Rethinking Credit Assignment in Cooperative MARL via Interventional Reward Response

Chamjin Joo[1] **Myoung Hoon Ha**[2] **Sang Wan Lee**[1,2,3,4]

[1]Kim Jaechul Graduate School of AI, [2]Center for Neuroscience-inspired AI,
[3]Department of Brain and Cognitive Sciences, [4]Graduate School of Data Science
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
chamjin.joo@kaist.ac.kr, mh.ha.soar@gmail.com, sangwan@kaist.ac.kr

Abstract

Credit assignment remains a central challenge in cooperative multi-agent reinforcement learning (MARL), especially under partial observability, where individual policy updates may not accurately reflect each agent's actual contribution to team outcomes. While policy-based methods such as MAPPO and IPPO provide strong optimization frameworks, their updates are typically derived from global rewards and observational value surrogates, without explicitly defining agent-specific credit. Misaligned credit signals can mislead individual policy improvement, resulting in inefficient coordination and weaker team performance. We address this challenge by formulating agent-level credit as an *interventional reward response* (IRR), using Proximal Causal Inference (PCI) to identify credit from observable proxies via an outcome bridge function. Building on this identification strategy, we design a practical credit-aligned update signal and integrate it into policy gradient methods. Empirical evaluations on diagnostic and benchmark tasks demonstrate that the proposed credit signal improves policy learning under partial observability, highlighting proximal identification as a promising foundation for designing credit-aware policy updates in cooperative MARL. To the best of our knowledge, this work presents the first PCI-based solution for online multi-agent cooperation. Our code is available at `https://github.com/chamjinJoo/credit_assign_IRR`.

1 Introduction

Cooperative multi-agent reinforcement learning (MARL) studies how multiple agents learn to coordinate toward a shared objective under local views of the environment. Such settings capture a core issue in the control of complex systems, including multi-robot coordination [41], traffic signal control [38], and large-scale resource allocation in communication networks [45], where multiple agents must collaborate despite limited information. Despite notable success across benchmark tasks, standard policy-based approaches still rely on update signals constructed from shared global rewards and observational value surrogates [5, 12]. These signals often fail to align with each agent's actual contribution to team outcomes [37, 43]. This issue becomes more pronounced under partial observability, where local observations alone may not sufficiently capture the decision context. Misaligned signals ultimately hinder efficient policy improvement, leading to suboptimal policies or slower convergence [11, 47]. Existing approaches enhance coordination through value decomposition [28, 34], centralized critics [42], or counterfactual baselines [11], but they do not fully accommodate agent-specific credit through an interventional reward target. That is, they do not explicitly capture the effect of changing a focal agent's action on the global reward.

This mismatch is especially pronounced when the reward-generation mechanism is not exchangeable across agents. In many cooperative settings, agents do not contribute symmetrically to team

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

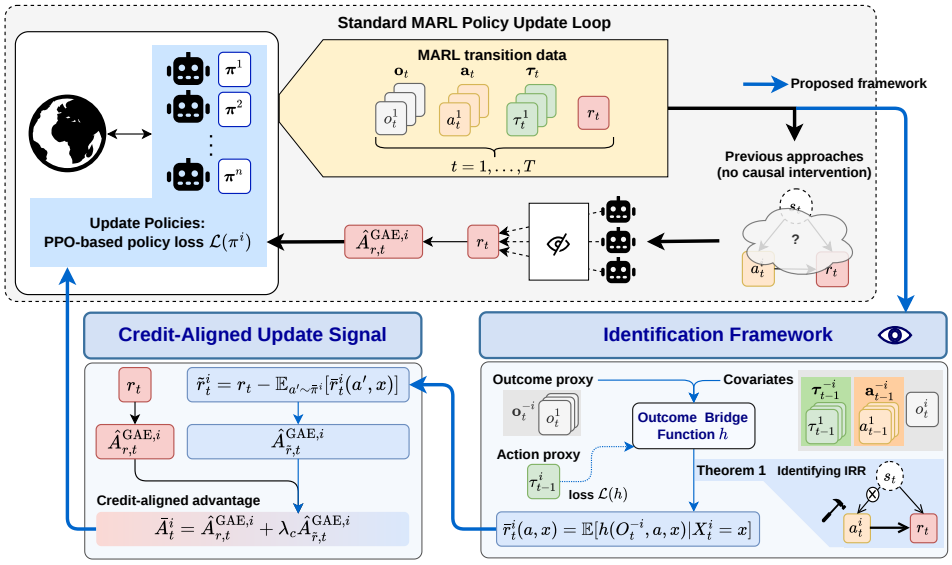


Figure 1: Overview of the proposed framework. The gray block illustrates the standard cooperative MARL policy-update loop without explicit consideration of the underlying causal structure. **Identification framework**: MARL transition data are organized into proxy variables and covariates to train an outcome bridge function, which identifies the IRR under Theorem 1 (Section 5). **Credit-aligned update signal**: The estimated IRRs are used to construct a counterfactual baseline and the credit-aligned reward $\tilde{r}_t^i$, which yields an agent-wise credit-aligned advantage for policy updates.

outcomes: their influence may depend on role, locality, admissible interaction sets, or coordination compatibility with others. In such cases, the shared global reward can remain informative about collective performance while still being highly ambiguous about which agent's action actually contributed to that outcome. This makes agent-wise credit assignment fundamentally interventional rather than something that can be resolved from shared reward alone. To resolve this issue, we define agent-specific credit in terms of intervention. In a decentralized partially observable Markov decision process (Dec-POMDP), the global reward is generated by the unobserved underlying state and the joint action. Consequently, the observed reward alone does not reveal how the team's outcome would change if a single agent's action were to change [25]. This motivates defining agent-wise credit via an interventional reward response (IRR)—how the global reward responds to intervening on the agent's action—rather than relying solely on observational reward associations. To operationalize this perspective, we leverage Proximal Causal Inference (PCI; [21, 35]), a principled framework for identifying IRRs from observable proxy variables when the reward-generating state is not observed. In the Dec-POMDP setting, we instantiate agent-specific proxies and identify the target interventional reward via an outcome-bridge function. Figure 1 describes the entire framework.

Our contributions are as follows:

- **Interventional Credit**: We formulate agent-specific credit under partial observability as an IRR and a corresponding contrastive credit quantity.
- **PCI-based identification**: We develop a PCI-based framework in Dec-POMDPs that identifies the IRR through an outcome-bridge function and operationalizes it as a practical credit-aligned update signal for online policy-based cooperative MARL.
- **Empirical Validation**: We empirically validate the method on diagnostic and benchmark domains, demonstrating improved policy learning over MAPPO/IPPO backbones and verifying the behavior of the learned outcome bridge and credit-aligned reward.

