Robust Multi-Agent Reinforcement Learning by Mutual Information Regularization

Simin Li, Ruixiao Xu, Jingqiao Xiu, Yuwei Zheng, Pu Feng, Yuqing Ma, Bo An, *Senior Member, IEEE*, Yaodong Yang, *Member, IEEE*, and Xianglong Liu, *Senior Member, IEEE*

***Abstract*—In cooperative multi-agent reinforcement learning (MARL), ensuring robustness against cooperative agents making unpredictable or worst-case adversarial actions is crucial for real-world deployment. In multi-agent settings, each agent may be perturbed or unperturbed, leading to an exponential increase in potential threat scenarios as the number of agents grows. Existing robust MARL methods either enumerate, or approximate all possible threat scenarios, leading to intense computation and insufficient robustness. In contrast, humans develop robust behaviors by maintaining a general level of caution rather than preparing for every possible threat. Inspired by human decision making, we frame robust MARL as a control-as-inference problem, and optimize worst-case robustness across all threat scenarios implicitly optimized through off-policy evaluation. Specifically, we introduce mutual information regularization as robust regularization (MIR3), which maximizes a lower bound on robustness during routine training, serving as a kind of caution for MARL without adversarial inputs. Further insights show that MIR3 acts as an information bottleneck, preventing agents from over-reacting to others and aligning policies with robust action priors. In the presence of worst-case adversaries, our MIR3 significantly surpasses baseline methods in robustness and training efficiency, and maintaining cooperative performance in StarCraft II, quadrotor swarm control, and robot swarm control. When deploying the robot swarm control algorithm in the real world, our method also outperforms the best baseline by 14.29% in reward. See code and demo videos at https://github.com/DIG-Beihang/MIR3**

Received 9 November 2024; revised 17 April 2025; accepted 4 June 2025. Date of publication 10 July 2025; date of current version 9 October 2025. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116405, in part by the State Key Laboratory of Complex and Critical Software Environment (CCSE), and in part by the Fundamental Research Funds for the Central universities. *(Corresponding author: Xianglong Liu.)*

Simin Li is with the State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China, and also with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798.

Ruixiao Xu, Yuwei Zheng, Pu Feng, and Yuqing Ma are with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China.

Jingqiao Xiu is with the School of Computing, National University of Singapore, Singapore 117417.

Bo An is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798.

Yaodong Yang is with the Institute of Artificial Intelligence, Peking University, Beijing 100871, China.

Xianglong Liu is with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China, also with Zhongguancun Laboratory, Beijing 100086, China, and also with the Institute of Data Space, Hefei Comprehensive National Science Center, Hefei 230031, China (e-mail: xlliu@buaa.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3577259, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3577259

***Index Terms*—Adversarial machine learning, deep reinforcement learning, robust control.**

I. Introduction

COOPERATIVE multi-agent reinforcement learning (MARL) [1], [2], [3], [4], [5], [6], [7], [8] has shown significant progress across various challenging scenarios, such as StarCraft [9]. In real-world applications, however, MARL algorithms often fall short when the actions of cooperative agents deviate from their intended policies due to numerous uncertainties during deployment. In such cases, cooperative agents may exhibit unpredictable behavior or even perform worst-case actions if being hacked by adversaries, [10], [11], [12], [13], [14], [15]. This vulnerability greatly limits the practical applicability of MARL in real-world scenarios, such as robot swarm control [16].

Research on robust MARL against action uncertainties primarily focuses on max-min optimization against worst-case adversaries [10], [11], [17], [18], [19]. This approach can be framed as a zero-sum game [17], [20], where defenders with fixed parameters during deployment aim to maximize performance despite unknown proportions of adversaries employing the worst-case, nonoblivious adversarial policies [12], [14]. However, in multi-agent scenario, each agent can be either perturbed or unperturbed, leading to an exponential increase in the number of potential threat scenarios, making max-min optimization against each threat intractable. To address this complexity, some methods [10], [11], [21] approximate the problem by treating all agents as adversaries. However, since not all agents are perturbed in reality, the learned policy can be overly pessimistic, making agents not cooperate at all. Others attempt to enumerate all threat scenarios [18], [19], [22], but often struggle to explore each threat scenario sufficiently during training, leaving defenders still vulnerable to worst-case adversaries. Consequently, max-min optimization provides limited defense capabilities in MARL and incurs high computational cost [23].

In daily life, humans make robust decisions without explicitly considering every possible threat, as in max-min optimization. This ability is explained by the theory of situational awareness [24], [25], where individuals maintain a general level of risk awareness and adaptively respond to a range of unforeseen threats. For example, while driving, people do not assess every possible scenario involving aggressive drivers, as max-min optimization would require. Instead, they maintain a general sense of caution, such as keeping a safe distance from other vehicles, to mitigate unexpected risks like sudden braking.

2162-237X © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

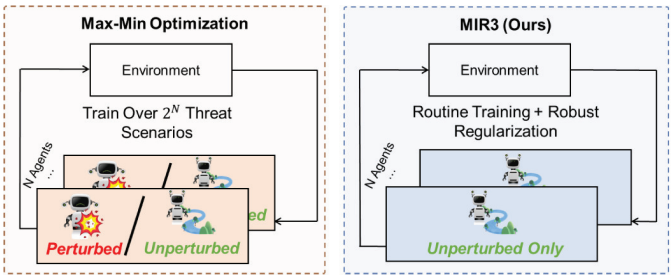

Fig. 1. Our policies are learned under routine scenarios but are provably robust against unseen worst-case adversaries through robust regularization, without experiencing all possible threat scenarios like existing approaches.

Inspired by the robust decision making process of human, we propose mutual information regularization as robust regularization (MIR3) for robust MARL. As illustrated in Fig. 1, our approach does not require exposure to all potential adversarial scenarios; instead, it trains policies in standard environments while ensuring provable robustness against unseen worst-case adversaries. Theoretically, we formulate this objective as a control-as-inference problem [26], a framework to derive optimal policies via probabilistic inference. Our objective is to optimize cooperative performance, and additionally maximize robustness across exponentially many threat scenarios via off-policy evaluation [27]. Within this framework, we show that, under specific conditions, regularizing the mutual information between histories and actions maximize a lower bound of our objective, enhancing robustness across all threat scenarios. This process serves as a general safeguard in MARL, akin to human caution in the face of diverse, unforeseen threats, without needing to model specific adversaries.

Beyond theoretical insights, MIR3 can be treated as an information bottleneck [28] or as learning a task-relevant robust action prior [29]. From the information bottleneck perspective, our goal is to learn a policy that solves the task using minimum sufficient information of current history. Thus, it suppresses false correlations in the policy created by action uncertainties and minimizes agents' overreactions to adversaries, fostering robust agent-wise interactions. From the view of robust action prior, we limit the policy from deviating from a prior action distribution which is not only generally favored by the task, but also maintains intricate tactics under attack. Experiments in StarCraft II, quadrotor swarm control, and rendezvous environments show MIR3 demonstrates higher robustness against worst-case adversaries on MADDPG, QMIX, and MAPPO backbones. When the magnitude of regularization is properly chosen, we find suppressing mutual information will not negatively affect cooperative performance, but even slightly enhance it. Finally, the superiority of MIR3 remains consistent when deployed in the real-world robot swarm control scenario, outperforming the best performing baseline by 14.29% in reward.

Our contributions can be summarized as follows.

1) Inspired by human caution, we propose MIR3 that serves as caution for MARL against diverse threat scenarios without adversarial input.
2) We theoretically frame robust MARL as a control-as-inference problem and optimize robustness via off-policy evaluation. In this framework, we prove that our MIR3 maximizes a lower bound of robustness reducing spurious correlations, and learning robust action prior.
3) Experiments on StarCraft, quadrotor swarm control, and robot swarm control show that our MIR3 surpasses baselines in robustness, while maintaining cooperative performance on MADDPG, QMIX and MAPPO backbones. This superiority is consistent when deploying the algorithm in real world.

II. Preliminaries

A. Cooperative MARL as Dec-POMDP

We formulate the problem of cooperative MARL as a decentralized partially observable Markov decision process (Dec-POMDP) [30], defined as a tuple

$$\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \mathcal{O}, O, \mathcal{A}, \mathcal{P}, R, \gamma \rangle. \tag{1}$$

Here, $\mathcal{N} = \{1, \ldots, N\}$ is the set containing N agents, $\mathcal{S}$ is the global state space, $\mathcal{O} = \times_{i \in \mathcal{N}} \mathcal{O}^i$ is the observation space, O is the observation emission function, $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}^i$ is the joint action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the shared reward function for cooperative agents, and $\gamma \in [0, 1)$ is the discount factor.

At each timestep, agent i observes $o_t^i = O(s_t, i)$ and add it to history $h_t^i = [o_0^i, a_0^i, \ldots, o_t^i]$ to alleviate partial observability issue [2], [30]. Then, it takes action $a_t^i \in \mathcal{A}^i$ using policy $\pi^i(a_t^i|h_t^i)$. The joint actions $\mathbf{a}_t$ leads to the next state s_{t+1} following state transition probability $P(s_{t+1}|s_t, \mathbf{a}_t)$ and shared global reward $r_t = R(s_t, \mathbf{a}_t)$. The objective for agents is to learn a joint policy $\pi(\mathbf{a}_t|\mathbf{h}_t) = \prod_{i \in \mathcal{N}} \pi^i(a_t^i|h_t^i)$ that maximize the value function $V_\pi(s) = \mathbb{E}_{s,\mathbf{a}} \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \mathbf{a}_t \sim \pi(\cdot|\mathbf{h}_t) \right]$. Many algorithms are proposed to solve MARL. Conventional methods includes MADDPG [1], QMIX [2], and MAPPO [3], while recent works focus on enhancing exploration [31], [32], interpretability [33], and policy transfer [34].

B. Robust MARL

Robust MARL aims to fortify against uncertainties in actions [10], states [35], [36], and rewards/environment [17], [37], [38]. Among these factors, action robustness have become a main focus due to the propensity for multiple agents to act unpredictably during deployment. Algorithms such as M3DDPG [10] and ROMAX [11] treat each agent as an adversary that deviates toward jointly worst-case actions [12]. However, in real world, since not all other agents are adversaries, such a policy can likely be overly pessimistic and make agents not cooperate at all. Later approaches attempt to directly train policies against these worst-case adversaries [18], [19], [22], [39]. However, as these methods must explore numerous distinct adversarial scenarios, each scenario may left insufficiently examined. As a consequence, attackers can be less powerful comparing with worst-case adversary, and defenders trained with such weaker attackers can still be vulnerable to worst-case adversaries at test time.

C. Robustness Without an Adversary

While it is tempting to directly train MARL policy against adversaries via max-min optimization, such process can be

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

overly pessimistic [10], unbalanced across threat scenarios [18], [19], and computationally demanding [23]. A parallel line of research in RL aims to achieve robustness without relying on adversaries. A2PD [40] shows a certain modification of policy distillation can be inherently robust against state adversaries. Through the use of convex conjugate, [41] has shown that max-entropy RL can be provably robust against uncertainty in reward and environment transitions. Derman et al. [23] further extended regularization to uncertainties in reward and transition dynamics under rectangular and ball constraints. The work most similar to ours is ERNIE [21], which minimize the Lipshitz constant of value function under worst-case perturbations in MARL. However, the method considers all agents as potential adversaries, thus inherits the drawback of M3DDPG, learning policy that can either be pessimistic or insufficiently robust.

