

Article

Adversarial Distributed Multi-Task Meta-Inverse Reinforcement Learning with Theory of Mind and Mean-Field Method

Li Song ^{1,2,*} , Kun Yang ³ and Chao Chen ⁴ ¹ School of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, China² College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China³ College of Mechanical Engineering, Taiyuan University of Technology, Taiyuan 030600, China; yangkun@tyut.edu.cn⁴ College of Materials Science and Engineering, Taiyuan University of Technology, Taiyuan 030600, China; chenchao@tyut.edu.cn

* Correspondence: slili516@zju.edu.cn

Abstract

Maximum entropy adversarial inverse reinforcement learning (ME-AIRL) has garnered widespread attention for its ability to learn rewards and optimize policies from expert demonstrations. In complex multi-task environments, applying meta-learning ME-AIRL to acquire rewards requires a substantial volume of homogeneous expert demonstrations across all tasks, which is often impractical in real-world scenarios. Moreover, interference between tasks further escalates computational complexity. To solve these challenges, this paper proposes a distributed multi-task meta ME-AIRL framework based on theory of mind and mean field, referred to as TMMF-MTAIRL. In TMMF-MTAIRL, the theory of mind is used to capture the relationships and representational information among multiple tasks. Furthermore, TMMF-MTAIRL integrates mean-field theory to transform interactions between complex tasks into interactions between the main task and the average of the remaining tasks. Furthermore, additional latent variables are introduced to enhance adaptation to novel tasks. We evaluate the proposed TMMF-MTAIRL on point-maze benchmarks and a real-world rolling bearing fault diagnosis dataset using metrics such as classification accuracy, mean rewards or cumulative rewards. TMMF-MTAIRL achieves the best performance across all tasks, with an average improvement of 0.16 in accuracy of fault classification over the strongest baseline.

Keywords: adversarial inverse reinforcement learning; partially heterogeneous multi-task; theory of mind; additional latent variables; mean-field method

MSC: 68Q25



Academic Editor: Jonathan Blackledge

Received: 10 January 2026

Revised: 10 February 2026

Accepted: 13 February 2026

Published: 15 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In the field of artificial intelligence, reinforcement learning (RL) has become one of the core technologies for solving complex decision-making problems by enabling interactive exploration between agents and their environments [1–5]. Traditional RL largely depends on pre-designed reward functions. However, crafting a reward function that both accurately captures task objectives and maintains generalization capability often proves highly challenging [6–8]. Therefore, inverse reinforcement learning (IRL) has been proposed to infer implicit reward functions and optimize policies by observing expert demonstrations, thereby avoiding the subjectivity and high costs associated with manual

reward design [9–12]. As real-world scenarios become increasingly complex, single-task IRL has begun to reveal significant limitations. In domains such as autonomous driving, medical robotics, and industrial automation, agents must simultaneously handle multiple interrelated tasks. However, conventional single-task IRL models struggle to capture shared knowledge among tasks and to resolve conflicting objectives [13,14]. These challenges have spurred the emergence of multi-task inverse reinforcement learning (MT-IRL), which integrates IRL's reward inference capability with knowledge-sharing mechanisms derived from multi-task learning, making it a prominent research focus in recent years [15–17].

The core objective of MT-IRLs is to infer either shared or task-specific latent reward functions across multiple tasks. Researchers have enabled cross-task knowledge transfer by introducing hierarchical Bayesian models and meta-learning framework [18–20]. For instance, IRL approaches based on probabilistic graphical models can map expert demonstrations from different tasks into a unified latent space, thereby extracting shared reward features [21,22]. By integrating deep neural networks (e.g., Transformer), researchers have proposed end-to-end multi-task IRL models that can automatically extract inter-task correlations from high-dimensional, homogeneous expert data [23–25]. For example, the dynamic properties of different tasks can be captured through multi-head attention mechanisms, and adaptive reward functions can be generated [26]. To solve the black box problem of traditional IRL in complex tasks, the causal graph models have been integrated into the multi-task framework. By analyzing causal relationships in expert demonstrations, the models can not only infer reward functions but also identify causal dependencies between tasks, thereby enhancing both policy interpretability and generalization performance [27]. Despite these advances, a significant challenge remains: limited or suboptimal expert demonstrations can adversely affect both reward learning and policy optimization.

To address this challenge in MT-IRLs, generative adversarial networks (GANs) have been integrated into MT-IRLs, aiming to resolve large-scale task problems in complex environments [28–30]. To tackle the challenge of autonomous reward acquisition in large-scale, high-dimensional environments with unknown dynamics, Finn et al. incorporated GANs into guided cost learning to learn reward functions that are robust to dynamic changes [31]. To address the poor transferability and robustness of recoverable rewards in varying environments, an adversarial IRL algorithm was proposed to learn hierarchically disentangled rewards through policy learning [32]. Additionally, semantic rewards were incorporated into the adversarial IRL to further enhance and stabilize its performance [33]. To make the graph computation end-to-end microscopic, the model-based adversarial IRL employed a dynamic self-attention model to optimize low-variance policies [34]. Moreover, to improve the adaptability of both adversarial IRL and multi-task adversarial IRL, meta-adversarial inverse reinforcement learning (Meta-AIRL) integrated meta-learning with adversarial IRL to learn adaptive policies [21,35].

However, existing adversarial IRL methods are often designed under the assumption of homogeneous expert presentations [35]. In real-world reward learning scenarios, however, updated mixed expert demonstrations for different tasks may include both human-generated experience data and algorithm-generated data, where state-action visitation frequencies can vary significantly across tasks. Such variations lead to heterogeneous modalities and distinct distribution frequencies. Moreover, these partially heterogeneous multi-task settings require a large number of expert demonstrations to learn rewards, which increases the computational complexity. To address these challenges, researchers have developed a mean-field RL model and a multi-agent IRL framework based on the mean-field game, which effectively mitigate high-dimensional state spaces and heavy computational burdens through simplified modeling approaches [36–40]. Moreover, prior studies in RL and IRL have incorporated the theory of mind (ToM) to enhance informa-

tion interaction, thus improving the understanding and prediction of other agents' behaviors and the intentions of others [41–44]. Inspired by the application of mean-field theory and ToM, and aiming to improve joint reward learning and knowledge transfer in multi-task IRL under complex partially heterogeneous environments, this paper proposes a meta-adversarial distributed multi-task inverse reinforcement learning based on the theory of mind (TMMF-MTAIRL). Given updated and mixed expert demonstrations, the proposed algorithm constructs the adversarial mean-field inverse reinforcement learning based on ToM and deep latent variables to enable explicit modeling of relationships between multiple tasks, enhancing information utilization among tasks, and facilitating rapid adaptation to novel tasks. Therefore, by simultaneously improving task interpretability and adaptability, TMMF-MTAIRL can more effectively infer potential reward functions, and optimize agents' policies.

In summary, the main contributions of this paper are as follows:

- (1) To enhance knowledge transfer between tasks, this paper proposes a meta-adversarial IRL framework based on theory of mind, which uses the theory of mind to make better use of the relationships between partially heterogeneous tasks and capture the representational information between tasks.
- (2) To reduce the computational complexity of reward and strategy optimization, the mean-field theory is introduced into TMMF-MTAIRL, which can transform interactions between complex tasks into the interaction between the main task and the rest of the average tasks.
- (3) By introducing additional latent variables to characterize modal information within mixed expert demonstrations, the TMMF-MTAIRL enables rapid adaptation to novel tasks during meta-adversarial training of the discriminator and generator, even with limited expert demonstration data.
- (4) The effectiveness of TMMF-MTAIRL is evaluated on both the maze-point benchmark and rolling bearing fault diagnosis tasks. Experimental results indicate that TMMF-MTAIRL outperforms conventional multi-task IRL methods, achieving state-of-the-art performance in both reward learning and policy optimization.

2. Related Works

Adversarial IRL and multi-task adversarial IRL: Adversarial IRL represents a state-of-the-art research direction that combines GANs with IRL, aiming to efficiently infer latent reward functions of expert behaviors through adversarial training mechanisms. Initially, Finn et al. introduced GANs into guided cost learning [31], and used the distribution of state-action pairs to optimize the overall goal of adversarial IRL: $\mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \right] = \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} f_{\theta}(s_t, a_t) - \log \pi(a_t | s_t) \right]$. In the adversarial IRL algorithm, sampling-based methods are used to calculate the gradient of the loss function with respect to θ , and the training is performed according to the trajectory distribution [32]. In practice, however, such estimates based on trajectory distributions exhibit high variance. Consequently, to cope with complex dynamic environments, researchers have proposed a series of extensions, including behavioral fusion IRL [45], model-based IRL [34], and semantic-based augmented adversarial IRL [33], etc. To address the challenge of inferring reward functions from expert trajectories in multi-agent environments, multi-agent generative adversarial IRL leverages diverse agents' behaviors to learn shared reward functions [46]. Furthermore, for multi-task scenarios, several approaches have been developed, such as hierarchical multi-task adversarial IRL [19], meta-adversarial IRL [20–22], variational multi-task IRL [23], and multi-task IRL based on self-attention dynamic model [26]. In addition, by incorporating causal concepts into reward learning and policy optimization for multi-task IRL, researchers have proposed causal entropy-based multi-task IRL [27].

To further enhance the adaptability of adversarial multi-task IRL, an IRL based on meta-learning is developed to learn adaptive policies through different update frequencies and meta-learning rate [35].

IRL based on mean field and ToM: In complex multi-task environments, agents must interact with highly intricate surroundings, resulting in substantial computational complexity that makes reward learning and policy optimization particularly challenging [20,23]. As the number of agents grows significantly, researchers have proposed mean-field-based RL algorithms to reduce computational complexity, which introduces mean-field theory into multi-agent RL. Typical examples include mean-field Q-learning, mean-field actor-critic algorithms [36], GAT-MF [39], and HMF-based DRL [46], which are representative RL methods based on mean field. Building on these studies, researchers have gradually introduced mean-field methods into IRL to better address the challenges of learning reward functions and optimizing policies [37,47]. Specifically, based on the adversarial IRL framework combined with meta-learning, researchers proposed MFG-AIRL based on the mean field [40]. Furthermore, to enhance belief modeling and intention inference in learning rewards, researchers have proposed ToM-based IRL [40], Bayesian MoT-based IRL [41], intrinsically motivated ToM-based IRL [42], and goal-oriented IRL based on ToM [43]. These approaches aim to predict an agent's future actions by analyzing its past trajectories. Overall, mean-field theory and ToM have achieved preliminary research progress in RL and IRL [44], effectively facilitating reward learning and policy optimization. These developments provide valuable inspiration for the theoretical framework proposed in this paper.

Table 1 shows the faced problems, advantages, and representative references of the related adversarial IRL methods based on mean field and ToM.

Table 1. Comparisons of related adversarial IRL methods based on mean field and ToM.

Methods	Faced Problems	Advantages	References
Behavioral fusion IRL Model-based IRL Semantic-based augmented adversarial IRL	Difficult to adapt to dynamic environments	Estimations based on trajectory distributions have small differences	[45] [34] [33]
Hierarchical multi-task adversarial IRL Meta-adversarial IRL Variational multi-task IRL Causal entropy-based multi-task IRL		Capturing information among multiple tasks in real time	[19] [20–22] [23] [27]
Mean-field actor-critic algorithms GAT-MF HMF-based DRL	Complexity of large-scale interactive computing and modeling	Reducing computational complexity, improving generalization ability	[36] [39] [46]
ToM-based IRL Bayesian MoT-based IRL Intrinsically motivated ToM-based IRL Goal-oriented IRL based on ToM		Enhanced cognitive depth and generalization ability of agents	[41] [42] [43]

3. Preliminaries

3.1. Markov Games

A Markov decision process (MDP) is formulated as $(\mathcal{S}, \mathcal{A}, R, \gamma, P)$, where $\mathcal{S}$ represents the set of states, $\mathcal{A}$ represents the set of actions, $R_s^a = \mathbb{E}[r_{t+1}|s_t = s, a_t = a]$ indicates the set of reward function, and $\gamma \in [0, 1]$ represents the discount factor. The function $P_{ss'}^a = \mathbb{P}(s_{t+1}|s_t, a_t)$ is the state transition distribution. The agent takes actions a_t from the state s_t to the next state s_{t+1} with probability $p(s_{t+1}|s_t, a_t)$. The agent aims to maximize the the expected cumulative discounted rewards $\mathbb{E}_\pi \left[\sum_{t=1}^T \gamma_t r_t(s_t, a_t) | s_0 = s, a_0 = a \right]$. Based on

the MDP in RL, if we cannot obtain the reward function of the IRL, the MDP in IRL is formulated as $(\mathcal{S}, \mathcal{A}, \gamma, P)$. Given the expert demonstrations $\mathcal{D}$, the main objectives of the IRL are as follows: (1) to recover the reward function, and (2) to realize policy optimization by making the learner's policy approach the expert's policy.

