DISTRIBUTIONAL INVERSE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a distributional framework for offline Inverse Reinforcement Learning (IRL) that jointly models uncertainty over reward functions and full distributions of returns. Unlike conventional IRL approaches that recover a deterministic reward estimate or match only expected returns, our method captures richer structure in expert behavior, particularly in learning the reward distribution, by minimizing first-order stochastic dominance (FSD) violations and thus integrating distortion risk measures (DRMs) into policy learning, enabling the recovery of both reward distributions and distribution-aware policies. This formulation is well-suited for behavior analysis and risk-aware imitation learning. Empirical results on synthetic benchmarks, real-world neurobehavioral data, and MuJoCo control tasks demonstrate that our method recovers expressive reward representations and achieves state-of-the-art imitation performance.

1 INTRODUCTION

Inverse Reinforcement Learning (IRL) aims to infer an expert's underlying reward function and policy from observed trajectories collected under unknown dynamics. IRL has been successfully applied in diverse domains, including robotics (Vasquez et al., 2014; Wu et al., 2024), animal behavior modeling (Ashwood et al., 2022; Ke et al., 2025), autonomous driving (Rosbach et al., 2019; Wu et al., 2020), and fine-tuning of large language models (Zeng et al., 2025). A pioneering work in this field, the Maximum Entropy IRL (MaxEntIRL) framework (Ziebart et al., 2008), formulates reward learning as a likelihood optimization problem and interprets expert policies as Boltzmann distributions over returns. Follow-up works have extended this framework to improve reward inference stability and generalization (Arora & Doshi, 2021; Garg et al., 2021; Zeng et al., 2022).

Despite these advances, most IRL methods assume that the expert's reward function is deterministic, thereby recovering only a point estimate, i.e., $r(s, a) \in \mathbb{R}$ for every state s and action a. This assumption, however, limits expressiveness in real-world settings where reward signals are inherently stochastic. For instance, in robotic manipulation tasks involving deformable or fragile objects (Yin et al., 2021), contact uncertainty introduces reward variability for identical state-action pairs—variability that directly influences the learned policy's robustness and safety. Similarly, in neuroscience, dopaminergic neuron activity has been shown, as reward signals, to drive animal behavior via RL policies (Markowitz et al., 2023b). Yet, dopamine signals exhibit significant trial-to-trial variations, suggesting that behavior may arise from an underlying stochastic reward distribution. These challenges are further amplified in offline IRL settings, where interaction with the environment is unavailable and the algorithm must fully rely on fixed demonstrations.

These examples highlight that in many real-world scenarios, demonstrations may be generated under stochastic reward functions, i.e., $r(s, a)$ is a random variable. This motivates the need to go beyond point estimates and instead recover the full distribution of rewards. Prior works such as Bayesian IRL (BIRL) methods infer a posterior over reward parameters using Markov chain Monte Carlo (MCMC) (Ramachandran & Amir, 2007), Maximum a posteriori (MAP) estimation (Choi & Kim, 2011), or variational inference (Chan & van der Schaar, 2021), but primarily capture uncertainty over the parameters of a deterministic reward function. More importantly, BIRL still optimizes the expected return, following the MaxEntIRL framework, failing to exploit the richer structure present in the full return distribution induced by stochastic rewards. In other words, if reward learning in IRL is based solely on maximizing expected return, then the resulting policy is influenced only by the mean and remains insensitive to the variance or higher-order moments of the reward. As a result, such an approach provides insufficient signal for accurately estimating the full reward distribution.

However, it remains unclear how to effectively learn reward distributions directly from expert demonstrations. Conventional MaxEntIRL fails to capture higher-order moments of the return, motivating the use of statistical distances between return distributions. Yet, such approaches introduce significant challenges for policy learning, the dual problem to reward inference, because most statistical distances couple the estimated return distribution with the (unknown) expert return distribution. This coupling exacerbates compounding errors and prevents leveraging established distributional RL techniques. Consequently, a principled framework is needed that enables reward distribution learning while simultaneously supporting return distribution estimation in the offline IRL setting.

To this end, we introduce *Distributional Inverse Reinforcement Learning* (DistIRL), a novel framework that explicitly models both the distributional nature of reward and the return. This allows us to capture stochasticity not only from transitions and policies but also from the reward function itself. Specifically, for reward learning, instead of matching expected returns as in MaxEntIRL, we propose to match the full return distribution using a First-order Stochastic Dominance (FSD) criterion. This allows us to capture not only the mean but also higher-order moments of the return distribution and thus capturing the full landscape of reward distributions, leading to a richer and more faithful estimate of the underlying reward structure. To the best of our knowledge, *this is the first work that learn the full distribution of the reward function in a principled manner.*

It is important to note that while our framework incorporates risk-sensitive policy learning, risk sensitivity primarily serves as a mechanism that enables robust reward distribution learning in the offline IRL setting. The connection is explained in detail in Sec. 4.2. Our contributions in this paper are summarized as follows:
(1) **Reward Distribution Learning.** We propose an intuitive framework for learning reward distributions in the offline IRL setting. With FSD objective emphasizing the match of the entire distribution, we are able to learning reward distributions beyond the first moment.
(2) **Distribution-aware Policy Learning.** Our algorithm learns the return distribution and recovers the distribution-aware policy, extending the modeling capability of IRL frameworks towards a broader range of behavior analyses and facilitating imitation learning in risk-sensitive scenarios.
(3) **Empirical Validation.** We demonstrate that our method recovers meaningful reward distributions on synthetic and real-world datasets, including neurobehavioral data (first-time studied for IRL). Our algorithm also achieves state-of-the-art performance on high-dimensional robotic control tasks in offline IRL settings.

2 Related Work

Inverse Reinforcement Learning Traditional offline IRL algorithms recover a reward function by matching expert feature expectations or maximizing an entropy-regularized likelihood. Apprenticeship learning (Abbeel & Ng, 2004) and MaxEntIRL (Ziebart et al., 2008; 2010) infer a deterministic reward whose induced policy reproduces expert behavior in expectation. Subsequent deep IRL variants incorporate neural network function approximators in the online setting (Ho & Ermon, 2016; Jeon et al., 2018; Wulfmeier et al., 2015; Ni et al., 2021; Garg et al., 2021; Zeng et al., 2022), which the policy further interacts with the environment but still match only the expected return. As a result, these approaches cannot capture risk preferences or higher-order statistics of the reward distribution present in many real-world tasks. In addition, online IRL methods require interactive access to a simulator during training, which is unsuitable for offline settings where reproducing the environment is undesirable or infeasible, e.g. modeling mouse behavior in a maze. Finally, while recent work has explored risk-aware policy learning within the IRL framework (Singh et al., 2018; Lacotte et al., 2019; Cheng et al., 2023), these approaches still assume a deterministic reward model, failing to capture the stochasticity of rewards in many real-world problems. We show a detailed comparison of IRL methods across modeling assumptions in Appendix A.

Bayesian Imitation Learning Bayesian IRL (BIRL) methods infer a posterior distribution over reward parameters to quantify uncertainty in reward estimation. Ramachandran and Amir (Ramachandran & Amir, 2007) introduces the first Bayesian IRL, using MCMC to sample from the reward posterior under a Boltzmann-rationality likelihood. Follow-up works use the same framework to handle larger state spaces and richer reward priors (Choi & Kim, 2011; Levine et al., 2011; Chan & van der Schaar, 2021; Li et al., 2023). Although these methods capture parameter uncertainty, they still rely on expected-return assumptions and do not exploit the full return distribution. Moreover, BIRL with a reward distribution fails to model continuous action spaces as obtaining the likelihood is computationally intractable for passing the gradient to the reward posterior. In this work, we propose a scalable algorithm framework for learning the full reward distributions.