D. Control-as-Inference Theory

Proposed by Levine [26], control-as-inference theory provides a principled way to infer the optimal decision policy via probabilistic inference. In RL, let $s \in \mathcal{S}$ be the states, $a \in \mathcal{A}$ be the actions, $p(s_{t+1}|s_t, a_t)$ be the environment dynamics, and $r_t = r(s_t, a_t)$ be the reward. Then, given a trajectory $\tau = [(s_1, a_1), \ldots, (s_t, a_t)]$, the trajectory distribution can be defined as

$$p(\tau) = p(s_1) \prod_{t=0}^{T} \left[\mathcal{P}(s_{t+1}|s_t, a_t)\pi(a_t|s_t)\right]. \tag{2}$$

Given trajectory distribution, Levine [26] propose a binary random variable $\mathcal{O}_t$ to denote if the policy at current timestep is optimal, $\mathcal{O}_t = 1$ and $\mathcal{O}_t = 0$ otherwise. The distribution of over $\mathcal{O}_t$ is given by the reward of current timestep, which ensures better state-action pairs are favored exponentially

$$p(\mathcal{O}_t = 1|s_t, a_t) = \exp(r_t). \tag{3}$$

Given our objective is to achieve optimal control $\mathcal{O}_t = 1$ at all timesteps $t \in \{1, \ldots, T\}$, we can write the probability of the optimal trajectory $p(\tau|o_{1:T})$ as

$$\begin{aligned} p(\tau|o_{1:T}) &\propto p(\tau, o_{1:T}) \\ &= p(s_1) \prod_{t=1}^{T} p(\mathcal{O}_t = 1|s_t, a_t) p(s_{t+1}|s_t, a_t) \\ &= \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1}|s_t, a_t)\right] \exp\left(\sum_{t=1}^{T} r_t\right). \end{aligned} \tag{4}$$

Given such formulation, the optimal policy trajectory can be solved by minimizing the KL divergence between policy trajectory $p(\tau)$ and the optimal trajectory $p(\tau|o_{1:T})$

$$\begin{aligned} &- D_{\mathrm{KL}}(p(\tau)\|p(\tau|o_{1:T})) \\ &= \mathbb{E}_{\tau\sim p(\tau)} \Bigg[\log p(s_1) + \sum_{t=1}^{T} (\log p(s_{t+1}|s_t, a_t) + r_t) \\ &\qquad - \log p(s_1) - \sum_{t=1}^{T} \big(\log p(s_{t+1}|s_t, a_t) \\ &\qquad + \log \pi(a_t, s_t)\big)\Bigg] \\ &= \mathbb{E}_{\tau\sim p(\tau)} \left[\sum_{t=1}^{T} r_t - \log \pi(a_t|s_t)\right] \\ &= \mathbb{E}_{\tau\sim p(\tau)}[r_t] + \mathbb{E}_{s_t\sim p(s_t)}[\mathcal{H}(a_t|s_t)] \end{aligned} \tag{5}$$

where $\mathcal{H}(a_t|s_t)$ is the entropy of the policy. As such, control-as-inference theory offers a principled explanation for RL policies that additionally maximize policy entropy [3] to encourage exploration. In our setting, we adopt the control-as-inference theory, but additionally consider adversarial transitions to get mutual information as a robust regularizer.

E. Mutual Information Estimation

Mutual information quantifies the dependency between two random variables. Formally, the mutual information between variables x and y is defined as

$$I(x; y) = \mathbb{E}_{p(x,y)} \left[\log \frac{p(x, y)}{p(x)p(y)}\right]. \tag{6}$$

However, when the joint distribution $p(x, y)$, is unknown, computing the exact value of mutual information becomes challenging, as it requires accurate estimation or sampling from the joint distribution. This difficulty is further exacerbated when the variables are high-dimensional. Since exact computation is generally infeasible in practice, a range of techniques have been proposed to approximate mutual information by estimating its upper and lower bounds. In this article, we first give a broad overview of the literature of mutual information estimation, then introduce several commonly used mutual information estimation methods as practical tools for approximation.

Early MI estimation techniques relied on histogram binning, k-nearest neighbors, or kernel density estimation [42], [43], [44], [45]. While theoretically grounded, these methods scale poorly to high-dimensional data and often suffer from bias-variance trade-offs that are difficult to control in practice. To address these limitations, neural MI estimators have gained popularity. These methods approximate MI through variational lower or upper bounds, parameterized by neural networks, and optimized using stochastic gradient descent. The key idea is to cast MI as a functional of the joint and marginal distributions, enabling learning-based approximations from samples alone. MINE [46] propose the first neural estimator for mutual information via the dual formulation of KL divergence using Donsker-Varadhan representation

$$I(x; y) \geq \sup_{T\in\mathcal{F}} \mathbb{E}_{p(x,y)} T(x, y) - \log \mathbb{E}_{p(x)p(y)}[\exp(T(x, y))] \tag{7}$$

where T is a function parameterized by a neural network. While MINE estimate the lower bound of mutual information, they are unable to estimate the upper bound of mutual information. Researchers first estimate the upper bound of mutual information in information bottleneck theory by defining a variational upper bound (VUB) [47]

$$I(x; y) \leq \mathbb{E}_{p(x,y)} \left[\frac{\log T(y|x)}{\mathcal{N}(y|0, I)}\right] \tag{8}$$

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

with $\mathcal{N}(y|0, I)$ a Gaussian distribution. Subsequent work replace $\mathcal{N}(y|0, I)$ in a leave-one-out manner (L1Out) [48]

$$I(x; y) \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\frac{T(y_i|x_i)}{1/N-1\sum_{j\neq i}T(y_i|x_j)}\right]\right]. \quad (9)$$

A recent advance made by CLUB [49] provides an tighter upper bound estimate of mutual information via variational inference

$$I(x; y) \leq \mathbb{E}_{p(x,y)}\left[\log T(y|x)\right] - \mathbb{E}_{p(x)p(y')}\left[\log T(y'|x)\right] \quad (10)$$

with y' the negative samples. CLUB gains wide attention as a simple yet effective method for general-purposed information upper bound estimation.

Since mutual information (MI) captures agent correlations, many MARL methods use MI regularization to promote collaboration. Existing approaches can be grouped into three categories: 1) social influence [50] and EITI [51] maximize MI between pairs of agents to enhance mutual influence; 2) MAVEN [52], SIC [53], and VM3-ac [54] maximize MI between each agent and a shared latent variable to improve coordination; and 3) PMIC [55] maximizes MI between states and joint actions to promote diverse yet predictable behaviors. However, these methods do not account for robustness against action perturbations—higher coordination in such cases may amplify cascading failures when individual agents falter.

III. Method

In this section, drawing inspiration from human cautious to unseen threats [24], we first formalize robust MARL as an action adversarial Dec-POMDP, aiming to maximize both cooperative and robust performance under all threat scenarios. Next, framed as an control-as-inference problem [26], policies are learned without attacks and adapt to worst-case scenarios using off-policy evaluation. We find minimizing mutual information between histories and actions maximizes lower bound of robustness, serving as cautious for MARL. Beyond theoretical insights, our method acts as an information bottleneck, reducing spurious correlations and learning robust action priors to maintain effective tactics under attack.

A. Problem Formulation

1) Action Adversarial Dec-POMDP: In this article, we consider action uncertainty as an unknown portion of agents taking unexpected actions. This can stem from robots losing control due to software/hardware error, or are compromised by an adversary [12], [17], [18], [19], [22]. Given a Dec-POMDP with action uncertainties, we define action uncertainties in MARL as an action adversarial Dec-POMDP (A2Dec-POMDP), which is written as

$$\hat{\mathcal{G}} = \langle \mathcal{N}, \Phi, \mathcal{S}, \mathcal{O}, O, \mathcal{A}, \mathcal{P}, R, \gamma \rangle. \quad (11)$$

Here, $\Phi = \{0, 1\}^N$ is a set containing *partitions* of agents into defenders and adversaries, with $\phi \in \Phi$ indicates a specific partition. For each agent i, $\phi^i = 1$ means the original policy of $\pi^i(\cdot|h_t^i)$ is replaced by a worst-case adversarial policy $\pi_\alpha^i(\cdot|h_t^i, \phi)$, while $\phi^i = 0$ means that the original policy is executed without change. We use α to denote adversarial policy throughout this article. In this way, Dec-POMDP is a special case of A2Dec-POMDP with $\phi = \mathbf{0}_N$.

2) Perturbed Policy: The perturbed joint policy is defined as $\hat{\pi}(\hat{\mathbf{a}}_t|\mathbf{h}_t, \phi) = \prod_{i\in\mathcal{N}}[\pi_\alpha^i(\cdot|h_t^i, \phi) \cdot \phi + \pi^i(\cdot|h_t^i) \cdot (1 - \phi)]$, with perturbed joint actions $\hat{\mathbf{a}}_t$ used for environment transition $\mathcal{P}(s_{t+1}|s_t, \hat{\mathbf{a}}_t)$, and reward $r_t = R(s_t, \hat{\mathbf{a}}_t)$. For each $\phi \in \Phi$, the value function is $V_{\pi,\pi_\alpha}(s, \phi) = V_{\hat{\pi}}(s, \phi) = \mathbb{E}_{s,\hat{\mathbf{a}}}\left[\sum_{t=0}^{\infty}\gamma^t r_t|s_0 = s, \hat{\mathbf{a}}_t \sim \hat{\pi}(\cdot|\mathbf{h}_t, \phi)\right]$.

3) Attacker's Objective: We assume the attack happens at test time, with parameters in defender's policy π fixed during deployment. Following the setting of [19], we assume the attacker has the same partial observation and same action space as defender. For a partition ϕ that indicates defenders and adversaries, the objective of a worst-case, zero-sum adversary aims to learn a joint adversarial policy $\pi_\alpha(\cdot|\mathbf{h}_t, \phi) = \prod_{i\in\{\phi^i=1\}}\pi_\alpha^i(\cdot|h_t^i, \phi) \in \times_{i\in\phi^i=1}\{\mathcal{A}^i\}$ that minimize cumulative reward [12]

$$\pi_\alpha^* \in \arg\min_{\pi_\alpha} V_{\pi,\pi_\alpha}(s, \phi). \quad (12)$$

Following [12], an optimal worst-case adversarial policy π_α^* always exists for all possible partitions $\phi \in \Phi$ and fixed π. Since the defender's policy π is held fixed during attack, we can view it as a part of environment transition, reducing the problem to a POMDP for one adversary or a Dec-POMDP for multiple adversaries. The existence of an optimal π_α^* is then a corollary of the existence of an optimal policy in POMDP [56] and Dec-POMDP [30]. In our experiments, the attacker always employs an optimal worst-case adversarial policy. This is achieved by first fixing the defender's policy and then training the adversary specifically to exploit it.

4) Defender's Objective: The objective of defenders is to learn a policy that maximize both normal performance and robust performance under attack, without knowing who is the adversary

$$\pi^* \in \arg\max_{\pi}\left[V_\pi(s) + \mathbb{E}_{\phi\sim p(\Phi^\alpha)}\left[\min_{\pi_\alpha} V_{\pi,\pi_\alpha}(s, \phi)\right]\right]. \quad (13)$$

We use $p(\Phi^\alpha)$ to denote the *distribution* of the partitions the defenders are facing. While existing max-min approaches requires explicitly training π with distribution $\phi \sim p(\Phi^\alpha)$, our method trains π with partition $\phi = \mathbf{0}_N$ only, but still capable of solving the max-min objective in (13). This is done by deriving a lower bound for objective $\min_{\pi_\alpha}\mathbb{E}_{\phi\sim p(\Phi^\alpha)}[V_{\pi,\pi_\alpha}(s, \phi)]$ as a regularization term.

