Situation-Dependent Causal Influence-Based Cooperative Multi-Agent Reinforcement Learning

Xiao Du, Yutong Ye, Pengyu Zhang, Yaning Yang, Mingsong Chen, Ting Wang[*]

Software Engineering Institute, East China Normal University

{52265902007, 52205902007, pengyu.zhang, 52215902011}@stu.ecnu.edu.cn, {mschen, twang}@sei.ecnu.edu.cn

Abstract

Learning to collaborate has witnessed significant progress in multi-agent reinforcement learning (MARL). However, promoting coordination among agents and enhancing exploration capabilities remain challenges. In multi-agent environments, interactions between agents are limited in specific situations. Effective collaboration between agents thus requires a nuanced understanding of when and how agents' actions influence others. To this end, in this paper, we propose a novel MARL algorithm named Situation-Dependent Causal Influence-Based Cooperative Multi-agent Reinforcement Learning (SCIC), which incorporates a novel Intrinsic reward mechanism based on a new cooperation criterion measured by situation-dependent causal influence among agents. Our approach aims to detect inter-agent causal influences in specific situations based on the criterion using causal intervention and conditional mutual information (CMI) . This effectively assists agents in exploring states that can positively impact other agents, thus promoting cooperation between agents. The resulting update links coordinated exploration and intrinsic reward distribution, which enhance overall collaboration and performance. Experimental results on various MARL benchmarks demonstrate the superiority of our method compared to state-of-the-art approaches.

Introduction

Along with the rapid development of deep reinforcement learning (RL), MARL has attracted increasing attention in recent years and witnessed significant in real-world problems such as traffic light control (Wu et al. 2020), coordination of autonomous vehicles (Kiran et al. 2022), and robotics control (Wang et al. 2023), which can be effectively modeled as multi-agent game models. The most original MARL training method is the complete independent training of each agent, in which other agents are treated as part of the environment without any coordinated interaction between the behaviors of agents. Nonetheless, this training method suffers from the dilemma within a non-stationarity environment.

The Centralized Training with Decentralized Execution (CTDE) emerges as a widely adopted solution for mitigating the challenges posed by non-stationary environments,

---

[*]Corresponding Author

where centralized training allows agents to access information about other agents. However, due to partial observability, during the execution phase, only local behavior-observation information can be relied upon. Both policy-based and value-based approaches have been introduced for CTDE, such as MADDPG (Lowe et al. 2017), COMA (Foerster et al. 2018), MAAC (Iqbal and Sha 2019), and QMIX (Rashid et al. 2018). Although global information can be accessed during the centralized training phase, these methods make an assumption that decentralized policies remain independent of each other such that the joint policy can be expressed as a product of independent policies. Thus, two pivotal challenges stem from the above issues. Firstly, the complete independence of agents' policies disregards the influence of other agents, thereby restricting the agents from learning coordinated behavior. Secondly, optimizing decentralized strategies for multiple agents solely based on task-dependent dense reward signals often proves inefficient, particularly when these reward signals are stochastic or sparse. Some studies (Liu et al. 2020; Wen et al. 2019) suggest that to alleviate the non-stationary issue and achieve collaboration between agents, it is imperative for agents to consider their impact on the behavior of other agents when making decisions. Other works (Chen et al. 2021; Mahajan et al. 2019; Kim et al. 2023; Li et al. 2022) propose to quantify the correlation of agent behaviors through mutual information (MI), so as to maximize the correlation of agent behaviors to enhance collaboration. MI has been identified as an effective intrinsic reward in promoting coordination in these works. Unfortunately, effectively coordinating the simultaneous actions of multi-agents in these approaches remains a challenging and intractable issue. To this end, the primary objective of this study is to address this challenge from a causal inference perspective.

In the context of control, if an agent can control other agents, it signifies that the agent can impact other agents through its actions. However, there is an underappreciated aspect hidden in this seemingly obvious observation, which is that the causal effects of behavior are state-dependent. Considering the multi-robot navigation scenario, when the current position of robot a is close to robot b and aligned with the current direction of movement of robot b, then robot a evidently exerts a strong influence on robot b, irrespective of the next action of robot a. Conversely, if robot a is situ-

ated far from robot b and not in the movement direction of robot b, then robot a has a comparatively weaker influence on robot b. Generally, there are situations in which agents have immediate causal influence, while in other situations, such influence is absent. The key intuition of this study lies in the recognition that the states that enable an agent to have the ability to influence other agents of interest are important from both exploration and collaboration perspectives, and we name these states of importance as *significant states*. Due to the fact that the initial states of an agent are rarely able to control the agents of interest, leading to inefficient training, these initial states are typically not regarded as *significant states*. Nonetheless, recognizing that the *significant states* are conducive to inter-agent collaboration, it becomes imperative for the training process to prioritize and proactively explore these *significant states*. This exploration impels agents to consider the state-dependent causal effects among agents during the exploration process.

In this work, we exploit situation-dependent nature to measure the causal influence between agents. We propose an algorithm designed to achieve coordination among agents and coordinate agents' exploration, which gives agents an intrinsic reward based on state-dependent causal influence. Specifically, at each time step, all agents measure the causal influence between their actions in the current state and the next state of other agents, which is used to quantify the extent of causality between the agent and other agents at the current moment. The mean of the causal influence originating from other agents is then taken as the agent's intrinsic reward. Furthermore, the computation of situation-dependent causal effects leverages CMI, which reliably identifies the *significant states*. This intrinsic reward mechanism effectively facilitates the detection of *significant states*, thereby enhancing the collaboration between agents. It's worth noting that calculating MI of continuous variables is known to be challenging. Compared to the variational information maximizing-based algorithms (Chalk, Marre, and Tkacik 2016; Kolchinsky, Tracey, and Wolpert 2019), MINE (Belghazi et al. 2018), which learns a neural estimate of the MI, has shown superior performance (Hjelm et al. 2019; Velickovic et al. 2019). Hence, we employ MINE to learn the CMI between agents by utilizing a forward dynamics model. To summarize, this paper makes the following three major contributions:

- We model multi-agent reinforcement learning as a causal graph model by explicitly modeling the influence of causal factors in a multi-agent setting.