2 Related Work

Multi-Agent Reinforcement Learning. Representative cooperative MARL methods include value-factorization approaches such as VDN [34] and QMIX [28], as well as PPO-based baselines such as MAPPO [42] and IPPO [7]. These methods achieve strong empirical performance, but still learn from shared team rewards and rollout-based value surrogates rather than explicitly defining agent-wise credit. COMA [11] introduces a counterfactual baseline for multi-agent credit assignment, yet it still relies on the observed global reward signal. In contrast, we formulate agent-wise credit as an interventional target and identify it through a proxy-based causal framework.

Causal Multi-Agent Reinforcement Learning. Recent research has sought to integrate causal inference methods with the powerful learning methods of RL, a line of work referred to as causal reinforcement learning (CRL) [2, 18]. CRL studies how intervention semantics, structural assumptions, or causal representations can improve policy learning, robustness, or interpretability in sequential decision-making. Most CRL studies focus on single-agent settings, including causal imitation and confounding-aware offline RL [44]. In multi-agent systems, however, explicit causal treatments of credit assignment remain relatively limited. The approaches closest to ours are deconfounded value decomposition (DVD) [17] and its follow-up studies [19, 40]. Our work differs in both target and methodology: rather than modifying the value decomposition, we focus on identifying agent-wise credit through an interventional reward target under partial observability.

Reinforcement Learning with Proximal Causal Inference. The most relevant prior works extend PCI to sequential decision-making within a single-agent domain. Bennett and Kallus [4] tackle off-policy evaluation in a partially observable MDP (POMDP) framework by developing proximal identification strategies for applications of offline RL to observational data. Hong et al. [13] propose a policy gradient method that identifies causal policy gradients under hidden confounders in offline settings. Shen and Cui [33] focus on the learning of optimal individualized treatment regimes (ITRs) for efficient estimation of a value function. These works predominantly focus on offline or off-policy settings motivated by observational domains such as healthcare, rather than cooperative MARL training. To the best of our knowledge, prior PCI-based sequential decision-making works have not treated practical cooperative MARL training as the primary setting. In contrast, we use PCI to identify agent-wise credit and turn it into a practical update signal for cooperative MARL training.

3 Background

Dec-POMDPs. A Dec-POMDP serves as a cooperative multi-agent decision-making framework. Formally, a Dec-POMDP is defined as a tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathbb{T}, \mathbb{O}, \mathcal{R}, \gamma \rangle$. $\mathcal{N} = \{1, \ldots, n\}$ is the set of agent indices and $\mathcal{S}$ is the discrete state space. $\mathcal{O} = \times_{i \in \mathcal{N}} \mathcal{O}^i$, and $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}^i$ are the joint observation and joint action spaces. At each timestep t, each agent $i \in \mathcal{N}$ chooses an action $a_t^i \in \mathcal{A}^i$ according to its local policy π^i, forming a joint action $\mathbf{a}_t = \{a_t^i\}_{i=1}^n$. The environment then transitions from the current state $s_t \in \mathcal{S}$ to the next state $s_{t+1} \in \mathcal{S}$ under the state transition probability function $\mathbb{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ specifying the probability $\mathbb{T}(s_{t+1}|s_t, \mathbf{a}_t)$. This transition also yields a global reward according to the reward function $\mathcal{R}$, where $r_t = \mathcal{R}(s_t, \mathbf{a}_t)$ denotes the global reward shared by all agents. $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ denotes the probability that a joint observation $\mathbf{o}_t = (o_t^1, \ldots, o_t^n) \in \mathcal{O}$ is observed by agents $1, \ldots, n$ respectively for a given state s_t. We assume conditional independence across agents, so that $\mathbb{O}(\mathbf{o}_t|s_t) = \prod_{i \in \mathcal{N}} \mathbb{O}(o_t^i|s_t)$. Since the environment state is not directly observed, decentralized policies are generally history-dependent. A local policy $\pi^i(\cdot|\tau_t^i)$ conditions on the local history $\tau_t^i = (o_1^i, a_1^i, \ldots, o_t^i)$. Finally, $\gamma \in (0, 1]$ is the discount factor. The objective is to maximize the expected return, $\mathbb{E}[\sum_{t=1}^T \gamma^{t-1} r_t]$.

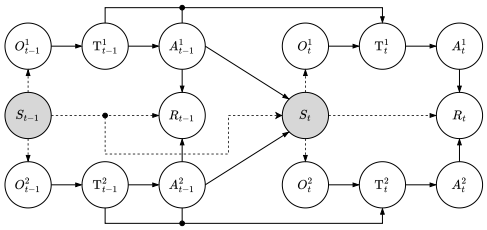


Figure 2: The causal graph of a Dec-POMDP with two agents. Shaded nodes indicate unobserved variables, and dashed edges represent causal relationships from unobservable variables. T_t^i denotes the random variable corresponding to the action-observation history of agent i up to timestep t, with a realization $\tau_t^i = (o_1^i, a_1^i, \ldots, o_t^i)$.

MAPPO and IPPO. As a standard policy gradient method, PPO ensures stable and efficient policy updates by approximating trust-region methods using a clipped surrogate loss [32]. MAPPO extends PPO to centralized-critic MARL settings [42], resulting in a shared advantage for all $i \in \mathcal{N}$. In contrast, IPPO uses local value networks conditioned solely on the encoded local history. Both optimize the PPO clipped surrogate objective of the joint policy $\boldsymbol{\pi}$,

$$\mathcal{L}(\boldsymbol{\pi}) = \mathbb{E}_{i,t}\Big[\min\big(\rho_t^i \hat{A}_t^i,\ \mathrm{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon)\hat{A}_t^i\big)\Big],\ \rho_t^i = \frac{\pi^i(a_t^i \mid \tau_t^i)}{\pi^i_{\mathrm{old}}(a_t^i \mid \tau_t^i)}. \tag{1}$$

$\mathbb{E}_{i,t}[\cdot]$ denotes the empirical mean over agents $i \in \mathcal{N}$ and batch $\mathcal{B}$. The $\mathrm{clip}(\cdot, 1 \pm \epsilon)$ operator clips the input to $[1-\epsilon, 1+\epsilon]$ if it is below/above the range. $\hat{A}_t^i$ denotes the advantage estimated via GAE [31]. The crucial difference lies in how $\hat{A}_t^i$ is estimated.

Intervention in Causal Inference. Causal inference is the process of understanding the independent and genuine impact of a particular variable within a complex system. The pivotal constructs of causal inference encompass association and intervention. An *association* is an observational relationship in which correlations between variables are observed without causal manipulation. An *intervention* involves deliberately manipulating a variable while severing all incoming pathways directed toward it, leaving the rest of the system unchanged, thereby facilitating the examination of its influence on an outcome. To introduce interventions within the Dec-POMDP causal graph (Figure 2), we adopt Pearl's do-operator [27, Sections 1.3.1 and 3.2], where an intervention on agent i's action is denoted as $do(A_t^i = a)$. The do-operator means that all arrows directed toward the action variable A_t^i are cut off, and A_t^i is fixed to the value a. In our Dec-POMDP setting, $do(A_t^i = a)$ denotes a one-step intervention that overrides only agent i's action at time t, while the environment dynamics and other agents' actions continue to follow their decentralized policies.

