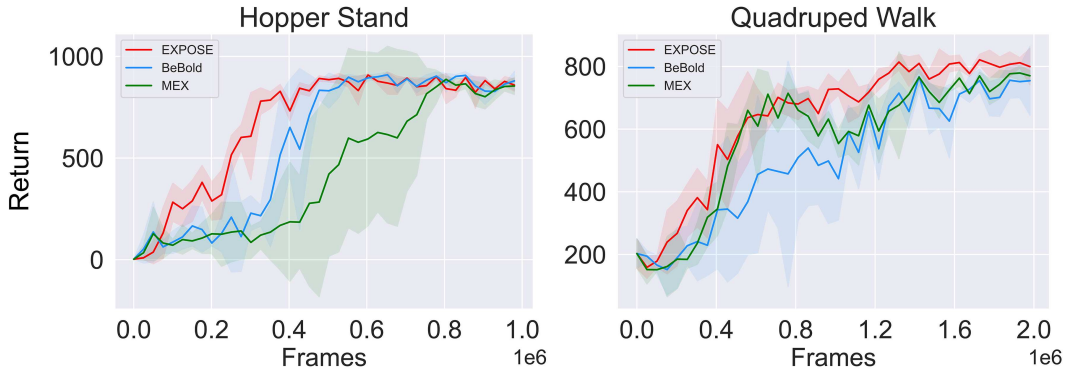


Figure 7 Comparison of different exploration methods on Hopper stand and Quadruped walk.

Furthermore, we delve into how p-RND stands out among other exploration techniques. Unlike p-RND, which stimulates exploration from any given state towards novel states, BeBold [40] is more restrictive, advocating for transitions from states that have been previously explored to those that are yet to be discovered. This targeted approach, while effective in certain situations, may result in a reduced incentive for the agent when moving between two areas that have received only cursory exploration, potentially slowing down the overall discovery process. On the other hand, MEX [41] simplifies things by combining the tasks of estimating and planning into one goal for the agent. While this makes the learning process less complicated, it might not help the agent as much in more complicated environments where it needs to think about many things at once.

Our results, shown in Figure 7, suggest that p-RND does a bit better than the other methods. We think this is because it works well together with e-RND. They help each other out, allowing the agent to explore more broadly and deeply, which is good for learning. This teamwork between p-RND and e-RND gives the agent a better chance of finding the best way to act in a situation.

5.4 Recognition ability of expert RND

We also conducted experiments to validate the effectiveness of e-RND in recognizing expert data, setting the parameter α in (7) to 0.98. In Figure 8, we demonstrate e-RND's capability for expert data recognition across four datasets: (a) **Expert**, expert data employed in the training of e-RND; (b) **Supp Expert**, supplementary expert data generated by the same expert policy; (c) **Buffer**, data sampled from the replay buffer during training; and (d) **Random**, data sampled by a random policy. The expert ratio in Figure 8 represents the percentage of expert samples filtered by e-RND. As shown in the figure, e-RND successfully recognizes a significant proportion of samples in both expert datasets (a) and (b). This underscores e-RND's ability to identify the expert data used in its training, while also highlighting its generalization capability to recognize additional expert data. As the policy's performance improves, the proportion of expert samples filtered by e-RND in (c) increases accordingly. Conversely, for the random dataset (d), e-RND consistently maintains a lower proportion of recognized samples. This demonstrates e-RND's effectiveness in distinguishing genuine expert data from random samples.

As shown in Figure 9, we monitor the single-step oracle rewards of samples throughout the training process. By comparing the rewards of true expert samples (Expert), policy samples accepted by the filter (Filter+), and samples rejected by the filter (Filter−), we observe that the accepted samples consistently maintain high rewards close to the expert baseline, while rejected samples stay at a significantly lower level throughout training. The slight gap between the Filter+ and Expert curves during the early stages of training is due to the fact that the policy has not yet learned to produce many high-quality actions; at the start, the policy produces a low density of high-quality samples, resulting in a higher proportion of medium-quality samples within the subset that meets the e-RND threshold. However, as the policy improves and begins generating more expert-like behaviors, the average reward of the filtered samples rapidly converges to the expert level. The persistent and substantial margin between the Filter+ and Filter− curves provides strong evidence that e-RND effectively identifies high-quality data for support expansion, ensuring the robustness of the learning process.