- We formalize the causal influence between agents as situation-dependent instead of action-dependent. Accordingly, a new Intrinsic Reward method with peer incentives is further proposed to promote the cooperation between agents using state-dependent causal influence, which is measured based on intervention and CMI.

- We conduct comprehensive experiments on various MARL benchmarks. The experimental results demonstrate that our approach outperforms other competitive methods, and the learned intrinsic reward proves to be conducive to learning better policies that achieve agents'

better cooperation in these complex tasks.

Related Work

As an important training paradigm for solving agent coordination problems, CTDE has been widely adopted in MARL in recent years, where VDN (Sunehag et al. 2018) and QMIX based on value function decomposition, and MADDPG and MAAC based on centralized critic are typical CTDE-based algorithms. Since our approach is designed to utilize causal influence as an intrinsic reward to enhance inter-agent collaboration in MARL, in this section we will briefly review related work in terms of Intrinsic Reward for MARL and causality in reinforcement learning.

Intrinsic Reward for MARL

Intrinsic rewards help agents learn useful policies across a wide variety of tasks and environments, even sometimes with sparse environmental rewards. At each time step, the agents receive not only environmental rewards but also intrinsic rewards. Intrinsic rewards can help agents improve their exploration ability and enhance social influence. In reinforcement learning, all kinds of information have various signals as intrinsic rewards, such as empowerment (Mohamed and Rezende 2015), model surprise (Blaes et al. 2019), information gain (Houthooft et al. 2016), and learning progress (Blaes et al. 2019). In MARL, some existing approaches utilize the correlation or influence of agents as intrinsic rewards to facilitate collaboration. EITI (Wang et al. 2020) leverages MI to capture the influence between an agent's current trajectory and the next states of other agents in the environment, which is used as an intrinsic reward to encourage agents to explore cooperatively. SI (Jaques et al. 2019) proposes a method that utilizes social influence as an intrinsic reward, measured by the MI between an agent's current action and the estimated next action of other agents, to achieve coordination and communication in MARL. VM3-AC (Kim et al. 2023) is similar to the idea of coordinating inter-agent behavior in SI-MOA, except that additional latent variables are introduced to induce nonzero MI between multi-agent actions, and the policy iteration algorithm is modified based on MI. PMIC (Li et al. 2022) utilizes the MI between global states and joint actions as a new standard. Based on this standard, it maximizes the MI related to behaviors that are conducive to collaboration and minimizes the MI related to behaviors that are not conducive to collaboration, breaking the current suboptimal collaboration and learning higher-level collaborative behaviors. Comparatively, our approach tackles the challenge of collaboration among agents from a causal inference perspective, which detects situation-dependent causal relationships among agents as intrinsic rewards to facilitate the exploration of inter-agent coordination.

Causality in Reinforcement Learning

The combination of reinforcement learning and causality has achieved some progress. The work (Schölkopf et al. 2021) utilizes causal modeling to achieve better state abstraction,

enabling the agent to concentrate on key aspects and indirectly improving sampling efficiency. The work (Seitzer, Schölkopf, and Martius 2021) proposes integrating measures of causal influence into reinforcement learning algorithms to address the problems of exploration and learning in the robot manipulation environment. The work (Lee et al. 2021) utilizes causal intervention to identify the most relevant state variables for completing a task, thereby reducing the dimensionality of the state space. The work (Foerster et al. 2018) utilizes influence detection to create counterfactual data to enhance the training of RL agents. (Maes, Meganck, and Manderick 2007) extends causal Bayesian networks to multi-agent models, which inspires us to develop a causal graph model for MARL.

Background

Preliminaries

In this work, we consider the fully cooperative multi-agent game in the partially observable setting, which can be modeled as a decentralized partially observable Markov Decision Process (Dec-POMDP) (Oliehoek, Amato et al. 2016). The Dec-POMDP is formally defined as a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{U}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{I} = \{1, 2, ...N\}$ denotes the finite set of agents, $s_t \in \mathcal{S}$ denotes the set of joint states that cannot be observed by agents at the time step t. At each time step t, each agent $i \in \mathcal{I}$ can only observe its local observation o_t^i from observation function $\mathcal{O}(s_t, i)$ and chooses an action $u_t^i \in \mathcal{U}^i$ according to its policy $\pi(\cdot|o_t^i)$, forming a joint action $u_t \in \mathcal{U}$. After executing u_t in environment, each agent i achieves a shared extrinsic reward r_t from the reward function $\mathcal{R}(s_t, u_t)$ with a discount factor $\gamma \in [0,1)$ and the next state s_{t+1} according to the transition function $\mathcal{P}(s_{t+1}|s_t, u_t)$. The goal of a collaborative team is to find a joint policy π^* that can maximize the expected extrinsic discount return $E[\sum_{k=1}^{\infty} \gamma^t r_t]$ under the setting of fully-cooperative MARL.

In our approach, we adopt a CTDE paradigm, which has been a widely considered training paradigm in recent efforts in MARL (Sunehag et al. 2018; Rashid et al. 2018; Lowe et al. 2017; Iqbal and Sha 2019). During training, each agent can access to full information including the states, actions, rewards, and actions of other agents, while decentralized execution is conditioned solely on individual observation o_t^i.