B. Theoretical Insights on Robustness of MIR3

We adopt a control-as-inference approach [26] to infer the defender's policy π. We first derive objectives for purely cooperative scenarios, then show objectives under attack by importance sampling. Let $\tau^0 = [(s_1, \mathbf{a}_1), (s_2, \mathbf{a}_2), \ldots, (s_t, \mathbf{a}_t)]$ denote the *optimal* trajectory of purely cooperative scenario generated on t consecutive stages, with superscript in τ^0 denotes $\phi = \mathbf{0}_N$. Following [26], the probability of τ being generated is:

$$p(\tau^0) = \left[p(s_1)\prod_{t=0}^{T}\mathcal{P}(s_{t+1}|s_t, \mathbf{a}_t)\right]\exp\left(\sum_{t=1}^{T} r_t\right) \quad (14)$$

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

with $\exp\left(\sum_{t=1}^{T} r_t\right)$ encourage trajectories with higher rewards to have exponentially higher probability [26]. The goal is to find the best approximation of joint policies $\pi(\mathbf{a}_t|\mathbf{h}_t) = \prod_{i\in\mathcal{N}} \pi^i(a_t^i|h_t^i)$, such that its induced trajectories $\hat{p}(\tau^0)$ match the optimal probability of $p(\tau^0)$

$$\hat{p}(\tau^0) = p(s_1)\left[\prod_{t=0}^{T} \mathcal{P}(s_{t+1}|s_t, \mathbf{a}_t)\pi(\mathbf{a}_t|\mathbf{h}_t)\right]. \tag{15}$$

Assume the dynamics is fixed, such that agents cannot influence the environment transition probability [57], the objective for purely cooperative scenario is derived as maximizing the negative of KL divergence between sampled trajectory $\hat{p}(\tau^0)$ and optimal trajectory $p(\tau^0)$

$$\begin{aligned} J^0(\pi) &= -D_{\mathrm{KL}}(\hat{p}(\tau^0)\|p(\tau^0)) \\ &= \textstyle\sum_{t=1}^{T} \mathbb{E}_{\tau^0\sim\hat{p}(\tau^0)}[r_t + \mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] \end{aligned} \tag{16}$$

where $\mathcal{H}(\cdot)$ is the entropy estimator.

As for scenarios with attack, for partition $\phi \sim p(\Phi^\alpha)$, let $\tau^\phi = [(s_1, \hat{\mathbf{a}}_1), (s_2, \hat{\mathbf{a}}_2), \ldots, (s_t, \hat{\mathbf{a}}_t)]$ denote the trajectories under attack. To evaluate the performance of π with partition ϕ, we can leverage importance sampling to derive an unbiased estimator $J^\phi(\pi)$ using τ^0, with $\rho_t = (\hat{\pi}(\mathbf{a}_t|\mathbf{h}_t, \phi)/\pi(\mathbf{a}_t|\mathbf{h}_t))$ the per-step importance sampling ratio

$$\begin{aligned} J^\phi(\pi) &= \textstyle\sum_{t=0}^{T} \mathbb{E}_{\tau^0\sim\hat{p}(\tau^0)}\left[\rho_t \cdot (r_t + \mathcal{H}(\mathbf{a}_t|\mathbf{h}_t))\right] \\ &= \textstyle\sum_{t=0}^{T} \mathbb{E}_{\tau^\phi\sim\hat{p}(\tau^\phi)}\left[r_t + \mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)\right]. \end{aligned} \tag{17}$$

By (13), the overall objective $J(\pi)$ for inference is

$$\begin{aligned} J(\pi) &= J^0(\pi) + \mathbb{E}_{\phi\sim p(\Phi^\alpha)}\left[\min_{\pi_\alpha} J^\phi(\pi)\right] \\ &= \textstyle\sum_{t=0}^{T} \mathbb{E}_{\tau^0\sim\hat{p}(\tau^0)}[r_t + \mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \\ &\quad \times \left[\min_{\pi_\alpha} \textstyle\sum_{t=0}^{T} \mathbb{E}_{\tau^\phi\sim\hat{p}(\tau^\phi)}\left[r_t + \mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)\right]\right]. \end{aligned} \tag{18}$$

Thus, our objective maximize cumulative reward in both cooperative task and across the distribution of defender-adversary partitions (i.e., threat scenarios) via off-policy evaluation. To do this, we assume the trajectories $p(\tau)$ under $p(\Phi^\alpha)$ follows the uniform coverage assumption [58], [59], [60], as commonly adopted by offline RL. Below are our main theoretical insights:

Proposition 1: $J(\pi) \geq \sum_{t=1}^{T} \mathbb{E}_{\tau^0\sim\hat{p}(\tau^0)}[r_t - \lambda I(\mathbf{h}_t; \mathbf{a}_t)]$, where $I(\mathbf{h}_t; \mathbf{a}_t)$ is the mutual information between joint histories and actions, and λ is a nonnegative hyperparameter to control the trade-off between cooperation and robustness.

Proof: [Proof sketch] The proof is constructed in three steps. First, since the defenders and adversary forms a zero-sum game, we show the log probability of optimal robust policy and optimal adversary differs by a constant. This allows us to transform attacker's policy to benign policy. Second, we derive a lower bound for all attack trajectories and partitions under the uniform coverage assumption. Third, we find the lower bound satisfy the definition of mutual information. □

Proof: Step 1: The first step we take from (18) is to transform the policy in adversarial trajectories $\pi(\mathbf{a}_t|\mathbf{h}_t)$ into $\hat{\pi}(\hat{\mathbf{a}}_t|\mathbf{h}_t, \phi)$, such that the policy meets the trajectory probability with adversary. Recall in probabilistic reinforcement learning [26], the optimal policy is defined via soft Bellman backup

$$\pi(a_t|s_t) = \frac{1}{Z}\exp(Q(s_t, a_t) - V(s_t)) \tag{19}$$

where Z is a normalizing constant. This is extended to MARL by marginalizing the actions of other agents [57]. In our case, we further add current partition ϕ to the objective, which is written as

$$\begin{aligned} \pi\left(a_t^i|h_t^i\right) &= \frac{1}{Z}\exp\left(Q\left(s_t, a_t^i, a_t^{-i}, \phi\right) - Q\left(s_t, a_t^{-i}, \phi\right)\right) \\ &= \frac{1}{Z}\exp\left(Q\left(s_t, a_t^i, a_t^{-i}, \phi\right)\right. \\ &\quad \left. - \log\int_{a_t^i} \exp\left(Q\left(s_t, a_t^i, a_t^{-i}, \phi\right)\right) \mathrm{d}a_t^i\right). \end{aligned} \tag{20}$$

Since the adversary is zero-sum, its objective is opposite to the objective of the defenders, which can be written as

$$\begin{aligned} \pi_\alpha\left(a_{t,\alpha}^i|h_t^i\right) &= \frac{1}{Z'}\exp\left(-Q\left(s_t, a_{t,\alpha}^i, a_t^{-i}, \phi\right) + Q\left(s_t, a_t^{-i}, \phi\right)\right) \\ &= \frac{1}{Z'}\exp\left(-Q\left(s_t, a_{t,\alpha}^i, a_t^{-i}, \phi\right) + \log\right. \\ &\quad \left. \times \int_{a_{t,\alpha}^i} \exp\left(Q\left(s_t, a_{t,\alpha}^i, a_t^{-i}, \phi\right)\right) \mathrm{d}a_{t,\alpha}^i\right). \end{aligned} \tag{21}$$

Next, we expand our objective in terms of history-action pairs, where history are added to meet the conditions of Dec-POMDP (i.e., policy always condition on current histories)

$$\begin{aligned} J(\pi) &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log\pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)}\left[\min_{\pi_\alpha} \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}\left[r_t - \log\pi(\mathbf{a}_t|\mathbf{h}_t)\right]\right]. \\ &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log\pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)}\left[\min_{\pi_\alpha} \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}\left[r_t\right.\right. \\ &\quad \left.\left. - \sum_{i=1}^{N} \log\pi^i(a_t^i|h_t^i)\right]\right]. \end{aligned} \tag{22}$$

Here, we cannot directly process the objective containing adversary since the trajectory is sampled using $\hat{\pi}(\hat{\mathbf{a}}_t|\mathbf{h}_t)$. However, since the objective of attacker is to minimize Q value, and the objective of defender is to maximize Q value, from the optimal policy defined in (20), we can compute the logarithm of the optimal robust policy $\pi(a_t^i|h_t^i)$ and optimal adversarial policy $\pi_\alpha(a_{t,\alpha}^i|h_t^i)$ as

$$\begin{aligned} \log\pi\left(a_t^i|h_t^i\right) &= -\log Z + Q\left(s_t, a_t^i, a_t^{-i}, \phi\right) \\ &\quad - \log\int_{a_t^i} \exp\left(Q\left(s_t, a_t^i, a_t^{-i}, \phi\right)\right) \mathrm{d}a_t^i \end{aligned} \tag{23}$$

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

for defenders and

$$\log \pi_\alpha \left(a^i_{t,\alpha}|h^i_t\right) = -\log Z' - Q\left(s_t, a^i_{t,\alpha}, a^{-i}_t, \phi\right) + \log \int_{a^i_{t,\alpha}} \exp\left(Q\left(s_t, a^i_{t,\alpha}, a^{-i}_t, \phi\right)\right) \mathrm{d}a^i_{t,\alpha} \quad (24)$$

for adversaries.

Thus, we have

$$\log \pi^\alpha \left(a^i_{t,\alpha}|h^i_t\right) = -\log \pi \left(a^i_t|h^i_t\right) + c \quad (25)$$

where $c = -\log Z + \log Z'$ is a constant. We ignore this in our subsequent derivations.

Plugging this into our objective, we get

$$\begin{aligned} J(\pi) &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \\ &\quad \left[\min_{\pi_\alpha} \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)} \left[r_t - \sum_{i=1}^{N} \log \pi^i(a^i_t|h^i_t)\right]\right] \\ &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)} \left[r_t - \sum_{i\in\{\phi^i=0\}} \log \pi(a_t|h^i_t) \right.\right. \\ &\quad \left.\left. + \sum_{i\in\{\phi^i=1\}} \log \pi_\alpha(a^i_{t,\alpha}|h^i_t, \phi)\right]\right]. \end{aligned} \quad (26)$$

Step 2: Next, we transform the objective containing adversarial rollouts into a regularization. Starting from our previous objective, we get

$$\begin{aligned} J(\pi) &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)} \left[r_t - \sum_{i\in\{\phi^i=0\}} \log \pi(a_t|h^i_t) \right.\right. \\ &\quad \left.\left. + \sum_{i\in\{\phi^i=1\}} \log \pi_\alpha(a^i_{t,\alpha}|h^i_t, \phi)\right]\right] \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)} \left[r_t + \sum_{i\in\{\phi^i=0\}} \log \pi(a_t|h^i_t) \right.\right. \\ &\quad \left.\left. + \sum_{i\in\{\phi^i=1\}} \log \pi_\alpha(a^i_{t,\alpha}|h^i_t, \phi)\right]\right] \\ &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}[r_t + \log \hat{\pi}(\hat{\mathbf{a}}_t|\mathbf{h}_t)]\right]. \end{aligned} \quad (27)$$