4 Credit Assignment Requires an Interventional View

4.1 Global Reward Conflates Agent-Specific Credits

In Dec-POMDPs, the global reward is determined by a function of the unobserved state and the joint action, and the same reward is given to all agents. Therefore, the realized global reward conflates the influences of the unobserved state and the actions of all agents, creating functional ambiguity for agent-wise credit assignment. In standard cooperative policy-based MARL, the shared global reward is used to construct individual update signals, such as value or advantage estimates. Although such shared update signals are effective for optimization, they do not explicitly isolate each agent's contribution to the team outcome [39]. Prior work has shown that learning from the global reward under partial observability can induce spurious rewards [11] and the lazy-agent problem [34], and more generally make it difficult for each agent to infer its own contribution to the team outcome [47].

4.2 Credit Assignment via Interventional Reward Response

What is missing, then, is not the team objective itself, but a principled agent-wise target for translating that shared reward into individual credit. The key question for agent-wise credit is how the global reward would change if the focal agent's action were changed under the same decision situation. We refer to the fixed information available before decision-making as the pre-action context. In other words, credit should be assigned to an action for its effect on the global reward, rather than for being merely associated with a realized reward. This is inherently a causal question, which motivates defining agent-wise credit through an interventional reward response rather than relying solely on realized global rewards and observational value surrogates.

Agent-specific interventional reward response. We first define the agent-specific interventional reward response (**IRR**) under a fixed $X_t^i = x$, where X_t^i denotes the pre-action context used to assign credit to agent i. In Section 5, we specify it as the focal current observation, other agents' histories, and other agents' previous actions.

$$\bar{r}_t^i(a, x) \triangleq \mathbb{E}\left[R_t \mid do(A_t^i = a), X_t^i = x\right]. \tag{2}$$

We refer to this quantity as the IRR, following the standard causal notion of a response function under intervention. Under partial observability, conditioning on such a history-based context is natural, since the current observation alone may not fully characterize the decision situation. It also allows us to define the reward response under a fixed pre-action context.

Interventional credit. Based on Eq. (2), we can further define the credit of agent i in the form of a treatment effect,

$$\delta_t^i(a, a', x) \triangleq \bar{r}_t^i(a, x) - \bar{r}_t^i(a', x), \tag{3}$$

where $a' \in \mathcal{A}^i \setminus \{a\}$. This contrast makes agent-wise credit explicit by comparing two candidate actions. However, $\bar{r}_t^i(a, x)$ and $\delta_t^i(a, a', x)$ are not directly recoverable from observational rollouts since they are defined through interventions on agent i's action under a latent reward-generation process. In cooperative MARL, trajectories provide only a partial proxy view of this process. This motivates proximal causal inference, which identifies the desired IRR from observable proxies via bridge functions. We develop this identification rationale in Section 5. The identified contrast (3) will serve as the basis for constructing a practical update signal in Section 6.

5 Proximal Identification of Interventional Credit under Dec-POMDP

The goal of proximal identification is to recover an interventional quantity of interest from observational data when the relevant reward-generation state is not directly observed. This is necessary because the interventional credit contrast in Eq. (3) is defined through the IRR $\bar{r}_t^i$, which is not directly available from rollout trajectories. To identify $\bar{r}_t^i$, we instantiate the PCI framework under the Dec-POMDP formulation. In essence, rather than requiring direct access to the hidden state, PCI uses observable proxies to identify causal quantities in the presence of unobserved confounding.

Table 1: Mapping PCI variables to Dec-POMDPs

PCI Concept	Notation	Dec-POMDP Variables
Hidden Confounder	U	S_t
Outcome	Y	R_t
Action/Treatment	A	A_t^i
Outcome-side proxy	W	O_t^{-i}
Action-side proxy	Z	T_{t-1}^i
Covariates	X	$(O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$

Action- and outcome-side proxies. Our interest is in how intervening on agent i's action A_t^i affects the corresponding outcome R_t while S_t is unobservable. For agent i, the observable variables around this intervention are $(O_t^i, O_t^{-i}, A_t^i, A_t^{-i}, \mathrm{T}_t^i, \mathrm{T}_t^{-i}, R_t)$ for $t \in \{1, \ldots, T\}$. Building on the causal relationships visualized in Figure 2, we consider the following conditional independencies for all $i \in \mathcal{N}$ and $t \in \{1, \ldots, T\}$. For $t = 1$, we set $\mathrm{T}_0^i \triangleq \emptyset$ and $A_0^i \triangleq \emptyset$ for all $i \in \mathcal{N}$.

$$\mathrm{T}_{t-1}^i \perp\!\!\!\perp R_t \mid S_t, A_t^i, X_t^i \quad \text{and} \quad O_t^{-i} \perp\!\!\!\perp (A_t^i, \mathrm{T}_{t-1}^i) \mid S_t, X_t^i. \tag{4}$$

These two properties formalize the validity of T_{t-1}^i and O_t^{-i} as the action-side (Z) and outcome-side (W) proxies, respectively, for all i and t, which serve as observable proxies for the unobserved state. The first condition states that T_{t-1}^i must not directly affect the outcome R_t other than through S_t, A_t^i, and X_t^i. Similarly, the second condition indicates that O_t^{-i} must not affect A_t^i or T_{t-1}^i other than through S_t and X_t^i. We define the pre-action context as $X_t^i \triangleq (O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$ and find that Eq. (4) holds under our definitions. Table 1 summarizes our definitions. Note that this partition of observed variables into action-/outcome-side proxies and X_t^i is not unique; rather, it is chosen to make the PCI identification assumptions more plausible in the Dec-POMDP setting. In addition to the proxy-validity conditions, we assume standard identification assumptions such as consistency, latent exchangeability, and positivity; their formal statements are deferred to Appendix C.1.

Outcome bridge function. Motivated by [15], we introduce an outcome bridge set $\mathbb{H}_t^i$ and its observable counterpart $\mathbb{H}_t^{obs,i}$ to identify the IRR. An outcome bridge function is not introduced to predict reward directly. Its role is to satisfy a proxy-based conditional moment restriction, thereby enabling identification of the IRR under partial observability.