Distributional Reinforcement Learning DistRL extends classical value-based methods by modeling the full distribution of returns rather than only their expectation. Early work, such as Categorical DQN (C51) (Bellemare et al., 2017) and Quantile Regression DQN (QR-DQN) (Dabney et al., 2018b), demonstrates that learning a distributional critic improves stability and sample efficiency. More recent advances include Implicit Quantile Networks (IQN) (Dabney et al., 2018a), Implicit Q-Learning (Kostrikov et al., 2021), Multivariate Distribution RL (Wiltzer et al., 2024), and Diffusion Process for RL (Hansen-Estruch et al., 2023; Li et al., 2024). Note that DistRL still inherently maximizes the expected return. Risk-sensitive extensions (Lim & Malik, 2022; Schneider et al., 2024) that optimize risk measures like CVaR, show that one can directly shape policies by tailoring decisions to specific regions of the return distribution. While these methods are widely adopted in RL, the IRL counterparts (Lee et al., 2022; Karimi & Ebadzadeh, 2025) with a distributional critic are limited in scope. These methods use a distributional critic to model return distributions and extract expert policies, but still assume deterministic reward functions, and take on MaxEntIRL as the blueprint, i.e., matching the mean of the return distribution.

3 Preliminaries

We model an environment as a discounted Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P(s'|s,a)$ the transition kernel, $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in [0,1)$ the discount factor. A policy $\pi(a|s)$ induces a return $Z^\pi = \sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t)$. The state-value and action-value functions under π are defined as

$$V^\pi(s) = \mathbb{E}\big[Z^\pi | s_t = s\big], \qquad Q^\pi(s,a) = \mathbb{E}\big[Z^\pi | s_t = s, a_t = a\big].$$

They satisfy the Bellman equations

$$V^\pi(s) = \mathbb{E}_{a\sim\pi, s'\sim P}\left[r(s,a) + \gamma V^\pi(s')\right], \quad Q^\pi(s,a) = \mathbb{E}_{s'\sim P}\big[\, r(s,a) + \gamma\, \mathbb{E}_{a'\sim\pi}[Q^\pi(s',a')]\big].$$

We also define the *occupancy measure* of π as $d^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\, \Pr(s_t = s)\,\pi(a|s)$, which satisfies $\sum_{s,a} d^\pi(s,a) = 1$ and characterizes the long-run state-action visitation probability.

3.1 Distributional RL and Risk-Sensitive Control

Rather than estimating only $\mathbb{E}[Z^\pi]$, distributional RL models the entire return distribution that obeys the *distributional Bellman operator* $\mathcal{T}^\pi$ (Bellemare et al., 2017):

$$Z^\pi(s,a) = \sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t),$$
$$\mathcal{T}^\pi Z(s,a) \overset{\mathrm{D}}{=} r(s,a) + \gamma\, Z\big(s', \pi(s')\big),$$

where $V \overset{\mathrm{D}}{=} U$ denotes equality of probability laws, indicating random variables $\{V, U\}$ are distributed according to the same law. A popular parameterization uses quantile regression: one approximates $Z^\pi(s,a)$ by N quantiles $\boldsymbol{\theta}(s,a) = [\theta_1(s,a), ..., \theta_N(s,a)] : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^N$ at fractions (quantile levels) $\tau_i = i/N$, for $i = 1, \ldots, N$. In other words, the quantile distribution of $Z^\pi(s,a)$ is represented a uniform probability distribution supported on $\{\theta_i(s,a)\}_{i=1}^N$: $Z^\pi(s,a) = \frac{1}{N}\sum_{i=0}^{N}\delta_{\theta_i}(s,a)$ where δ_{θ_i} denotes a Dirac at θ_i. An example of quantile functions is illustrated in Fig. 1, with θ and τ indicated.

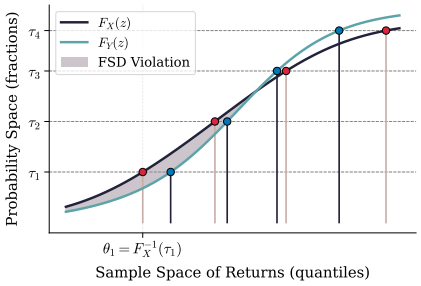

Figure 1: Illustration of quantile functions and first-order stochastic dominance (FSD).

To update the critic, instead of formulating the TD error, one can minimize the quantile Huber loss (Dabney et al., 2018b) with threshold $\kappa > 0$:

$$\rho_\tau^\kappa(\delta) = \big|\tau - \mathbf{1}\{\delta < 0\}\big|\, H_\kappa(\delta), H_\kappa(\delta) = \begin{cases} \frac{1}{2}\,\delta^2, & |\delta| \le \kappa, \\ \kappa\,|\delta| \;-\; \frac{1}{2}\,\kappa^2, & |\delta| > \kappa. \end{cases} \tag{1}$$

In distributional RL with N quantile fractions $\{\tau_i\}$, the loss for the critic is defined as

$$\min_\theta \mathcal{L}_{\mathrm{QR}}(\theta) = \min_\theta \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{\tau_i}\left(\delta_{ij}\right), \delta_{ij} = r + \gamma\,\theta_j(s',a') - \theta_i(s,a). \tag{2}$$

Once the return distribution is learned, one can optimize risk measures M, e.g. Conditional Value at Risk (CVaR) (Rockafellar et al., 2000), by maximizing $\mathrm{CVaR}\big(Z^\pi\big)$ rather than $\mathbb{E}[Z^\pi]$, yielding risk-sensitive policies.

3.2 Maximum Entropy Inverse Reinforcement Learning

Given demonstrations $\{(s_t, a_t)\}_{t\geq 1}$ collected by an unknown expert policy π^E, MaxEntIRL (Ziebart et al., 2008) aims to recover the unknown policy, and the corresponding reward function r which the policy is optimized to. Specifically, we consider the following formulation (Ho & Ermon, 2016):

$$\max_{\pi} \min_{r} \mathbb{E}_{d^\pi}[r(s,a)] - \mathbb{E}_{d^{\pi^E}}[r(s,a)] + \mathcal{H}(\pi) + \psi(r), \tag{3}$$

where $\mathcal{H} := \mathbb{E}_{d\pi}[-\log \pi(a|s)]$ denotes the entropy, and ψ is a general convex regularizer. This formulation reduces to MaxEntIRL if $\psi = 0$. If $\psi = \mathrm{KL}(q(r)||p_0(r))$, it can be seen as a BIRL framework, since the optimal policy follows a Boltzmann distribution of the action-values[1].

4 Distributional Inverse Reinforcement Learning Framework

In our model, we treat the reward as a *distribution* rather than a deterministic function. During optimization, the first two terms in Eq. 3, $\mathbb{E}_{d^\pi}[r(s,a)] - \mathbb{E}_{d^{\pi^E}}[r(s,a)]$, enforce *mean dominance*—that is, the learned reward should yield a higher expected return for the expert policy than for any arbitrary policy. At optimality, this difference becomes zero, indicating *mean matching* between expert and agent returns. However, if the reward is inherently a distribution, mean matching alone fails to capture the relationship between the expert's return distribution and the agent's in its entirety. This leads to a loss of higher-order information in the reward. To accurately model the full reward distribution, we must impose a *distributional form of dominance* during optimization, ensuring that the entire return distribution is aligned at optimality, not just the mean.

Let's consider a notion of order in term of the entire distributions.

Definition 4.1 (First-Order Stochastic Dominance (FSD) (Hadar & Russell, 1969))**.** Let X and Y be real-valued integrable random variables with cumulative distribution functions F_X and F_Y. We say that X *first-order stochastically dominates* Y, written as $X \succeq_{\mathrm{FSD}} Y$, if $F_X(z) \leq F_Y(z), \forall\, z \in \mathbb{R}$.

The concept of FSD is illustrated in Fig. 1. If we aim for $X \succeq_{\mathrm{FSD}} Y$, then the shaded region indicates a violation of this condition. FSD has an equivalent definition relating to utility functions, which further implies mean dominance.

Proposition 4.2 (Theorem 1-2 (Hadar & Russell, 1969))**.** *For real-valued X and Y, the following are equivalent:*

1. $F_X(z) \leq F_Y(z)$ *for all* $z \in \mathbb{R}$.
2. $\mathbb{E}[\,u(X)\,] \geq \mathbb{E}[\,u(Y)\,]$ *for every non-decreasing utility function* $u : \mathbb{R} \to \mathbb{R}$.

Corollary 4.3 (Mean Dominance). If $X \succeq_{\mathrm{FSD}} Y$, it follows that $\mathbb{E}[X] \geq \mathbb{E}[Y]$, as the identity utility $u(x) = x$ is non-decreasing.