Causal Graphical Models

A causal graphical model (CGM) (Shanmugam 2001) is commonly represented as directed cyclic graphs (DAGs) $\mathcal{G} = (V, E)$ and is defined by a distribution $\mathcal{P}$ over the set of random variables $\mathcal{X}$. The graph $\mathcal{G}$ consists of nodes V and edges $\mathcal{E} \in V^2$ with (v, v) for any $v \in V$. Each node v_i is associated with a random variable $\mathcal{X}^i$ and each edge $(v_i \rightarrow v_j)$ represents that $\mathcal{X}^i$ is a direct cause of $\mathcal{X}^j$, i.e., $\mathcal{X}^i$ is called a parent of $\mathcal{X}^j$. The set of parents of $\mathcal{X}^j$ is denoted by $\mathbf{PA}_j^{\mathcal{G}}$. The distribution $\mathcal{P}$ can be represented as

$$p(X^1, ..., X^v) = \prod_{i=1}^{v} p_i(X^i | \mathbf{PA}_j^{\mathcal{G}}), \quad (1)$$

where $\mathbf{PA}_j^{\mathcal{G}} \subset \{X^1, ..., X^v\} \setminus \{X^j\}$ denotes the set of parents of X^j. The CGM models the structure of the causality. To reveal the causal structure from the data distribution, we assume that CGM satisfies the Markov condition and the faithfulness assumption, which makes the independence consistent between the joint distribution $P(X^1, ..., X^v)$ and the graph $\mathcal{G}$.

Intervention

Intervention sampling is a typical operation in causal discovery. Different from standard sampling, it sets the distribution of variables that require intervention in the causal graph model to a uniform distribution or fixed value, and then conducts sampling to obtain intervention data. Intervention data has more causal information than observational data, which is beneficial to measure causal relationships between random variables. In this work, we need to measure the influence of the current time step's action of an agent on the state of other agents in the next time step, which can be obtained by intervening in the current behavior distribution.

Our Approach

In this section, we present the design of our SCIC approach, in which the agents simultaneously learn a policy and an intrinsic reward function by maximizing the causal influence between agents, as illustrated in Figure 1. Our SCIC approach detects the inter-agent causal influences in particular situations based on causal intervention and CMI, facilitating agents to explore states that can affect other agents, thereby promoting cooperation among agents.

Causal Influence-based Intrinsic Reward Design

A key component of SCIC is the intrinsic reward mechanism that brings each agent to a state where it can influence interested agents as much as possible. Intuitively, an agent is more likely to influence interested agents in certain states, and making agents reach these states as much as possible is more likely to enhance the causal influence between agents. We represent agent j being able to be influenced by agent i as "agent i being able to take control of agent j". Through mutual incentives between agents, agents can achieve friendly interaction and efficient cooperation with their peers. Thus, the causal influence of other agents on an agent can be viewed as a necessity for learning intrinsic reward mechanisms. We detect the causal influences between agents in particular situations using causal intervention and CMI. It is worth noting that the computation of causal influence is only executed during intensive training, where each agent can know the policies and actions of other agents.

Multi-agent Causal Graph Model We extend the CGM to the case of decentralized multi-agents as Multi-agent Causal Graph Model (MACGM), where agents share an environment and have access to private and/or public variables of interest during centralized training. The one-step transition dynamics of MACGM at time step t is modeled a causal graphical $\mathcal{G}$ (see Figure 2) over the set of random variables $\mathcal{V} = \{S_t, S_{t+1}, S_t^1, ..., S_t^N, A_t^1, ..., A_t^N, S_{t+1}^1, ..., S_{t+1}^N\}$, consisting of a conditional distribution $P(V_i|PA(V_i))$, where
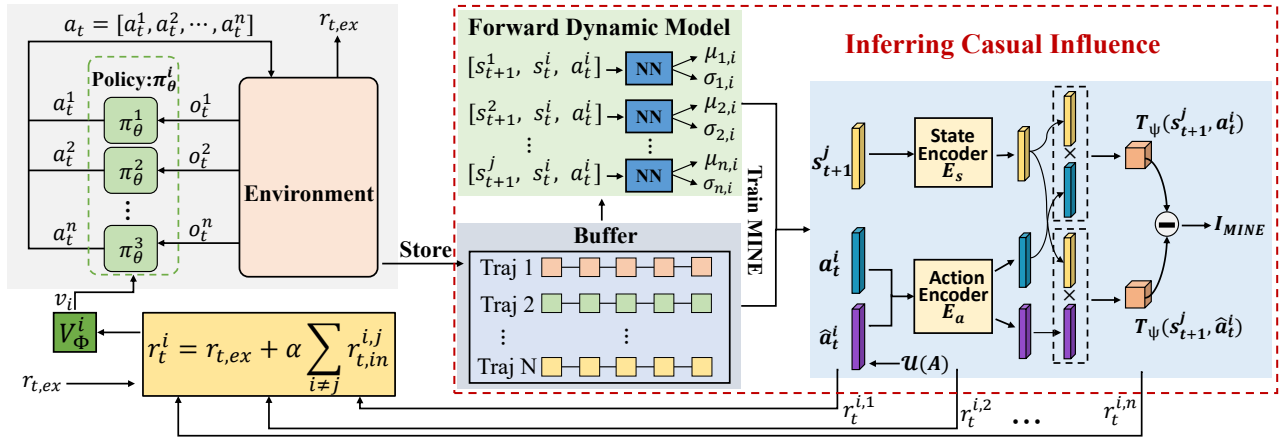
Figure 1: An overview of SCIC framework.