3.2. Adversarial Inverse Reinforcement Learning

To solve the reward functions under expert demonstrations $\mathcal{D}$, Ziebart et al. [42] proposed maximum entropy inverse reinforcement learning (MaxEnt IRL), which formulates constrained optimization problems as maximum entropy models [17,35]:

$$p(\xi_j|\theta) = \frac{1}{z(\theta)} e^{\sum_{s_t \in \xi_j} \theta^T \phi_{s_t}} \prod_{(s_t, a_t, s_{t+1}) \in \xi_j} p(s_{t+1}|s_t, a_t) \quad (1)$$

where ϕ_{s_t} represents the feature function of each reward, the partition function $z(\theta) = \sum_{\xi_j} e^{\sum_{s_t \in \xi_j} \theta^T \phi_{s_t}} \prod_{(s_t, a_t, s_{t+1}) \in \xi_j} p(s_{t+1}|s_t, a_t)$ is a normalized constant, and $p(s_{t+1}|s_t, a_t)$ represents the state transition probability. In Equation (1), the trajectory ξ_j that the expert demonstrates is exponentially proportional to the cumulative rewards along that trajectory, implying that trajectories with higher rewards are more likely to be preferred by the agent. Furthermore, the MaxEnt IRL algorithm is extended to the multi-task environment. Given the expert demonstrations $\mathcal{D} = \{\xi_j\}_{j=1}^M$, the agent can obtain the reward function and optimize the policy by imitating the expert strategy π_E , aiming to maximize the objective function:

$$\max_{\theta} J(\theta) = \max_{\theta} E_{\xi_j \sim \mathcal{D}} J[\log p_{\theta}(\xi_j)] \quad (2)$$

where the likelihood probability distribution $p_{\theta}(\xi_j)$ of the expert trajectory ξ_j is parametrized as $p_{\theta}(\xi_j) \propto p(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) e^{r_{\theta}(s_t, a_t)}$, which explains the likelihood probability of the observed data in terms of the reward function; that is, the probability distribution of the expert trajectory ξ_j is induced by r .

The reward function is obtained by minimizing $D_{KL}(p_{\pi_E}(\xi)||p_{\theta}(\xi))$:

$$\begin{aligned} \arg \min_{\theta} D_{KL}[p_{\pi_E}(\xi_j)||p_{\theta}(\xi_j)] &= \arg \min_{\theta} \mathbb{E}_{\xi_j \sim p_{\pi_E}(\xi_j)} [\log p_{\theta}(\xi_j)] \\ &= \arg \min_{\theta} \mathbb{E}_{\xi_j \sim p_{\pi_E}(\xi_j)} \left[\sum_{t=1}^T r_{\theta}(s_t, a_t) \right] - \log Z_{\theta} \end{aligned} \quad (3)$$

However, the partition function Z_{θ} is difficult to calculate. To solve this problem, adversarial IRL has been proposed as an efficient framework for MaxEnt IRL. In adversarial IRL, a discriminator $D_{\theta}(s, a)$ is trained:

$$D_{\theta}(s, a) = \frac{e^{r_{\theta}(s, a)}}{e^{r_{\theta}(s, a)} + \pi(a|s)} \quad (4)$$

To make policy π_{θ} more closely approximate the expert's policy, the discriminator minimizes the cross-entropy loss between expert demonstration and generated samples:

$$L(\theta) = \sum_{t=0}^T -E_D[\log D_{\theta}(s_t, a_t)] - E_{\pi_t}[\log(1 - D_{\theta}(s_t, a_t))] \quad (5)$$

The optimal objective of the discriminator is achieved when $\pi = \pi_E$, where π denotes the policy learned by the generator, and π_E is the expert's policy. At this point, the optimal output of the discriminator is 0.5, indicating that $e^{r_{\theta}(s, a)} = \pi_E(a|s)$.

To learn reward functions that can adapt to different tasks, meta-learning is introduced within the adversarial maximum entropy IRL (AMEIRL) to infer task characteristics from

expert policy samples and learn rewards [29,31]. The likelihood function distribution of samples with different trajectories is formulated as

$$p(\zeta = \{s_{0:T}, a_{0:T}\} | v) = \frac{1}{Z_\theta} \eta(s_0) \prod_{t=0}^T p(s_{t+1} | s_t, a_t) e^{\sum_{t=0}^T r_\theta(s_t, a_t, v)} \quad (6)$$

where v represents the parameter of meta learning, and $r_\theta(s_t, a_t, v)$ denotes the reward function.

However, in the complex multi-task learning environment, tasks may interfere with each other, which brings challenges to learning rewards and optimizing policies for the agents. Effectively leveraging the relationships between tasks, capturing the representational information between tasks, and reducing the complexity of optimizing rewards and policies are thus important problems to address. Motivated by mean-field theory in IRL, which primarily addresses interactions and computational challenges among large-scale agents, we incorporate mean-field approximation into multi-task IRL to reduce the complexity of learning rewards and optimizing policies.

4. Multi-Task Meta-Adversarial Mean-Field IRL Based on Theory of Mind

In complex, partially heterogeneous multi-task environments, using meta-learning adversarial IRL to learn reward functions traditionally requires a large set of homogeneous expert demonstrations, and the presence of multiple tasks may lead to mutual interference, which significantly increase the algorithm's computational complexity. However, in practical task environments, ensuring that the acquired expert demonstrations across different tasks are homogeneous remains a significant challenge. To address these challenges, this paper presents a meta-adversarial mean-field IRL framework based on the context-conditional policy to learn rewards and optimize policy from the constantly updated mixed expert demonstrations. The algorithm incorporates theory of mind to enhance knowledge transfer between multiple tasks, thereby capturing inter-task relationships and fully leveraging environmental information. Furthermore, the mean field is used to model the interactions between the main task and the average of the remaining tasks, which reduces computational complexity. Additionally, latent task variables are accurately inferred from sampled trajectories for the new task, ultimately deriving task-specific adaptive reward functions and optimal policies under the maximum entropy framework. Moreover, the overall generative adversarial framework can continuously be updated and optimized to improve the mixed expert demonstrations.

Different from existing mean-field IRL and ToM-based IRL, we introduce a ToM-conditioned mean field that aggregates cross-task interactions into an embedding of the averaged state of mind, and derive a mutual-information regularized adversarial learning objective with tractable gradient estimators, which can quickly adapt to new tasks with latent task variables and continuously improve mixed expert demonstrations.

4.1. Adversarial Mean-Field IRL Based on Probabilistic Context Variable

In complex, partially heterogeneous multi-task environments, the context-conditional meta-adversarial mean-field IRL framework first samples M expert demonstration trajectories ζ_j from the update mixed expert demonstrations $\mathcal{D}$, then learns reward functions and optimizes policies under each trajectory ζ_j . Each reward and policy learning process with an expert demonstration trajectory is treated as a task. These mixed demonstration trajectories exhibit diverse behavioral patterns and varying distribution frequencies. Therefore, the

reward learning and policy optimization task can be viewed as M partially heterogeneous distributed learning tasks, aiming to infer latent reward functions across multiple tasks.

Using the update mixed expert demonstrations, the adversarial mean-field multi-task IRL (MMF-MTAIRL) introduces the probabilistic contextual variable $v \in \mathcal{U}$ to describe the distribution of task characteristics, where $\mathcal{U}$ is the value space that can enhance the learning precision of rewards for partially heterogeneous multiple tasks. In this framework, both rewards and policies π depend on the probabilistic implicit variable v . We make the following definitions: the expert policy $\pi_E(a|s, v)$ and corresponding reward functions $R(s, a, \eta, v)$ that depend on mean field and state of mind for different tasks, η -conditional mean-field flow $\eta_E(\cdot|v)$ and policy flow $\pi_E(\cdot|v)$, and marginal distribution of expert policy $p_{\eta_E, \pi_E}(\tau) = \int p_{\eta_E, \pi_E}(\tau|v)p(v)dv$, where $p(v)$ indicates a prior distribution. MMF-MTAIRL with probabilistic contextual variable aims to use meta-learning to learn an inference model $q(v|\tau)$ and rewards with samples of the $p_{\pi_E}(\tau)$.

4.2. The Embedding of Theory of Mind for Improved MMF-MTAIRL

In complex distributed and partially heterogeneous multi-task environments, the interconnections among multiple partially heterogeneous tasks significantly increase computational complexity. To facilitate knowledge transfer across multiple tasks, we incorporate theory of mind into MMF-MTAIRL (TMMF-MTAIRL), which explicitly models the latent relationships among tasks.

First, following the task partitioning in Section 4.1 under partially heterogeneous expert demonstrations, we assume there are M tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_M\}$. The assumption that there exist M tasks reflects a common practical setting in which expert demonstrations are collected from multiple sources or under different operating conditions, leading to partially heterogeneous data. In such cases, demonstrations can be naturally partitioned into task-specific subsets, either explicitly (e.g., by data source) or implicitly through latent context inference. Modeling these subsets as distinct tasks enables a principled multi-task IRL formulation and provides a foundation for scalable task interaction modeling using the mean-field approximation. Importantly, this assumption does not require precise task boundaries or manual task labeling, and it mainly serves as a modeling abstraction to facilitate tractable learning and analysis.

For the improved MMF-MTAIRL with theory of mind, each task corresponds to rewards $r_i(s, a, s_i^{ToM})$ and policies $\pi_i(a|s, s_i^{ToM})$ that depend on the state of mind. Then the state of mind is encoded using an encoder network. Before defining the state of mind for ToM, we provide an intuitive interpretation of its components. In our multi-task setting, each task not only has its own learning status but also maintains an internal belief about other tasks, which determines how knowledge should be transferred across tasks. Therefore, we define the state of mind s_i^{ToM} as a compact representation consisting of three parts: (i) a loss function g_i , (ii) belief variables b_i^j over other tasks, and (iii) a confidence term σ_i . Specifically, the loss function g_i reflects the learning status of task $\mathcal{T}_i$ and helps the model capture task-specific dynamics. The belief term b_i^j represents the task i 's belief over other tasks j to model inter-task relations. And the confidence term σ_i measures the confidence of such inter-task influences and controls the strength of cross-task influence, thereby reducing negative transfer when task relationships are weak or noisy. Based on these considerations, the state of mind for task $\mathcal{T}_i$ is defined as $s_i^{ToM} = [g_i, b_i^1, \dots, b_i^{M-1}, \sigma_i]$. In this encoder, an LSTM is used to extract state of mind s_i^{ToM} from historical data ξ_i of the task.

The architecture of the ToM encoder is as follows. For each task $\mathcal{T}_i$, we construct a historical observation sequence from the sampled trajectory $\xi_i = \{(s_t, a_t)\}_{t=1}^T$. At each time step t , the state-action pair (s_t, a_t) is first mapped into a feature function ϕ_t through a

learnable embedding (a fully connected layer). The feature sequence $\{\phi_t\}_{t=1}^T$ is then fed into an LSTM encoder to capture temporal dependencies and behavioral patterns in ζ_i :

$$h_t, c_t = \text{LSTM}(\phi_t, h_{t-1}, c_{t-1}) \quad (7)$$

where h_t and c_t denote the hidden state and cell state, respectively. The final hidden representation h_T is used as the summary of the task's historical behavior and is projected via a multilayer perceptron to obtain the embedding for the state of mind:

$$s_i^{ToM} = f_{ToM}(h_T) \quad (8)$$

The output s_i^{ToM} is designed to encode the loss function g_i , belief b_i^j , and the confidence σ_i . During training, the ToM state is dynamically updated, and the average state of mind s_{avg}^{ToM} is recomputed by the LSTM every K steps based on the most recent mixed expert demonstration trajectories. We use a 2-layer LSTM with hidden size H , and the output dimension of s_i^{ToM} is d_{ToM} .