We model the reward as a conditional distribution, $r_t \sim q(\cdot|s_t, a_t)$, and define the random return for a trajectory $(s_0, a_0, \dots)$ sampled from policy π as $Z^\pi = \sum_{t=0}^{\infty} \gamma^t r_t$. We now introduce the distributional counterpart to Eq. 3, the objective for distributional IRL, expressed as

$$\max_{\pi} \min_{r} \mathcal{L}(\pi, r) := \max_{\pi} \min_{r} \int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ dz + \mathcal{H}(\pi) + \psi(r), \tag{4}$$

where Z^E denotes the return distribution of the expert policy.

4.1 Learning Reward Distribution through Stochastic Dominance

From Eq. 4, the objective of the reward function is

$$\min_{r} \mathcal{L}_{\mathrm{FSD}}(\pi, r) + \psi(r) = \min_{r} \int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ dz + \psi(r). \tag{5}$$

This objective minimizes the violation of FSD, drawing inspiration from the Kolmogorov-Smirnov (K-S) test (Massey Jr, 1951). To model the reward distribution in a principled manner, we adopt a Bayesian learning framework. In particular, we define a likelihood function over the expert demonstrations $\mathcal{D}$ using the Energy-Based Model (EBM) formulation (LeCun et al., 2006): $p(\mathcal{D}|r) \propto \exp\left(-\mathcal{L}_{\mathrm{FSD}}(\pi, r)\right)$. We also introduce a *prior distribution* $p_0(r)$, which reflects our initial

[1]The Kullback-Leibler divergence is convex in its first argument when the second argument is fixed.

belief before observing any data. The goal is to infer the posterior distribution $p(r|\mathcal{D})$ using Bayes' rule. As direct inference under the EBM formulation is generally intractable, we adopt the variational inference framework (Blei et al., 2017) by introducing a *variational distribution* $q_\phi(r|s,a)$, parameterized by ϕ, to approximate the posterior and optimize the *evidence lower bound (ELBO)*:

$$\text{ELBO} = \mathbb{E}_{q_\phi(r|s,a)}\left[\log p(\mathcal{D}|r)\right] - \text{KL}\left(q_\phi(r|s,a) \,\|\, p_0(r)\right). \tag{6}$$

Substituting the energy-based likelihood into the ELBO yields:

$$\min_\phi \mathcal{L}_r(\phi) := \min_\phi \mathbb{E}_{q_\phi(r|s,a)}\left[\mathcal{L}_{\text{FSD}}(\pi, r)\right] + \text{KL}\left(q_\phi(r|s,a) \,\|\, p_0(r)\right). \tag{7}$$

Notice the natural relationship between KL and ψ. Formally, we learn the reward distribution by solving Eq. 7. To compute the gradient of the first term, we apply the Inverse Transform Sampling technique (Devroye, 2006). We use the empirical quantile to approximate the quantile of the return. Specifically, using the change of variable formula, and the relation between CDF and quantile, we have

$$\int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ dz = \int_0^1 \left[F^{-1}_{Z^\pi}(v) - F^{-1}_{Z^E}(v)\right]_+ \mathrm{d}v. \tag{8}$$

We provide a short proof of the above relation in Appendix B.1. To approximate F^{-1}_π, we draw N samples $\{z_n\}$ by Monte Carlo sampling $z_n = \sum_0^\infty \gamma^t r_t, r_t \sim q_\phi(\cdot|s_t, a_t)$, and form the empirical quantile using its order statistics $F^{-1}_{Z^\pi} \approx (z_{(-N)}, \ldots, z_{(1)})$. As a result, minimizing $\mathcal{L}_r(\phi)$ generalizes the usual IRL objective of matching expected returns by aligning higher-order moments beyond matching the mean.

4.2 Risk-aware Policy Learning

Once the inner minimization over r yields a fixed reward distribution, the policy, parameterized by φ, is updated by maximizing the following objective:

$$\max_\varphi \mathcal{L}_\pi(\varphi) = \max_\varphi \int_0^1 [F^{-1}_{Z^{\pi_\varphi}}(v) - F^{-1}_{Z^E}(v)]_+ dv + \mathcal{H}(\pi_\varphi). \tag{9}$$

Let's define $\mathcal{I}(v) := \mathbb{1}_{F^{-1}_{Z^{\pi_\varphi}}(v) \geq F^{-1}_{Z^E}(v)}$. Fig. 1 shows that $\mathcal{I}(v)$ takes the value 1 in regions where FSD is violated (shaded area), and 0 otherwise. We then rewrite the objective in Eq. 9 as

$$\int_0^1 \left(F^{-1}_{Z^{\pi_\varphi}}(v) - F^{-1}_{\pi^E}(v)\right) \mathcal{I}(v) dv + \mathcal{H}(\pi_\varphi). \tag{10}$$

Note that the indicator function $\mathcal{I}$ depends on the current policy, the expert policy, and the quantile level v. Conceptually, $\mathcal{I}$ assigns weight only to regions of the return distribution where FSD is violated. The policy now aims to increase these FSD violations—encouraging the agent to obtain higher return samples in those regions. This leads to a maximization scheme that is inherently risk-aware, as it requires reasoning over the full return distribution rather than just its expectation.

Unfortunately, directly optimizing Eq. 9 is intractable, as the indicator function $\mathcal{I}$ is not observable during training. To address this, we take a broader perspective on risk-aware policy learning and propose replacing $\mathcal{I}(v)$ with a risk measure that retains the goal of encouraging risk-sensitive behavior while yielding a tractable objective. Furthermore, we show that the resulting surrogate objective provides a weaker form of optimality, but under certain conditions, it can theoretically achieve the same optimum as Eq. 9. To present our new objective, we need a few essential concepts.

Definition 4.4 (Distortion function)**.** A distortion function ξ is a non-decreasing function $\xi : [0,1] \to [0,1]$ such that $\xi(0) = 0, \xi(1) = 1$.

Definition 4.5 (Distortion Risk Measure (DRM) (Dhaene et al., 2012))**.** For an integrable random variable X, and a distortion function ξ, a Distortion Risk Measure M_ξ is defined as

$$M_\xi(X) = \int_0^1 F^{-1}_X(v) d\tilde{\xi}(v), \tag{11}$$

where $\tilde{\xi} = 1 - \xi(1-v) \geq 0$ is the dual distortion function.

Common examples of DRMs and distortion functions are listed in Table 1. These measures offer various ways to quantify risk based on the return distribution. Intuitively, when $\tilde{\xi}$ is concave, it places greater emphasis on lower returns, thereby encouraging risk-averse behavior. To induce risk-aware policies using distortion $\xi(v)$, we need to maximize the DRM defined in Eq. 11.

Building on the above definitions, we propose replacing $\mathcal{I}(v)$ with $\tilde{\xi}(v)$ in Eq. 10, resulting in:

Table 1: Examples of distortion risk measures.

Risk Measure	$\xi(v)$	Interpretation
CVaR_α	$\min(v/\alpha, 1)$	Average of worst α-fraction of outcomes
Wang's Transform	$\Phi(\Phi^{-1}(v) + \lambda)$	$\lambda > 0$ implies risk-aversion, $\lambda < 0$ risk-seeking

$$\max_{\varphi} \int_0^1 \left(F^{-1}_{Z^\pi}(v) - F^{-1}_{Z^E}(v)\right) d\tilde{\xi}(v) + \mathcal{H}(\pi) = \max_{\varphi} \int_0^1 F^{-1}_{Z^\pi}(v) d\tilde{\xi}(v) + \mathcal{H}(\pi). \tag{12}$$

The equality is obtained as the expert policy does not depend on φ. We denote the final objective as

$$\max_{\varphi} \mathcal{L}_\pi(\varphi) := \max_{\varphi} M_\xi(Z^{\pi_\varphi}) + \mathcal{H}(\pi_\varphi) = \max_{\varphi} \int_0^1 F^{-1}_{Z^{\pi_\varphi}}(v) d\tilde{\xi}(v) + \mathcal{H}(\pi_\varphi), \tag{13}$$

where M_ξ is a chosen DRM with a distortion function ξ.

Relation to Eq. 9. Additionally, we know that $X \succeq_{\text{FSD}} Y \Rightarrow M_\xi(X) \geq M_\xi(Y)$ (Sereda et al., 2010). Then naturally one wonders what's the sufficient condition for FSD? We observe that the converse implication requires a stronger condition.