Definition 1. *(Outcome bridge set and observed outcome bridge set). Define*

$$\mathbb{H}_t^i = \{h : \mathbb{E}\left[R_t - h(O_t^{-i}, A_t^i, X_t^i) \mid S_t, A_t^i, X_t^i\right] = 0\}, \tag{5}$$

where h is a square-integrable scalar-valued function. The observed outcome bridge set is

$$\mathbb{H}_t^{obs,i} = \{h : \mathbb{E}\left[R_t - h(O_t^{-i}, A_t^i, X_t^i) \mid \mathrm{T}_{t-1}^i, A_t^i, X_t^i\right] = 0\}, \tag{6}$$

where $\mathbb{H}_t^{obs,i}$ replaces the unobserved-state restriction in (5) *with a moment condition involving only proxies and observed variables* (6)*, making bridge learning possible from observational trajectories.*

Given (4) and the law of iterated expectations, any outcome bridge is also an observed bridge: $\mathbb{H}_t^i \subseteq \mathbb{H}_t^{obs,i}$ [15]. To ensure that learning from the observed-data restriction yields a *valid* outcome bridge, we impose a **completeness** assumption: For any $g \in L^2$ and for all i and $t \geq 2$,

$$\mathbb{E}\big[g(S_t) \mid \mathrm{T}_{t-1}^i, A_t^i = a, X_t^i = x\big] = 0 \text{ a.s.} \iff g(S_t) = 0 \; a.s., \; \forall a \in \mathcal{A}^i, \; \forall x \in \mathrm{supp}(X_t^i). \tag{7}$$

At $t = 1$, it is not generally guaranteed, so $t = 1$ is excluded (see Appendix C.2). As a standard identifiability condition in causal inference, the completeness condition characterizes the informativeness of the auxiliary variable about the confounder and its ability to recover the latent-state-induced bias [22]. The burden of completeness in our setting falls on the informativeness of T_{t-1}^i beyond $X_t^i = (O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$. Intuitively, it requires the action-side proxy to preserve enough S_t-related information beyond X_t^i to distinguish reward-relevant latent structure. While we do not attempt to verify completeness for arbitrary Dec-POMDPs, this assumption is adopted in various works, such

as [4, 13]. Moreover, Hong et al. [13] mention that many commonly used conditional models like exponential families [23] and location-scale families [14] satisfy completeness, and in finite/discrete settings it reduces to familiar full-rank conditions on conditional probability operators. We discuss this in Appendix C.2.

Under this completeness condition (Eq. (7)), the observed bridge restriction is sufficient to recover valid outcome bridges. Proposition 1 states this formally.

Proposition 1 (Validation of observed bridges under completeness). *For any agent $i \in \mathcal{N}$ and timestep $t \in \{2, \dots, T\}$, suppose Assumptions C.1–C.5 hold. If $\mathbb{H}_t^i \neq \emptyset$, then the observed outcome-bridge set $\mathbb{H}_t^{\mathrm{obs},i}$ coincides with $\mathbb{H}_t^i$, i.e., $\mathbb{H}_t^{\mathrm{obs},i} = \mathbb{H}_t^i$.*

In particular, any $h \in \mathbb{H}_t^{\mathrm{obs},i}$ satisfies $\mathbb{E}\big[R_t - h(O_t^{-i}, A_t^i, X_t^i) \mid S_t, A_t^i, X_t^i\big] = 0$, so learning from the observed-data restriction yields a valid outcome bridge. The proof is given in Appendix D. Proposition 1 ensures that, under completeness, a function learned from the observed-data restriction is a valid outcome bridge. We can therefore identify the IRR introduced in Section 4 via the bridge functional, as stated in the following theorem.

Theorem 1 (Identification of the interventional reward response). *For any agent $i \in \mathcal{N}$ and timestep $t \in \{2, \dots, T\}$, suppose Assumptions C.1–C.5 hold. If the outcome-bridge set $\mathbb{H}_t^i$ is nonempty, then for any $a \in \mathcal{A}^i$, any $x \in \mathrm{supp}(X_t^i)$, and any $h \in \mathbb{H}_t^{\mathrm{obs},i}$,*

$$\bar{r}_t^i(a, x) = \mathbb{E}\big[h(O_t^{-i}, a, x) \mid X_t^i = x\big]. \tag{8}$$

The proof is given in Appendix D. Theorem 1 identifies the IRR from observational trajectories for $t \geq 2$ using proxy variables; at $t = 1$, we use the bridge-based signal as a practical approximation without a formal guarantee. This formulation grounds agent-wise credit in the IRR rather than in rewards merely observed together with an action, providing a principled basis for the credit-aligned update signals in Section 6.

6 From Proximal Identification to Credit Assignment

Building on the bridge-based identification result in Theorem 1, we now show how to (1) learn an outcome bridge function and (2) convert the identified IRR into a practical update signal, using MAPPO and IPPO as the backbone MARL frameworks.

6.1 Training the Outcome Bridge Function

A centralized outcome-bridge network h_ϕ is trained to satisfy the observed-data moment restriction in Eq. (6) using the minimax approach of Dikkala et al. [9] and Kallus et al. [15]. Following the variational formulation of conditional moment problems [3, 16, 46], this restriction can be rewritten as a tractable minimax objective over the bridge function and an adversarial critic,

$$\hat{h} \in \arg\min_h \max_{\nu \in \mathcal{H}} \mathbb{E}_{i,t}\big[\nu(\tau_{t-1}^i, a_t^i, x_t^i)\big(h(o_t^{-i}, a_t^i, x_t^i) - r_t\big) - \lambda\nu^2(\tau_{t-1}^i, a_t^i, x_t^i)\big], \tag{9}$$

where $\nu \in \mathcal{H}$ is an adversarial critic function taken from a *Reproducing Kernel Hilbert Space* (RKHS), which is induced by an RBF kernel. The quadratic penalty $\lambda\nu^2$ is a stabilizer to improve training stability and convergence rates without introducing estimation bias [9, 15]. A derivation of this minimax form and additional discussion of the stabilizer are deferred to Appendix C.3.

With this RKHS choice for the critic and a Tikhonov regularization coefficient $\gamma_{st} > 0$, the inner maximization admits a closed-form solution, yielding the following outer objective in matrix form [15, 20]:

$$\hat{h} \in \arg\min_h \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \frac{1}{|\mathcal{B}|} \mathbf{e}_i^\top \boldsymbol{K}_i^{1/2} \left(\gamma_{st} \boldsymbol{I} + \lambda \boldsymbol{K}_i\right)^{-1} \boldsymbol{K}_i^{1/2} \mathbf{e}_i, \tag{10}$$

where $\mathbf{e}_i \in \mathbb{R}^{|\mathcal{B}|}$ is the residual vector, whose k-th component is $h(o_{t_k}^{-i}, a_{t_k}^i, x_{t_k}^i) - r_{t_k}$. Here, the coefficient λ is set to 1 for all experiments in Section 7. To construct the Gram matrix, we use a product of RBF kernels over the three inputs $(\tau_{t-1}^i, a_t^i, x_t^i)$. Specifically, τ_{t-1}^i and x_t^i are normalized block-wise, while a_t^i is represented by a one-hot vector. We then form separate Gram matrices $\boldsymbol{K}_{\tau,i}$, $\boldsymbol{K}_{a,i}$, and $\boldsymbol{K}_{x,i}$, and define the final Gram matrix as

$$\boldsymbol{K}_i = \boldsymbol{K}_{\tau,i} \odot \boldsymbol{K}_{a,i} \odot \boldsymbol{K}_{x,i}. \tag{11}$$

Note that $\boldsymbol{K}_{x,i}$ is decomposed again since X_t^i is composed of three features, O_t^i, T_{t-1}^{-i}, and A_{t-1}^{-i}. The bandwidth parameters are set according to the corresponding input dimensions (see Appendix F.1).