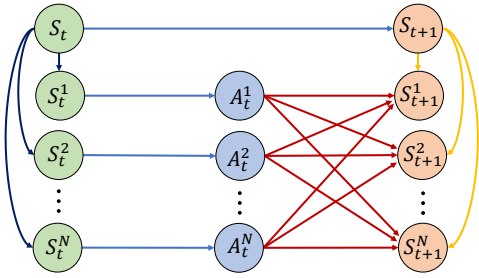Figure 2: Multi-agent Causal Graph Model.

V_i represents an agent's state component e.g. S_t, or the agent's action component e.g. A_t^1. Apart from the actions computed by the policy $\pi(A_t|S_t)$, within a time step, there are no edges, i.e. no transient effects. In the MACGM, the actions A_t at the t time step are intended to affect S_{t+1} by influencing the environment. To better express the causal relationship between actions and states, considering that the influence of A_t on S_{t+1} is actually the influence on the environment at time $t+1$, thus the influence on the environment can be differentiated into the influence on the state S_{t+1}^j of each agent j at $t+1$ time step. Nevertheless, at most time steps, there should be no instantaneous effects between agents in the world. In particular, an agent's sphere of influence is limited, i.e., its action A can only affect other agents sparsely, which rests on two basic assumptions about the causal structure of the world. First, the multi-agent environment is composed of independent agents, according to the principle of independent causal mechanism (ICM) (Parunak 2018), which states that the generative process of the world is composed of autonomous modules. The second assumption is that latent influences between entities are spatially local and temporally sparse. We can view this as explaining the sparsity mechanism shift hypothesis, which suggests that natural distribution shift will be caused by changes in local mechanisms (Schölkopf et al. 2021). This can usually be traced back to the ICM principle, which states that intervention in one mechanism does not affect other mech-

anisms (Parascandolo et al. 2017). We believe that this is also due to the limited scope of intelligent agent intervention, which limits the breadth and frequency of mechanism changes. Therefore, in this work, we are interested in inferring the influence of the action of agent i over other agents in a particular situation, i.e. a local inter-agent causal model. Next, we provide the following definitions:

Definition 1 *(Controllable State Variable) If the edge $A_t^i \to S_{t+1}^j$ in the graph is "active", S_{t+1}^j is a controllable state variable of A_t^i.*

Definition 2 *(Uncontrollable State Variable) If the edge $A_t^i \to S_{t+1}^j$ in the graph is "inactive", S_{t+1}^j is a uncontrollable state variable of A_t^i.*

Given these definitions, in this work, our aim is to detect whether the state is a "Controllable State Variable" for other agents in a particular situation, i.e. whether the presence of red arrows in Figure 2 is "active".

The Cause of an Influence When is action $A = a$ the cause of outcome $S = s$? Inspired by the "but-for" test, i.e. "Without $A = a$, $S = s$ would not have happened.", we can derive that $A = a$ is a necessary condition for $S = s$ to happen, and when A changes, S also gets a different value. This fits with the algorithmic view of causality: if the value of S is determined by the value of A, then A is the cause of S. The "but-for" test yields potentially counter-intuitive assessments. Considering the influence between two agents located close to each other, the behavior of agent a is regarded as the cause of the influence on another agent b, because different behaviors of agent a will lead to different behavior choices of agent b, thus affecting the next state of b. Algorithmically, the behavior of the agent a needs to be known in order to determine the effect on agent b - all possible behaviors of agent a are considered as causes. This means that we cannot distinguish whether the agent's behavior is the cause, but only whether the agent has a causal influence on other agents in the current situation.

Intervention for Causal Inference As discussed above, the causal relationship between agents depends on the cur-

rent situation rather than the behavior chosen by the agents. In order to detect whether there is a causal relationship between agents, we utilize an intervention method to achieve it. Formally, we define "agent i can causally affect agent j in the current situation" (or "agent i takes control of agent j") if there exists an edge $A_t^i \to S_{t+1}^j$ in causal graph (as shown in Figure 2) under all interventions $do(A_t^i = \pi(a_t^i|s_t^i))$ with π having full support. According to Markov property and the faithfulness assumption of the causal graph model, if $A_t^i \not\perp\!\!\!\perp S_{t+1}^j|S_t^i = s_t^i$, then there must be an unblocked edge from A_t^i to S_{t+1}^j in a causal graph $\mathcal{G}$. Since the path over S_{t+1}^j is blocked by observing S_t^i, while assuming no instantaneous influences, the direct path $A_t^i \to S_{t+1}^j$ is the only possible path. Therefore, in causal graph $\mathcal{G}$, there is an edge $A_t^i \to S_{t+1}^j$ under the intervention $do(A_t^i = \pi(a_t^i|s_t^i))$ if $A_t^i \not\perp\!\!\!\perp S_{t+1}^j|S_t^i$. The following lemma shows that a conclusion taken from an intervention generalizes to numerous interventions with π having full support. The detailed proof is available in our arxiv version (Du et al. 2023). According to the lemma, the conclusion from an intervention can determine whether there is a causal influence between agents.