Proposition 4.6. *$M_\xi(X) \geq M_\xi(Y)$ for every distortion function ξ implies $X \succeq_{\text{FSD}} Y$.*

The proof is straightforward by observing that $M_\xi(X) - M_\xi(Y) = \int_0^1 (F^{-1}_X(v) - F^{-1}_Y(v)) d\tilde{\xi}(v)$ and the fact that $\tilde{\xi}(v) \geq 0$. We present a short proof in Appendix B. This implies that if we solve $\max_{\pi_\varphi} \int_0^1 \left(F^{-1}_{Z^{\pi_\varphi}}(v) - F^{-1}_E(v)\right) d\tilde{\xi}(v) + \mathcal{H}(\pi_\varphi)$ for every distortion function, we obtain the solution to Eq. 9. However, since optimizing over all utility conditions is intractable, our proposed objective serves as an approximation using a specific DRM. Nonetheless, under the conditions of the proposition, this surrogate objective can theoretically achieve the same optimality as Eq. 9.

4.3 Practical Algorithm

Algorithm 1: A DistIRL method with FSD objective

Input: Expert data $\mathcal{D} = \{(s_t^E, a_t^E)\}$, prior $p_0(r)$, risk measure ξ, step sizes $\eta^\theta, \eta^\varphi, \eta^\phi$
Output: Reward distribution $q_\phi(r|s,a)$; policy $\pi_\varphi(a|s)$

```
Initialize parameters of reward network φ, policy ϕ, and critic θ;
for k = 1 to K do
    Sample a mini-batch {(s_t^E, a_t^E)} from D;
    foreach (s_t^E, a_t^E) in mini-batch do
        For each s_t^E, sample a_t ~ π_ϕ(·|s_t^E), r_t ~ q_φ(·|s_t^E, a_t), r_t^E ~ q_φ(·|s_t^E, a_t^E);
    Compute return samples Z^{π_k}, Z^E;
    Critic update via quantile regression (Eq. 2): θ_{k+1} ← θ_k − η^θ ∇L_QR(θ_k);
    Policy update with distortion risk measure (Eq. 13): ϕ_{k+1} ← ϕ_k − η^ϕ ∇L_π(ϕ_k);
    Reward distribution update via FSD loss (Eq. 7): φ_{k+1} ← φ_k − η^φ ∇L_r(φ_k).
```

To enable tractable and expressive modeling of reward uncertainty, we parameterize the reward distribution $q_\phi(r|s,a)$, for example, using Azzalini's skew-normal distribution (Azzalini & Valle, 1996): $q_\phi(r|s,a) = \mathcal{SN}(\mu_\phi(s,a), \sigma^2_\phi(s,a); \alpha_\phi(s,a))$, where the mean $\mu_\phi(s,a)$, standard deviation $\sigma_\phi(s,a)$ and the skew parameter $\alpha_\phi(s,a))$ are outputs of a neural network with parameters ϕ. This choice allows for efficient sampling and computing regularization when using a standard normal prior. During training, for each state-action pair, we sample rewards $r_t \sim q_\phi(\cdot|s_t, a_t)$ to construct return samples for both the expert and the current policy.

Note that the choice of prior depends heavily on the task domain and the type of variability we expect in the reward signal. For example, skew-normal distributions can capture asymmetric reward uncertainty in tasks with systematic biases (e.g., contact-rich manipulation), whereas heavy-tailed priors may be more suitable when outliers or rare but significant events dominate the return structure. In contrast, the broader statistical learning community often defaults to Gaussian priors, primarily because of their analytical tractability, conjugacy with many likelihood models, and well-understood concentration properties. That said, DistIRL does not rely on a fixed distributional assumption. Any parameterized distribution p_θ whose log-density or quantile function is differentiable in θ is compatible with our framework, since the algorithm requires only gradient updates for learning.

To estimate the spectral risk measure $M_\xi(Z^\pi)$ for the policy, we follow an offline approach: we use states s_t drawn from the expert demonstration dataset, but sample actions $a_t^\pi \sim \pi_\theta(\cdot|s_t)$

from the current policy, and a reward $r_t \sim q_\phi(\cdot|s_t, a_t)$. Then we compute the return Z^π by taking the sum. For policy update, we first learn the critic by Off-policy Evaluation (OPE) (Sutton et al., 1998) on $(s_t, a_t, r_t, s_{t+1}, a^\pi_{t+1})$ where we use Quantile Regression with the Quantile Huber loss $\mathcal{L}_{QR}$ as in Eq. 2. We then update the risk-aware policy by solving $\min_\pi \mathrm{KL}\left(\pi(\cdot \mid s) \,\middle\|\, \frac{1}{Z}\exp\{M_\xi(Z^\pi(\cdot \mid s))\}\right)$, which corresponds to the KKT solution to Eq. 9, as originally introduced by Ziebart et al. (2008). We summarize the full procedure in Alg. 1. Notebly, the algorithm scheme follow the Two Timescale Stochastic Approximation (TTSA) scheme used in prior works (e.g. as in Zeng et al. (2022) and Wu et al. (2023)) to enable scalable and efficient learning. Regrettably, a rigorous convergence analysis is beyond the scope of this paper. But due to the contraction property of the distributional Bellman operator (Bellemare et al., 2017) with mild regularity conditions of Eq. 7, bounding the iteration of policy optimization may be achievable.

5 Experiment

5.1 Gridworld

We begin with a 5×5 gridworld environment where the agent is trained to navigate from the starting state $(2, 0)$ (left-center) to rewarding goal locations. Two high-reward states are placed at $(0, 4)$ (top-right) and $(4, 4)$ (bottom-right), with the top-right reward modeled as a stochastic outcome drawn from $\mathcal{N}(1, 1)$. The first column of Fig. 2 illustrates the ground-truth reward mean and variance.

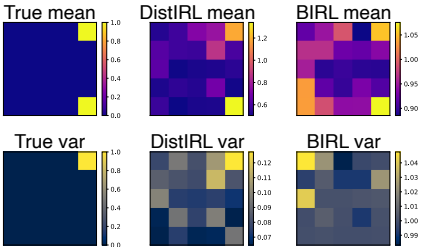

Figure 2: Inferring reward mean and variance in the gridworld example with 10 demonstrations.

This setup mimics an animal exploring an arena with two reward ports. In such compact environments, animals often display risk-averse behavior, i.e., avoiding locations where rewards have previously failed to appear (Mobbs et al., 2018; Daw et al., 2006). To model this, we collect 10 trajectories from a risk-averse agent trained under stochastic rewards. In 9 out of 10 episodes, the agent chooses the more reliable bottom-right goal. We then apply our DistIRL method to recover the full reward distribution. As shown in Fig. 2, using a symmetric Gaussian reward estimator combined with risk-averse policy learning, our approach not only identifies both high-reward states but also captures the variance at the top-right goal. This highlights the model's ability to infer higher-order moments of the reward from expert demonstrations.

As a baseline, we evaluate Bayesian IRL (BIRL) (Chan & van der Schaar, 2021; Mandyam et al., 2023; Bajgar et al., 2024). BIRL is a widely used framework that assumes a reward distribution but learns it by matching only the mean, without capturing the full distributional structure. We select BIRL because it is the method most comparable to ours in its ability to recover a reward distribution. BIRL reasonably recovers the mean reward but produces spurious high estimates in the lower-left corner. Furthermore, it fails to capture reward variance, emphasizing the need to enforce distance over the full distribution. Simply specifying a reward distribution, without integrating distribution-aware learning, fails to capture the true variance of the rewards.

5.2 Mouse Spontaneous Behavior

We apply our framework to a neuroscience dataset in which mice freely explore an arena without explicit rewards (Markowitz et al., 2023a). Behavior was recorded using a depth camera, and the raw trajectories were converted into sequences of discrete syllables (e.g., grooming, sniffing). We model these trajectories with an MDP, treating each syllable as a state and the next syllable as the action, yielding ten states and ten actions. In total, we analyzed 159 such state-action sequences. The dataset also includes a time-aligned one-dimensional trace of dopamine fluctuations from the dorsolateral striatum. Prior work (Markowitz et al., 2023a) showed that using dopamine as a reward enabled a simulated RL agent to reproduce observed transitions, suggesting IRL should recover a reward pattern resembling dopamine. Since dopamine varies even within the same state-action pair, the prior study used only its mean for simplicity. Here, we compare rewards learned under deterministic vs. distributional assumptions to assess how well they capture both the mean and the full distribution of dopamine signals.