Plugging in the uniform coverage assumption [58], [59], [60], i.e., $\log p(\mathbf{h}_t) = 1/c$, we get

$$\begin{aligned} &J(\pi) \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t - \log \pi(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}[r_t + \log \hat{\pi}(\hat{\mathbf{a}}_t|\mathbf{h}_t)]\right] \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \\ &\quad \times \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}[r_t] + \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)}[\log \pi(\mathbf{a}_t|\mathbf{h}_t)]\right] \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \\ &\quad \times \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\hat{\mathbf{a}}_t\sim p(\tau^\phi)} \log \pi(\mathbf{a}_t|\mathbf{h}_t) + \log p(\mathbf{h}_t) - \log p(\mathbf{h}_t)\right] \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] + \mathbb{E}_{\phi\sim p(\Phi^\alpha)} \\ &\quad \times \left[\sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t\sim p(\tau^\phi)}[\log \pi(\mathbf{a}_t|\mathbf{h}_t) + \log p(\mathbf{h}_t)]\right] + c \\ &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad - \sum_{t=0}^{T} [\mathcal{H}(\pi(\mathbf{a}_t, \mathbf{h}_t))] + c \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] \\ &\quad - \sum_{t=0}^{T} \mathcal{H}(\pi(\mathbf{a}_t)) + \mathcal{H}(p(\mathbf{h}_t)) + c \\ &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t)] - \mathcal{H}(\pi(\mathbf{a}_t)). \end{aligned} \quad (28)$$

Step 3: Finally, from information theory, we have

$$I\left(\mathbf{h}_t; \mathbf{a}_t\right) = \mathcal{H}\left(\pi\left(\mathbf{a}_t\right)\right) - H\left(\mathbf{a}_t \mid \mathbf{h}_t\right) \quad (29)$$

plugging in our derivations above, we get

$$\begin{aligned} J(\pi) &\geq \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[\mathcal{H}(\mathbf{a}_t|\mathbf{h}_t) - \mathcal{H}(\pi(\mathbf{a}_t))] \\ &= \sum_{t=0}^{T} \mathbb{E}_{\mathbf{h}_t,\mathbf{a}_t\sim p(\tau^0)}[r_t] + \mathbb{E}_{\mathbf{h}_t\sim p(\tau^0)}[-I(\mathbf{h}_t; \mathbf{a}_t)] \\ &= \sum_{t=0}^{T} \mathbb{E}_{\tau^0\sim p(\tau^0)}[r_t - I(\mathbf{h}_t; \mathbf{a}_t)]. \end{aligned} \quad (30)$$

This completes the proof. □

Remark 1: In control-as-inference theory, λ is not needed in theoretical derivation since we can linearly scale the reward

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 MIR3 Defense With MADDPG Backbone

Input: Policy network of agents $\{\pi_1, \pi_2, \ldots \pi_N\}$, value function network $Q_i^{\pi}(s, a_1, \ldots, a_N)$, mutual information estimation network based on CLUB [49]: $CLUB(h_t^i, a_t^i)$, hyperparameter λ for mutual information regularization.

Output: Trained robust policy networks $\{\pi_1, \pi_2, \ldots \pi_N\}$.

```
for episode = 0, 1, 2,...K do
    Perform rollout using current policy, save trajectory in buffer D.
    Update CLUB(h_t^i; a_t^i) using D.
    I(h_t, a_t) ← Σ_{i∈N} CLUB(h_t^i, a_t^i).
    r_t^MI ← r_t − λ · I(h_t; a_t).
    Update critic {Q_i} of each agents using r_t^MI.
    Update parameters of each agents using MADDPG.
    // To implement MIR3 on other backbones,
    just change the way of parameter update.
end for
```

by $r' = r/\lambda$ and absorb λ in reward function. Here, we make it explicit to represent the trade-off between reward and mutual information. See Haarnoja et al. [61] for more details.

Remark 2: Minimizing the objective $I(\mathbf{h}_t; \mathbf{a}_t)$ and enhancing robustness can be explained as follows: when some agents fail due to uncertainties, their erroneous actions will alter the global state, affecting future observations and histories of other benign agents. Compared to the intuitive approach of minimizing the mutual information between agents' actions, our objective also accounts for environmental transitions under the control-as-inference framework.

Remark 3: We acknowledge that in real-world settings, the uniform coverage assumption might not hold. While undesirable, the assumption is indispensable in many offline RL articles [58], [59], [60]. Similar to their settings, we derive theoretical insights based on this assumption, and evaluated MIR3's performance under individual threat scenarios involving one or two adversaries, which violates the uniform coverage assumption. Despite this, MIR3 demonstrated improved robustness consistently. In the context of robustness, the assumption is in fact favorable, since it ensures all possible scenarios and trajectories are considered, which eliminates corner cases.

Finally, all we need is to add the mutual information between histories and joint actions $-\lambda I(\mathbf{h}_t; \mathbf{a}_t)$ as a robust regularization term to reward r_t. Since our MIR3 is only an additional reward, it can be optimized by *any* cooperative MARL algorithms. Technically, the exact value of $I(\mathbf{h}_t; \mathbf{a}_t)$ is intractable to calculate, so we estimate its upper bound as a lower bound for $-I(\mathbf{h}_t; \mathbf{a}_t)$. We use CLUB [49], [55], an off-the-shelf mutual information upper bound estimator, to estimate this information. The pseudo code of our MIR3 on MADDPG backbone is given in Algorithm 1. The algorithm first collects the history and actions of each agent, and estimate the upper bound of mutual information $I(h_t^i; a_t^i)$ by training the CLUB estimator. Next, we added the upper bound of mutual information $I(\mathbf{h}_t^i; \mathbf{a}_t^i)$ to reward, and update the policy using standard MADDPG algorithm. Notably, MIR3 is architecture-agnostic: to apply MIR3 to other MARL frameworks such as QMIX or MAPPO, one only needs to replace the MADDPG-specific policy update step with the corresponding optimization procedure from the desired backbone.

1) Convergence: Our MIR3 introduces mutual information as a regularization term directly into the reward function, without altering the policy space, the transition dynamics, or the observation structure of the original problem. As a result, the policy is still learned in a Dec-POMDP with a shaped reward. As a result, the convergence of our algorithm can be established via standard proofs of Q-learning [62].

Proposition 2: Define the Bellman equation used to update the value function as

$$\mathcal{B}V^i(s) = \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}|\mathbf{h}) \Bigg[r - I(\mathbf{h};\mathbf{a}) + \gamma \sum_{s'\in\mathcal{S}} p\left(s'^i|s^i,\mathbf{a}\right) V^i\left(s'^i\right) \Bigg].$$

With finite joint action space $\mathcal{A}$, state space $\mathcal{S}$, and assume each state-action pair is visited infinitesimally often, updating value function by Bellman operator $\mathcal{B}$ converge to the optimal value $V^*(s)$.

Proof: Since our mutual information estimation is not related to value function learning step, the mutual information term cancels out just like the reward term in standard proof of convergence of value function. Specifically, define $V_1^i, V_2^i \in \mathbb{R}^{|\mathcal{S}|}$, we have

$$\begin{aligned}
&\left|\mathcal{B}V_1^i(s) - \mathcal{B}V_2^i(s)\right| \\
&= \left| \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}|\mathbf{h}) \left[r - I(\mathbf{h};\mathbf{a}) + \gamma \sum_{s'\in\mathcal{S}} p(s'^i|s^i,\mathbf{a}) V_1^i(s'^i) \right] \right. \\
&\quad \left. - \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}|\mathbf{h}) \left[r - I(\mathbf{h};\mathbf{a}) + \gamma \sum_{s'\in\mathcal{S}} p(s'^i|s^i,\mathbf{a}) V_2^i(s'^i) \right] \right| \\
&= \left| \gamma \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}|\mathbf{h}) \sum_{s'\in\mathcal{S}} p(s'^i|s^i,\mathbf{a}) (V_1^i(s) - V_2^i(s)) \right| \\
&\le \gamma \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}|\mathbf{h}) \sum_{s'\in\mathcal{S}} p(s'^i|s^i,\mathbf{a}) \left| V_1^i(s) - V_2^i(s) \right| \\
&= \gamma \left| V_1^i(s) - V_2^i(s) \right|. \qquad (31)
\end{aligned}$$

Thus, $\mathcal{B}$ is a contraction operator. Finally, by Banach's fixed point theorem, with finite joint action space $\mathcal{A}$, state space $\mathcal{S}$, and assume each state-action pair is visited infinitesimally often, updating $V^i(s)$ by Bellman operator $\mathcal{B}$ will converge to the optimal value function $V^i(s)$. Note that the guaranteed convergence happens in tabular case. In practice, we use different MARL algorithms parameterized by nonconvex-nonconcave neural networks for better representation learning capabilities. □

2) Computational Complexity: We now analyze the computational complexity introduced by MIR3. During training, the primary overhead arises from the estimation of mutual information. To ensure scalability, we adopt a common practice in centralized training with decentralized execution (CTDE): we approximate the upper bound of $I(\mathbf{h}_t; \mathbf{a}_t)$ using the mutual information between the global state and the joint actions, $I(s_t; \mathbf{a}_t)$. This approximation remains an upper bound, as the

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

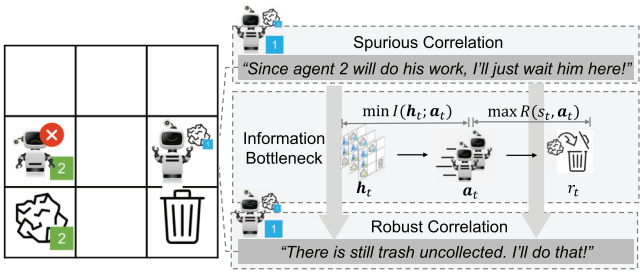

Fig. 2. MIR3 as an information bottleneck for robust multi-agent coordination. By minimizing mutual information between histories and actions, MIR3 reduces spurious agent-to-agent correlations. In the top example, Agent 1 waits based on an incorrect assumption about Agent 2's action, leading to failure. As illustrated below, MIR3 instead promotes robust, task-relevant decisions, where Agent 1 independently takes action when needed, improving reliability under uncertainty.

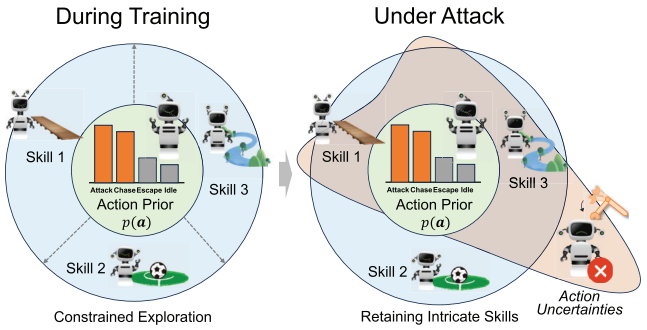

Fig. 3. MIR3 as a robust action prior. During training, MIR3 constrains exploration to a learned action prior $p(a)$, promoting task-relevant behaviors such as skill execution in diverse situations. Under attack, this prior anchors the agent's behavior in a reliable action distribution, allowing it to retain essential skills while remaining robust to action uncertainties.

global state s_t contains more information than the individual agent observation histories $\mathbf{h}_t$. This approximation enables MIR3 to estimate mutual information for the entire team with a single call to the mutual information estimator per training step, ensuring scalability to large multi-agent systems. In contrast, existing max-min optimization-based approaches [10], [11], [63] often require multiple gradient backward steps to compute worst-case joint actions or to optimize worst-case actions for each individual agent [18], resulting in significantly higher computational cost. At test time, MIR3 introduces no additional overhead, as all policies are executed using the same neural network architectures as in standard MARL algorithms. Thus, the runtime complexity of MIR3 matches that of the baseline policies during execution. Further empirical results on training time comparisons are reported in Section IV-B7.

C. Understandings and Discussions

Beyond theoretical insights, our MIR3 can be understood as an information bottleneck that reduce unnecessary correlations between agents, or as learning a robust action prior that favors effective actions in the environment. These discussions provide explanations for the success of our approach.