In Figure 10, we compare the discriminator outputs of EXPOSE and PUGAIL, both designed to address the issue of discriminator overfitting. Both EXPOSE and PUGAIL enable the discriminator to assign high rewards to additional expert data. However, PUGAIL also tends to assign relatively high rewards to random samples, which can lead to an overestimation of the critic for these samples and contribute to suboptimal performance on certain tasks. In contrast, the discriminator trained with EXPOSE effectively differentiates between expert and random

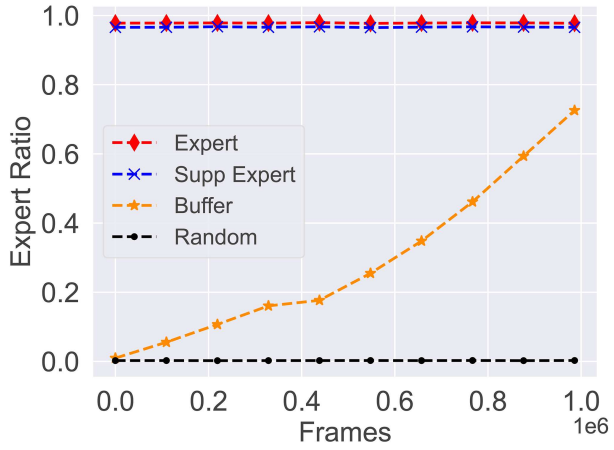


Figure 8 Proportions of samples accepted by e-RND across four datasets.

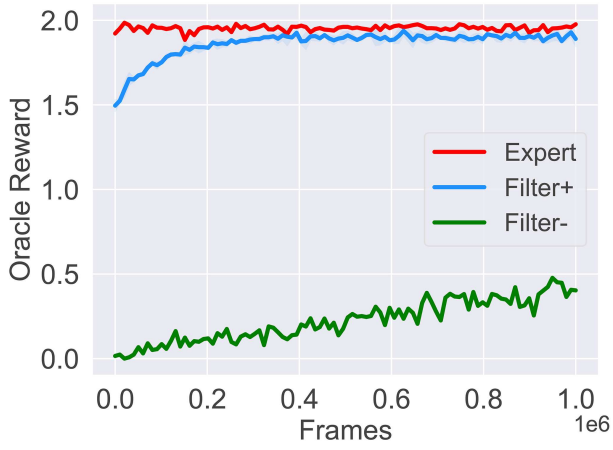


Figure 9 Oracle reward analysis of filtered samples. We compare the single-step oracle rewards of expert data (Expert), policy samples accepted by the filter (Filter+), and samples rejected by the filter (Filter−).

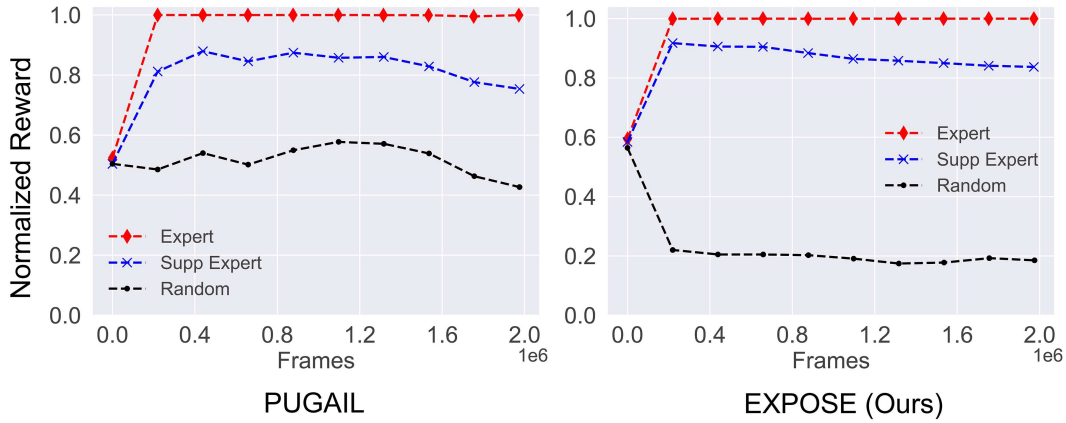


Figure 10 Normalized outputs of the discriminator trained with PUGAIL and EXPOSE across various datasets.

samples, avoiding overfitting to the limited set of expert training data.