We use both Azzalini's skew-normal distribution (denoted "S-") and the symmetric Gaussian as reward models for both DistIRL and BIRL. Fig. 3A) and B) show two example state-action pairs, illustrating the true dopamine fluctuation distribution alongside the estimated reward distributions from four methods. For each case, we display both the probability density function and the CDF, along with the corresponding means. Deterministic rewards (Det) are shown as pink dashed lines

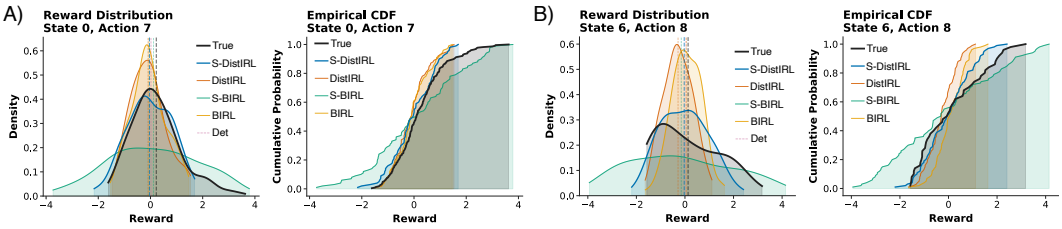

Figure 3: Learned reward distribution versus recorded dopamine signals and their empirical CDFs. in the density plots. Among all methods, S-DistIRL most accurately recovers the shape of the dopamine distribution, which is often right-skewed and multimodal. Its estimated mean also closely matches both the true mean and the deterministic estimate.

We also quantify the similarity between estimated rewards and actual dopamine distributions. In Fig. 4A), we report the correlation between the mean of dopamine fluctuations and the mean of the estimated reward across all mice and trajectories. Deterministic reward models yield moderate correlation, while DistIRL improves upon this, with S-DistIRL achieving the highest correlation overall. This finding indicates that incorporating full reward distributions, using suitable skewed distributional models, is essential for IRL to capture biologically meaningful reward signals. Fig. 4B) shows that, compared to BIRL, S-DistIRL also achieves a lower Wasserstein-1 distance between the estimated reward distribution and the actual dopamine distribution, indicating better alignment of the shape. Taken together, both qualitative examples and quantitative metrics support that modeling skewed reward distributions significantly enhances the ability to track dopamine fluctuations.

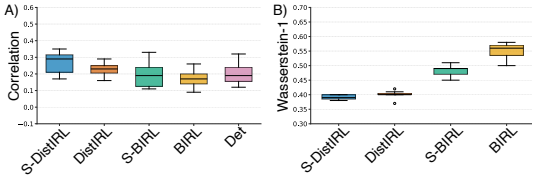

Figure 4: Left: Pearson correlation of the reward mean and dopamine level. Right: W-1 loss between learned distribution and dopamine level.

This is a scientifically interesting result showing that we can infer the reward structure directly from behavior data. While it is known that dopamine neurons encode reward-related signals (Schultz et al., 1997; Markowitz et al., 2023a), this is the first demonstration that not only is there a nontrivial correlation between the inferred and measured mean rewards (with a correlation around 0.3), but also that the full reward distribution recovered from behavior reasonably resembles the distribution of dopamine fluctuations. This suggests that detailed features of neuromodulatory signals, such as the variability in dopamine release, can be decoded from behavior alone, highlighting the potential of inverse modeling to uncover internal motivational states and their neural substrates.

5.3 MuJoCo Benchmarks

Risk-sensitive D4RL. In earlier experiments, we applied DistIRL to discrete state-action MDPs and compared it with BIRL. Here we extend the study to continuous MDPs to demonstrate DistIRL's scalability and generalizability. We evaluate our method on Risk-sensitive D4RL benchmarks, following the reward formulations introduced in recent robustness studies (Urpí et al., 2021). Specifically, the reward functions incorporate stochastic penalties triggered by safety-related conditions:
(1) **Half-Cheetah:** $R_t(s,a) = \bar{r}_t(s,a) - 70\mathbb{I}_{\nu>\bar{\nu}} \cdot \mathcal{B}_{0.1}$, where $\bar{r}_t(s,a)$ is the environment reward, ν is the forward velocity, and $\bar{\nu}$ is a velocity threshold ($\bar{\nu} = 4$ for the medium variant and $\bar{\nu} = 10$ for the easy variant). This penalty models rare but catastrophic robot failures at high speed.
(2) **Walker2D/Hopper:** $R_t(s,a) = \bar{r}_t(s,a) - p\mathbb{I}_{|\theta|>\bar{\theta}} \cdot \mathcal{B}_{0.1}$, where $\bar{r}_t(s,a)$ is the environment reward, θ is the pitch angle, $\bar{\theta}$ is a task-dependent threshold (0.5 for Walker2D-M/E and 0.1 for Hopper-M/E), and p is the penalty magnitude (30 for Walker2D and 50 for Hopper).

We train expert agents on these stochastic reward formulations using Risk-averse Distributional SAC, a variant of DSAC (Duan et al., 2021) with CVaR objective, and collect 10 demonstration trajectories. We then evaluate DistIRL against several state-of-the-art baselines. Results are averaged over 5 random seeds. We use a standard normal as the prior due to its general applicability, in the setting of not knowing the underlying true reward distribution.

Table 2 shows that our method consistently outperforms other **offline** IRL baselines under stochastic reward settings. Notice popular online methods such as GAIL (Ho & Ermon, 2016) are not applicable in this setting. **Offline ML-IRL** (Zeng et al., 2023) is a model-based MaxEntIRL method that relies on a separately trained transition model using additional non-expert data. Its poor performance

here is expected: the transition model was pretrained under risk-neutral rewards and does not align with the new expert data generated under risk-sensitive objectives, leading to severe distribution mismatch. **ValueDICE** (Kostrikov et al., 2019), a model-free offline MaxEntIRL baseline, also underperforms since it optimizes with respect to expected risk-neutral returns, while our experts follow risk-averse behavior. **Behavior Cloning (BC)** achieves moderately strong results, as it simply mimics the demonstrated actions without explicitly optimizing for either risk-neutral or risk-sensitive objectives. However, its performance is limited as the model overfit the limited demonstration data.

Table 2: Performance averaged over 5 seeds on Risk-sensitive D4RL.

Environment	IPMD (ours)	Offline ML-IRL	ValueDICE	BC	Expert
HalfCheetah	$\mathbf{3469 \pm 59}$	826 ± 231	1259 ± 78	2828 ± 281	3540 ± 44
Hopper	$\mathbf{886 \pm 1}$	192 ± 56	260 ± 10	346 ± 1	892 ± 3
Walker2d	$\mathbf{1526 \pm 148}$	240 ± 50	798 ± 311	1321 ± 26	1478 ± 200

To further validate the fidelity of our inferred return distributions from DistIRL and compare with the BIRL framework that only matches the mean, we collect 200 trajectories and sample its learned return distribution for each learned policy, plot against the expert's return distribution in Fig. 5. This shows that DistIRL's reward and policy model better align with the expert. We also report a Pearson correlation coefficient of 0.92 between the mean estimated by DistIRL and the mean of the true return. This indicates strong agreement and demonstrates that our inferred reward is an accurate proxy for the true reward model.

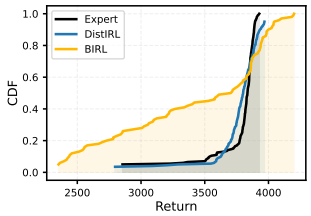

Figure 5: Return distributions comparison in HalfCheetah.

Risk-neutral D4RL. We also test our algorithm in conventional deterministic reward settings using D4RL's `medium-expert` trajectories (Fu et al., 2020). Table 3 shows our method achieves competitive or superior performance even without tailoring to deterministic assumptions, underscoring the generality of DistIRL. We want to emphasize that Offline ML-IRL requires additional data[2].