Lemma 1 *If $A_t^i \not\perp\!\!\!\perp S_{t+1}^j$ holds under an intervention $do(A_t^i := \pi(a_t^i|s_t^i))$, then the dependence holds and the edge $A_t^i \to S_{t+1}^j$ exits under all interventions with π having full support. If $A_t^i \perp\!\!\!\perp S_{t+1}^j$ holds under an intervention $do(A_t^i := \pi(a_t^i|s_t^i))$ with π having full support, then the independence holds and the edge $A_t^i \to S_{t+1}^j$ does not exit under all interventions.*

Causal Influence Detection between Agents   Our goal is to measure state-dependent causal influence between agents, which is linked to the independence $A_t^i \perp\!\!\!\perp S_{t+1}^j|S_t^i = s_t^i$ or dependence $A_t^i \not\perp\!\!\!\perp S_{t+1}^j|S_t^i = s_t^i$. CMI is a well-known measure of dependence, which is proposed to be utilized as a measure of causal influence (CI) between agents. If CMI >0, it suggests that A_t^i is necessary to predict S_{t+1}^j, $A_t^i \not\perp\!\!\!\perp S_{t+1}^j|S_t^i = s_t^i$ is true, and the causal path $A_t^i \to S_{t+1}^j$ exists. However, MI of continuous variables is notoriously difficult to compute in real-world settings. Compared to the variational inference-based approaches, MINE-based algorithms have shown superior performance. Motivated by MINE, our approach learns a neural estimate of MI, which utilizes a lower bound to approximate the MI.

$$CI^{ij} := I(S_{t+1}^j; A_t^i|S_t^i) \tag{2}$$

$$= KL\Big(P_{S_{t+1}^j, A_t^i|s_t^i} \big\| P_{S_{t+1}^j|s_t^i} \otimes P_{A_t^i|s_t^i}\Big) \tag{3}$$

$$= \sup_{T:\Omega \to R} \mathbb{E}_{p(S_{t+1}^j, A_t^i|s_t^i)}[T] - log(\mathbb{E}_{p(S_{t+1}^j|s_t^i)p(A_t^i|s_t^i)}[e^T]) \tag{4}$$

$$\geq \sup_{\psi \in \Psi} \mathbb{E}_{p(S_{t+1}^j, A_t^i|s_t^i)}[T_\psi] - log(\mathbb{E}_{p(S_{t+1}^j|s_t^i)p(A_t^i|s_t^i)}[e^{T_\psi}]). \tag{5}$$

First, the CMI formulation is rewritten as Equation (3) using the Donsker-Varadhan representation (Donsker and Varadhan 2010). The input space Ω is a domain of $\mathbb{R}_d$. The upper bound holds for all functions T such that two expectations are finite. Then, the CMI in the Donsker-Varadhan representation is derived with a lower bound using the compression lemma in the PAC-Bayes literature in Equation (5) (Banerjee 2006). The statistical model T is parameterized by a deep neural network with parameter ψ. In addition, since the data in the off-policy reinforcement learning algorithm stems from a mixture of different policies, the agent's sampling strategy cannot be used for intervention. Fortunately, Lemma 1 has shown that a single policy is sufficient to demonstrate (in-)dependence. Therefore, we select a uniform distribution $\mathcal{U}(\mathcal{A})$ over the action space as the intervention policy.

One thing to note is that the forward dynamics model $p(s_{t+1}^j|s_t^i, a_t^i)$ and the distribution $p(s_{t+1}^j|s_t^i)$ need to be calculated. The following will elaborate on how to calculate these two distributions. Computing these two distributions involves representing complex distributions, computing high-dimensional integrals, and only limited data. In addition, each state s_t^i actually can only be seen once in continuous space. Since methods based on non-parametric estimation do not scale well to high dimensions, we address this issue by learning neural network models with appropriately simplifying assumptions. As shown in Figure 1, we first utilize the sampled data from the buffer to estimate the transition distribution $p(s_{t+1}^j|s_t^i, a_t^i)$, assuming that the transition distribution is a normal distribution. Then, the forward dynamics model is utilized to compute the transition marginal distribution $p(s_{t+1}^j|s_t^i)$ by marginalizing out the actions $p(s_{t+1}^j|s_t^i) = \int \pi(a_t^i|s_t^i)p(s_{t+1}^j|s_t^i, a_t^i)$. Actually, we utilize Monte-Carlo to approximate the mixture $p(s_{t+1}^j|s_t^i) \approx \frac{1}{K}\sum_{k=1}^K p(s_{t+1}^j|s_t^i, a_t^{i,(k)})$, instead of integrals. During the training process, the dynamic model $p(s_{t+1}^j|s_t^i)$ is trained simultaneously with the statistical network T and agents' policy network.

Training with Causal Influence as Intrinsic Reward

With the causal influence estimations introduced in the previous subsection, the goal of the following is to learn the joint policy that maximizes the expected discounted reward, using causal influence between agents as an intrinsic reward. Specifically, agent i receives a joint reward, combining the extrinsic team reward and the intrinsic reward from peer causal influence, that can be represented as

$$r_{t,total}^i = r_{t,ex} + \alpha \sum_{j \neq i} CI^{i,j}, \tag{6}$$

where $CI^{i,j}$ represents the causal influence of A_i^t on S_{t+1}^j. α is a hyper-parameter, which is utilized to balance the intrinsic reward and the extrinsic reward. The intrinsic reward $r_{t,in}^i$ of agent i is represented as $r_{t,in}^i = \alpha \sum_{j \neq i} CI^{i,j}$. In particular, each agent needs to learn a policy to maximize the conventional objective $J(\pi): J(\pi_\theta^i) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_{t,total}^i]$. By maximizing the conventional objective, agent i can take control of other interested agents. In principle, our proposed in-

---

Algorithm 1: Training algorithm

Initialize: The critic networks $\phi = (\phi_1, ..., \phi_n)$, the actor networks $\theta = (\theta_1, ..., \theta_n)$, experience replay buffer $\mathcal{D}$ and target networks $\phi^{'}$, $\theta^{'}$, the parameters of the statistic networks $\mathcal{T}$, the forward dynamic models $f = \{f_i\}_{i=1}^n$ and the state encoder networks e.