TMMF-MTAIRL, based on theory of mind, simplifies multi-task interactions by modeling each task in interaction with an average task $\tilde{\mathcal{T}}$. The average state of mind is calculated as:

$$s_{avg}^{ToM} = \mathbb{E}_{i \sim u(\{1, \dots, M\} \setminus \{j\})} [s_i^{ToM}] \approx \frac{1}{M-1} \sum_{i \neq j} s_i^{ToM} \quad (9)$$

where s_i^{ToM} represents the state of mind of the expert demonstration trajectory ζ_i , and u indicates the uniform distribution. Each task $\mathcal{T}_i$ interacts only with the average task $\tilde{\mathcal{T}}$, rather than with all other tasks, thereby reducing computational complexity. Note that s_i^{ToM} represents continuous embedding produced by the ToM encoder; thus, the averaging in Equation (9) is implemented as an element-wise mean over the embedding dimensions, which can be interpreted as an empirical expectation under the mean-field approximation.

In MMF-MTAIRL with ToM, according to the Bayes' theorem, these probabilistic contextual variables follow a prior distribution $p(v)$. In this complex partially heterogeneous multi-task environment, the mean-field context-conditional policy based on theory of mind with different v represents the task preference with shared structures. The policy, mean field, mind, and reward function all depend on v , which are defined as: $\pi : \mathcal{S} \times \mathcal{U} \rightarrow \Delta(\mathcal{A})$, $\eta : \mathcal{U} \rightarrow \Delta(\mathcal{S})$, $R : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \times \mathcal{U} \rightarrow \mathbb{R}$. For simplicity, we assume that the state space, action space, state transition function, and initial state distribution associated with mean field and mind are independent of v . Based on the above definition, the mind-based context-conditional trajectory distribution induced by a policy π is formulated as:

$$p_{\eta, \pi}(\zeta|v, \theta) = \eta_1(s_1) \prod_{t=1}^T \pi(a_t | (s_t, s_i^{ToM}), v) P(s_{t+1} | (s_t, s_i^{ToM}), a_t, \eta_t(\cdot|v)) \quad (10)$$

In Equation (10), θ denotes the parameters of the reward $r_\theta(\cdot)$ obtained by Equation (6). Therefore, the trajectory distribution depends on θ through the reward function r_θ and the corresponding MaxEnt-induced policy.

With the prior distribution $p(v)$ of the latent mind-based context variable and the conditional distribution in Equation (10), the optimal mind-based mean-field context-conditional policy aims to maximize the entropy-regularized expected rewards:

$$\pi^* = \arg \max_{\eta, \pi} \mathbb{E}_{v \sim p(v), \tau \sim p_{\eta, \pi}(\zeta|v)} \left[\sum_{t=1}^T r((s_t, s_i^{ToM}), a_t, \eta_t, v) - \log \pi_t(a_t | (s_t, s_i^{ToM}), v) \right] \quad (11)$$

where $\eta_t(s_t, s_i^{ToM}|v)$ follows the consistency between the mean-field flow and policy flow.

Let $R(s, a, \eta, v)$ indicate the reward function of expert, $\pi_E(a_t|s_t, v)$ represent the expert policy corresponding to the optimal solution of Equation (11), $\eta_E(\cdot|v)$ indicate v -conditional mind flow, and ξ indicate the expert demonstration sampled from the induced marginal distribution $p_{\eta_E, \pi_E}(\xi) = \int p_{\eta_E, \pi_E}(\xi|v, \theta)p(v, \theta)dv$. Within the framework of TMMF-MTAIRL, given the updated mixed expert demonstrations, the meta-learned inference model $q(v|\xi_E)$ first infers the task-specific latent probability variable $\tilde{v} \sim q(v|\xi_E)$. This enables the corresponding reward function $r(s, a, \eta, \tilde{v})$ for tasks to guide the learning of an optimal policy that approximates the policies of expert's rewards $R(s, a, \eta, \tilde{v})$.

TMMF-MTAIRL explicitly models latent relationships between multiple tasks by embedding the ToM, which enhances both reward learning and policy optimization through cross-task observation and utilization, while also providing more comprehensive environmental information. This improves interpretability and adaptability in multi-task learning.

4.3. TMMF-MTAIRL with Mutual Information Regularization

Under TMMF-MTAIRL, the context variable inference model is parametrized as $q_\psi(v|\xi)$, and the reward function is parametrized as $r_\theta((s_t, s_i^{ToM}), a, \eta, v)$, where v is inferred from $q_\psi(v|\xi)$. Let $(\eta_\theta(\cdot|v), \pi_\theta(\cdot|(s_t, s_i^{ToM}), v))$ represent the mind-based mean field induced by r_θ . Therefore, under the $(\eta_\theta(\cdot|v), \pi_\theta(\cdot|(s_t, s_i^{ToM}), v))$, the trajectory distribution $p_\theta(\xi|v)$ induced by the reward function under different potential variables can be expressed as:

$$p_\theta(\xi|v) = \frac{1}{Z(\theta)} \left[\prod_{t=1}^T \eta_\theta^t(s_t, s_i^{ToM}) \right] e^{\sum_{t=1}^T r_\theta((s_t, s_i^{ToM}), a_t, \eta_\theta^t, v)} \quad (12)$$

where $Z(\theta)$ represents the partition function, and η_θ denotes the induced mean-field flow consistent with π_θ .

The maximum likelihood function of the conditional distribution $p_\theta(\xi|v)$ in Equation (12) is taken to learn rewards corresponding to multiple tasks using the proposed TMMF-MTAIRL. Because v is inferred from $q_\psi(v|\xi)$, if the probabilistic context variable v is not further restricted, v will be inferred in advance using $q_\psi(v|\xi)$ in the process of solving the maximum likelihood function, resulting in rewards that are independent of ξ . Therefore, it is necessary to introduce an implicit latent context variable and its relationship with the rewards. We first consider the mutual information between the probabilistic contextual variable and the trajectory ξ under the joint distribution $p_\theta(v, \xi) = p(v)p_\theta(\xi|v)$, which is defined as

$$I_{p_\theta}(v; \xi) = \mathbb{E}_{v \sim p(v), \tau \sim p_\theta(\xi|v)} [\log p_\theta(v|\xi) - \log p(v)] \quad (13)$$

where $p_\theta(v|\xi)$ is the posterior distribution corresponding to the conditional distribution $p_\theta(\xi|v)$ calculated by in Equation (12); the trajectory ξ is sampled from a trajectory distribution induced by the parameterized reward function.

Since we cannot directly obtain a prior distribution $p(v)$, a posterior distribution $p_\theta(v|\xi)$, v -induced conditional mean field $\eta_E(\cdot|v)$, or the conditional mind $\mu_E(\cdot|v)$, optimizing the solution of interactive information in Equation (13) becomes challenging. To address this challenge, following the InfoGAIL's formula reasoning in [48], $q_\psi(v|\xi)$ is first used to estimate experience value of $\eta_\theta(\cdot|v)$ from the updated mixed expert demonstration, thereby converting the v -induced conditional mean dependence on v into dependence on ψ . Subsequently, the inference network $q_\psi(v|\xi)$ can serve as a variational approximation of the posterior distribution $p_\theta(v|\xi)$, allowing both the inference of task uncertainty and the approximation of sampling from $p(v)$. The detailed estimation process will be elaborated later. Let $p_{\pi_E}(\xi)$ represent the expert trajectory distribution.

In addition to the above mutual information constraints, the learning of rewards and inference network must satisfy two objectives: Goal 1) $\arg \min \mathbb{E}_{v \sim p_v} [D_{KL}(p_{\pi_E}(\xi|v)) || p_{\theta}(\xi|v)]$; Goal 2) $\arg \min \mathbb{E}_{\xi \sim p_{\theta}(\xi)} [D_{KL}(p_{\theta}(v|\xi)) || q_{\psi}(v|\xi)]$. Goal 1 aims to learn a reward function r_{θ} that induces data optimally explaining the expert policy. This involves taking mean over the prior for a potential variable v , ensuring both that the trajectory likelihood distributions are as similar as possible, and that $p_{\pi_E}(\xi)$ closely matches $p_{\theta}(\xi)$. When the first objective is satisfied, $p_{\theta}(\xi|v)$ can be considered to represent $p_{\pi_E}(\xi|v)$. Goal 2 requires that the inference network $p_{\theta}(v|\xi)$ approximates the true posterior distribution $q_{\psi}(v|\xi)$ of latent variables, enabling $q_{\psi}(v|\xi)$ to infer latent context variables from new expert presentations. With the objective of the mutual information in Equation (13) and the constraints of Goal 1 and Goal 2, the mind-based meta IRL with mutual information regularization can be formulated as a constrained optimization problem. We consider using the Lagrangian function to solve this optimal problem,

$$\min_{\theta, \psi} -I_{p_{\theta}}(v; \xi) + \alpha \mathbb{E}_{p(v)} [D_{KL}(p_{\pi_E}(\xi|v) || p_{\theta}(\xi|v))] + \beta \cdot \mathbb{E}_{p_{\theta}(\xi)} [D_{KL}(p_{\theta}(v|\xi) || q_{\psi}(v|\xi))] \quad (14)$$

In Equation (14), let $\alpha = 1$ and $\beta = 1$; the Lagrangian function can be rewritten as

$$\begin{aligned} \min_{\theta, \psi} & \mathbb{E}_{p(v)} [D_{KL}(p_{\pi_E}(\xi|v) || p_{\theta}(\xi|v))] + \mathbb{E}_{p_{\theta}(v, \xi)} \left[\log \frac{p(v)}{p_{\theta}(v|\xi)} + \log \frac{p_{\theta}(v|\xi)}{p_{\psi}(v|\xi)} \right] \\ &= \min_{\theta, \psi} -\mathbb{E}_{p(v)} [D_{KL}(p_{\pi_E}(\xi|v) || p_{\theta}(\xi|v))] + \mathbb{E}_{v \sim p(v), \xi \sim p_{\theta}(\xi|v)} [\log q_{\psi}(v|\xi)] \\ &= \min_{\theta, \psi} -\mathbb{E}_{p(v)} [D_{KL}(p_{\pi_E}(\xi|v) || p_{\theta}(\xi|v))] + \mathcal{K}_{mut}(\theta, \psi) \end{aligned} \quad (15)$$

Because the negative term $-H_p(v) = \mathbb{E}_{p_{\theta}(v, \xi)} [\log p(v)] = \mathbb{E}_{p(v)} [\log p(v)]$ can be regarded as a constant during the optimization of the parameters θ and ψ , the $-H_p(v)$ is therefore omitted.