6 Related work

IL is a paradigm where agents enhance their capabilities by observing and replicating the actions of experts, without relying on explicit rewards from the environment. This field is primarily divided into two approaches: behavioral cloning (BC) [42, 43] and inverse reinforcement learning (IRL) [44]. BC is a straightforward method that trains agents on a static set of demonstrations, treating the learning process as a supervised task. While effective for tasks with limited variations, BC may struggle in situations that deviate from the training data, leading to suboptimal performance on out-of-distribution samples [45]. In contrast, IRL aims to recover the expert's policy by inferring the expert's reward function, cultivating agents capable of thriving in unprecedented situations. However, IRL often requires substantial interaction with the environment to learn an effective reward function, which may be impractical in scenarios with limited samples.

GAIL [17] is a popular AIL method inspired by maximum entropy IRL [44] and generative adversarial networks (GANs) [46]. GAIL utilizes the framework of GANs to learn from expert demonstrations without the need for an explicit reward function. The core mechanism of GAIL involves training a discriminator to distinguish between the behaviors of an expert policy and those generated by the learning policy. The learning policy is then trained to maximize the discriminator's confusion by producing behaviors that mimic the expert's demonstrations. A considerable body of work has focused on improving GAIL's stability [18, 47] and enhancing its sample efficiency by porting it to an off-policy setting [24]. However, the majority of prior AIL algorithms [17, 18, 24] were initially designed with

low-dimensional tasks in mind, neglecting the unique challenges and potential pitfalls that arise when scaling these algorithms to high-dimensional environments. Recent investigations have illuminated the challenges AIL algorithms encounter in such settings. To address these issues, several studies have focused on learning representations that effectively reduce the dimensionality of discriminator inputs. For example, VAIL [48] enhances feature representation by constraining the mutual information between the representation and the discriminator input. TRAIL [49] guides the discriminator away from using task-irrelevant features for distinguishing between expert and generated samples, ensuring informative rewards from the discriminator. Meanwhile, HashReward [27] employs a strategy involving supervised hashing and contrastive learning to achieve dimensionality reduction.

However, none of these approaches tackle the fundamental issue of discriminator overfitting, which is a key factor contributing to the failure of the AIL algorithm in high-dimensional environments. PUGAIL [29] highlights that the discriminator often provides a low reward signal regardless of the generator's sample quality. Ref. [26] also demonstrated that the generated successful trajectories treated as negative samples during discriminator training can have a detrimental effect on GAIL training. The root cause of the overfitting problem lies in the fact that the high-quality samples in the generated samples are not trained as positive samples, resulting in low rewards for these samples. Ref. [30] pointed out that an excessively accurate discriminator may pose a challenge to agent training. To mitigate these issues, Ref. [30] conducted experiments incorporating common AIL regularizers such as gradient penalty, spectral norm, mixup, etc. Nevertheless, as TRAIL emphasizes, unstructured regularization may not effectively prevent the discriminator from fitting to visual features that differ systematically between the agent and the demonstrator. Additionally, employing strong regularization might exacerbate the problem, leading the discriminator to exploit spurious features rather than task-relevant ones. It is important to note that the overfitting challenge becomes more pronounced in high-dimensional environments.

The overfitting issue stems from the adversarial training process, prompting several IL methods to construct reward functions by estimating expert support, thus avoiding adversarial training. Notably, SQIL [50] trains a policy by assigning a reward of 1 to samples from expert data and 0 to samples in the replay buffer. RED [32] leverages RND to estimate expert policy support, yielding a continuous reward function. Ref. [21] also utilized RND to estimate expert support and augment expert data, thereby mitigating the compound error problem of BC. Additionally, DRIL [51] employs ensemble learning principles to construct a reward function based on policy disagreement among models trained through BC.

While certain non-adversarial approaches—such as RED [32] and the framework by [21]—leverage RND to estimate expert support and circumvent adversarial training, they frequently encounter challenges with reward misspecification. In contrast, EXPOSE retains the adversarial framework, integrating RND as an auxiliary robustness mechanism rather than a replacement. By employing e-RND to filter positive samples, we strategically expand the expert support, enabling the discriminator to learn a more robustly generalized decision boundary. This synergy addresses the core overfitting challenges in high-dimensional settings while preserving the foundational advantages of distribution matching inherent in AIL.