1: **while** episode $<M$ **do**
2:     **for** t = (1,...,T) **do**
3:         Collecting $s_{t+1} = \{s_{t+1}^i\}_{i=1}^n$ and extrinsic reward $r_{t,ex}$ by executing joint actions a_t via collecting $a_t^i \sim \pi_{\theta_i}(s_t^i)$.
4:     **end for**
5:     Store episode trajectory $\{s_t, a_t, s_{t+1}, r_{t,ex}\}$ from the multi-agent environment to replay buffer $\mathcal{D}$.
6:     Sample a batch data of transition form buffer $\mathcal{D}$.
7:     Update forward dynamic model for each agent i.
8:     Update $\mathcal{T}$ via Equation 5.
9:     Compute intrinsic reward $r_{t,in}^i$ for each agent i via $r_{t,in}^i = \sum_{j \neq i} CI^{i,j}$.
10:     Update critic networks via Equation 7 and intrinsic reward.
11:     Update actor networks via Equation 8.
12: **end while**

---

trinsic reward mechanism can be combined with different CTDE-based MARL algorithms. In this work, our method is built upon MADDPG. The joint action-value function of each agent i is approximated by the joint action-value network $\mathcal{Q}_\phi$ by minimizing the following loss

$$\mathcal{L}_\mathcal{Q}(\phi_i) = \mathbb{E}_{s_t,a_t,r_t,s_{t+1} \sim D}\big[(\hat{y}_i - \mathcal{Q}_{\phi_i}(s_t,a_t))^2\big], \quad (7)$$

where $\hat{y}_i = r_{t,ex} + r_{t,in}^i + \gamma\mathcal{Q}_{\phi_i'}(s_{t+1}, \pi_{\theta'}(s_{t+1}))$, ϕ_i indicates the parameters of critic network, and ϕ_i' and θ' are the parameters of target networks. Then, the policy of agent i is updated by minimizing the loss

$$\mathcal{L}_\pi(\theta_i) = \mathbb{E}_{s_t \sim \mathcal{D}}\big[-\mathcal{Q}_{\phi_i}(s_t, \pi_\theta(\cdot|s_t))\big], \quad (8)$$

where $\theta = (\theta_1, ..., \theta_n)$ are the parameters of actor networks. The training algorithm is described in Algorithm 1.

Performance Evaluation

To verify the effectiveness of our proposed method, in this section, we conduct extensive experiments to evaluate SCIC on various multi-agent tasks and compare SCIC with state-of-the-art methods. Moreover, we also evaluate the effectiveness of the components of the proposed method by ablation experiments.

Multi-Agent Task Benchmark

We evaluate our proposed approach on three benchmark multi-agent tasks: Partial Observation Cooperative Predator Prey, Cooperative Navigation, and Cooperative Line Control. The benchmarks' environment is implemented in a Multi-Agent Particle Environment ((Lowe et al. 2017)),

where agents can follow a dual integrator dynamic model to move in 2D space. Like Cooperative Navigation, Cooperative Predator Prey is a well-known evaluation task for MARL. The Cooperative Line environment is a complex task environment, where there are M agents and 2 targets, with the goal of allowing agents to evenly distribute themselves on the line between two targets. All algorithms are trained in a Linux server with a 2.30 GHz Xeon(R) CPU and two Nvidia 4090 graphics cards. The learning rates of the critic network and the actor network are set to 0.001. The discount factor γ is set to 0.95. Each episode lasts up to 25 timesteps. To estimate the transition marginal distribution $p(s_{t+1}^j|s_t^i)$, the number K of per Monte-Carlo sample is set to 64.

Baselines

To verify the superiority of our method, we employed three baseline algorithms for comparison, namely PMIC, MAD-DPG and SI, with a relaxation that all SI agents are equipped with the social influence reward. Among them, PMIC and SI are currently state-of-the-art MARL algorithms using intrinsic rewards in Multi-Agent Particle Environment. Since our proposed algorithm is based on MADDPG, we name it SCIC-MADDPG in the experiments and MADDPG is also utilized as a baseline algorithm for comparison. Additionally, to verify the effectiveness of the components of the proposed method, some ablation studies also will be provided. Furthermore, considering the significant role played by the temperature parameters of intrinsic reward in balancing the relative importance between extrinsic and intrinsic reward, we also provide an experimental study on the temperature parameters.

Experimental Results

First, we evaluate the performance of our approach and other baseline approaches on Cooperative Predator Prey with 3, 4, and 5 predators, in which the policy of predators needs to be trained and the policy of prey is fixed. Figure 3(a), Figure 3(b), and Figure 3(c) illustrate the comparison results of rewards for SCIC-MADDPG, PMIC, SI, and MD-DDPG on Predator-Prey tasks with 3, 4, 5 predators, respectively. We can observe that our SCIC-MADDPG demonstrates better performance compared to other algorithms, although there are relatively minor advantages when there are 4 predators in the environment. SCIC-MADDPG converges to a better reward than MADDPG, which indicates that the intrinsic reward is indeed conducive to improving the cooperation among agents. Both PMIC and SI receive lower rewards, suggesting that promoting agents to reach *significant states* is more beneficial for cooperation between agents than merely focusing on behavior coordination. From another perspective, enhancing the causal influence between agents is evidently more conducive to cooperation between agents than solely focusing on enhancing their correlation.

Moreover, we further perform evaluations on the Cooperative Navigation task with 3, 4, and 5 robots, which requires each agent to reach a distinct landmark to achieve the shortest total distance traveled. From Figure 3(d), Fig-

(a) Predator Prey (3 agents).   (b) Predator Prey (4 agents)

(c) Predator Prey (5 agents).   (d) Cooperative Navi. (3 agents).