For calculating Equation (15), the unknown $\eta_{\theta, t}(\cdot|v)$ makes it difficult to estimate the KL divergence between the empirical expert distribution and the energy-based trajectory distribution $p_{\theta}(\xi|v)$. To calculate the $\eta_{\theta, t}(\cdot|v)$, we first consider using the estimate $\hat{\eta}_{E, t}(\cdot|v)$ to implement the calculation of $\eta_{\theta, t}(\cdot|v)$. However, the dependence of $\hat{\eta}_{E, t}(\cdot|v)$ on v makes it difficult to obtain the posterior distribution $p(v|\xi)$. Fortunately, the adversarial maximum entropy-based IRL provides an efficient sampling method for estimating the maximum likelihood function. Consequently, $\eta_{\theta, t}(\cdot|v)$ can be used to approximate the $p(v|\xi)$, and $\hat{\eta}_{E, t}(\cdot|v)$ can be estimated:

$$\hat{\eta}_{\psi, t} \left((s, s_i^{ToM}) | v \right) = \mathbb{E}_{\xi_E \sim p_{\eta_E, \pi_E}(\xi)} \left[q_{\psi}(v|\xi_E) \cdot \mathbb{I}_{s_t, s_i^{ToM} = s, s_i^{ToM}} \right] \quad (16)$$

When $\hat{\eta}_{\psi, t}(s, s_i^{ToM}|v)$ replaces $\eta_{\theta, t}(\cdot|v)$, $p_{\theta}(\xi|v)$ additionally depends on ψ . Therefore, based on Equation (16), the conditional distribution can be rewritten as

$$p_{\theta, \psi}(\xi|v) = \frac{1}{Z(\theta, \psi)} \left[\prod_{t=1}^T \hat{\eta}_{\psi, t} \left(s_t, s_i^{ToM} \right) \right] \cdot e^{\sum_{t=1}^T r_{\theta}((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi, t}, v)} \quad (17)$$

Lemma 1. In the adversarial framework of the mind-based meta IRL with mutual information regularization, the adaptive samplers are defined as $\pi_{\theta, t}(a|(s, s_i^{ToM}), v)$ and the discriminator is written as $D_{\theta}((s, s_i^{ToM}), a, \hat{\eta}_{\psi, t}, v) = \frac{e^{r_{\theta}((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi, t})}}{e^{r_{\theta}((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi, t})} + \pi_{\theta}(a|(s_t, s_i^{ToM}), v)}$. During adversarial training process with deterministic dynamics, the adaptive samplers π_{θ} eventually converge to the optimal value with the reward signal $(\log D_{\theta} - \log(1 - D_{\theta}))$. At this point, the trajectory

distribution induced by $(\hat{\eta}_\psi, \pi_\theta^*)$ corresponds to the maximum entropy conditional trajectory distribution $p_{\theta,\psi}(\xi|v)$:

$$p_{(\hat{\eta}_\psi, \pi_\theta^*)}(\xi|v) = \frac{1}{Z(\theta, \psi)} \left[\kappa(s_1) \prod_{t=1}^T \hat{\eta}_{\psi,t}(s_t, s_i^{ToM}) \right] \cdot e^{\sum_{t=1}^T r_\theta((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v)} = p_{\theta,\psi}(\xi|v) \quad (18)$$

Proof. See Appendix A $\square$

Lemma 2. The gradient of $\mathcal{K}_{mut}(\theta, \psi)$ with respect to θ can be estimated by Equation (19):

$$\begin{aligned} \mathbb{E}_{v \sim p(v), \xi \sim p_{\pi_\omega^*}(\xi|v)} \left[\log q_\psi(v|\xi) \left[\sum_{t=1}^T \frac{\partial}{\partial \theta} r_\theta((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v) \right. \right. \\ \left. \left. - \mathbb{E}_{\xi' \sim p_{\pi_\omega^*}(\xi'|v)} \frac{\partial}{\partial \theta} r_\theta((s_{t+1}, s_i^{ToM}), a_{t+1}, \hat{\eta}_{\psi,t+1}, v) \right] \right] \end{aligned} \quad (19)$$

Proof. See Appendix B and Appendix C and Appendix D. $\square$

Using Lemma 2, v can be sampled through the generative process and the trajectories are sampled using the conditional trajectory distribution $p_\theta^*(\xi|v)$ to estimate $\frac{\partial}{\partial \theta} \mathcal{K}_{mut}(\theta, \psi)$. Therefore, the whole training goal is:

$$\begin{aligned} \min_{\omega} \max_{\theta, \psi} \mathbb{E}_{\xi_E \sim p_{\pi_E}(\xi), v \sim q_\psi(v|\tau_E), (s, a) \sim p_{\pi_\omega}(s, a|v)} \log(1 - D_\theta((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v_t)) \\ + \mathbb{E}_{\xi_E \sim p_{\pi_E}(\xi), v \sim q_\psi(v|\tau_E), (s, a) \sim p_{\pi_\omega}(s, a|v)} \log(D_\theta((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v_t)) + \mathcal{K}_{mut}(\theta, \psi) \end{aligned} \quad (20)$$

$$\text{where } D_\theta((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v_t) = \frac{e^{r((s_t, s_i^{ToM}), a_t, v_t)}}{e^{r((s_t, s_i^{ToM}), a_t, v_t)} + \pi(a_t|(s_t, s_i^{ToM}), v_t)}.$$

In the discriminator, the inputs consist of both states s_t and the average state of mind s_{avg}^{ToM} , while the outputs are task-specific rewards $\hat{r}_i(s_t, a_t, s_i^{ToM})$. Incorporating of average state of mind introduces additional dependency conditions for task-specific rewards, ensuring that multiple learned rewards do not correspond to identical policies, thereby resolving the problem of reward ambiguity.

The pseudo-code of the multi-task meta-adversarial mean-field IRL based on theory of mind in the training process is as shown in Algorithm 1. In the proposed TMMF-MTAIRL, all neural network parameters are initialized using standard He initialization. Training is terminated when the maximum number of training iterations is reached. The same stopping criterion is applied across all methods to ensure fair comparison.

As well, the overall structural framework of the multi-task meta-adversarial mean-field IRL based on theory of mind is shown in Figure 1. As depicted in Figure 1, the original expert demonstration data is first combined with the data generated by the generator to create mixed expert demonstration data, from which M expert demonstration trajectories are sampled. The learning of rewards and policies under these M mixed expert demonstration trajectories is treated as M correlated tasks. Next, the states of mind are initialized, and an LSTM is employed to encode them, obtaining the task-specific state of mind. The averaged state of mind across other tasks is computed. Furthermore, a multi-task adversarial IRL framework is constructed. This framework extends adversarial IRL to average theory of mind and introduces probabilistic context variables to characterize task feature distributions. This enables simultaneous learning of multi-task reward functions and average state of mind. During training, the discriminator takes states s_t , actions a_t , and average state of mind s_{avg}^{ToM} as inputs, and outputs both task-specific rewards $\hat{r}_i(s_t, a_t, s_i^{ToM})$ and predicted average state of mind $\hat{s}_{avg}^{ToM}$. To achieve

dynamic updates of state of mind, the average state of mind $\hat{s}_{avg}^{ToM}$ is recomputed using the LSTM every K steps based on the most recent expert demonstration trajectory data. Simultaneously, the framework minimizes the discriminator's adversarial loss function $\mathcal{L}_{AIAL} = \mathbb{E}_{\xi \sim \pi} [\log D_{\theta}(s_t, a_t, v_t)] + \mathbb{E}_{\xi \sim \pi^*} [\log(1 - D_{\theta}(s_t, a_t, v_t))] + \lambda \|\hat{s}_{avg}^{ToM} - s_{avg}^{ToM}\|^2$. During the training of the generator, the generator takes states s_t and states of mind s_i^{ToM} as inputs, and outputs actions $a_t^i \sim \pi_i(a_t^i | s_t, s_i^{ToM})$, aiming to maximize the cumulative rewards $\mathbb{E}_{\pi_i} [\sum_t r_i(s_t, a_t, s_i^{ToM})]$. Finally, the generator's mixed data is utilized to update the mixed expert demonstration dataset, enhancing the quality of expert demonstrations and yielding superior demonstration data. Subsequently, reward learning is performed on the newly acquired expert demonstrations in preparation for the next training iteration, and this process continues until convergence is achieved.

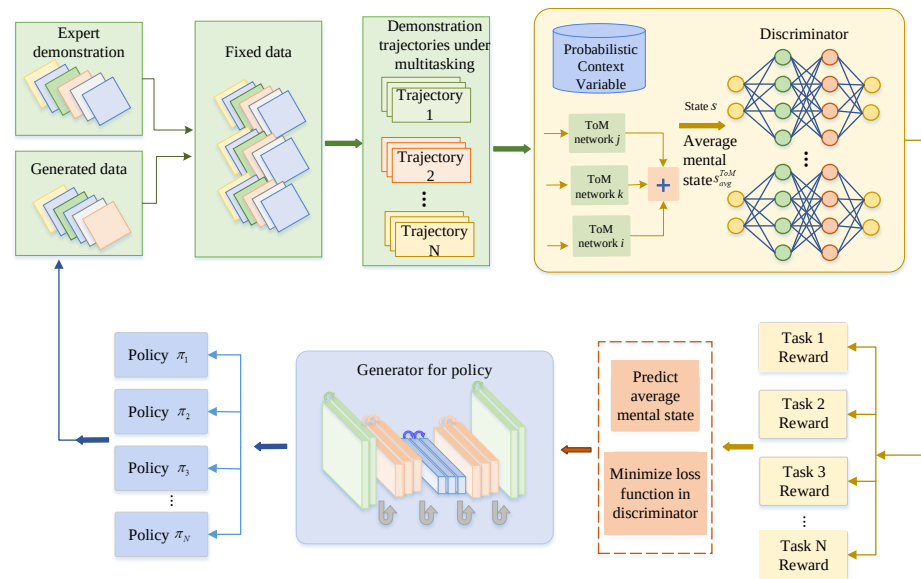


Figure 1. The overall structure framework of the multi-task meta-adversarial mean-field IRL based on theory of mind.

We analyze the computational complexity of the proposed TMMF-MTAIRL with respect to the number of tasks M , the trajectory length T (for $\xi = \{(s_t, a_t)\}_{t=1}^T$), and the number of inner updates K in Algorithm 1. In each update, for each task $\mathcal{T}_i$, we sample trajectories $\xi \in \mathcal{D}$ and perform: (i) context inference of the latent task via $q_{\psi}(v|\xi)$ (and the mutual-information-related updates), (ii) reward update of the discriminator D_{θ} , and (iii) policy update for π_{ω} . Let c_{ψ} , c_{θ} , and c_{ω} denote the per-step (forward/backward) computational costs of q_{ψ} , D_{θ} , and π_{ω} , respectively. Then the per-task cost of one inner update scales as $\mathcal{O}(T(c_{\psi} + c_{\theta} + c_{\omega}))$, and the overall cost per outer iteration (over all M tasks and K inner updates) is $\mathcal{O}(MK \cdot T(c_{\psi} + c_{\theta} + c_{\omega}))$. The key to reducing complexity through mean-field lies in the modeling approach for cross-task interactions. If traditional pair-wise task interaction is adopted, the interaction-related computations usually increase quadratically with the number of tasks, and the complexity is $\mathcal{O}(M^2)$. In contrast, our method aggregates the influence of other tasks through the averaged state of mind, so that the main task interacts only with this mean-field representation instead of all $M - 1$ tasks individually. As a result, the interaction-related complexity is reduced from quadratic to linear, that is, $\mathcal{O}(M^2KT) \rightarrow \mathcal{O}(MKT)$. Moreover, since s_{avg}^{ToM} is recomputed by the ToM encoder every K steps, the ToM recomputation cost is amortized over K inner updates, further improving practical efficiency. Therefore, the proposed algorithm reduces the computational complexity.

Algorithm 1 Multi-task meta-adversarial mean-field IRL based on theory of mind

Input: Expert trajectories $D_E = \{\xi_j^E\}$, initial parameters of $r_\theta, \pi_\theta, q_\psi$

for iteration = $1, \dots, N$ **do do**

 Sample training tasks $\mathcal{T}_i \sim p(\mathcal{T}), i \in \{1, 2, \dots, M\}$;

for all tasks $\mathcal{T}_i$ **do do**

 Obtain expert demos $\mathcal{D}_{E, \mathcal{T}_i}$;

for For update = $1, \dots, K$ **do do**

 Obtain generated demos $\mathcal{D}_{G, \mathcal{T}_i}$ and mixed demons $\mathcal{D}_{\mathcal{T}_i}$;

 Infer state of mind and a batch of latent context variables from the sampled demonstrations: s_{avg}^{ToM} and $v \sim q_\psi(v|\xi), \xi \in \mathcal{D}_{\mathcal{T}_i}$;

 Update ψ to increase $\mathcal{K}_{mut}(\theta, \psi)$ with gradients with samples from $\mathcal{D}_{\mathcal{T}_i}$;

 Update θ to increase $\mathcal{K}_{mut}(\theta, \psi)$ with gradients with samples from $\mathcal{D}_{\mathcal{T}_i}$;

 Update θ to minimize the loss function of the discriminator: $\mathcal{L}_{AIAL} = \mathbb{E}_{\xi \sim \pi} [\log D_\theta(s_t, a_t, \hat{\eta}_{\psi, t}, v_t)] + \mathbb{E}_{\xi \sim \pi^*} [\log(1 - D_\theta(s_t, a_t, \hat{\eta}_{\psi, t}, v_t))] + \lambda \|s_{avg}^{ToM} - s_{avg}^{ToM}\|^2$;

 Update ω with TRPO to increase the following objective: $\mathbb{E}_{(s_t, a_t, \hat{\eta}_{\psi, t}, v_t) \sim \mathcal{D}_{\mathcal{T}_i}} [\log D_\theta(s_t, a_t, \hat{\eta}_{\psi, t}, v_t)]$;

end for

end for

Output: Learned inference model $q_\psi(v|\xi)$, reward function $r_\theta(s, a, v)$ and policies $\pi_\theta(a|s, v)$;

5. Experiments and Simulations

In this section, we use the point-maze environment and the rolling bearing fault diagnosis experiment to represent the performance of the proposed TMMF-MTAIRL algorithm for learning rewards and optimizing policies.