1) MIR3 as Information Bottleneck: Our mutual information minimization objective can be seen as an information bottleneck, which encourage policies to eliminate spurious correlations among agents. This concept, initially introduced by [28], seeks to identify a compressed representation that retains the maximum relevant information with the label [64], [65], [66]. In MARL, as depicted in Fig. 2, our objective $\max_\pi \mathbb{E}_{\tau^0 \sim p(\tau^0)}[r_t - \lambda I(\mathbf{h}_t; \mathbf{a}_t)]$ functions as an information bottleneck, considering history as input, actions as an intermediate representation, and reward as the final label. The aim is to find a set of actions employing *minimal sufficient information* from the current history, which is maximally relevant for solving the task and getting higher reward. This behavior can be further understood through the lens of the exploration-exploitation trade-off. On one hand, MIR3 encourages the policy to extract minimal information from history, promoting compact representations and thus favoring exploitation. On the other hand, MIR3 still requires the agent to achieve high rewards, which inherently demands sufficient exploration.

The objective is crucial for eliminating spurious correlation between agents, which helps handling action uncertainties in MARL. For example, robot swarms trained in simulation environment assumes each agent to be optimally cooperative to enable best performance. As shown in Fig. 2, this objective can form a spurious correlation that encourage robots to *overly* rely on others. In reality, individual robots can malfunction due to software/hardware errors, execute suboptimal actions, or send erroneous signals, which is reflected in histories. As such, information bottleneck encourage agents not to overly rely on current history, and form a loose cooperation with others only in case of need. Therefore, even if some agents falter, our objective enables the remaining agents to fulfill their tasks independently without overreacting or being swayed by failed agents.

2) MIR3 as a Robust Action Prior: Minimizing mutual information can be interpreted as establishing a robust action prior, which favors actions that are both useful for the current task and resilient under action uncertainties through exploration. In information theory, the relationship $-I(\mathbf{h}; \mathbf{a}) = \mathbb{E}_{\mathbf{h} \sim p(\mathbf{h})}[-D_{\mathrm{KL}}(\pi(\mathbf{a}|\mathbf{h}) \| p(\mathbf{a}))]$ ensures that policy exploration does not diverge significantly from the marginal distribution $p(\mathbf{a})$. This aligns with the concept of an *action prior* [29], [67]. Similarly, the commonly used maximum entropy reinforcement learning objective [41], $\mathcal{H}(\mathbf{a}|\mathbf{h})$, can be viewed as employing a uniform action prior ($\mathcal{H}(\mathbf{a}|\mathbf{h}) = \mathbb{E}_{\mathbf{h} \sim p(\mathbf{h})}[-D_{\mathrm{KL}}(\pi(\mathbf{a}|\mathbf{h}) \| \mathcal{U}(a))] - c$), where $\mathcal{U}(a)$ is a uniform action distribution and c is a constant. In this way, our MIR3 replaces the random exploratory prior $\mathcal{U}(a)$ with a robust exploratory prior $p(\mathbf{a})$, which is implicitly learned from the environment. This enables enhanced exploration in regions that are more robust and task-relevant. This concept parallels the human cautious in decision making found in situational awareness theory [24]. In real-world scenarios like driving, individuals do not generate specific responses for each potential threat. Instead, they rely on general "robust action priors" such as reducing speed or changing lanes when confronted with uncertain conditions, ensuring adaptable yet cautious behavior.

As shown in Fig. 3, the benefit of constraining a policy to $p(\mathbf{a})$ on robustness can be interpreted from two aspects. First,

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

$p(\mathbf{a})$ can be viewed as a set of task-relevant actions consistently favored by the environment, independent of current histories. For example, in StarCraft II, actions directed at moving toward and attacking enemies are usually preferred for victory. More intricate tactics, such as kiting or focused fire, are optional and depend on current histories [68]. Thus, if certain actions are broadly effective within the environment, the policy is prone to succeed in accomplishing the task by leaning on these actions, even when confronted with action uncertainties. Second, keeping the policy near $p(\mathbf{a})$ fosters exploration in its vicinity. Therefore, even if some agents deviate from the optimal policy, the enhanced exploration around $p(\mathbf{a})$ encourages the policy to identify diverse methods for handling the task, preserving some intricate tactics for the task to succeed.

IV. Experiments

A. Experiment Settings

1) Environments: We evaluated our result on six tasks in StarCraft multi-agent challenge (SMAC) [68], quadratic swarm control (Quads) [69], and a continuous robot swarm control task with ten agents performing *rendezvous*, where agents are randomly placed in the arena and learn to gather together. We use SMAC and Quads to evaluate the performance of MIR3 on both discrete and continuous control. In all tasks, agents are required to complete the task with worst-case adversaries during testing, which differs from the standard cooperative MARL setting. For SMAC, we find having an adversary controlling one agent makes the environment unsolvable. We address this by allowing algorithms to control over additional agents to ensure fair evaluation.

2) Compared Methods: We implement MIR3 on MADDPG [1] and QMIX [2] backbones. The compared methods include M3DDPG [10], ROMAX [11], and ERNIE [21], which consider all other agents as adversaries; ROM-Q [18], which considers one or more agents as adversaries. Note that the design of M3DDPG [10] and ROMAX [11] relies on the central critic of MADDPG, so we do not evaluate it on the QMIX backbone. To comprehensively evaluate MIR3's performance across diverse backbones and environments, we benchmark MIR3 against ERNIE, ROM-Q, and EIR [19] on Quads environment that requires continuous control, and MAPPO backbone that yields stochastic policy. Under MAPPO backbone, we add EIR [19] as a new baseline that identifies unreliable agents and enables other agents to act optimally based on inferred reliability. We do not evaluate EIR with other backbone as its method relies on stochastic control, making it incompatible with deterministic backbones like MADDPG and QMIX. All methods are compared based on the same network architecture, hyperparameters, and tricks. *We leave hyperparameters and implementation details in Appendix I. See code and demo videos at https://github.com/DIG-Beihang/MIR3*

3) Evaluation Protocol: For environments with N agents, all methods to be attacked were trained using five random seeds. During attack, we fix the parameters in defender's policy, and train a worst-case adversary against current policy [12]. For scenarios with one agent as adversary, we average the attack result of each N agents using the same five seeds, reporting results averaged over $5 * N$ seeds. For scenarios with more than one adversary, we report the result with five attack scenarios sampled randomly. All results are reported with 95% confidence interval.

B. Simulation Results

We first present our results on six SMAC tasks. Experiments show our MIR3 significantly surpasses baselines in robustness and training efficiency, while maintaining cooperative performance. Next, we evaluate the performance of MIR3 on Quads environment with MAPPO backbone. The result of real-world multi-agent rendezvous will be discussed in Section IV-C.

1) MIR3 Is More Robust: We evaluate the defense capability of MIR3 against worst-case attacks, with one agent as an adversary. Experiments involving more adversaries will be discussed later. As shown in Fig. 4, although MIR3 does not encounter adversaries during training, it demonstrates superior defense capabilities across six tasks and two backbones, consistently outperforming even the best-performing baselines that directly consider adversaries.

The improved performance of MIR3 over baselines can be explained as follows. Compared to M3DDPG, ERNIE, and ROMAX, which assume all other agents as potential adversaries, MIR3 avoids learning overly pessimistic or less effective policies. Compared to ROM-Q, which prepares for each threat scenario, its adversaries and defenders cannot adequately explore or respond to the myriad threat scenarios during training. Thus, the adversaries and defenders remain weak at test time. Additional experiments in Appendix II prove that while baselines are effective against adversaries in training, their defenses can be easily compromised by the learned optimal adversary at test time. In contrast, MIR3, without exploring any threat scenarios, implicitly maximizes the lower bound performance under any threat scenario.

2) MIR3 Does Not Harm Cooperative Performance: We further show our MIR3 maintains cooperative performance while enhancing robustness. This is achieved by minimizing mutual information as an information bottleneck, which has been reported to enhance task performance in computer vision tasks [47]. Additionally, this is supported by the objective in (13), where defenders maximize both cooperative and robust performance.

3) MIR3 Learns Robust Behaviors: Next, we show MIR3 learns distinct robust behaviors. As illustrated in Fig. 5, for MADDPG backbone, raw MADDPG can be easily swayed by adversaries, causing agents to move forward without attacking and getting killed by enemies. Other robust baselines are rarely swayed but fail to retain cooperative behaviors (e.g., no focused fire on enemies), eventually losing the game. In contrast, by reducing mutual information, MIR3 ensures that agents are not only unswayed but also maintain focused fire behavior under attack. We attribute the improved performance of MIR3 to its ability to reduce spurious correlations between agents. In cooperative tasks that require agents to focus fire on enemies, baseline methods often lead agents to overfit to their teammates' behaviors. As a result, when some agents become unreliable—such as when compromised by an attacker—the

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

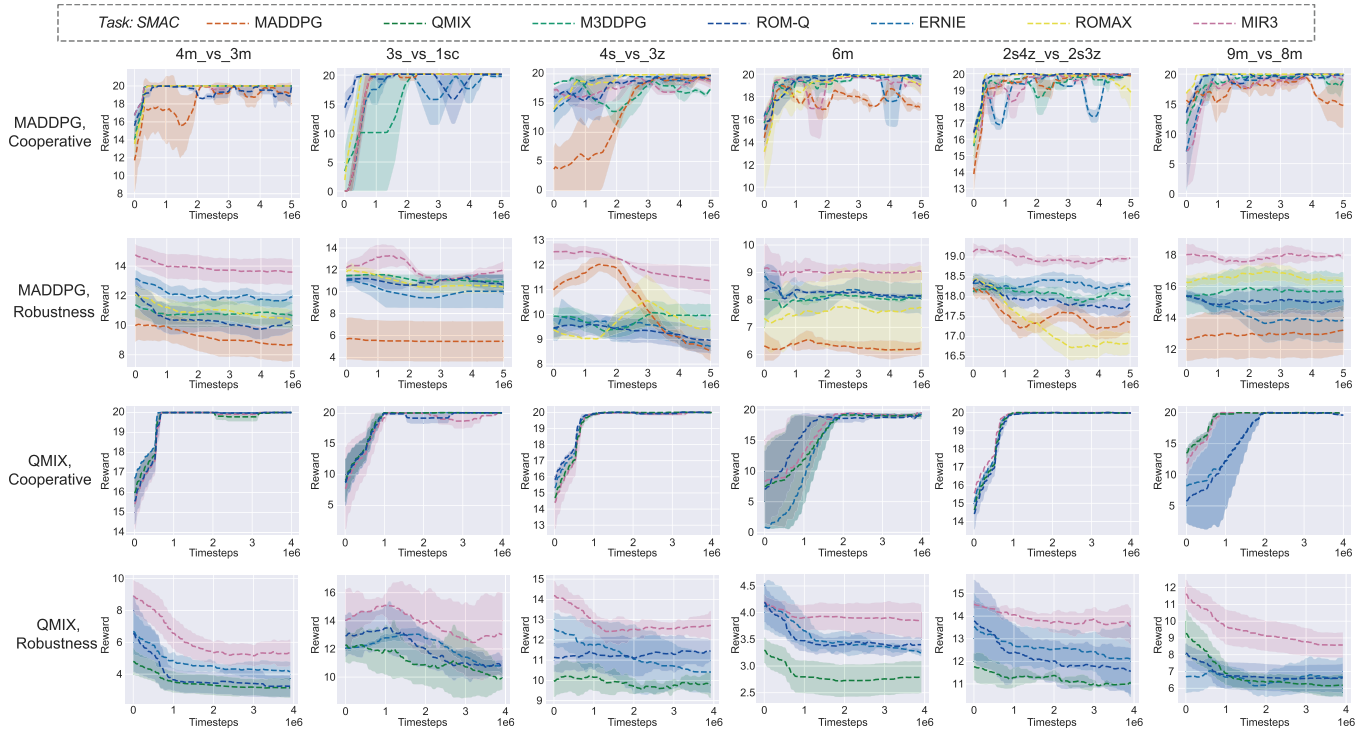

Fig. 4. Cooperative and robust performance under worst-case adversarial attacks. The four rows correspond to cooperative performance without attack and robustness under worst-case attack, under MADDPG and QMIX backbone. Results are reported across six SMAC tasks with 95% confidence intervals. Despite being trained without adversarial exposure, MIR3 consistently outperforms baselines—including those trained with adversaries—in terms of robustness, while also maintaining strong cooperative performance.