(e) Cooperative Navi. (4 agents). (f) Cooperative Navi. (5 agents).

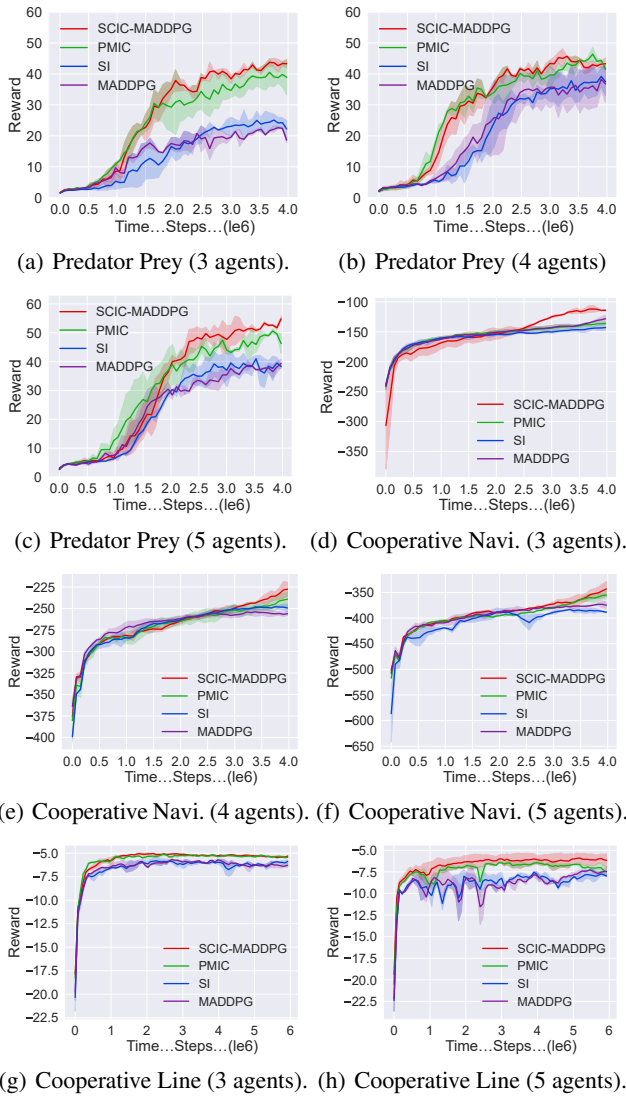(g) Cooperative Line (3 agents). (h) Cooperative Line (5 agents).

Figure 3: Performances comparison of SCIC-MADDPG with the other three approaches in various multi-agent tasks.

ure 3(e), and Figure 3(f), we can observe that our SCIC-MADDPG consistently achieves higher rewards than the other baseline methods in the Cooperative Navigation tasks with varying numbers of agents. Furthermore, we evaluate the performance of SCIC-MADDPG on more challenging Control Line tasks involving 3 and 5 agents, representing an exceptionally difficult task scenario. The experimental findings illustrated in Figure 3(h) reveal that SCIC-MADDPG outperforms other methods in the context of a 5-agents scenario. Additionally, Figure 3(g) illustrates that SCIC-MADDPG achieves superior performance in comparison to SI and MADDPG, while achieving comparable performance to PMIC. In summary, the results of these experiments collectively indicate that SCIC-MADDPG proves to be an effective approach, consistently outperforming other methods across various tasks.
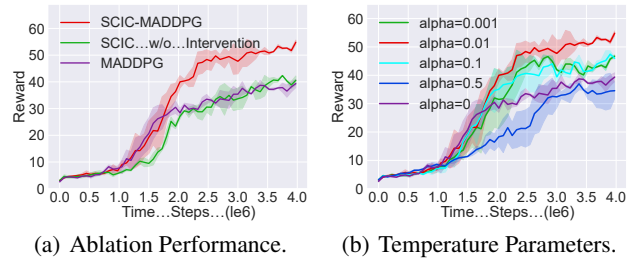

(a) Ablation Performance.   (b) Temperature Parameters.

Figure 4: Ablation Study.

Ablation Study

To further evaluate our proposed approach, we further conducted a series of ablation experiments in this subsection. Specifically, to evaluate the effectiveness of intervention sampling, we implemented the SCIC w/o Intervention method, which obtains the action set when estimating causal effects between agents, not through intervention with a uniform distribution, but by sampling from the replay buffer. The ablation study was conducted on the Predator Prey task with 5 agents. As shown in Figure 4(a), SCIC-MADDPG performs significantly better than MADDPG and SCIC w/o Intervention. The key reason behind this achievement can be attributed to the fact that the sampling actions in the off-policy reinforcement learning algorithm originate from a mixture of different policies, which cannot be utilized to estimate causality. The results reveal two facts: 1) Using causal influence between agents as a Reward Bonus contributes to enhancing the performance of MARL algorithms by facilitating cooperation among agents; 2) The adoption of intervention leads to more accurate estimates of causal influence.

Temperature Parameters α The role of temperature parameters α is to control the relative importance between intrinsic and extrinsic reward. We evaluate SCIC-MADDPG by varying α values within the range of [0, 0.001,0.01,0.1,0.5] in the Predator Prey task with 5 predators. As illustrated in Figure 4(b), SCIC-MADDPG with a temperature value of 0.01 outperforms its performance with other temperature values.