Ablation studies. We evaluate the contribution of different design choices by ablating our model under the HalfCheetah setting with right-skewed normal ($\mathcal{SN}_\alpha, \alpha > 0$) stochastic rewards and risk-averse expert policy, indicating the expert prefers conservative actions that yield more consistent rewards. Variants include: **Dis/Det**: Distributional or Deterministic rewards; **QR/TD**: Quantile Regression or TD-based critic; **FSD/Mean**: FSD loss or Mean matching. As shown in Table 4, which scales the performance between worst and best, using distributional rewards with FSD loss significantly outperforms mean-matching alternatives. Additionally, deterministic TD-learning with mean-matching (**Det-TD-Mean**) underperforms in learning risk-averse policies due to a lack of distributional supervision. This confirms the effectiveness of FSD-based reward learning and risk-sensitive policy optimization. Note that the BIRL framework aligns with our **Dis-TD-Mean** configuration; RIZE Karimi & Ebadzadeh (2025) aligns with **Det-Qt-Mean**, which performs the worst; **Det-TD-Mean** aligns with ValueDice but with an explicit reward estimation. Thus, in this ablation study, we treat them as a specific setting within our framework when benchmarking against other IRL approaches.

Table 3: Performance on deterministic reward settings (D4RL).

Environment	DistIRL (Ours)	Offline ML-IRL	ValueDICE	BC	Expert
HalfCheetah	7779 ± 228	$\mathbf{11231 \pm 585}$	4935 ± 2836	623 ± 56	12175 ± 91
Hopper	$\mathbf{3411 \pm 42}$	3347 ± 238	3073 ± 539	3236 ± 46	3512 ± 22
Walker2d	$\mathbf{4570 \pm 305}$	4201 ± 638	3191 ± 1888	2822 ± 979	5384 ± 52

Table 4: Ablation study on model setting. Performance scaled for clarity.

DistIRL (Ours)	Dis-Qt-Mean	Det-Qt-Mean	Dis-TD-FSD	Dis-TD-Mean	Det-TD-Mean
$\mathbf{1.0 \pm 0.02}$	0.22 ± 0.02	0.00 ± 0.01	0.67 ± 0.31	0.33 ± 0.01	0.22 ± 0.00

6 Conclusion

We introduce a distributional framework for inverse reinforcement learning that jointly models reward uncertainty and return distributions. Our method enables risk-aware policy learning and accurate inference of high-order structure in demonstrations. We validate the framework on stochastic control tasks, deterministic settings, and real neural datasets, demonstrating state-of-the-art performance and strong generalization across domains.

[2]For HalfCheetah, with the same amount of data as Offline ML-IRL, DistIRL can reach 11239 ± 539.

Ethics Statement

IRL enables powerful tools for understanding behavior, with positive applications in neuroscience, animal modeling, and AI alignment. However, it also raises ethical concerns. IRL could be misused in military settings to model or mimic adversarial behavior, or in surveillance contexts to infer personal goals without consent, posing risks to privacy and autonomy. These concerns highlight the need for careful oversight and responsible deployment.

Reproducibility Statement

We list parameter choice in Table. 6. The implementation will be made publicly available following the paper decision.

References

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

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. *Advances in neural information processing systems*, 35:29663–29676, 2022.

Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4): 715–726, 1996.

Ondrej Bajgar, Alessandro Abate, Konstantinos Gatsis, and Michael A Osborne. Walking the values in bayesian inverse reinforcement learning. *arXiv preprint arXiv:2407.10971*, 2024.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Alex J Chan and Mihaela van der Schaar. Scalable bayesian inverse reinforcement learning. *arXiv preprint arXiv:2102.06483*, 2021.

Ziteng Cheng, Anthony Coache, and Sebastian Jaimungal. Eliciting risk aversion with inverse reinforcement learning via interactive questioning. *arXiv preprint arXiv:2308.08427*, 2023.

Jaedeug Choi and Kee-Eung Kim. Map inference for bayesian inverse reinforcement learning. *Advances in neural information processing systems*, 24, 2011.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2:319–328, 2012.

Jingliang Duan, Yang Guan, Shengbo Eben Li, Yangang Ren, Qi Sun, and Bo Cheng. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. *IEEE transactions on neural networks and learning systems*, 33(11):6584–6598, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34: 4028–4039, 2021.

Allan Gut and Allan Gut. *Probability: a graduate course*, volume 200. Springer, 2006.

Josef Hadar and William R Russell. Rules for ordering uncertain prospects. *The American economic review*, 59(1):25–34, 1969.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Christopher Heil. *Introduction to real analysis*, volume 280. Springer, 2019.

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

Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A bayesian approach to generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.

Adib Karimi and Mohammad Mehdi Ebadzadeh. Rize: Regularized imitation learning via distributional reinforcement learning. *arXiv preprint arXiv:2502.20089*, 2025.

Jingyang Ke, Feiyang Wu, Jiyi Wang, Jeffrey Markowitz, and Anqi Wu. Inverse reinforcement learning with switching rewards and history dependency for characterizing animal behaviors. *arXiv preprint arXiv:2501.12633*, 2025.

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

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Jonathan Lacotte, Mohammad Ghavamzadeh, Yinlam Chow, and Marco Pavone. Risk-sensitive generative adversarial imitation learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2154–2163. PMLR, 2019.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006.

Keuntaek Lee, David Isele, Evangelos A Theodorou, and Sangjae Bae. Risk-sensitive mpcs with deep distributional inverse rl for autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7635–7642. IEEE, 2022.

Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. *Advances in neural information processing systems*, 24, 2011.

Mengdi Li, Xufeng Zhao, Jae Hee Lee, Cornelius Weber, and Stefan Wermter. Internally rewarded reinforcement learning. In *International Conference on Machine Learning*, pp. 20556–20574. PMLR, 2023.

Yangming Li, Chieh-Hsin Lai, Carola-Bibiane Schönlieb, Yuki Mitsufuji, and Stefano Ermon. Bellman diffusion: Generative modeling as learning a linear operator in the distribution space. *arXiv preprint arXiv:2410.01796*, 2024.

Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, 2022.

Aishwarya Mandyam, Didong Li, Diana Cai, Andrew Jones, and Barbara E Engelhardt. Kernel density bayesian inverse reinforcement learning. *arXiv preprint arXiv:2303.06827*, 2023.

Jeffrey E Markowitz, Winthrop F Gillis, Maya Jay, Jeffrey Wood, Ryley W Harris, Robert Cieszkowski, Rebecca Scott, David Brann, Dorothy Koveal, Tomasz Kula, Caleb Weinreb, Mohammed Abdal Monium Osman, Sandra Romero Pinto, Naoshige Uchida, Scott W Linderman, Bernardo L Sabatini, and Sandeep Robert Datta. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614(7946):108–117, January 2023a.

Jeffrey E Markowitz, Winthrop F Gillis, Maya Jay, Jeffrey Wood, Ryley W Harris, Robert Cieszkowski, Rebecca Scott, David Brann, Dorothy Koveal, Tomasz Kula, et al. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614(7946):108–117, 2023b.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Dean Mobbs, Pete C Trimmer, Daniel T Blumstein, and Peter Dayan. Foraging for foundations in decision neuroscience: insights from ethology. *Nature Reviews Neuroscience*, 19(6):419–427, 2018.

Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pp. 529–551. PMLR, 2021.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Sascha Rosbach, Vinit James, Simon Großjohann, Silviu Homoceanu, and Stefan Roth. Driving with style: Inverse reinforcement learning in general-purpose planning for automated driving. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2658–2665. IEEE, 2019.

Lukas Schneider, Jonas Frey, Takahiro Miki, and Marco Hutter. Learning risk-aware quadrupedal locomotion using distributional reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11451–11458. IEEE, 2024.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Ekaterina N Sereda, Efim M Bronshtein, Svetozar T Rachev, Frank J Fabozzi, Wei Sun, and Stoyan V Stoyanov. Distortion risk measures in portfolio optimization. *Handbook of portfolio construction*, pp. 649–673, 2010.

Sumeet Singh, Jonathan Lacotte, Anirudha Majumdar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via semi-and non-parametric methods. *The International Journal of Robotics Research*, 37(13-14):1713–1740, 2018.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

Dizan Vasquez, Billy Okal, and Kai O Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1341–1346. IEEE, 2014.