6.2 Agent-Specific Credit Assignment

How to operationalize the interventional targets for policy update. Based on the identified $\bar{r}_t^i(a,x)$ and $\delta_t^i(a,a',x)$, we construct practical bridge-based signals for policy optimization, rather than plug-in estimators of the full theoretical contrast. Replacing the unknown valid bridge in Eq. (8) with the learned network h_ϕ, the IRR is approximated as

$$\bar{r}_t^i(a,x) \approx \mathbb{E}\big[h_\phi(O_t^{-i}, a, x) \mid X_t^i = x\big]. \tag{12}$$

The right-hand side is interpreted as a bridge-implied estimate of the IRR from observational data. Directly computing the conditional expectation in Eq. (12) would be computationally prohibitive in online RL, since it requires repeated averaging over O_t^{-i} under a fixed $X_t^i = x$. We therefore use the following single-sample Monte Carlo estimator: $h_\phi(o_t^{-i}, a, x)$. This replaces the conditional expectation in Theorem 1 by a single observed draw of the outcome-side proxy, yielding a practical online approximation. For fixed h_ϕ, this is an unbiased stochastic estimator of the bridge-implied target in Eq. (12), corresponding to the one-sample case of simple Monte Carlo estimation [26]. Note that the unbiasedness claim applies only to the single-sample bridge estimate for fixed h_ϕ, not to the full downstream policy update after additional approximations. To connect this estimator to the interventional credit contrast in Eq. (3), we consider the following policy-averaged counterfactual baseline over counterfactual actions:

$$b_t^i \triangleq \mathbb{E}_{a'\sim\bar{\pi}^i(\cdot|\tau_t^i)}\big[h_\phi(o_t^{-i}, a', x)\big], \quad \bar{\pi}^i(a' \mid \tau_t^i) = \tfrac{\pi^i(a'|\tau_t^i)\,\mathbf{1}(a'\neq a_t^i)}{1-\pi^i(a_t^i|\tau_t^i)}. \tag{13}$$

We average the counterfactual baseline under a renormalized policy $\bar{\pi}^i$, rather than under the unmodified policy or a uniform distribution over non-factual actions. This ensures that all non-factual actions remain represented in the baseline even when the policy becomes highly concentrated, while still assigning a larger weight to counterfactual actions that are more plausible under the current policy. Eq. (13) yields the following credit-aligned reward:

$$\tilde{r}_t^i \triangleq r_t - b_t^i, \tag{14}$$

which is used as a practical credit-aligned correction signal for downstream advantage estimation, not as an exact estimator of the theoretical contrast $\delta_t^i(a,a',x)$. We retain the realized global reward as the factual term. This keeps the factual term tied to the original global reward, while limiting bridge-based approximation to the counterfactual baseline. As a result, the practical signal remains anchored to the team objective even though it is a surrogate for the theoretical interventional contrast.

PCI-IPPO/MAPPO. The credit-aligned reward (14) is now integrated into standard policy-gradient MARL pipelines, such as IPPO and MAPPO. Specifically, we augment the standard advantage $\hat{A}_{r,t}^{\mathrm{GAE},i}$ with a PCI-based correction term $\hat{A}_{\tilde{r},t}^{\mathrm{GAE},i}$, yielding the combined advantage $\bar{A}_t^i = \hat{A}_{r,t}^{\mathrm{GAE},i} + \lambda_c \hat{A}_{\tilde{r},t}^{\mathrm{GAE},i}$, where $\hat{A}_{\tilde{r},t}^{\mathrm{GAE},i}$ is computed from $\tilde{r}_t^i$. To estimate it, we introduce a separate critic network (delta-critic) and train it on returns constructed from $\tilde{r}_t^i$. This network follows the same architecture, conditioning structure, and hyperparameters as the backbone critic. Algorithm 1 presents the details. The policy is then updated with

$$\mathcal{L}(\boldsymbol{\pi}) = \mathbb{E}_{i,t}\big[\min\big(\rho_t^i \bar{A}_t^i,\ \mathrm{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon)\bar{A}_t^i\big)\big], \tag{15}$$

where the coefficient λ_c is tuned for each domain introduced in the next section.

In contrast to action-independent baselines used for unbiased variance reduction, Eq. (14) deliberately changes the policy gradient through the counterfactual baseline term. Appendix E shows how it actually induces a bias in the gradient, which systematically displaces the standard policy-gradient contribution. Also, we quantify this effect as a change in the magnitude and direction of the policy gradient, which is geometrically interpreted as an agent-wise **steering**.

7 Experiments

7.1 Oracle Validation of the Identified Reward Response

The identification assumptions of Section 5 cannot be verified in common Dec-POMDPs. We therefore first validate the estimand itself in a controlled model where those assumptions hold by construction and the true contrast is available in closed form.

An unobservable state S and a context X are independent and uniform on $\{\pm 1\}$; the proxies Z and W agree with S with probabilities p_z and p_w; the focal action is a noisy threshold on Z, and the other agent acts as $A^- = \mathbf{1}(W = 1)$. The reward $R = \beta SA + \mathbf{1}(A = A^-)/2 + XA + \epsilon$ combines a confounded term, a coordination term, and the credit term of interest. For binary actions, the individual contrast (3) reduces to the single pair $\delta(x) = \delta(1, 0, x)$, whose true value is x: the confounded term vanishes because $\mathbb{E}[S] = 0$, and the coordination term cancels between the two arms. Since $W \perp\!\!\!\perp A \mid Z$, the bridge equation $\mathbb{E}[R - h(W, A, X) \mid Z, A, X] = 0$ reduces, for each (a, x), to the 2×2 linear system $\boldsymbol{M} h(\cdot, a, x) = r_{a,x}$ with $\boldsymbol{M}[z, w] = P(W = w \mid Z = z)$ and $r_{a,x}[z] = \mathbb{E}[R \mid A = a, Z = z, X = x]$, after which the interventional response follows by averaging h over the marginal of W, see Eq. (23). The system has a unique solution if and only if $\det \boldsymbol{M} = (2p_z - 1)(2p_w - 1) \neq 0$. Here, $p_z \neq 0.5$ corresponds to the completeness condition, whereas $p_w \neq 0.5$ ensures that the outcome-side proxy remains informative enough to admit a unique solution. We evaluate $\boldsymbol{M}$ and $r_{a,x}$ by exact enumeration over the generative parameters, so this *oracle* carries no finite-sample error.

We compare four estimators of δ on identical data: (1) a bridge oracle, (2) its learned counterpart, (3) an observational oracle that reads the contrast off $\mathbb{E}[R \mid W, A, X]$, treating the observational action contrast as interventional, and (4) its learned counterpart. We train the learned counterparts with the bridge objective (10) and a regression loss, respectively. The two networks share the same architecture and differ only in their objectives. Over 10 seeds at $n = 5000$ with $\beta = 1$ and $p_z = p_w = 0.85$, the bridge oracle recovers δ exactly and its learned counterpart attains $\max_x |\hat{\delta}(x) - \delta(x)| = 0.099 \pm 0.043$, against 0.399 and 0.422 ± 0.031 (mean $\pm$ SD over 10 seeds for the learned estimators) for the observational oracle and its learned counterpart—about four times more accurate under a matched budget.