Conclusion

To promote coordination between agents and encourage exploration, we propose the SCIC approach, which incorporates a new intrinsic reward mechanism based on a new cooperation criterion measured by situation-dependent causal influence between agents. SCIC encourages agents to explore states that positively affect other agents by detecting inter-agent causal influences and utilizing them as intrinsic rewards, thereby enhancing collaboration and overall performance. We conduct comprehensive experiments to evaluate the performance of our proposed approach across various cooperative MARL tasks. Extensive experimental results prove the effectiveness of our SCIC. In the future, we expect to further extend SCIC to decentralized training-based MARL algorithms and model-based MARL algorithms.

Acknowledgments

References

Banerjee, A. 2006. On bayesian bounds. In *Proceedings of the 23rd international conference on Machine learning*, 81–88.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Blaes, S.; Pogancic, M. V.; Zhu, J.; and Martius, G. 2019. Control What You Can: Intrinsically Motivated Task-Planning Agent. In *NeurIPS*, 12520–12531.

Chalk, M.; Marre, O.; and Tkacik, G. 2016. Relevant sparse codes with variational information bottleneck. In *NIPS*, 1957–1965.

Chen, L.; Guo, H.; Du, Y.; Fang, F.; Zhang, H.; Zhang, W.; and Yu, Y. 2021. Signal Instructed Coordination in Cooperative Multi-agent Reinforcement Learning. In *DAI*, volume 13170 of *Lecture Notes in Computer Science*, 185–205. Springer.

Donsker, M. D.; and Varadhan, S. R. S. 2010. Asympototic evalution of certain Markov process expectations for large time, IV. *Communications on Pure & Applied Mathematics*, 36(2): 183–212.

Du, X.; Ye, Y.; Zhang, P.; Yang, Y.; Chen, M.; and Wang, T. 2023. Situation-Dependent Causal Influence-Based Cooperative Multi-agent Reinforcement Learning. arXiv:2312.09539.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual Multi-Agent Policy Gradients. In *AAAI*, 2974–2982. AAAI Press.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*. OpenReview.net.

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; Turck, F. D.; and Abbeel, P. 2016. VIME: Variational Information Maximizing Exploration. In *NIPS*, 1109–1117.

Iqbal, S.; and Sha, F. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2961–2970. PMLR.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gülçehre, Ç.; Ortega, P. A.; Strouse, D.; Leibo, J. Z.; and de Freitas, N. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 3040–3049. PMLR.

Kim, W.; Jung, W.; Cho, M.; and Sung, Y. 2023. A Variational Approach to Mutual Information-Based Coordination for Multi-Agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, 40–48. ACM.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A. A. A.; Yogamani, S. K.; and Pérez, P. 2022. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.*, 23(6): 4909–4926.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear Information Bottleneck. *Entropy*, 21(12): 1181.

Lee, T. E.; Zhao, J. A.; Sawhney, A. S.; Girdhar, S.; and Kroemer, O. 2021. Causal Reasoning in Simulation for Structure and Transfer Learning of Robot Manipulation Policies. In *ICRA*, 4776–4782. IEEE.

Li, P.; Tang, H.; Yang, T.; Hao, X.; Sang, T.; Zheng, Y.; Hao, J.; Taylor, M. E.; Tao, W.; and Wang, Z. 2022. PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 12979–12997. PMLR.

Liu, M.; Zhou, M.; Zhang, W.; Zhuang, Y.; Wang, J.; Liu, W.; and Yu, Y. 2020. Multi-agent interactions modeling with correlated policies. *arXiv preprint arXiv:2001.03415*.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6382–6393.

Maes, S.; Meganck, S.; and Manderick, B. 2007. Inference in multi-agent causal models. *International Journal of Approximate Reasoning*, 46(2): 274–299.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. MAVEN: Multi-Agent Variational Exploration. In *NeurIPS*, 7611–7622.

Mohamed, S.; and Rezende, D. J. 2015. Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. In *NIPS*, 2125–2133.

Oliehoek, F. A.; Amato, C.; et al. 2016. *A concise introduction to decentralized POMDPs*, volume 1. Springer.

Parascandolo, G.; Rojas-Carulla, M.; Kilbertus, N.; and Schölkopf, B. 2017. Learning Independent Causal Mechanisms. *CoRR*, abs/1712.00961.

Parunak, H. V. D. 2018. Elements of causal inference: foundations and learning algorithms. *Computing reviews*, 59(11): 588–589.

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 4295–4304. PMLR.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *CoRR*, abs/2102.11107.

Seitzer, M.; Schölkopf, B.; and Martius, G. 2021. Causal Influence Detection for Improving Efficiency in Reinforcement Learning. In *NeurIPS*, 22905–22918.

Shanmugam, R. 2001. Causality: Models, Reasoning, and Inference : Judea Pearl; Cambridge University Press, Cambridge, UK, 2000, pp 384, ISBN 0-521-77362-8. *Neurocomputing*, 41(1-4): 189–190.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V. F.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; and Graepel, T. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *AAMAS*, 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM.

Velickovic, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2019. Deep Graph Infomax. In *ICLR (Poster)*. OpenReview.net.

Wang, T.; Du, X.; Chen, M.; and Li, K. 2023. Hierarchical Relational Graph Learning for Autonomous Multirobot Cooperative Navigation in Dynamic Environments. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(11): 3559–3570.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2020. Influence-Based Multi-Agent Exploration. In *ICLR*. OpenReview.net.

Wen, Y.; Yang, Y.; Luo, R.; Wang, J.; and Pan, W. 2019. Probabilistic Recursive Reasoning for Multi-Agent Reinforcement Learning. In *ICLR (Poster)*. OpenReview.net.

Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; and Wu, D. O. 2020. Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks. *IEEE Transactions on Vehicular Technology*, 69(8): 8243–8256.