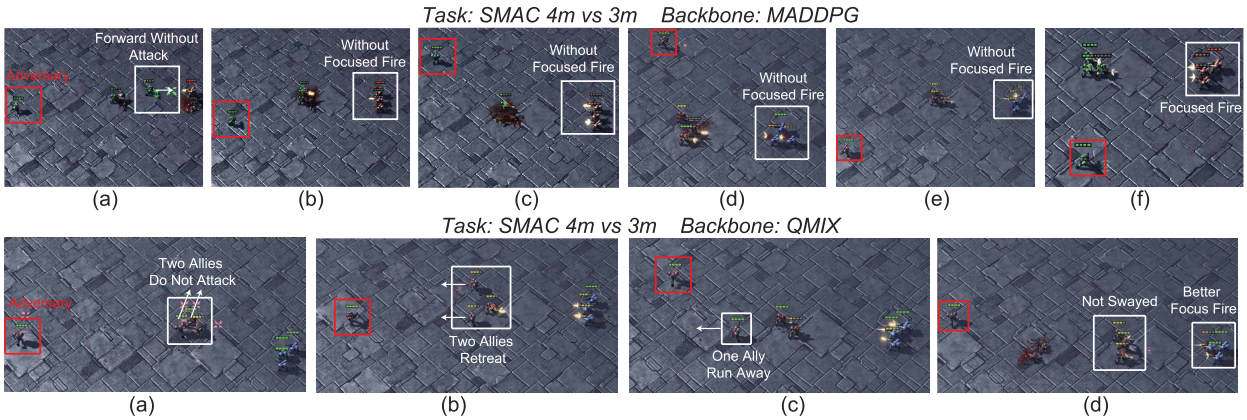

Fig. 5. Agent behaviors under attack in task 4 versus 3 m, adversary denoted by red square. Under MADDPG backbone, our MIR3 is not swayed by adversary and preserves focused fire behavior. Under QMIX backbone, baselines agents are frequently swayed back and forth. In contrast, our MIR3 is less swayed by adversary. (a) MADDPG; (b) M3DDPG; (c) ROM-Q; (d) ERNIE; (e) ROM-Q; and (f) MIR3 (Ours) (first row). (a) QMIX; (b) ROM-Q; (c) ERNIE; and (d) MIR3 (Ours) (second row).

remaining agents may overreact or lose coordination, leading to a breakdown in focused fire strategies. In contrast, MIR3 learns to execute such coordinated behaviors during training without relying on attacker presence, and is inherently less sensitive to erroneous behaviors from compromised agents. This enables it to maintain robust performance under attack.

Under QMIX backbone, benign agents in all baselines are frequently swayed by the adversary, moving randomly without attacking. In contrast, MIR3 agents are less swayed by the adversary. We explain by MIR3 learning a robust action prior. In QMIX, the underlying assumption is that all agents contribute positively toward team performance. However, this assumption breaks down in the presence of adversaries, rendering baseline methods vulnerable—even those explicitly designed for robustness. MIR3, by contrast, learns a prior centered around broadly effective actions (e.g., advancing and attacking the enemy), which remain valid regardless of teammate behavior. This grounding allows MIR3 to maintain effective performance even when facing adversarial perturbations. *Videos available at https://github.com/DIG-Beihang/MIR3*

4) MIR3 Is Robust With Many Adversaries: In extreme situations, there could be more than one adversaries. We examined this by adding an extra adversarial agent in map 4

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

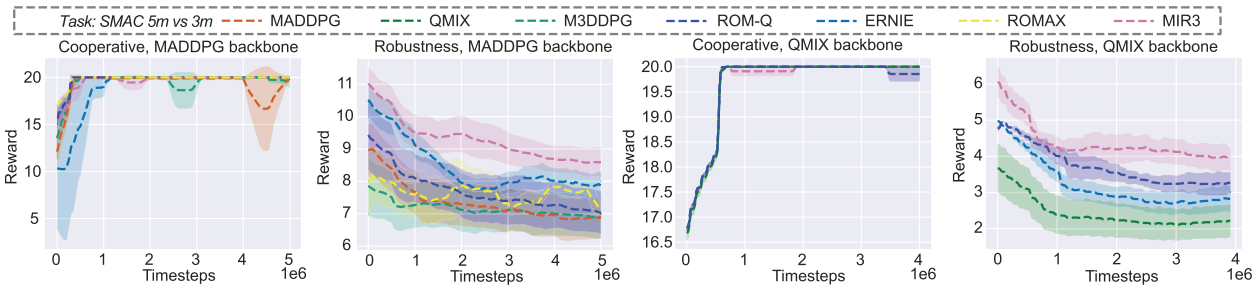

Fig. 6. Defense with two adversaries, evaluated in SMAC *5 versus 3 m*. In this challenging scenario, our MIR3 consistently exhibits stronger defense capability than all baselines, in both MADDPG and QMIX backbones. This demonstrates the potential of our MIR3 to be applied in more complex scenarios with many adversaries. *See results of another five SMAC tasks in Appendix III.*.

versus 3 m in SMAC, creating a map 5 versus 3 m with two adversaries. As illustrated in Fig. 6, in this challenging scenario, our MIR3 consistently exhibits stronger defense capability than all baselines, in both MADDPG and QMIX backbones. This demonstrates the potential of our MIR3 to be applied in more complex scenarios with many adversaries. *See results of another five SMAC tasks in Appendix III.*

5) MIR3 Is Robust Against Nonadversarial Disturbances: We further demonstrate that MIR3 handles common nonadversarial disturbances more effectively than traditional max-min approaches. To illustrate this, we design three typical types of disturbances affecting observations, actions, and environments. For observation failure, we introduce Gaussian noise $\mathcal{N}(0, \epsilon)$ with $\epsilon = 0.3$ to the observations of all agents at every timestep, simulating sensor inaccuracies or noisy observations. For action failure, each agent has a probability of 0.3 at every timestep to repeat the action from the previous timestep instead of executing the action based on the current observation, mimicking delayed observation scenarios. For environment uncertainty, we increase the difficulty of the opponent rule-based agents from level 7 to level 9, forcing MIR3 to adapt to stronger opponents and consequently increasing the uncertainty in environment transitions. All methods are evaluated using the MADDPG backbone on the SMAC tasks 4 versus 3 m and 3 versus 1 s.

As demonstrated in Table I, MIR3 effectively handles nonadversarial disturbances better than the MADDPG baseline and traditional max-min approaches. We attribute this difference to the fact that max-min methods explicitly consider worst-case attackers, making them particularly robust against perturbations defined within their uncertainty sets, but less capable of generalizing to typical, nonadversarial uncertainties encountered in practice. Consequently, while max-min approaches still achieve greater robustness than the nonrobust MADDPG baseline, they do not significantly enhance robustness against common, realistic disturbances. In contrast, MIR3 not only provides stronger resilience against worst-case scenarios but also maintains superior general-purpose robustness by minimizing mutual information, which acts as both an information bottleneck and a robust action prior. Thus, MIR3 agents effectively learn a generalized notion of caution, enabling better performance when faced with unforeseen uncertainties.

TABLE I

ROBUSTNESS EVALUATION OF MIR3 AGAINST NONADVERSARIAL DISTURBANCES. WE ASSESS ROBUSTNESS UNDER THREE DIFFERENT DISTURBANCE CONDITIONS: OBSERVATION FAILURE (OBS. FAIL.), WHERE GAUSSIAN NOISE IS ADDED TO AGENTS' OBSERVATIONS TO MIMIC HARDWARE NOISE; ACTION FAILURE (ACT. FAIL.), WHERE AGENTS REPEAT THEIR PREVIOUS ACTIONS TO SIMULATE DELAYED OBSERVATIONS; AND ENVIRONMENT UNCERTAINTY (ENV. UNCERT.), WHERE OPPONENT POLICIES ARE MODIFIED TO SIMULATE UNCERTAINTY IN ENVIRONMENTAL TRANSITIONS

Method	Obs. Fail.	Act. Fail.	Env. Uncert.
Map: SMAC *4m_vs_3m*			
MADDPG	10.96 ± 0.25	10.50 ± 0.35	11.55 ± 0.20
M3DDPG	13.37 ± 0.01	13.17 ± 0.25	13.78 ± 0.21
ROM-Q	12.16 ± 0.80	11.94 ± 0.13	12.40 ± 0.14
ROMAX	12.54 ± 0.04	12.21 ± 0.04	12.78 ± 0.03
ERNIE	14.95 ± 0.17	14.54 ± 0.10	15.07 ± 0.02
MIR3 (ours)	**16.93 ± 0.22**	**16.31 ± 0.18**	**17.00 ± 0.14**
Map: SMAC *3s_vs_1sc*			
MADDPG	7.10 ± 1.42	7.07 ± 2.15	7.25 ± 2.00
M3DDPG	12.16 ± 0.23	11.80 ± 0.17	11.71 ± 0.28
ROM-Q	12.13 ± 0.04	11.47 ± 0.44	11.85 ± 0.44
ROMAX	11.66 ± 0.04	11.27 ± 0.07	11.56 ± 0.10
ERNIE	11.05 ± 0.42	11.13 ± 0.50	11.25 ± 0.33
MIR3 (ours)	**13.74 ± 0.32**	**13.79 ± 0.03**	**13.30 ± 0.05**

6) MIR3 Generalizes to Continuous Control With Stochastic Policies: Previously, we have demonstrated that MIR3 achieves robust performance in discrete-control tasks within the SMAC environment using deterministic policy backbones (MADDPG and QMIX). To comprehensively assess the generalization capability of MIR3, we further evaluate it in the Quads environment, which requires continuous control with a stochastic policy backbone (MAPPO). Specifically, we compare MIR3 and baseline methods on two tasks: *static_same_goal* and *swarm_vs_swarm*, examining both cooperative performance and robustness under worst-case adversarial scenarios involving one or two adversaries. As shown in Fig. 7, MIR3 consistently maintains robust performance one or more adversaries in continuous control settings with stochastic policies, without compromising cooperative outcomes. These results are consistent with our earlier findings in discrete control settings with deterministic policies.