Furthermore, we observe that most existing AIL algorithms concentrate exclusively on replicating expert behaviors, frequently overlooking the necessity of active exploration. In complex, high-dimensional tasks, focusing solely on distribution matching often leads to mode collapse, where the agent's behavioral patterns become trapped in a narrow set of states. EXPOSE introduces p-RND to encourage the agent to explore diverse behavioral patterns through intrinsic motivation. This creates a critical synergy: while e-RND ensures the reward signal remains informative beyond narrow expert trajectories, p-RND provides the agent with the drive to actively explore these generalized regions, preventing premature convergence to local optima.

7 Conclusion

In this work, we establish a connection between discriminator overfitting in high-dimensional environments and the suboptimal performance that results from underestimating the expert support set. Motivated by this insight, we present EXPOSE, a practical AIL algorithm designed to mitigate overfitting issues by expanding the expert support set. Specifically, EXPOSE employs two RND components: e-RND, which identifies and relabels samples within the expert policy support as positive during the discriminator training process, and p-RND, which encourages exploration to provide more candidate samples for e-RND to filter. We demonstrate that our method achieves state-of-the-art performance on two pixel-based benchmarks, DMControl and Meta-World. Notably, our approach significantly improves the asymptotic performance, convergence speed, and stability of AIL algorithms in challenging high-dimensional environments. We also conducted ablation experiments to assess the impact of e-RND and p-RND on performance, confirming the effectiveness of e-RND in identifying and filtering expert data.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62276126), "111 Center" (Grant No. B26023), and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM118).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 2018
2 Wu C, Zhang Z. Surfing information: the challenge of intelligent decision-making. Intell Comput, 2023, 2: 0041
3 Luo F M, Xu T, Lai H, et al. A survey on model-based reinforcement learning. Sci China Inf Sci, 2024, 67: 121101
4 Yuan L, Jiang T, Li L H, et al. Robust cooperative multi-agent reinforcement learning via multi-view message certification. Sci China Inf Sci, 2024, 67: 142102
5 Qin R J, Yu Y. Learning in games: a systematic review. Sci China Inf Sci, 2024, 67: 171101
6 Cao J H, Yuan L, Wang J H, et al. LINDA: multi-agent local information decomposition for awareness of teammates. Sci China Inf Sci, 2023, 66: 182101
7 Qin R, Chen F, Wang T, et al. Multi-agent policy transfer via task relationship modeling. ArXiv:2203.04482
8 Hussein A, Gaber M M, Elyan E, et al. Imitation learning: a survey of learning methods. ACM Comput Surv, 2017, 50: 1–35
9 Osa T, Pajarinen J, Neumann G, et al. An algorithmic perspective on imitation learning. ArXiv:1811.06711
10 Liu X, Xu F, Zhang X, et al. How to guide your learner: imitation learning with active adaptive expert involvement. In: Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, 2023. 1276–1284
11 Zheng B, Verma S, Zhou J, et al. Imitation learning: progress, taxonomies and challenges. IEEE Trans Neural Netw Learn Syst, 2024, 35: 6322–6337
12 Chu W J, Yu A, Zhang W, et al. Perception field based imitation learning for unlabeled multi-agent pathfinding. Sci China Inf Sci, 2024, 67: 152107
13 Codevilla F, Müller M, López A, et al. End-to-end driving via conditional imitation learning. In: Proceedings of the International Conference on Robotics and Automation, 2018. 4693–4700
14 Giusti A, Guzzi J, Cireşan D C, et al. A machine learning approach to visual perception of forest trails for mobile robots. IEEE Robot Autom Lett, 2016, 1: 661–667
15 Shi J, Yu Y, Da Q, et al. Virtual-taobao: virtualizing real-world online retail environment for reinforcement learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019. 4902–4909
16 Fei C, Wang B, Zhuang Y, et al. Triple-GAIL: a multi-modal imitation learning framework with generative adversarial nets. In: Proceedings of International Joint Conference on Artificial Intelligence, 2020. 2929–2935
17 Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 4565–4573
18 Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2018
19 Lin J, Zhang Z. ACGAIL: imitation learning about multiple intentions with auxiliary classifier GANs. In: Proceedings of Pacific Rim International Conference on Artificial Intelligence, 2018. 321–334
20 Ghasemipour S K S, Zemel R, Gu S. A divergence minimization perspective on imitation learning methods. In: Proceedings of Conference on Robot Learning, 2020. 1259–1277
21 Han F, Zhang Z. Expert data augmentation in imitation learning (student abstract). In: Proceedings of AAAI Conference on Artificial Intelligence, 2023. 16220–16221
22 Rajaraman N, Yang L F, Jiao J, et al. Toward the fundamental limits of imitation learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 2914–2924
23 Xu T, Li Z, Yu Y. Error bounds of imitating policies and environments. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 15737–15749
24 Kostrikov I, Agrawal K K, Dwibedi D, et al. Discriminator-actor-critic: addressing sample inefficiency and reward bias in adversarial imitation learning. In: Proceedings of International Conference on Learning Representations, 2019
25 Kostrikov I, Nachum O, Tompson J. Imitation learning via off-policy distribution matching. In: Proceedings of International Conference on Learning Representations, 2020
26 Zolna K, Saharia C, Boussioux L, et al. Combating false negatives in adversarial imitation learning. ArXiv:2002.00412
27 Cai X Q, Ding Y X, Jiang Y, et al. Imitation learning from pixel-level demonstrations by HashReward. In: Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, 2021. 279–287
28 Wang T, Karnwal N, Atanasov N. Latent policies for adversarial imitation learning. ArXiv:2206.11299
29 Xu D, Denil M. Positive-unlabeled reward learning. In: Proceedings of Conference on Robot Learning, 2021. 205–219
30 Orsini M, Raichuk A, Hussenot L, et al. What matters for adversarial imitation learning? In: Proceedings of Advances in Neural Information Processing Systems, 2021. 14656–14668