Figure 3 reports more validations. Varying the confounding strength (left), the true contrast is invariant to β because the latent term contributes equally to both arms; the bridge class tracks it, whereas the observational error grows linearly. Varying the proxy quality (right), conditioning on W—what the regression does—fails to block the backdoor path as p_w decreases toward 0.5: the bridge oracle stays exact, and the trained bridge degrades far more slowly. However, the regressor class suffers from rapidly growing bias. As $p_w \to 0.5$, $\det \boldsymbol{M} \to 0$ and the advantage disappears since W cannot act as a proxy. Full specifications are given in Appendix G.1.

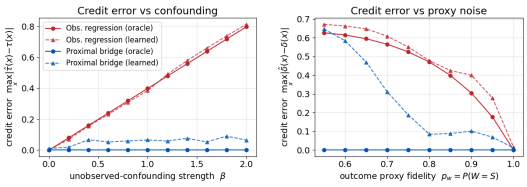
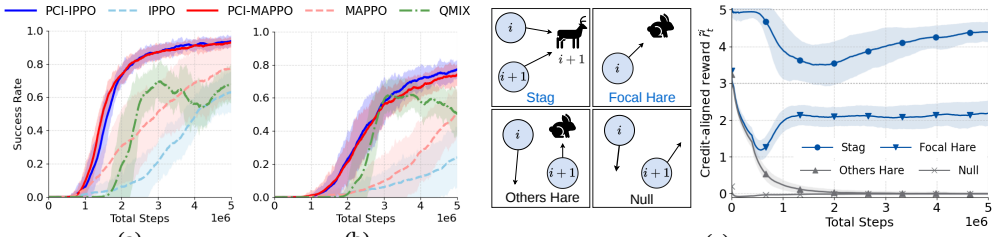


Figure 3: (Left) Bridge class shows robust credit estimation in severe confounding. (Right) Noisy W affects both classes, but the learned outcome bridge degrades more slowly than observational estimators. The point $p_w = 0.5$—resulting in singular M—is excluded.

7.2 Benchmark Evaluation of the Proposed Method

In this section, we report the training results and performance of the proposed methods compared to their backbone models (IPPO and MAPPO, respectively) and another MARL baseline, QMIX, over the three MARL benchmarks.

Proposed diagnostic benchmark: Stag-Hunt. We introduce a partially observable Stag-Hunt benchmark with equal numbers of agents, hares, and stags. Hares can be captured individually for $+3$, whereas stag i requires coordination between a designated agent pair $(i, i + 1)$ and yields $+5$. This creates a non-exchangeable contribution structure under a shared global reward. Details are in Appendix G. Figure 4a shows the training curves. As shown in the figure, PCI-IPPO and PCI-MAPPO reach nearly 90% success, compared with roughly 70% for IPPO and 80% for MAPPO. PCI-IPPO also outperforms MAPPO despite the latter's centralized critic, suggesting that global critic information alone may not resolve agent-wise credit. Under observation noise (Figure 4b), PCI



(a) (b) (c)

Figure 4: Average success rates in the Stag-Hunt environment (a) without and (b) with observation noise $\epsilon \sim N(0, 0.1^2)$. (c) Representative cases—*Stag* (catch a stag with the partner), *Focal Hare* (catch a hare), *Others Hare* (a hare is caught by another agent), and *Null* (others)—with the corresponding $\tilde{r}_t^i$.

variants remain near 80% while the baselines fall near or below 50%. Finally, Figure 4c shows that $\tilde{r}_t^i$ stays positive and approaches the hunting reward for contributing cases, while converging toward zero for no-contribution cases, consistent with the intended counterfactual separation.

Evaluation on SMAX. We also evaluate on **SMAX** [29], adding target-allocation constraints that preserve the cooperative objective while making agent-wise credit contrast more explicit. Across six maps, *Hard*, *Medium*, and unmodified *Vanilla* settings yield 15 tasks. PCI variants outperform their backbones on average in most scenarios, often more under stronger constraints. However, gains are not uniform, and several IPPO settings have high seed variance. Improvements on multiple dense-reward *Vanilla* maps suggest that the method is not limited to sparse or highly stylized coordination tasks. Per-map results are in Appendix G.2, Table 5.

Evaluation on LBF. Another benchmark is **Level-Based Foraging** (LBF; [6]), where agents collect food when their combined levels meet the food level. We use five distinct tasks with a variable world size, number of agents and foods, and cooperation settings indicating whether all agents are required to load a food item or not. Results in Appendix G, Table 4 show small, mixed gains and no reliable degradation.

Aggregate statistical evaluation. To provide a rigorous and reliable comparison, we follow the statistical evaluation recommendations by Agarwal et al. [1] and aggregate over runs.

Table 2 provides a cross-domain aggregate comparison with two complementary quantities. The first is the IQM of per-run success rates, which is not dominated by unlucky seeds—a failure mode seen in several SMAX tasks. We report the interval on the difference in IQM. The second is the probability that a random PCI run beats a random baseline run on the same task, averaged over tasks (ties count one half). It is insensitive to the scale of the success rate and captures how consistently the PCI variant improves. Both intervals come from stratified cluster bootstraps that resample tasks and then seeds within tasks (10,000 replicates). Pooling all 22 tasks, IQM rises from 58.6 to 75.9 for PCI-IPPO and from 70.7 to 83.8 for PCI-MAPPO, with improvement probabilities of 0.76 and 0.78; all four intervals exclude their nulls (zero and 0.5, respectively), and the per-task mean gain is positive on 20 and 19 of 22 tasks. The effect concentrates on Stag-Hunt and SMAX, where the improvement probability reaches 0.79–1.00, and the IQM gain exceeds 14 points. On LBF, none of the intervals exclude the null, and the IQM difference stays within 3 points either way.

Figure 5 reports performance profiles for each domain, including QMIX. The curve shows the fraction of runs whose test success rate exceeds a given threshold; shaded regions are 95% stratified cluster-bootstrap intervals, resampling tasks and then seeds within tasks. On Stag-Hunt, our methods' profiles lie above those of the baselines at every threshold, implying that our methods stochastically dominate baselines. On SMAX, PCI-MAPPO shows statistical dominance over its baselines; on LBF, the profiles overlap.

Ablation study. Appendix G.3 separately replaces the bridge with an observational regressor and the renormalized counterfactual weighting with ordinary policy averaging. The ablation results in Appendix G.3 show that the renormalization is important for preserving a useful counterfactual contrast: removing it brings both PCI variants close to backbone performance on `3m:hard`. In contrast, the incremental benefit of the outcome bridge over an observational regressor is smaller and model-dependent in this setting, indicating that the contribution of bridge-based identification beyond the renormalized counterfactual construction remains less clearly resolved.