5.1. Point-Maze Environments

This environment is refactored from the D4RL repository, as introduced in [49]. The task involves a 2-DoF ball that is actuated by forces in the cartesian directions x and y , and must navigate toward a target goal within a closed maze. The environment can be initialized with different maze configurations of varying complexity. In the MuJoCo simulation, the target goal can be visualized as a red static ball, while the actuated ball is green. The control frequency of the ball is $f = 10$ Hz. Table 2 shows the values of control parameters used during the collection of the expert demonstrations. The values of control parameters used in the training process of TMMF-MTAIRL are listed in Table 3. Most hyperparameters listed in Tables 2 and 3 were selected based on the successful configuration of similar environments and algorithms. Unless otherwise stated, we use the same set of hyperparameters across all tasks and environments. Empirically, we observe that the performance of TMMF-MTAIRL is robust to moderate variations of these parameters, which facilitates practical deployment.

Table 2. Values of control parameters in data collection.

Control Parameters	Values
Number of iterations	2000
Batch size	20,000
Maximum path length	500
Discount factor	0.99
Step size	0.01

Table 3. Values of control parameters in TMMF-MTAIRL.

Control Parameters	Values
Number of iterations	3000
Batch size	16
Meta batch size	50
Maximum path length	100
Discount factor	0.99
Step size	0.01
Number of pre-training epochs	1000
Entropy weight	1.0
Information coefficient	0.1

For the convenience of experimental comparison, we make the following definitions. TaskRmean denotes the mean episodic rewards averaged over all tasks and runs; Mulinfo denotes the average mutual information between multiple tasks in Equation (13); MeanKL denotes the average KL divergence used for evaluating policy distribution shift across tasks; and dLoss denotes the discriminator's loss function $\mathcal{L}_{AIAL}$ averaged over training iterations; Rmean denotes the mean episodic rewards averaged over all runs; times indicates the running times.

To verify the performance of the proposed algorithm, we compared TMMF-MTAIRL with PEMIRL, and Meta-IRLSOT. All experimental results were obtained from five independent runs. Meta-inverse reinforcement learning for mean-field games via probabilistic context variables (PEMIRL) [20]: This method provides a deep latent variable MFG-IRL model for IRL in mean-field games, which can infer reward functions from different but structurally similar expert demonstration tasks. Importantly, it does not require prior knowledge of the underlying context or modifications to the MFG model itself. A meta-inverse reinforcement learning method for fast adaptation of trajectory prediction networks (Meta-IRLSOT) [18]: This Meta-learning-based RL method incorporates an IRL framework to explore trajectory-scene relationships, achieving task-level scene understanding. By introducing meta-learning for co-training, the model enhances its generalization to novel scenarios.

Figure 2a–f show the visualization of the learned reward function using the proposed TMMF-MTAIRL at iteration 0, 20, 40, 60, 80, 99, respectively. The horizontal and vertical axes now denote the spatial coordinates of the point-maze environment, and the colorbar indicates the magnitude of the learned reward values. From Figure 2, it can be seen that the positional intuitive schematic diagrams of the target goal and the agent point are obtained during the learning process of rewards. The target goal's position changes dynamically, and the agent aims to reach it.

Figure 3a–f show the results of loss 1, loss 2, loss 3, policy likelihood, KL, total loss for PEMIRL, Meta-IRLSOT, and TMMF-MTAIRL in the maze left-v0 environment. From these results, we can obtain that our proposed TMMF-MTAIRL has lower loss 1, loss 2, loss 3, policy likelihood, KL, and total loss than those of PEMIRL and Meta-IRLSOT during the pre-training process in the point-maze left-v0 environment. These results demonstrate that TMMF-MTAIRL outperforms both PEMIRL and Meta-IRLSOT.

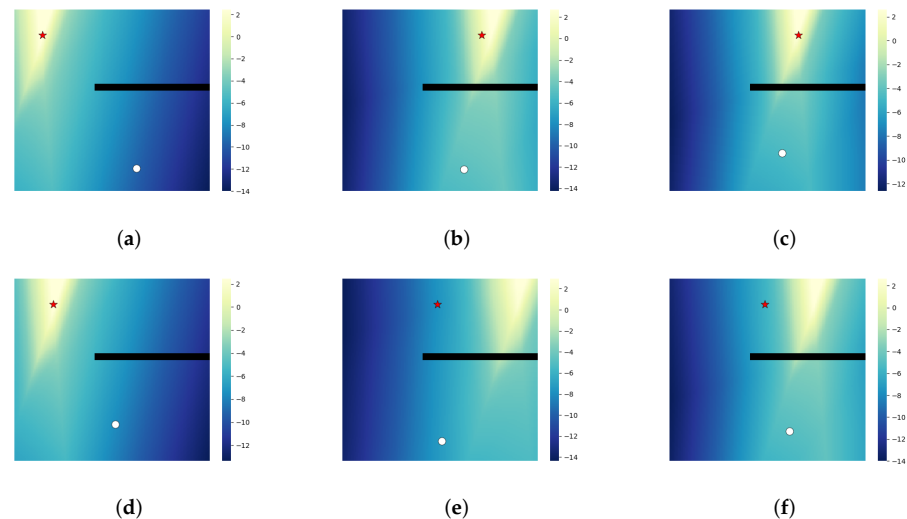


Figure 2. Visualization of the learned rewards: (a–f) are obtained at iteration 0, 20, 40, 60, 80, 99, respectively.

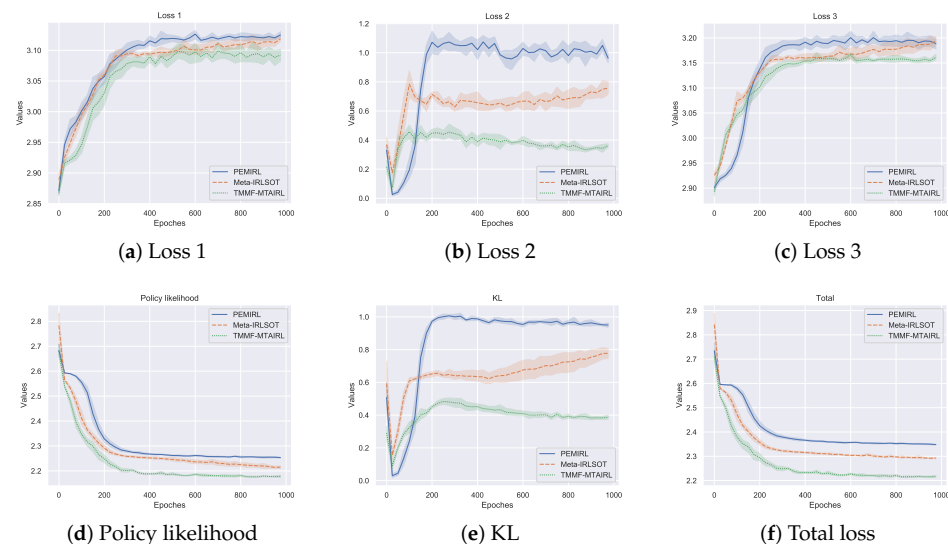


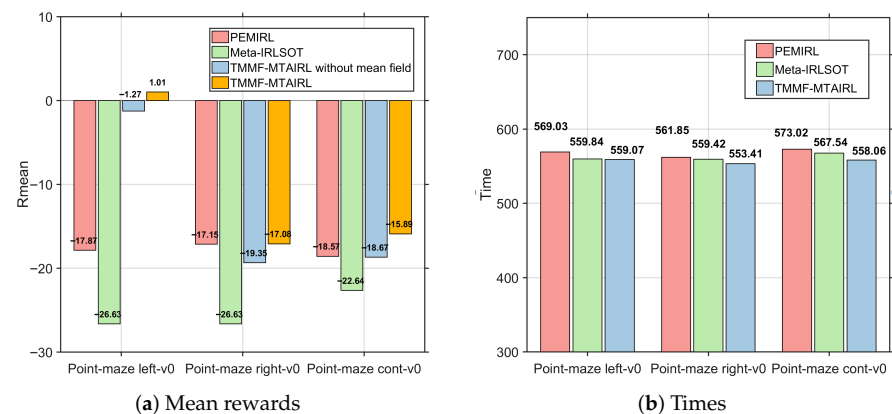
Figure 3. The results of PEMIRL, Meta-IRLSOT, and TMMF-MTAIRL in the point-maze left-v0 environment.

Table 4 shows the training data obtained in point-maze left-v0, point-maze right-v0, and point-maze cont-0 for TMMF-MTAIRL, PEMIRL, and Meta-IRLSOT. From the results of point-maze left-v0 and point-maze right-v0, it can be seen that our proposed TMMF-MTAIRL has higher TaskRmean and Mulinfo, while exhibiting lower MeanKL, and dLoss, than PEMIRL and Meta-IRLSOT. Although some metrics obtained by our proposed TMMF-MTAIRL algorithm are not optimal, overall performance analysis indicates that it outperforms the other algorithms. To further illustrate the effectiveness of the proposed algorithm, we also collected the testing data in point-maze cont-v0 using TMMF-MTAIRL, PEMIRL, and Meta-IRLSOT to test invisible target points. The results of Rmean, TaskRmean, Mulinfo, MeanKL, and dLoss indicate that our proposed TMMF-MTAIRL consistently performs better than the other algorithms during testing.

Figure 4a,b illustrate that TMMF-MTAIRL achieves higher Rmean and lower times than PEMIRL, Meta-IRLSOT, and TMMF-MTAIRL without mean field in point-maze environments. Therefore, TMMF-MTAIRL demonstrates superior performance relative to the other algorithms used in this paper.

Table 4. The training data obtained in point-maze environments.

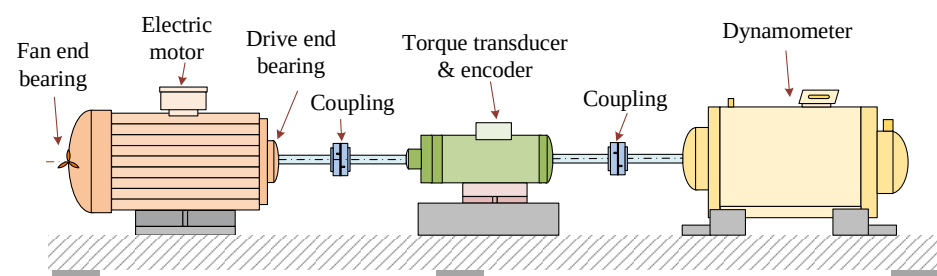
Environments	Methods	TaskRmean	Mulinfo	MeanKL	dLoss
Point-maze left-v0	PEMIRL	-0.143 ± 0.022	-0.275 ± 0.026	0.106 ± 0.003	0.002 ± 0.001
	Meta-IRLSOT	-0.128 ± 0.010	-0.248 ± 0.016	0.317 ± 0.007	$1.3 \times 10^{-5} \pm 0.000$
	TMMF-MTAIRL	-0.125 ± 0.009	-0.112 ± 0.013	0.079 ± 0.002	0.0004 ± 0.000
Point-maze right-v0	PEMIRL	-0.171 ± 0.018	-0.420 ± 0.028	0.068 ± 0.004	0.0007 ± 0.000
	Meta-IRLSOT	-0.187 ± 0.013	-0.432 ± 0.020	0.037 ± 0.006	0.0009 ± 0.000
	TMMF-MTAIRL	-0.171 ± 0.008	-0.069 ± 0.019	0.029 ± 0.001	0.0002 ± 0.000
Point-maze cont-v0	PEMIRL	-0.179 ± 0.021	-0.384 ± 0.024	0.105 ± 0.003	0.005 ± 0.001
	Meta-IRLSOT	-0.195 ± 0.015	-0.367 ± 0.020	0.075 ± 0.005	0.007 ± 0.001
	TMMF-MTAIRL	-0.115 ± 0.004	-0.283 ± 0.017	0.083 ± 0.002	0.003 ± 0.000

**Figure 4.** Visualization of the learned rewards in point-maze environments.

5.2. Rolling Bearing Fault Diagnosis Experiments

For comprehensive performance validation of the proposed method, the TMMF-MTAIRL algorithm is first used to learn the reward function, while simultaneously optimizing expert demonstrations throughout the training phase. Subsequently, RL is applied to classify bearing fault diagnosis data based on the learned reward function. All experimental results were obtained by averaging over 10 independent runs.