Ran Wei, Siliang Zeng, Chenliang Li, Alfredo Garcia, Anthony D McDonald, and Mingyi Hong. A bayesian approach to robust inverse reinforcement learning. In *Conference on Robot Learning*, pp. 2304–2322. PMLR, 2023.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 101297–101336, 2024.

Feiyang Wu, Jingyang Ke, and Anqi Wu. Inverse reinforcement learning with the average reward criterion. *Advances in Neural Information Processing Systems*, 36:69117–69129, 2023.

Feiyang Wu, Zhaoyuan Gu, Hanran Wu, Anqi Wu, and Ye Zhao. Infer and adapt: Bipedal locomotion reward learning from demonstrations via inverse reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16243–16250. IEEE, 2024.

Zheng Wu, Liting Sun, Wei Zhan, Chenyu Yang, and Masayoshi Tomizuka. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5355–5362, 2020.

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

Hang Yin, Anastasia Varava, and Danica Kragic. Modeling, learning, perception, and control methods for deformable object manipulation. *Science Robotics*, 6(54):eabd8803, 2021.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 36:65531–65565, 2023.

Siliang Zeng, Yao Liu, Huzefa Rangwala, George Karypis, Mingyi Hong, and Rasool Fakoor. From demonstrations to rewards: Alignment without explicit human preferences. *arXiv preprint arXiv:2503.13538*, 2025.

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.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *International conference on machine learning*. Carnegie Mellon University, 2010.

A Related work comparison

Table 5: Comparison of IRL methods under various settings

Reference	Model reward dist.?	Infer risk aware policy?	Recover reward dist.?	Learn return dist.?
(Wulfmeier et al., 2015; Ziebart et al., 2008) (Garg et al., 2021; Ni et al., 2021) (Zeng et al., 2022; 2023; Wei et al., 2023)	✗	✗	✗	✗
(Ramachandran & Amir, 2007; Choi & Kim, 2011) (Chan & van der Schaar, 2021; Lee et al., 2022)	✓	✗	✗	✗
(Karimi & Ebadzadeh, 2025)	✗	✗	✗	✓
(Singh et al., 2018; Lacotte et al., 2019) (Cheng et al., 2023)	✗	✓	✗	✗
This work	✓	✓	✓	✓

In Table A, we compare DistIRL with existing IRL methods along four key dimensions. The first column, *Model reward distribution*, asks whether a method explicitly represents the reward as a random variable rather than as a fixed deterministic function. For example, Bayesian IRL methods place a prior over reward parameters, thereby modeling uncertainty, but they do not recover the actual shape of the underlying distribution. This is distinct from *Recover reward distribution*, which requires learning the full distribution of rewards themselves, including higher-order statistics such as variance and skewness, rather than just a posterior over parameters.

The third column, *Infer risk-aware policy*, evaluates whether a method incorporates risk measures into policy inference. Methods in this category optimize beyond expected return, often capturing aversion or preference to variability in outcomes. The final column, *Learn return distribution*, indicates whether a method leverages distributional reinforcement learning (DistRL) techniques to estimate the full distribution of returns, rather than only their expectation. Unlike reward distributions, which describe stochasticity at the immediate reward level, return distributions capture the cumulative effect of randomness from rewards, transitions, and policies over trajectories.

As shown in the table, most prior IRL methods either assume deterministic rewards or restrict themselves to expectation-based inference. In contrast, DistIRL is the first framework that simultaneously models stochastic rewards, learns full reward distributions, integrates distributional return estimation, and supports risk-aware policy learning, thereby unifying these capabilities in a principled way.

B Proofs

We first wish to show that

$$\int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ dz = \int_0^1 \left[F_{Z^\pi}^{-1}(v) - F_{Z^E}^{-1}(v)\right]_+ dv. \tag{14}$$

Proposition B.1. *Let Z^π and Z^E be two real-valued integrable random variables with cumulative distribution functions F_{Z^π} and F_{Z^E}, and corresponding quantile functions $F_{Z^\pi}^{-1}$ and $F_{Z^E}^{-1}$. Then we have*

$$\int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ \, dz = \int_0^1 \left[F_{Z^\pi}^{-1}(v) - F_{Z^E}^{-1}(v)\right]_+ \, dv,$$

where $[x]_+ := \max(x, 0)$.

Proof. Note that

$$\begin{aligned}\int_{-\infty}^{\infty} [F_{Z^\pi}(z) - F_{Z^E}(z)]_+ \, dz &= \int_{-\infty}^{\infty} \int_0^1 \mathbb{1}_{F_{Z^\pi}(z) \geq v, v \geq F_{Z^E}(z)} dv dz \\ &= \int_0^1 \int_{-\infty}^{\infty} \mathbb{1}_{F_{Z^\pi}(z) \geq v, v \geq F_{Z^E}(z)} dv dz \\ &= \int_0^1 \int_{-\infty}^{\infty} \mathbb{1}_{F^{-1}_{Z^\pi}(v) \geq z, z \geq F^{-1}_{Z^E}(v)} dv dz \\ &= \int_0^1 \left[F^{-1}_{Z^\pi}(v) - F^{-1}_{Z^E}(v)\right]_+ dv\end{aligned}$$

The interchange of integrals are permitted by the Theorem of Fubini-Tonelli as everything is positive (Heil, 2019). Note that the definition of the quantile function (Gut & Gut, 2006) is:

$$F^{-1}(v) := \inf_{z \in \mathbb{R}} \{F(z) \geq v\}.$$

□

Proposition 4.6. *$M_\xi(X) \geq M_\xi(Y)$ for every distortion function ξ implies $X \succeq_{\text{FSD}} Y$.*

Proof. Define the difference in quantile functions:

$$h(v) := F_X^{-1}(v) - F_Y^{-1}(v).$$

Suppose for contradiction that the set

$$A := \{v \in [0,1] \mid h(v) < 0\}$$

has positive Borel measure, i.e., $\mu(A) > 0$. Let's define a distortion function $\tilde{\xi}_A$ whose derivative is:

$$\tilde{\xi}'_A(v) = \begin{cases} \frac{1}{\mu(A)} & \text{if } v \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\tilde{\xi}_A$ is a valid distortion function and satisfies $\int_0^1 d\tilde{\xi}_A(v) = 1$. Note that

$$\mathcal{M}_{\xi_A}(X) - \mathcal{M}_{\xi_A}(Y) = \int_0^1 h(v) \, d\tilde{\xi}_A(v) = \int_A h(v) \cdot \frac{1}{\mu(A)} \, dv < 0.$$

This contradicts the assumption that $\mathcal{M}_{\tilde{\xi}}(X) \geq \mathcal{M}_{\tilde{\xi}}(Y)$ for all distortion functions $\tilde{\xi}$. Therefore, the set where $F_X^{-1}(v) < F_Y^{-1}(v)$ must have measure zero. Thus we have

$$F_X^{-1}(v) \geq F_Y^{-1}(v) \quad \text{for } v \in [0,1] \text{ almost everywhere (a.e.)}$$

which implies

$$F_X(z) \leq F_Y(z) \quad \text{for all } z \in \mathbb{R},$$

since

$$\begin{aligned}F_X(z) = P_X(X < z) &= \mu\left(\{v \in [0,1] | F_X^{-1}(v) \leq z\}\right) \\ &\leq \mu\left(\{v \in [0,1] \cap A^c | F_X^{-1}(v) \leq z\}\right) + \mu\left(\{v \in [0,1] \cap A | F_X^{-1}(v) \leq z\}\right) \\ &= \mu\left(\{v \in [0,1] \cap A^c | F_X^{-1}(v) \leq z\}\right) \\ &\leq \mu\left(\{v \in [0,1] \cap A^c | F_Y^{-1}(v) \leq z\}\right) \\ &\leq \mu\left(\{v \in [0,1] | F_Y^{-1}(v) \leq z\}\right) \\ &= F_Y(z)\end{aligned}$$

The second inequality is due to the fact that for any z,

$$\{v \in [0,1] \cap A^c | F_X^{-1}(v) \leq z\} \subseteq \{v \in [0,1] \cap A^c | F_Y^{-1}(v) \leq z\}$$

Hence,

$$X \succeq_{\text{FSD}} Y.$$

□

C Model Architecture and Hyper-parameters

Throughout this paper, we use the following model architecture for all the experiments.