We note that although EIR demonstrates strong performance, it remains inferior to MIR3 under adversarial

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

TABLE II

PER-EPOCH TRAINING TIME (IN SECONDS) OF MIR3 AND BASELINES ACROSS FIVE ENVIRONMENTS. MIR3 INTRODUCES MINIMAL OVERHEAD TO MADDPG AND QMIX WHILE MUCH FASTER THAN METHODS THAT CONSIDERS THREAT SCENARIOS EXPLICITLY

Method	**MADDPG** ***4m vs 3m***	**QMIX** ***4m vs 3m***	**MADDPG** ***9m vs 8m***	**MADDPG** ***2s4z vs 2s3z***	**Rendezvous**
MADDPG	0.28 ± 0.11	–	0.42 ± 0.20	0.72 ± 0.27	0.61 ± 0.04
QMIX	–	0.69 ± 0.17	–	–	–
M3DDPG	0.41 ± 0.12	–	1.90 ± 0.50	1.54 ± 0.35	2.16 ± 0.29
ROM-Q	0.42 ± 0.15	1.01 ± 0.08	1.63 ± 0.25	1.04 ± 0.31	2.43 ± 0.29
ROMAX	0.48 ± 0.14	–	2.43 ± 0.54	1.39 ± 0.29	2.82 ± 0.40
ERNIE	0.40 ± 0.14	0.98 ± 0.08	1.66 ± 0.25	1.00 ± 0.28	1.57 ± 0.12
MIR3 (Ours)	$\mathbf{0.31} \pm 0.16$	$\mathbf{0.81} \pm 0.09$	$\mathbf{0.65} \pm 0.21$	$\mathbf{0.79} \pm 0.28$	$\mathbf{0.63} \pm 0.04$

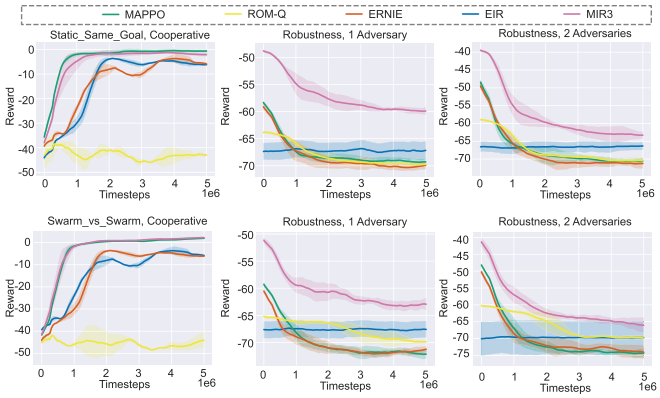

Fig. 7. Performance and robustness of MIR3 and baselines in continuous Quads environment (MAPPO backbone), under cooperative conditions and worst-case attacks with 1 and 2 adversaries. MIR3 generalizes effectively beyond SMAC tasks and deterministic backbones (QMIX, MADDPG) to the stochastic MAPPO backbone.

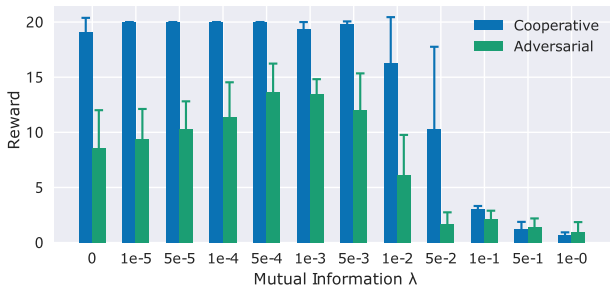

Fig. 8. Ablations on hyperparameter λ, showing a trade-off between policy effectiveness and limiting information flow. Evaluated on SMAC 4 versus 3 m.

conditions. This performance gap arises primarily due to EIR's reliance on explicitly detecting unreliable agents and subsequently selecting optimal actions based on inferred agent reliability. As the number of agents increases, the complexity and number of potential threat scenarios grow substantially. While EIR accurately identifies unreliable agents in practice, the expanding scenario space results in insufficient coverage of each potential threat during training. Consequently, EIR's defense remains susceptible to worst-case adversarial attacks during testing.

7) MIR3 Requires Less Training Time: We also demonstrate that our MIR3 method is computationally more efficient than baselines that explicitly consider threat scenarios. Following [40], we report the average training time per epoch over 50 episodes. All statistics are obtained based on one Intel Xeon Gold 5220 CPU and one NVIDIA RTX 2080 Ti GPU, using task 4 versus 3 m, 9 versus 8 m, and *2s4z versus 2s3z* for SMAC and ten *agents* for rendezvous (rendezvous requires agents to gather together. We train rendezvous in simulation for real-world experiment discussed later). We include 9 versus 8 m to evaluate our method with large number of agents, and *2s4z versus 2s3z* to evaluate our method with heterogeneous agents. As shown in Table II, our MIR3 only requires moderately more training time than backbones without considering robustness (+10.71% in MADDPG 4 versus 3 m, +17.39% in QMIX 4 versus 3 m, +54.76% in MADDPG 9 versus 8 m, +9.72% in MADDPG *2s4z versus 2s3z*, +3.28% in rendezvous), showing our defense can be added at low cost. In contrast, considering threat scenarios involves the costly approach of approximating an adversarial policy, resulting in significantly higher training times compared to our MIR3 approach (+29.03% in MADDPG 4 versus 3 m, +20.99% in QMIX 4 versus 3 m, +74.74% in MADDPG 9 versus 8 m, +26.58% in MADDPG 9 versus 8 m, and +149.21% in rendezvous).

8) Ablations on Hyperparameters: We next study the effect of hyperparameter λ in penalizing mutual information between histories and actions, which can be seen as an information bottleneck. We set λ in $\{0, 10^{-5}, \ldots, 1\}$ and evaluate the results on task 4 versus 3 m in SMAC with MADDPG backbone. The results are illustrated in Fig. 8. Note that with $\lambda = 0$, our MIR3 reduces to MADDPG.

For relatively small λ (i.e., $\lambda \leq 5 \times 10^{-4}$), the policy is steered to focus less, but more relevant information in the current history. This efficiently suppresses unnecessary agent-wise interactions, leading to more robust policies and even slightly enhancing cooperative performance, which is also evident in computer vision tasks using information bottleneck as regularizer [47]. Conversely, when $\lambda > 5 \times 10^{-4}$, the policy is restricted from utilizing any information from the current history, resulting in a collapse of both cooperative and robust performance. As a consequence, we select $\lambda = 5 \times 10^{-4}$ for an optimal trade-off between limiting information flow and maintaining policy effectiveness.

9) Ablations on Mutual Information Estimation Methods: We further evaluate the impact of different mutual information upper bound estimation methods beyond CLUB [49]. Specifically, we consider VUB [47], L1Out [48], and CLUB-Sample [49] as alternative estimators, each tuned for optimal

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

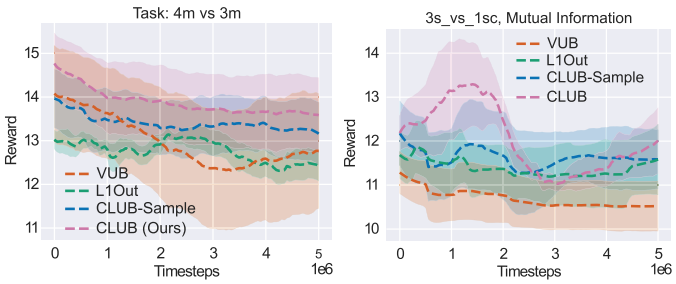

Fig. 9. Ablations on mutual information estimation methods. The robustness is evaluated with one worst-case adversary. While using CLUB as mutual information upper bound estimation method empirically shows the best result, our MIR3 also yields reasonable robustness when using other mutual information estimation methods. Evaluated on SMAC 4 versus 3 m and 3 versus 1 s.

performance. An introduction of these methods are available at preliminaries. Experiments are conducted on the SMAC 4 versus 3 m and 3 versus 1 s scenarios, each involving one worst-case adversary.

As shown in Fig. 9, while CLUB yields the strongest empirical performance, MIR3 achieves competitive robustness across all evaluated MI estimators. These results suggest that MIR3 is not tightly coupled to a specific MI estimator and can maintain robust performance even when the estimator is replaced by a less accurate one.

C. Real-World Experiments

In this section, we evaluate the robustness of our MIR3 under action uncertainties in real-world robot swarm control. This setting presents three major challenges. First, consistent with our simulation setup, we introduce an adversarial agent executing a worst-case policy; the remaining agents must maintain robust behavior despite this adversary's presence. Second, physical parameters in the real world can differ from those in simulation due to variations in friction, mass, and other factors. As a result, policies trained in simulation may struggle to maintain effective cooperation when deployed in real-world settings. Third, system-level noise further disrupts control accuracy: for instance, robot localization may be imprecise, and hardware imperfections may cause actions to deviate from those prescribed by the policy. In this challenging scenario, following the widely accepted Sim2Real paradigm in the reinforcement learning community [70], we directly transfer the policies for both defenders and adversaries trained in simulation environments, to our robots in the real world.

We first show simulation results. As illustrated in Fig. 10(a), our MIR3 consistently outperforms baselines in robustness, without sacrificing cooperative performance. It is interesting to note that in our simulation, while only trained on *rendezvous* task, our MIR3 agents show an emergent pursuit-evade behavior when facing an adversary running away. *See analysis of this behavior in Appendix IV.*

Next, following the Sim2Real paradigm, we deploy our trained algorithm in a 2×2 m indoor arena with ten e-puck2 robots [71], see Fig. 10(b) for depictions of our arena. We run each algorithm in the real-world arena ten times, with all algorithms following the same initialization. The results are

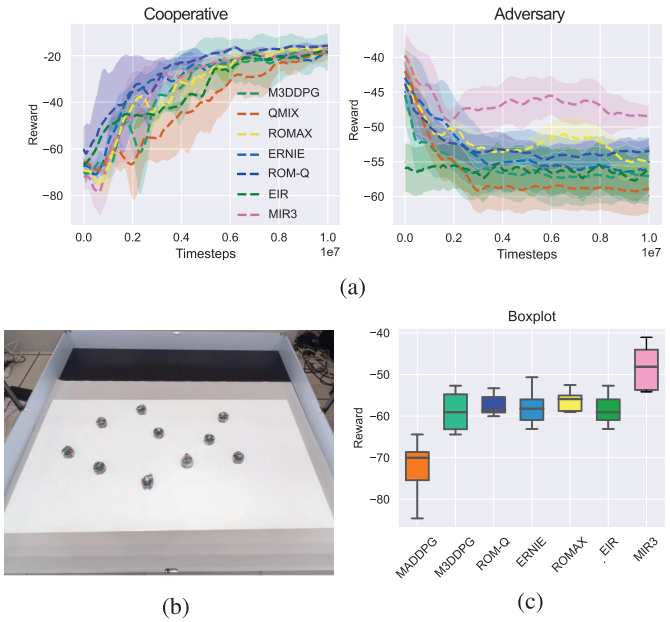

Fig. 10. Illustration of our real-world rendezvous environment. Our MIR3 obtains stronger robustness against action uncertainties in both simulation and real-world deployment. (a) Simulation results. (b) Arena. (c) Real-world results.