Figure 5 shows the rolling bearing fault test rig from Case Western Reserve University, which consists of a fan-end bearing, a motor, a drive-end bearing, a torque transducer/encoder, and a power meter [50]. The experiment collected normal bearing signals and (single-point) drive-end fault signals at sampling rates of 12,000 samples/second and 48,000 samples/second, respectively, whereas all fan-end fault signals were acquired at 12,000 samples/second. To validate the proposed method, the acquired vibration data were divided into four datasets (A, B, C, and D). Each dataset contains 10 distinct classes, with no unseen classes among these four datasets. On average, each class within a given dataset contains approximately 480,000 samples, indicating that the laboratory equipment was operated under each class condition to capture an average of 480,000 observations from vibration sensors.

**Figure 5.** Rolling bearing fault experiment rig.

In the fault classification process, the design of the reward function plays a crucial role in determining the accuracy of fault diagnosis. Therefore, we first employ TMMF-MTAIRL to learn the reward function from expert demonstrations, and then utilize the deep Q-Learning (DQN) algorithm to perform fault diagnosis and classification. The main framework for fault diagnosis based on the TMMF-MTAIRL and DQN algorithms is illustrated in Figure 6. First, the dataset is preprocessed using cross-validation to obtain the training set and validation set. Subsequently, a stacked autoencoder (SAE) is employed to extract the latent features of the data while simultaneously updating and preserving optimal parameters. We then select a portion of the training data as expert demonstration data and sample expert demonstration trajectories from it. Based on the varying preferences among these expert demonstration trajectories, the proposed TMMF-MTAIRL algorithm is employed to learn rewards and optimize the expert demonstration data, treating each trajectory's reward learning as an independent task. Finally, a DQN algorithm that incorporates Bayesian search methodology and learned reward functions is implemented to perform fault classification.

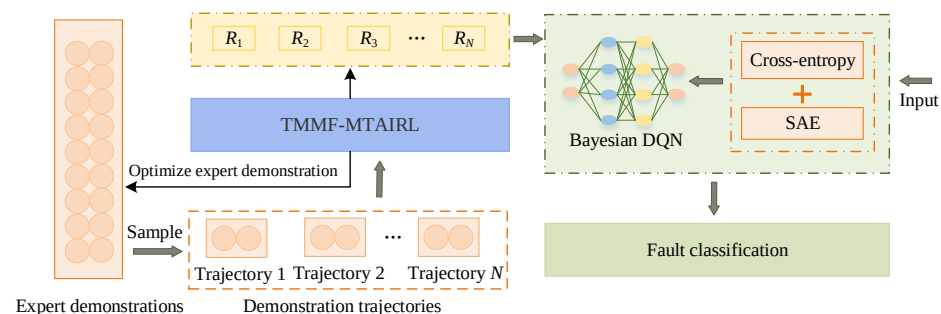


Figure 6. Framework of the fault diagnosis with TMMF-MTAIRL and DQN.

Expert demonstrations $\mathcal{D}$ are constructed by sampling labeled trajectories ξ from the training set, which refer to the high-quality labeled vibration samples collected under controlled operating conditions from the Case Western Reserve University bearing test rig. A trajectory corresponds to a sequential decision process formed by sliding-window segmentation of a vibration time series. Therefore, a trajectory $\xi = \{(s_t, a_t)\}_{t=1}^T$ represents a sequence of vibration windows and predicted diagnostic decisions. Additionally, each fault type can be viewed as a task because it defines a distinct diagnostic objective and induces a different state distribution.

In the bearing fault diagnosis experiment, to evaluate the performance of the TMMF-MTAIRL algorithm, the proposed TMMF-MTAIRL integrated with cross-validated Bayesian DQN (DQNB-TMMF-MTAIRL-5) was compared with the following four algorithms: (1) traditional deep Q-Learning algorithm (DQN) [51]; (2) Bayesian-based deep Q-Learning algorithm (DQN-Bayesian) [52]; (3) the algorithm combining TMMF-MTAIRL with Bayesian DQN (DQNB-TMMF-MTAIRL); and (4) a deep Q-learning algorithm incorporating Bayesian inference and cross-validation (DQN-Bayesian-5).

To further demonstrate the performance of the proposed algorithm, DQN, DQN-Bayesian, DQNB-TMMF-MTAIRL, DQN-Bayesian-5, and DQNB-TMMF-MTAIRL-5 are applied to fault diagnosis using the bearing fault diagnosis dataset B.

During the training process, the complexity of classification data and the design of reward function constrain the accuracy of fault classification using RL. Therefore, we analyze the cumulative rewards obtained during training for fault classification using DQN, DQN-Bayesian, DQNB-TMMF-MTAIRL, DQN-Bayesian-5, and DQNB-TMMF-MTAIRL-5, as shown in Figure 7. During the training process for fault diagnosis, the cumulative rewards learned by DQNB-TMMF-MTAIRL and DQNB-TMMF-MTAIRL-5 converge to

200 more rapidly than those achieved by DQN, DQN-Bayesian, and DQN-Bayesian-5 with manually designed reward functions. To further enhance model performance, the bearing fault diagnosis data was split into five folds via five-fold cross-validation, where four folds were used for training and one fold for validation in each iteration. As a result, DQNB-TMMF-MTAIRL-5 demonstrates the fastest convergence of cumulative rewards, indicating its superior capability in learning reward functions.

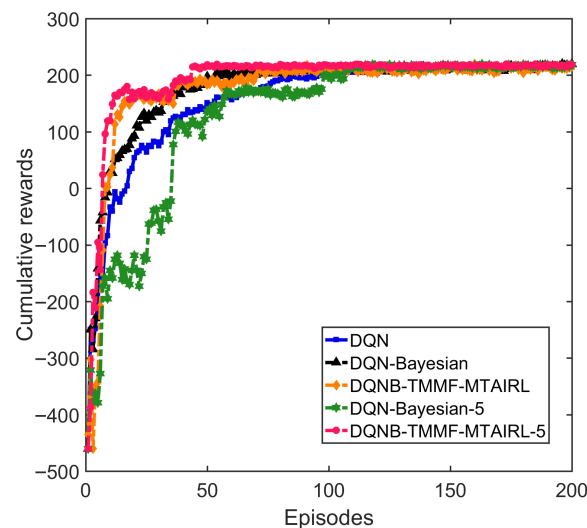


Figure 7. Cumulative rewards of DQN, DQN-Bayesian, DQNB-TMMF-MTAIRL, DQN-Bayesian-5, and DQNB-TMMF-MTAIRL-5. Different line styles and markers are used to distinguish methods.

Figure 8 shows the accuracy of fault classification during the training process for the DQN, DQN-Bayesian, DQNB-TMMF-MTAIRL, DQN-Bayesian-5, and DQNB-TMMF-MTAIRL-5. Figure 9 shows the validation accuracy of the fault classification obtained using DQN-Beyasian, DQNB-TMMF-MTAIRL, DQN-Beyasian-5, and DQNB-TMMF-MTAIRL-5 algorithms. The proposed TMMF-MTAIRL, which explicitly models inter-task relationships, learns reward functions that mitigate the difficulty of manual reward design in RL. In Figures 8 and 9, compared with DQN, DQN-Bayesian, DQN-Bayesian-5, and DQNB-AMEIRL-5, DQNB-TMMF-MTAIRL and DQNB-TMMF-MTAIRL-5 achieve better fault classification, resulting in higher accuracy in both fault classification and verification. The comparison with DQN and DQN-Bayesian is intended to demonstrate the advantage of learned rewards over manually engineered rewards in RL-based fault diagnosis. Furthermore, to improve the utility of bearing fault data, we adopted five-fold cross-validation for data preprocessing. This approach maximizes data utilization, reduces the risks of both overlearning and underlearning, and enhances the credibility of experimental findings. DQNB-TMMF-MTAIRL-5 without mean-field algorithm suffers from the lack of mean field, resulting in low exploration efficiency. Therefore, DQNB-TMMF-MTAIRL-5 algorithm achieves the highest fault classification accuracy on both the test set and the verification set.

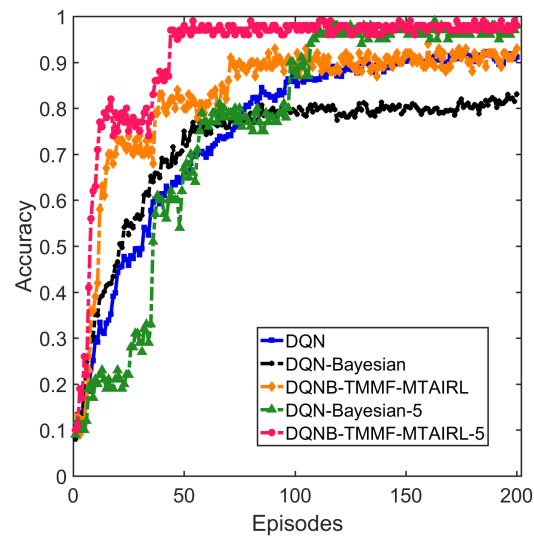


Figure 8. Accuracy of fault classification. Different line styles and markers are used to distinguish methods.

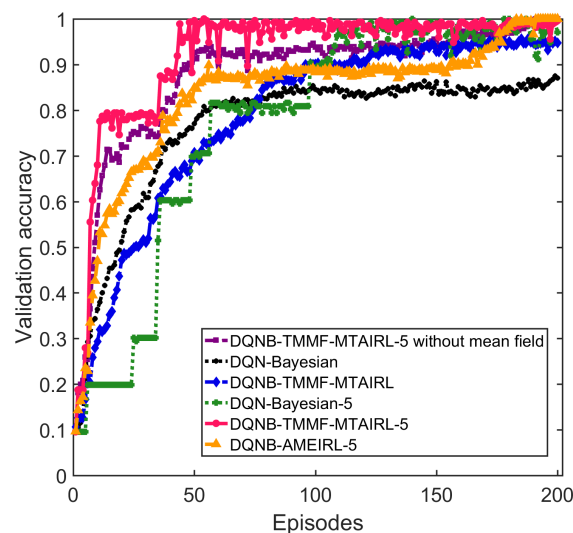


Figure 9. Validation accuracy of fault classification. Different line styles and markers are used to distinguish methods.

To verify the robustness of the proposed TMMF-MTAIRL, we obtained the accuracy of fault classification as 0.97, 0.98, and 0.98, respectively when the multi-task scale M was 10, 15, and 20. The accuracy of this algorithm remains basically stable at around 0.97~0.98 as the scale of multiple tasks changes, which indicates that TMMF-MTAIRL has good robustness.

The proposed TMMF-MTAIRL is optimized via an adversarial minimax objective. Under standard assumptions commonly adopted in adversarial IRL—such as sufficient capacity of discriminator network, accurate inner-loop optimization, and appropriate learning rates—the training dynamics admit stationary points where the induced policy distribution matches the expert trajectory distribution. Furthermore, the mutual-information regularization and mean-field approximation improve stability by preventing degenerate latent inference and reducing high-variance cross-task interactions. In addition, the results in Figures 7–9 show that the proposed DQNB-TMMF-MTAIRL can converge. Therefore, the proposed framework enjoys conditional convergence properties and demonstrates stable empirical behavior.