Table 6: Model Parameters for DistIRL

Parameter	Value
Training Parameters	
Learning Rate	3×10^{-4}
Batch Size	512
Total Iterations	5,000
Entropy Coefficient	0.1
Risk Measure	CVaR
Risk Parameter	0.05
Reward Regularization	0.01
Network Architecture	
Policy Network	[256, 128]
Distribution Type	Skew Gaussian
Reward Range	[-5.0, 5.0]
Number of Quantiles	200
Reward Hidden Features	128

For gridworld, we specify the reward range as $[0, 2]$. For MuJoCo tasks, $[-10, 10]$. This is achieved by applying a (scaled) tanh function.

D Additional Results on Dopamine Level

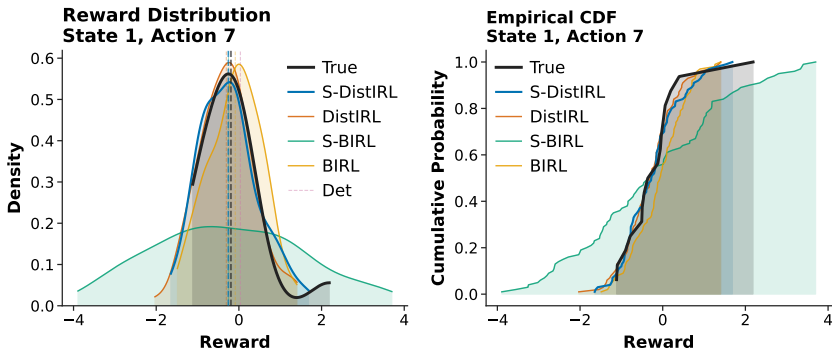

Figure 6: Reward recovery for state 1 action 7

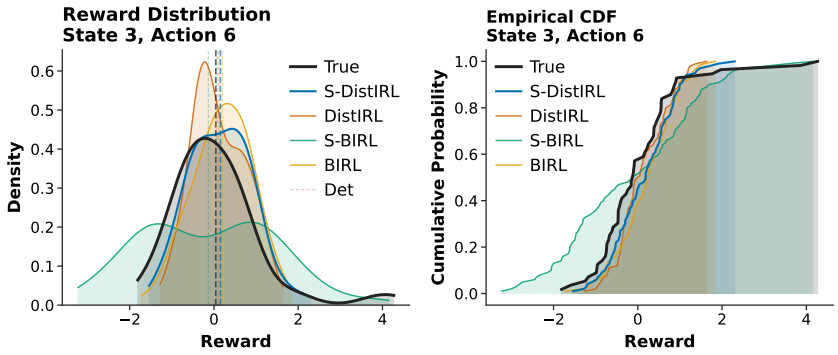

Figure 7: Reward recovery for state 3 action 6

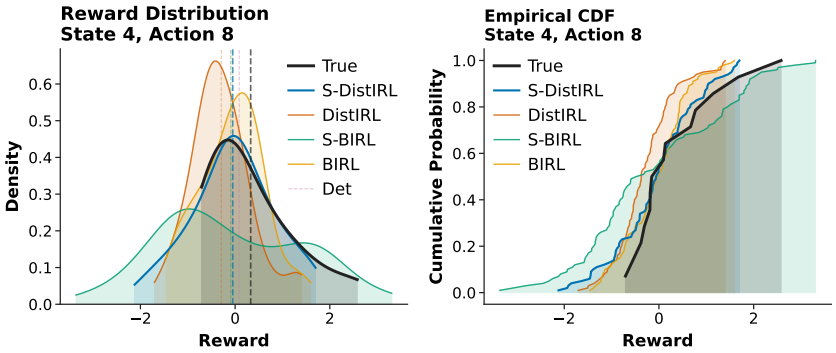

Figure 8: Reward recovery for state 4 action 8

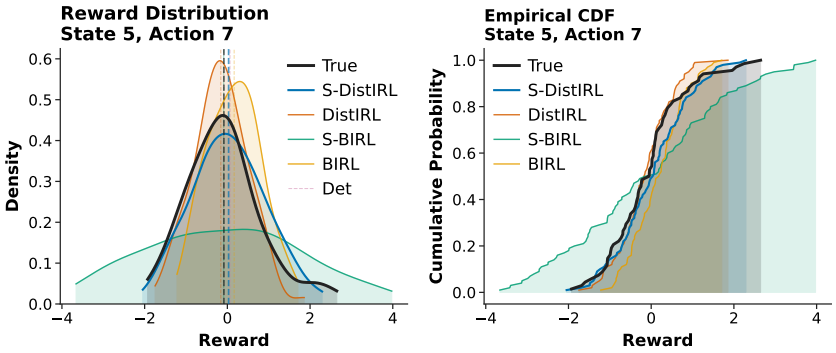

Figure 9: Reward recovery for state 5 action 7

E LLM USAGE AND REPRODUCIBILITY

We use LLM to aid or polish writings only. Research ideation, retrieval and discovery (e.g., finding related work) are conducted by ourselves.

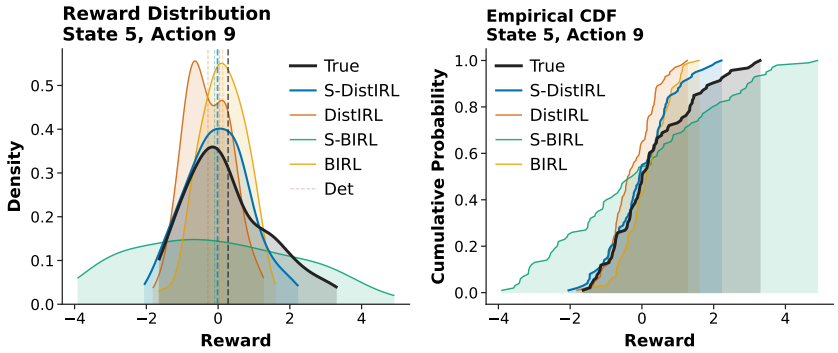

Figure 10: Reward recovery for state 5 action 9

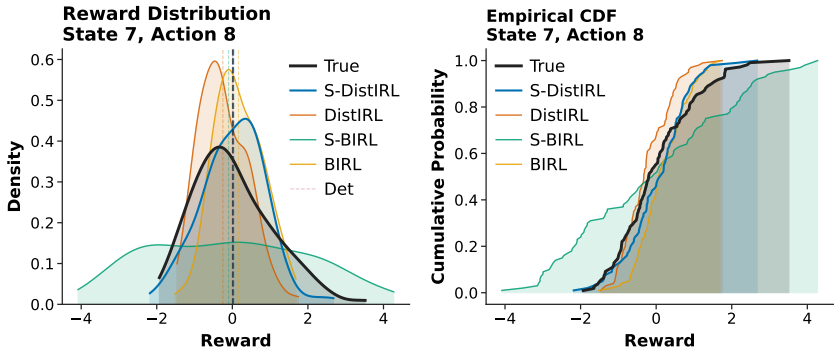

Figure 11: Reward recovery for state 7 action 8

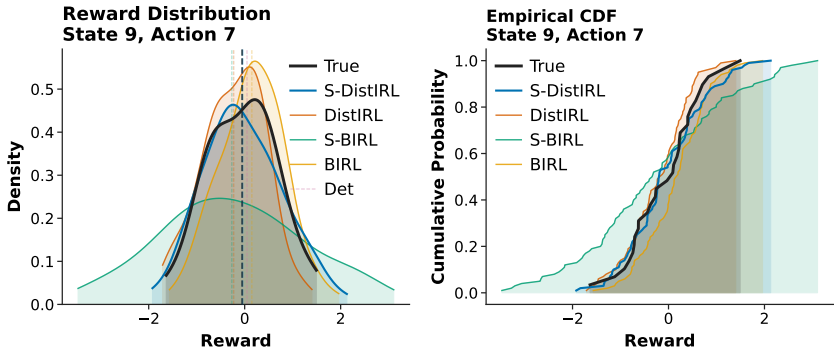

Figure 12: Reward recovery for state 9 action 7