31 Burda Y, Edwards H, Storkey A J, et al. Exploration by random network distillation. In: Proceedings of the International Conference on Learning Representations, 2019

32 Wang R, Ciliberto C, Amadori P V, et al. Random expert distillation: imitation learning via expert policy support estimation. In: Proceedings of International Conference on Machine Learning, 2019. 6536–6544

33 Tassa Y, Doron Y, Muldal A, et al. DeepMind control suite. ArXiv:1801.00690

34 Yu T, Quillen D, He Z, et al. Meta-World: a benchmark and evaluation for multi-task and meta reinforcement learning. In: Proceedings of Conference on Robot Learning, 2020. 1094–1100

35 Puterman M L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Hoboken: John Wiley & Sons, 2014

36 Syed U, Bowling M, Schapire R E. Apprenticeship learning using linear programming. In: Proceedings of International Conference on Machine Learning, 2008. 1032–1039

37 Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 2126–2135

38 Yarats D, Fergus R, Lazaric A, et al. Mastering visual continuous control: improved data-augmented reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2022

39 Liu M, He T, Zhang W, et al. Visual imitation learning with patch rewards. In: Proceedings of the International Conference on Learning Representations, 2022

40 Zhang T, Xu H, Wang X, et al. BeBold: exploration beyond the boundary of explored regions. ArXiv:2012.08621

41 Liu Z, Lu M, Xiong W, et al. Maximize to explore: one objective function fusing estimation, planning, and exploration. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 22151–22165

42 Pomerleau D A. ALVINN: an autonomous land vehicle in a neural network. In: Proceedings of Advances in Neural Information Processing Systems, 1988. 305–313

43 Torabi F, Warnell G, Stone P. Recent advances in imitation learning from observation. In: Proceedings of International Joint Conference on Artificial Intelligence, 2019. 6325–6331

44 Ziebart B D, Maas A L, Bagnell J A, et al. Maximum entropy inverse reinforcement learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2008. 1433–1438

45 Ross S, Gordon G J, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of Artificial Intelligence and Statistics, 2011. 627–635

46 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680

47 Dadashi R, Hussenot L, Geist M, et al. Primal Wasserstein imitation learning. In: Proceedings of International Conference on Learning Representations, 2021

48 Peng X B, Kanazawa A, Toyer S, et al. Variational discriminator bottleneck: improving imitation learning, inverse RL, and GANs by constraining information flow. In: Proceedings of International Conference on Learning Representations, 2019

49 Zolna K, Reed S, Novikov A, et al. Task-relevant adversarial imitation learning. In: Proceedings of Conference on Robot Learning, 2021. 247–263

50 Reddy S, Dragan A D, Levine S. SQIL: imitation learning via reinforcement learning with sparse rewards. In: Proceedings of International Conference on Learning Representations, 2020

51 Brantley K, Sun W, Henaff M. Disagreement-regularized imitation learning. In: Proceedings of International Conference on Learning Representations, 2020