From an application perspective, it is also important to distinguish the proposed approach from alternative fault diagnosis paradigms. Deep transfer learning methods for fault diagnosis typically rely on labeled data and supervised objectives, achieving strong performance when sufficient annotations are available. In contrast, the proposed TMMF-MTAIRL focuses on learning reward functions and policies from heterogeneous demonstrations without requiring explicit fault labels. These two paradigms address complementary problem settings, and integrating transfer learning representations into the proposed IRL framework is a promising direction for future research.

The proposed TMMF-MTAIRL can explicitly model the relationship between tasks and thus improve the explainability and adaptability across tasks, enabling the algorithm to learn more effective rewards and optimization policies compared to the other algorithms discussed in this paper.

6. Conclusions and Future Works

To overcome the challenges of learning rewards in the complex multi-task environment, a multi-task meta-adversarial mean-field IRL framework based on the theory of mind is proposed to enhance information interaction between partially heterogeneous tasks. In TMMF-MTAIRL, the mean-field theory reduces the computational complexity of learning rewards from fixed expert demonstrations in multi-task settings; the theory of mind improves the information sharing and transferability between multiple tasks, and the deep potential variables enhance the adaptability to novel tasks. These improvements jointly increase learning accuracy and reduce computational complexity in learning reward functions. Experiments in a maze-point environment and a rolling bearing fault diagnosis task demonstrate that TMMF-MTAIRL can learn higher-quality reward functions and policies through continuous iterative updates, thereby outperforming the original expert demonstrations.

While the experimental evaluation focuses on representative benchmark environments and a real-world fault diagnosis dataset, the proposed framework is not limited to these settings. The modular design of the ToM encoder, mean-field approximation, and adversarial IRL objective allows straightforward extension to other domains. Evaluating the method on a wider range of environments and large-scale benchmarks will be explored in future work. In the future work, we will focus on the following three aspects: (1) Extending the proposed TMMF-MTAIRL algorithm to real-world environments for policy optimization, beyond the point-maze and rolling bearing fault experiments used for initial validation; (2) Further enhancing the reward learning performance of TMMF-MTAIRL in ultra-large-scale multi-agent and highly complex multi-task environments; and (3) Deep transfer learning methods and the proposed TMMF-MTAIRL address complementary problem settings, and integrating transfer learning representations into the proposed IRL framework is a promising direction for future research.

Author Contributions: Conceptualization, L.S. and C.C.; methodology, L.S.; validation, C.C. and K.Y.; formal analysis, C.C. and K.Y.; investigation, C.C. and L.S.; resources, L.S. and K.Y.; writing original draft preparation, L.S.; writing-review and editing, C.C. and K.Y.; visualization, L.S.; supervision, K.Y. and C.C.; project administration, L.S. and K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Postdoctoral Research Startup Fund of the Big Picture Center of Hangzhou City University (No.201000-584105/002).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: Thanks to “the Research Center for High-Performance Computing of Ultra-Large-Scale Graph Data at Hangzhou City University”, “the Zhejiang Provincial Engineering Research Center for Real-Time Digital and Intelligent Technology in Urban Safety Governance”, and “the Supercomputing Center of Hangzhou City University” for providing the research start-up funds and some experimental hardware conditions for the research.

Conflicts of Interest: All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Appendix A

Ziebart conducted extensive and in-depth research on MaxEnt IRL, publishing multiple articles in this area and completing his doctoral dissertation on MaxEnt IRL. Based on the theoretical derivations of MaxEnt IRL presented by Ziebart in Sections 2 and 3 of reference [20], it can be concluded that the policies inducing the trajectory distribution in Equation (12) conform to the following energy-based form of MaxEnt IRL:

$$\pi_{\theta}\left(a_t \mid \left(s_t, s_i^{ToM}\right), v\right)=e^{\left(Q_{soft}\left(\left(s_t, s_i^{ToM}\right), a_t, v\right)-V_{soft}\left(\left(s_t, s_i^{ToM}\right), v\right)\right)} \quad (A1)$$

$$Q_{soft}\left(\left(s_t, s_i^{ToM}\right), a_t, v\right)=r_{\theta}\left(\left(s_t, s_i^{ToM}\right), a_t, v\right)+\log \mathbb{E}_{s_{t+1} \sim P\left(\cdot \mid\left(s_t, s_i^{ToM}\right), a_t, v\right)}\left[e^{\left(V_{soft}\left(\left(s_{t+1}, s_i^{ToM}\right), v\right)\right)}\right] \quad (A2)$$

$$V_{soft}\left(\left(s_t, s_i^{ToM}\right), v\right)=\log \int_{\mathcal{A}} e^{Q_{soft}\left(\left(s_t, s_i^{ToM}\right), a', v\right)} d a' \quad (A3)$$

The optimal solution of the entropy-regularized IRL problem in Equation (11) can be obtained by solving for the optimal policy in Equation (10).

$$\max _{\pi} \mathbb{E}_{\pi}\left[\sum_{t=1}^T r_{\theta}\left(\left(s_t, s_i^{ToM}\right), a_t, \hat{\eta}_{\psi, t}, v\right)-\log \pi\left(a_t \mid\left(s_t, s_i^{ToM}\right), \hat{\eta}_{\psi, t}, v\right)\right] \quad (A4)$$

With the basic knowledge of the maximum entropy adversarial IRL in Section 3.2, Equation (A4) can be interpreted as the training objective of the adaptive sampler π_{θ} in adversarial IRL. Accordingly, we can obtain the optimal policy π_{θ}^{*} corresponding to $p_{\theta, \psi}(\xi|v)$ in Equation (12).

Next, we need to compute $\mathcal{K}_{mut}(\theta, \psi)$ with respect to the parameters in the main objective function defined in Equation (16). The gradient of $\mathcal{K}_{mut}(\theta, \psi)$ with respect to the parameters is as follows:

$$\frac{\partial}{\partial \psi} \mathcal{K}_{mut}(\theta, \psi)=\mathbb{E}_{v \sim p(v), \xi \sim p_{\theta}(\xi|v)} \frac{1}{q(v|\xi, \psi)} \frac{\partial q(v|\xi, \psi)}{\partial \psi} \quad (A5)$$

Appendix B

The gradient of $\mathcal{K}_{mut}(\theta, \psi)$ with respect to θ is as follows:

$$\frac{\partial}{\partial \theta} \mathcal{K}_{mut}(\theta, \psi)=\mathbb{E}_{v \sim p(v), \xi \sim p_{\theta}(\xi|v)} \log q(v|\xi, \psi) \frac{\partial}{\partial \theta} \log p_{\theta}((\xi|v)) \quad (A6)$$

From Equation (12), $p_\theta(\xi|v)$ is the energy-based distribution. The gradient of $\log p(\xi|v, \theta)$ with respect to θ is obtained:

$$\begin{aligned} \frac{\partial \log p(\xi|v, \theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \left[\log \left(\eta(s_1) \prod_{t=1}^T P\left((s_{t+1}, s_i^{ToM}) | (s_t, s_i^{ToM}), a_t, \eta_t(\cdot|v)\right) \right) \right. \\ &\quad \left. + \prod_{t=1}^T r_\theta\left((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v\right) - \log Z(\theta) \right] \\ &= \sum_{t=1}^T \frac{\partial}{\partial \theta} r_\theta\left((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v\right) - \frac{\partial}{\partial \theta} \log Z(\theta) \\ &= \sum_{t=1}^T \frac{\partial}{\partial \theta} r_\theta\left((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v\right) - \mathbb{E}_{\xi \sim p(\xi|v, \theta)} \log q(v|\xi, \psi) \frac{\partial}{\partial \theta} \log p_\theta(\xi|v) \end{aligned} \quad (A7)$$

Substituting the derivation result of $\frac{\partial \log p(\xi|v, \theta)}{\partial \theta}$ into Equation (A7),

$$\begin{aligned} \mathbb{E}_{v \sim p(v), \tau \sim p_\theta(\tau|v)} &\left[\log q_\psi(v|\tau) \left[\sum_{t=1}^T \frac{\partial}{\partial \theta} r_\theta\left((s_t, s_i^{ToM}), a_t, \hat{\eta}_{\psi,t}, v\right) \right. \right. \\ &\quad \left. \left. - \mathbb{E}_{\tau' \sim p_\theta(\tau|v)} \sum_{t=1}^T \frac{\partial}{\partial \theta} r_\theta\left((s_{t+1}, s_i^{ToM}), a_{t+1}, \hat{\eta}_{\psi,t+1}, v\right) \right] \right] \end{aligned} \quad (A8)$$

With Lemma 1, when θ is trained to get the optimal value, an unbiased estimate can be constructed using the samples sampled from $p_\theta^*(\xi|v)$.

Appendix C

Table A1 systematically lists all primary variables, operators, and functions used throughout the paper, along with their definitions and domain descriptions.

Table A1. Summary of key mathematical notations.

Notation	Meaning
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
R	Reward space
γ	Discount factor
$P_{ss'}^a$	State transition distribution
$\mathcal{D}$	Expert demonstrations
ϕ_{s_i}	Feature function of each reward
$z(\theta)$	Partition function
$p(s_{t+1} s_t, a_t)$	State transition probability
M	The number of expert demonstration trajectories
π_E	Expert strategy
ξ_j	Expert demonstration trajectories
$D_\theta(s, a)$	Discriminator
$\mathcal{U}$	Value space
$R(s, a, \eta, v)$	Reward functions of expert different tasks
$r(s, a, \eta, v)$	Corresponding reward functions for different tasks
$r_\theta((s_t, s_i^{ToM}), a, \eta, v)$	Parametrized reward function
$\eta_E(\cdot v)$	η -conditional mean-field flow
$\eta_E(\cdot v)$	v -conditional mind flow
$\pi_E(\cdot v)$	Policy flow
$p_{\eta_E, \pi_E}(\tau)$	Marginal distribution of expert policy
$p(v)$	A prior distribution

Table A1. Cont.

Notation	Meaning
$\pi_i(a s, s_i^{ToM})$	Policies
s_i^{ToM}	State of mind
$\mathcal{T}_i$	Tasks
$\tilde{\mathcal{T}}$	Average task
s_{avg}^{ToM}	Average state of mind
$p_{\eta, \pi}$	The mind-based context-conditional trajectory distribution induced by a policy π
$q_{\psi}(v \xi)$	Context variable inference model
$p_{\theta}(\xi v)$	Conditional distribution
$p_{\theta}(v \xi)$	Posterior distribution
ϕ_{s_t}	Feature function of each reward
K	The number of inner updates
T	Trajectory length

Appendix D

The Meta-adversarial mean-field IRL with context-conditional policy is proposed to learn rewards and optimize policy from constantly updated mixed expert demonstrations, improving reward learning precision in partially heterogeneous multi-task environments. The specific contributions and innovative points are shown in Table A2.

Table A2. Summary of contributions and novelty of the proposed TMMF-MTAIRL.

Contribution	Novelty/Key Idea
Embedding ToM for enhanced multi-task knowledge transfer	Introduces ToM-based task representations to explicitly model inter-task relationships and better leverage environmental/task interaction information, enabling effective knowledge transfer across tasks
Mean-field modeling of task interactions to reduce computational complexity	Approximates multi-task coupling by modeling interactions between the main task and the average of remaining tasks (mean-field approximation), significantly reducing complexity caused by pair-wise task interactions
Latent task variable inference for task-specific adaptation under the maximum entropy framework	Accurately infers latent task variables from sampled trajectories of novel tasks and derives task-specific adaptive reward functions and optimal policies under the maximum entropy IRL formulation
Continuously updatable adversarial generative learning framework	Constructs an iterative adversarial training procedure that continuously updates and optimizes the generator and discriminator, improving the quality and utility of mixed expert demonstrations

References

1. Zhang, X.Y.; Cai, X.Y.; Liu, B.; Huang, W.D.; Zhu, S.C.; Qi, S.Y.; Yang, Y.D. Differentiable Information Enhanced Model-Based Reinforcement Learning. In Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 1–9.
2. Mohsin, M.A.; Rizwan, H.; Umer, M.; Bhattacharya, S.; Bilal, A.; Cioffi, J.M. Hierarchical Deep Reinforcement Learning for Adaptive Resource Management in Integrated Terrestrial and Non-Terrestrial Networks. In Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 1–8.
3. Li, D.D.; Xie, L.; Wang, Z.; Yang, H. Brain Emotion Perception Inspired EEG Emotion Recognition with Deep Reinforcement Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 12979–12992.
4. Lv, C.L.; Lv, X.L.; Wang, Z.Y.; Zhao, T.Q.; Tian, W.; Zhou, Q.Q.; Zeng, L.; Wan, M.; Liu, C.G. A focal quotient gradient system method for deep neural network training. *Appl. Soft Comput.* **2025**, *184*, 113704.