TABLE III

SUMMARY STATISTICS OF REAL-WORLD PERFORMANCE FOR EACH METHOD, WITH EACH METHOD DEPLOYED TO REAL WORLD IN TEN RUNS. OUR MIR3 ACHIEVES +14.29% PERFORMANCE INCREASE IN AVERAGE (I.E., MEAN PERFORMANCE)

Method	Mean	Median	IQR
MADDPG	-72.64	-70.04	6.76
M3DDPG	-58.83	-59.08	8.39
ROMQ	-57.30	-58.56	3.69
ERNIE	-57.79	-58.23	4.93
ROMAX	-56.27	-55.91	3.72
EIR	-58.36	-59.08	4.93
MIR3 (Ours)	**-48.23**	**-48.15**	9.67

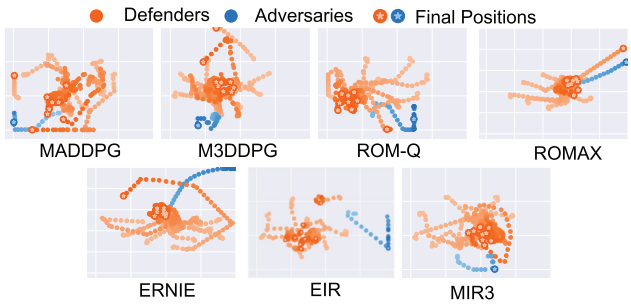

Fig. 11. Real-world trajectories of robot swarm control, with defenders in orange and adversaries in blue. We use a star to denote the final position. Our MIR3 act reliably without being swayed by adversary.

summarized in Fig. 10(c), with statistics shown in Table III. Our MIR3 achieves +14.29% average reward improvement compared to the best-performing baseline. Moreover, as shown in Fig. 11, a detailed examination of the trajectories reveals that MIR3 successfully learn robust behaviors. In contrast to the simulation, MADDPG completely failed to handle

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

real-world uncertainties, leading to multiple agents malfunctioning and failing to gather, underscoring the necessity of evaluating robustness in the real world. M3DDPG, ROMAX, ERNIE, EIR, and ROMAX perform substantially better than MADDPG, although one or several agents are still misled by the adversary. While EIR has the potential to find the adversary and learn the optimal equilibrium, it struggles to find the best equilibrium with large number of agents. Conversely, our MIR3 can group together without deviation and maintain consistent behavior throughout the evaluation. *Videos available at https://github.com/DIG-Beihang/MIR3*

V. Conclusion

In this article, we introduce MIR3, a novel regularization-based approach for robust MARL. Motivated by robust decision making of humans, MIR3 does not require training with adversaries, yet is provably robust against cooperative agents deviate from their policy and executing worst-case actions. Theoretically, we formulate robust MARL as an control-as-inference problem, which implicitly optimize worst-case robustness through off-policy evaluation. Under this formulation, we prove that minimizing mutual information serves as a lower bound for robustness. This objective can further be interpreted as suppressing spurious correlations through an information bottleneck, or as learning a robust action prior that encourages actions favored by the environment. In line of our theoretical findings, empirical results demonstrate that our MIR3 surpass baselines in robustness and training efficiency in StarCraft II, quadrotor swarm control and robot swarm control, and consistently exhibits superior robustness when deployed in real world.

References

[1] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 6382–6393.

[2] T. Rashid, M. Samvelyan, C. S. d. Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2018, pp. 4295–4304.

[3] C. Yu et al., "The surprising effectiveness of PPO in cooperative, multi-agent games," 2021, *arXiv:2103.01955*.

[4] Y. Hu, J. Fu, and G. Wen, "Graph soft actor–critic reinforcement learning for large-scale distributed multirobot coordination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 665–676, Jan. 2025.

[5] J. Hao et al., "Exploration in deep reinforcement learning: From single-agent to multiagent domain," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 8762–8782, Jul. 2024, doi: 10.1109/TNNLS.2023.3236361.

[6] T. Zhang, Z. Liu, J. Yi, S. Wu, Z. Pu, and Y. Zhao, "Multiexperience-assisted efficient multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 12678–12692, Sep. 2024, doi: 10.1109/TNNLS.2023.3264275.

[7] Y. Xie, S. Mou, and S. Sundaram, "Communication-efficient and resilient distributed Q-learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3351–3364, Mar. 2024, doi: 10.1109/TNNLS.2023.3292036.

[8] B. Chen, Z. Cao, and Q. Bai, "SATF: A scalable attentive transfer framework for efficient multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 6627–6641, Apr. 2025.

[9] O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019.

[10] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 33. Palo Alto, CA, USA: AAAI Press, Feb. 2019, pp. 4213–4220.

[11] C. Sun, D. K. Kim, and J. P. How, "ROMAX: Certifiably robust deep multiagent reinforcement learning via convex relaxation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 5503–5510.

[12] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," 2019, *arXiv:1905.10615*.

[13] B. Ly and R. Ly, "Cybersecurity in unmanned aerial vehicles (UAVs)," *J. Cyber Secur. Technol.*, vol. 5, no. 2, pp. 120–137, Apr. 2021.

[14] L. C. Dinh, D. Mguni, T. A. Han, J. Wang, and Y. Yang, "Online Markov decision processes with non-oblivious strategic adversary," *Auto. Agents Multi-Agent Syst.*, vol. 37, no. 1, p. 15, Jan. 2023.

[15] S. Li et al., "Attacking cooperative multi-agent reinforcement learning by adversarial minority influence," 2023, *arXiv:2302.03322*.

[16] M. Hüttenrauch, S. Adrian, and G. Neumann, "Deep reinforcement learning for swarm systems," *J. Mach. Learn. Res.*, vol. 20, no. 54, pp. 1–31, 2019.

[17] K. Zhang, T. Sun, Y. Tao, Ş. Genç, S. Mallya, and T. BaŞar, "Robust multi-agent reinforcement learning with model uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jan. 2020, pp. 10571–10583.

[18] E. Nisioti, D. Bloembergen, and M. Kaisers, "Robust multi-agent Q-learning in cooperative games with adversaries," in *Proc. AAAI Workshop Reinforcement Learn. Games*, Feb. 2021. [Online]. Available: https://aaai.org/ocs/index.php/WS/AAAIW21/paper/view/17731

[19] S. Li et al., "Byzantine robust cooperative multi-agent reinforcement learning as a Bayesian game," 2023, *arXiv:2305.12872*.

[20] C. Tessler, Y. Efroni, and S. Mannor, "Action robust reinforcement learning and applications in continuous control," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6215–6224.

[21] A. Bukharin et al., "Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 68121–68133.

[22] L. Yuan et al., "Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2023, pp. 11753–11762.

[23] E. Derman, M. Geist, and S. Mannor, "Twice regularized MDPs and the equivalence between robustness and regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 22274–22287.

[24] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 37, no. 1, pp. 32–64, Mar. 1995.

[25] M. R. Endsley et al., "Situation awareness in aviation systems," in *Handbook of Aviation Human Factors*, D. J. Garland, J. A. Wise, and V. D. Hopkin, Eds., Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1999, pp. 257–276.

[26] S. Levine, "Reinforcement learning and control as probabilistic inference: Tutorial and review," 2018, *arXiv:1805.00909*.

[27] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[28] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*.

[29] J. Grau-Moya, F. Leibfried, and P. Vrancx, "Soft Q-learning with mutual-information regularization," in *Proc. Int. Conf. Learn. Represent.*, Sep. 2018. [Online]. Available: https://openreview.net/pdf?id=HyEtjoCqFX

[30] F. A. Oliehoek and C. Amato, *A Concise Introduction To Decentralized POMDPs*, vol. 1. Cham, Switzerland: Springer, 2016.

[31] J. Li et al., "Two heads are better than one: A simple exploration framework for efficient multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 20038–20053.

[32] K. Lin et al., "DCIR: Dynamic consistency intrinsic reward for multi-agent reinforcement learning," 2023, *arXiv:2312.05783*.

[33] Z. Liu, Y. Zhu, Z. Wang, Y. Gao, and C. Chen, "MIXRTs: Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 4090–4107, May 2025, doi: 10.1109/TPAMI.2025.3540467.

[34] B. Du et al., "Safe adaptive policy transfer reinforcement learning for distributed multiagent control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1939–1946, Jan. 2025.

[35] S. Han et al., "What is the solution for state-adversarial multi-agent reinforcement learning?," 2022, *arXiv:2212.02705*.

[36] S. He, S. Han, S. Su, S. Han, S. Zou, and F. Miao, "Robust multi-agent reinforcement learning with state uncertainty," *Trans. Mach. Learn. Res.*, Jan. 2023. [Online]. Available: https://openreview.net/forum?id=CqTkapZ6H9

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.

[37] E. KardeŞ, F. Ordóñez, and R. W. Hall, "Discounted robust stochastic games and an application to queueing control," *Oper. Res.*, vol. 59, no. 2, pp. 365–382, Apr. 2011.
[38] S. He, Y. Wang, S. Han, S. Zou, and F. Miao, "A robust and constrained multi-agent reinforcement learning electric vehicle rebalancing method in AMoD systems," 2022, *arXiv:2209.08230*.
[39] T. Phan et al., "Learning and testing resilience in cooperative multi-agent systems," in *Proc. 19th Int. Conf. Auto. Agents MultiAgent Syst.*, May 2020, pp. 1055–1063.
[40] X. Qu, A. Gupta, Y.-S. Ong, and Z. Sun, "Adversary agnostic robust deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6146–6157, Sep. 2023, doi: 10.1109/TNNLS.2021.3133537.
[41] B. Eysenbach and S. Levine, "Maximum entropy RL (provably) solves some robust RL problems," 2021, *arXiv:2103.06257*.
[42] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A, Gen. Phys.*, vol. 33, no. 2, p. 1134, 1986.
[43] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.
[44] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.
[45] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E-Stat., Nonlinear, Soft Matter Phys.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.
[46] M. I. Belghazi et al., "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.
[47] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," 2016, *arXiv:1612.00410*.
[48] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5171–5180.
[49] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2020, pp. 1779–1788.
[50] N. Jaques et al., "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2018, pp. 3040–3049.
[51] T. Wang, J. Wang, Y. Wu, and C. Zhang, "Influence-based multi-agent exploration," 2019, *arXiv:1910.05512*.
[52] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "MAVEN: Multi-agent variational exploration," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 7613–7624.
[53] L. Chen et al., "Signal instructed coordination in cooperative multi-agent reinforcement learning," in *Proc. Int. Conf. Distrib. Artif. Intell.*, Shanghai, China, Dec. 2021, pp. 185–205.
[54] W. Kim, W. Jung, M. Cho, and Y. Sung, "A maximum mutual information framework for multi-agent reinforcement learning," 2020, *arXiv:2006.02732*.
[55] P. Li et al., "PMIC: Improving multi-agent reinforcement learning with progressive mutual information collaboration," 2022, *arXiv:2203.08553*.
[56] K. J. Åström, "Optimal control of Markov processes with incomplete state information," *J. Math. Anal. Appl.*, vol. 10, no. 1, pp. 174–205, Feb. 1965.
[57] Y. Wen, Y. Yang, R. Luo, J. Wang, and W. Pan, "Probabilistic recursive reasoning for multi-agent reinforcement learning," 2019, *arXiv:1901.09207*.
[58] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 2052–2062.
[59] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 1042–1051.
[60] P. Liao, Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy, "Batch policy learning in average reward Markov decision processes," *Ann. Statist.*, vol. 50, no. 6, p. 3364, Dec. 2022.
[61] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2017, pp. 1352–1361.
[62] F. S. Melo, "Convergence of Q-learning: A simple proof," Inst. Syst. Robot., Instituto Superior Técnico, Lisbon, Portugal, 2001, pp. 1–4.
[63] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
[64] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.
[65] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.
[66] A. Saxe et al., "On the information bottleneck theory of deep learning," *J. Stat. Mech., Theory Exp.*, vol. 2019, no. 12, Dec. 2019, Art. no. 124020.
[67] K. Pertsch, Y. Lee, and J. J. Lim, "Accelerating reinforcement learning with learned skill priors," in *Proc. Conf. robot Learn.*, Jan. 2020, pp. 188–204.
[68] M. Samvelyan et al., "The StarCraft multi-agent challenge," 2019, *arXiv:1902.04043*.
[69] S. Batra, Z. Huang, A. Petrenko, T. Kumar, A. Molchanov, and G. S. Sukhatme, "Decentralized control of quadrotor swarms with end-to-end deep reinforcement learning," in *Proc. Conf. Robot Learn.*, 2022, pp. 576–586.
[70] S. Höfer et al., "Sim2Real in robotics and automation: Applications and challenges," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 398–400, Apr. 2021.
[71] F. Mondada et al., "The e-puck, a robot designed for education in engineering," in *Proc. 9th Conf. Auto. Robot Syst. Competitions*, Jan. 2009, vol. 1, no. 1, pp. 59–65.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on November 10,2025 at 09:29:31 UTC from IEEE Xplore. Restrictions apply.