Table 2: Cross-domain aggregate comparison of each PCI variant against its base, over all tasks $\times$ five seeds per benchmark. Parentheses give 95% intervals on the *difference* in IQM; brackets give 95% intervals on P(PCI $>$ base). Entries whose interval excludes the null (0 for ΔIQM, 0.5 for P(PCI $>$ base)) are in blue; the last column counts tasks with a positive mean gain. Larger IQM gains and P(PCI $>$ base) above 0.5 both indicate that the PCI variant outperforms its base; an interval excluding the null marks the improvement as reliable.

Benchmark	Tasks	Comparison	IQM (base $\rightarrow$ PCI)	P(PCI $>$ base)	Tasks improved
SH	2	PCI-IPPO vs IPPO	53.9 $\rightarrow$ 85.9 (+15.1, +45.3)	1.00 [1.00, 1.00]	2/2
		PCI-MAPPO vs MAPPO	72.1 $\rightarrow$ 86.7 (-2.3, +29.9)	0.79 [0.44, 1.00]	2/2
SMAX	15	PCI-IPPO vs IPPO	59.5 $\rightarrow$ 76.9 (+6.9, +26.9)	0.79 [0.68, 0.90]	15/15
		PCI-MAPPO vs MAPPO	74.0 $\rightarrow$ 88.7 (+5.0, +24.6)	0.85 [0.76, 0.94]	14/15
LBF	5	PCI-IPPO vs IPPO	59.3 $\rightarrow$ 62.0 (-3.1, +9.3)	0.58 [0.40, 0.77]	3/5
		PCI-MAPPO vs MAPPO	58.5 $\rightarrow$ 59.6 (-5.0, +8.7)	0.55 [0.33, 0.78]	3/5
All	22	PCI-IPPO vs IPPO	58.6 $\rightarrow$ 75.9 (+7.8, +25.9)	0.76 [0.66, 0.86]	20/22
		PCI-MAPPO vs MAPPO	70.7 $\rightarrow$ 83.8 (+5.8, +20.8)	0.78 [0.67, 0.87]	19/22

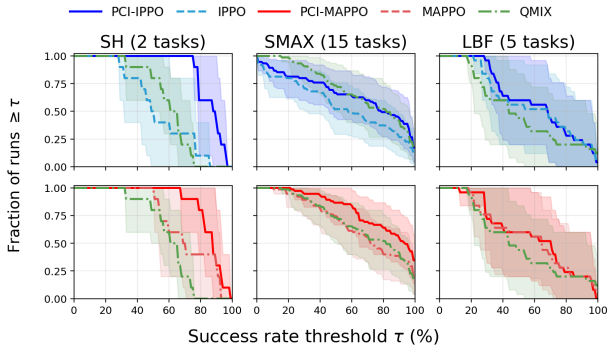


Figure 5: Performance profile over all domains. Shaded bands are 95% stratified cluster-bootstrap CIs.

8 Conclusion and Limitations

Summary. We studied agent-wise credit assignment in cooperative MARL through an interventional lens and argued that global rewards and observational value surrogates do not directly capture individual contribution under partial observability. To this end, we formulated agent-wise credit using an *interventional reward response* and showed that it can be identified from observational trajectories via PCI. Building on this, we developed a practical method that uses a bridge-based counterfactual baseline for credit-aligned policy updates. Empirically, our method improves policy learning in the diagnostic Stag-Hunt domain and SMAX benchmark, suggesting that the resulting credit-aligned update provides useful agent-wise credit signals.

Limitations. From a *theoretical perspective*, the identification results rely on valid proxy variables and the existence of a well-approximated outcome bridge. Violations of these assumptions, finite-sample effects, optimization error, or model misspecification may degrade bridge estimation and bias the resulting credit signal, especially with weak proxies. From a *practical perspective*, the method depends on the backbone RL framework and the quality of RNN-encoded histories, while the one-sample estimator may affect training stability. The RKHS-based bridge objective also incurs additional computational cost through Gram matrix computation, limiting scalability to larger multi-agent domains; empirically, each PCI update takes approximately 1.8–2.7 times as long as its backbone counterpart (Appendix G.5). In our *empirical results*, large CIs in some SMAX maps limit statistical conclusions and motivate evaluation with more seeds and stable benchmark configurations. Finally, while the ablation supports the role of renormalization, the incremental benefit of bridge-based estimation is not uniformly resolved and calls for broader evaluation.

Acknowledgments and Disclosure of Funding

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00233251, System3 reinforcement learning with high-level brain functions; No. RS-2026-25522672, Development of Unified Reasoning Technology Mimicking Human Cognition for Hierarchical Understanding and Unbounded Problem Solving; No. RS-2019-II190075, Artificial Intelligence Graduate School Support Program (KAIST)), Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [N01250565, Development of Beyond X-verse Core Technology for Hyper-realistic interactions by Synchronizing the Real World and Virtual Space], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00341805). **Competing interests.** The authors declare no competing interests.

References

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

[2] Elias Bareinboim. Causal reinforcement learning. ICML 2020, 2020. URL `https://icml.cc/virtual/2020/tutorial/5752`.

[3] Andrew Bennett and Nathan Kallus. The variational method of moments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):810–841, 2023.

[4] Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086, 2024.

[5] Yu-Han Chang, Tracey Ho, and Leslie Kaelbling. All learning is local: Multi-agent learning in global reward games. *Advances in neural information processing systems*, 16, 2003.

[6] Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[8] Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 165–172, 2014.

[9] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33: 12248–12262, 2020.

[10] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:37567–37593, 2023.

[11] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[12] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

[13] Mao Hong, Zhengling Qi, and Yanxun Xu. A policy gradient method for confounded pomdps. 2024 International Conference on Learning Representations (ICLR 2024), 2024.

[14] Yingyao Hu and Ji-Liang Shiu. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2018.

[15] Nathan Kallus, Xiaojie Mao, Masatoshi Uehara, and Zhiheng Zhang. Causal inference under unmeasured confounding with negative controls: A minimax learning approach, 2026. URL `https://arxiv.org/abs/2103.14029`.

[16] Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments, 2018. URL `https://arxiv.org/abs/1803.07164`.

[17] Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Changjie Fan, Fei Wu, and Jun Xiao. Deconfounded value decomposition for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 12843–12856. PMLR, 2022.

[18] Jin Li, Ye Luo, and Xiaowei Zhang. Causal reinforcement learning: An instrumental variable approach. *Available at SSRN 3792824*, 2022.

[19] Zichuan Liu, Yuanyang Zhu, and Chunlin Chen. N$\mathbf{A}^2$Q: Neural Attention Additive Model for Interpretable Multi-Agent Q-Learning. In *International Conference on Machine Learning*, pages 22539–22558. PMLR, 2023.

[20] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

[21] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[22] Wang Miao, Wenjie Hu, Elizabeth L Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023.

[23] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[24] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

[25] Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

[26] Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. Chapter 2: Simple Monte Carlo.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

[28] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[29] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garðar Ingvarsson, Timon Willi, Ravi Hammond, Akbir Khan, Christian S de Witt, et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951, 2024.