5. Wu, X.; Zhang, Y.T.; Lai, C.K.; Yang, M.Z.; Yang, G.L.; Wang, H.H. A Novel Centralized Federated Deep Fuzzy Neural Network with Multiobjectives Neural Architecture Search for Epistatic Detection. *IEEE Trans. Fuzzy Syst.* **2025**, *33*, 94–107.
6. Li, W.L.; Qiu, F.; Li, L.X.; Zhang, Y.N.; Wang, K. Simulation of Vehicle Interaction Behavior in Merging Scenarios: A Deep Maximum Entropy-Inverse Reinforcement Learning Method Combined with Game Theory. *IEEE Trans. Intell. Veh.* **2024**, *9*, 1079–1093.
7. Nan, J.F.; Deng, W.W.; Zhang, R.Z.; Zhao, R.; Wang, Y.; Ding, J. Car-Following Behavior Modeling with Maximum Entropy Deep Inverse Reinforcement Learning. *IEEE Trans. Intell. Veh.* **2024**, *9*, 3998–4010.
8. Li, J.H.; Wu, H.; He, Q.; Zhao, Y.J.; Wang, X. Dynamic QoS Prediction with Intelligent Route Estimation Via Inverse Reinforcement Learning. *IEEE Trans. Serv. Comput.* **2024**, *17*, 509–523.
9. Song, L.; Li, D.Z.; Wang, X.; Xu, X. AdaBoost Maximum Entropy Deep Inverse Reinforcement Learning with Truncated Gradient. *Inform. Sci.* **2022**, *602*, 328–350.
10. Wang P.; Li, H.H.; Chan C.Y. Meta-adversarial Inverse Reinforcement Learning for Decision-making Tasks. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 12632–12638.
11. Chen, J.B.; Yu, T.; Pan, Z.N.; Zhang, M.Y.; Lu, G.H.; Zhu K.D. Stochastic Dynamic Power Dispatch with Human Knowledge Transfer Using Graph-GAN Assisted Inverse Reinforcement Learning. *IEEE Trans. Smart Grid* **2024**, *15*, 3303–3315.
12. Wang, Y.J.; Niu, Y.C.; Zhu, W.Y.; Chen, W.Q.; Li, Q.; Wang T. Predicting Pedestrian Crossing Behavior at Unsignalized Mid-Block Crosswalks Using Maximum Entropy Deep Inverse Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 3685–3698.
13. Bighashdel, A.; Meletis, P.; Jancura, P.; Dubbelman G. Deep Adaptive Multi-intention Inverse Reinforcement Learning. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2021; pp. 206–221.
14. Krishnan, S.; Garg, A.; Liaw, R.; Miller, L.; Pokorny, F.T.; Goldberg, K. Hirl: Hierarchical Inverse Reinforcement Learning for Long-horizon Tasks with Delayed Rewards. *arXiv* **2016**, arXiv:1604.06508.
15. Zhang, N.; Zhao, Y.J.; Yang, M.G.; Dai, S.L. Hierarchical Reinforcement Learning with Demonstration for Long-Horizon Robotic Manipulation. *Arab. J. Sci. Eng.* **2025**. <https://doi.org/10.1007/s13369-025-10499-x>.
16. Glazer, N.; Navon, A.; Shamsian, A.; Fetaya, E. Multi Task Inverse Reinforcement Learning for Common Sense Reward. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025; pp. 1–13.
17. Tang, Q.H.; Guo, H.Y.; Chen, Q.X. Bidding Strategy Evolution Analysis Based on Multi-task Inverse Reinforcement Learning. *Electr. Power Syst. Res.* **2022**, *212*, 108286.
18. Yang, B.; Lu, Y.N.; Wan, R.; Hu, H.Y.; Yang, C.C.; Ni, R.R. Meta-IRLSOT++: A meta-Inverse Reinforcement Learning Method for Fast Adaptation of Trajectory Prediction Networks. *Expert Syst. Appl.* **2024**, *240*, 122499.
19. Chen, J.Y.; Tamboli, D.; Lan, T.; Aggarwal, V. Multi-task Hierarchical Adversarial Inverse Reinforcement Learning. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 4895–4920.
20. Chen, Y.; Lin, X.; Yan, B.; Zhang, L.B.; Liu, J.M.; Tan N.O.; Witbrock, M. Meta-Inverse Reinforcement Learning for Mean Field Games via Probabilistic Context Variables. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 11407–11415.
21. Ghasemipour, S.K.S.; Gu, S.X.; Zemel, R. SMILe: Scalable Meta Inverse Reinforcement Learning through Context-conditional Policies. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1–11.
22. Yu, L.T.; Yu, T.H.; Finn, C.; Ermon S. Meta-Inverse Reinforcement Learning with Probabilistic Context Variables. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 11772–11783.
23. Yoo, S.W.; Seo, S.W. Learning Multi-Task Transferable Rewards via Variational Inverse Reinforcement Learning. In Proceedings of the 2022 International Conference on Robotics and Automation, Pennsylvania, PA, USA, 23–27 May 2022; pp. 434–440.
24. Gleave, A.; Habryka, O. Multi-task Maximum Entropy Inverse Reinforcement Learning. *arXiv* **2018**, arXiv:1805.08882.
25. Baert, M.; Mazzaglia, P.; Leroux, S.; Simoens, P. Maximum Causal Entropy Inverse Constrained Reinforcement Learning. *Mach. Learn.* **2025**, *114*, 103.
26. Cheng, G.R.; Dong, L.; Cai, W.Z.; Sun C.Y. Multi-Task Reinforcement Learning with Attention-Based Mixture of Experts. *IEEE Robot Autom. Lett.* **2023**, *8*, 3812–3819.
27. Vazquez-Chanlatte, M.; Seshia, S.A. Maximum Causal Entropy Specification Inference from Demonstrations. *arXiv* **2020**, arXiv:1907.11792.
28. Nishi, K.; Shimosaka, M. Fine-Grained Driving Behavior Prediction via Context-Aware Multi-Task Inverse Reinforcement Learning. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation, Paris, France, 31 May–4 June 2020; pp. 2281–2287.

29. Song, L.; Li, D.Z.; Xu, X. Adaptive Generative Adversarial Maximum Entropy Inverse Reinforcement Learning. *Inform. Sci.* **2025**, *695*, 121712.
30. Wu, K.Y.; Wu, F.G.; Lin, Y.J.; Zhao, J.S. Stable Control Policy and Transferable Reward Function via Inverse Reinforcement Learning. In Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence, Tianjin, China, 17–20 March 2023; pp. 733–742.
31. Finn, C.; Levine, S.; Abbeel, P. Guided cost learning: Deep Inverse Optimal Control via Policy Optimization. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 49–58.
32. Venuto, D.; Chakravorty, J.; Boussiou, L.; Wang, J.H.; McCracken, G.; Precup, D. Oirl: Robust Adversarial Inverse Reinforcement Learning with Temporally Extended Actions. *arXiv* **2002**, arXiv:2002.09043.
33. Wang, P.; Liu, D.P.; Chen, J.Y.; Li, H.H.; Chan, C.Y. Decision Making for Autonomous Driving via Augmented Adversarial Inverse Reinforcement Learning. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 1036–1042.
34. Sun, J.K.; Yu, L.T.; Dong, P.Q.; Lu, B.; Zhou B.L. Adversarial Inverse Reinforcement Learning with Self-attention Dynamics Model. *IEEE Rob. Autom. Lett.* **2021**, *6*, 1880–1886.
35. Sestini, A.; Kuhnle, A.; Bagdanov, A.D. Demonstration Efficient Inverse Reinforcement Learning in Procedurally Generated Environments. *arXiv* **2020**, arXiv:2012.02527.
36. Yang, Y.D.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.N.; Wang, J. Mean Field Multi-Agent Reinforcement Learning. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5571–5580.
37. Wang, X.Q.; Ke, L.J.; Zhang, G.W.; Zhu, D.P. Adaptive Mean Field Multi-agent Reinforcement Learning. *Inform. Sci.* **2024**, *669*, 120560.
38. Yu C. Hierarchical Mean-Field Deep Reinforcement Learning for Large-Scale Multiagent Systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February, 2023; pp. 11744–11752.
39. Hao, Q.Y.; Huang, W.Z.; Feng, T.; Yuan, J.; Li, Y. GAT-MF: Graph Attention Mean Field for Very Large Scale Multi-Agent Reinforcement Learning. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 6–10 August 2023; pp. 685–697.
40. Chen, Y.; Zhang, L.B.; Liu, J.M.; Witbrock, M. Adversarial Inverse Reinforcement Learning for MeanField Games. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, London, UK, 29 May–2 June 2023; pp. 1088–1096.
41. Wu, H.C.; Sequeira, P.; Pynadath, D.V. Multiagent Inverse Reinforcement Learning via Theory of Mind Reasoning. In Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, London, UK, 29 May–2 June 2023; pp. 708–716.
42. Wei, R.; Zeng, S.; Li, C.L.; Garcia, A.; McDonald, A.; Hong, M.Y. A Bayesian Approach to Robust Inverse Reinforcement Learning. In Proceedings of the 7th Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023; pp. 1–19.
43. Oguntola I.; Campbell, J.; Stepputtis, S.; Sycara, K. Theory of Mind as Intrinsic Motivation for Multi-Agent Reinforcement Learning. In Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, HI, USA, 10 July 2023; pp. 1–8.
44. Wang, Y.F.; Zhong, F.W.; Xu, J.; Wang, Y.Z. Tom2c: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022; pp. 1–17.
45. Shi, H.B.; Li, J.C.; Chen S.C.; Hwang, K.S. A behavior Fusion Method based on Inverse Reinforcement Learning. *Inform. Sci.* **2022**, *609*, 429444.
46. Alsaleh, R.; Sayed, T. Markov-game Modeling of Cyclist-pedestrian Interactions in Shared Spaces: A Multi-agent Adversarial Inverse Reinforcement Learning Approach. *Transp. Res. Part C Emerg. Technol.* **2021**, *128*, 103191.
47. Chen, Y.; Zhang, L.B.; Liu, J.M.; Hu, S.Y. Individual-Level Inverse Reinforcement Learning for MeanField Games. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, 9–16 May 2022; pp. 253–262.
48. Li, Y.Z.; Song, J.M.; Ermon, S. InfoGAIL: Interpretable Imitation Learning From Visual Demonstrations. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 3–9 December 2017; pp. 3815–3825.
49. Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; Levine, S. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *arXiv* **2021**, arXiv:2004.07219.
50. Fan, S. Xinmin Zhang, Zhihuan Song. Imbalanced Sample Selection with Deep Reinforcement Learning for Fault Diagnosis. *IEEE Trans. Ind. Inf.* **2022**, *18*, 2518–2527.

51. Mismar, F.B.; Evans, B.L. Deep Q-Learning for Self-organizing Networks Fault Management and Radio Performance Improvement. In Proceedings of the 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 28–31 October 2018; pp. 1457–1461.
52. Zhou, H.; Aral, A.; Brandić, I.; Erol-Kantarç, M. Multiagent Bayesian Deep Reinforcement Learning for Microgrid Energy Management under Communication Failures. *IEEE Internet Things J.* **2022**, *9*, 11685–11698.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.