[30] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2186–2188, 2019.

[31] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[33] Tao Shen and Yifan Cui. Optimal treatment regimes for proximal causal learning. *Advances in Neural Information Processing Systems*, 36:47735–47748, 2023.

[34] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

[35] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

[36] Kale-ab Abebe Tessera, Arrasy Rahman, Amos Storkey, and Stefano V Albrecht. Hypermarl: Adaptive hypernetworks for multi-agent rl. *arXiv preprint arXiv:2412.04233*, 2024.

[37] Ziyan Wang, Yali Du, Yudi Zhang, Meng Fang, and Biwei Huang. Macca: Offline multi-agent reinforcement learning with causal credit assignment. *arXiv preprint arXiv:2312.03644*, 2023.

[38] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. Colight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1913–1922, 2019.

[39] David H Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(02n03):265–279, 2001.

[40] Qinyu Xu, Yuanyang Zhu, Xuefei Wu, and Chunlin Chen. High-order interactions modeling for interpretable multi-agent q-learning. *arXiv preprint arXiv:2510.20218*, 2025.

[41] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.

[42] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

[43] Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. *Purdue AI Lab, West Lafayette, IN, USA, Tech. Rep*, 2016.

[44] Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33:12263–12274, 2020.

[45] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[46] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 3(4), 2020.

[47] Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:11853–11864, 2020.

Appendix

A Future Work

Several directions remain for future work. First, the current empirical study should be extended to more challenging benchmark domains to better understand when PCI-based credit assignment provides the greatest benefit. Second, richer proxy choices and more expressive bridge-learning strategies may improve robustness in high-dimensional domains. Finally, it would be valuable to study tighter integration between bridge learning and policy optimization, as well as broader applications of proximal identification ideas to multi-agent decision making beyond the current cooperative setting.

B Impact Statements

Our research offers a methodological perspective on cooperative MARL by framing credit assignment in terms of causal inference. The work is primarily foundational and evaluated on diagnostic and benchmark simulation environments; it does not involve human subjects, personal data, surveillance data, or protected attributes. Therefore, we do not anticipate direct negative societal impacts from the present study.

At the same time, MARL methods can be applied in real-world multi-agent control settings such as robotics, traffic signal control, communication networks, and resource allocation. If deployed in such settings, inaccurate credit assignment or insufficiently validated coordination policies could lead to unsafe, inefficient, or unfair outcomes. We therefore recommend that downstream applications include task-specific safety validation, robustness testing under distribution shift, and fairness or allocation audits where human-facing decisions are involved. We also recommend a security review before deployment in safety-critical or networked control systems, since multi-agent control failures or adversarial manipulation could lead to unintended system-level behavior. The proposed method is not intended for weapons, surveillance, deception, harassment, or applications that restrict individual rights. The environmental impact of the experiments in this paper is limited to the reported computational resources, although larger-scale extensions should consider computational efficiency and carbon costs.

C Further Discussion on Proximal Causal Inference

C.1 Discussion on PCI Proxies and Standard Assumptions

In Section 5, we define the time- and agent-dependent PCI variables under Dec-POMDP as shown in Table 1. Motivated by the Dec-POMDP causal structure depicted in Figure 2, we use the focal agent's previous local history T^i_{t-1} as an action-side proxy (Z) and the other agents' current observations as outcome-side proxies (W), while $X^i_t = (O^i_t, \mathrm{T}^{-i}_{t-1}, A^{-i}_{t-1})$ defines the attribution context. This choice of X^i_t is intended to capture the pre-action information most directly relevant to focal credit assignment: the focal agent's current local observation together with the other agents' recent histories and actions. Thus, X^i_t absorbs shared observable information relevant to both action selection and reward generation, while action- and outcome-side proxies provide additional proxy variation on the treatment and outcome sides.

These variable choices follow the roles required by PCI and are made for identification rather than uniqueness. The focal history T^i_{t-1} is a natural action-side proxy because it summarizes the information available to agent i immediately before observing the current local context O^i_t, and therefore retains action-relevant latent-state information beyond the attribution context X^i_t. By contrast, using T^i_t itself as the action-side proxy would introduce substantial overlap with X^i_t, since T^i_t is generated from T^i_{t-1}, O^i_t, and past actions, while O^i_t already appears in X^i_t. This would blur the distinction between context and treatment-side proxy, and may weaken the additional proxy variation needed for the PCI assumptions, especially completeness. In contrast, the other agents' current observations O^{-i}_t are used as an outcome-side proxy because they provide a contemporaneous footprint of S_t that is informative about reward generation at time t. We do not include O^{-i}_t in X^i_t, because X^i_t is used to define the pre-action attribution context, whereas O^{-i}_t is intended to supply additional observable variation on the outcome side through the bridge restriction. Accordingly,

the partition of observed variables into proxies and X_t^i should be understood as an identification-oriented modeling choice that distributes Dec-POMDP variables so as to make the PCI assumptions (Assumptions C.1–C.4) and completeness condition more plausible.

The validity of our approach relies on standard assumptions within the PCI framework. Let $R_t(a) \triangleq R_t(do(A_t^i = a))\ \forall a \in \mathcal{A}^i$, which denotes the potential team reward realized by intervening on agent i to set $A_t^i = a$ at time t, while the environment dynamics and other agents' policies remain unchanged. We assume the following hold for all agents $i \in \mathcal{N}$ and timesteps $t \in \{1, \dots, T\}$. For $t = 1$, we set $\mathrm{T}_0^i \triangleq \emptyset$ and $A_0^i \triangleq \emptyset$ for all $i \in \mathcal{N}$, i.e., an empty history and a null previous action, implemented as the zero initial hidden state and a zero vector in place of the one-hot action encoding, respectively.

Assumption C.1. *(Consistency). $R_t = R_t(A_t^i)$ a.s.; equivalently, if $A_t^i = a$, then $R_t = R_t(a)$ a.s.*

Assumption C.2. *(Latent Exchangeability). $R_t(a) \perp\!\!\!\perp A_t^i \mid S_t, X_t^i \quad \forall a \in \mathcal{A}^i$.*

Assumption C.3. *(Latent Positivity). $P(A_t^i = a \mid S_t, X_t^i) > 0\ a.s.\ \forall a \in \mathcal{A}^i$.*

Assumption C.4. *(Conditional Independence of Proxies).*

$$\mathrm{T}_{t-1}^i \perp\!\!\!\perp R_t \mid S_t, A_t^i, X_t^i, \quad O_t^{-i} \perp\!\!\!\perp (A_t^i, \mathrm{T}_{t-1}^i) \mid S_t, X_t^i.$$

Interpretations for each assumption are given as follows:

- Assumption C.1 is a standard consistency condition ensuring the well-definedness of action A_t^i. It states that when agent i's factual action is a, the observed reward exactly matches the interventional reward under $do(A_t^i = a)$.
- Assumption C.2 says that the state S_t captures all common causes of A_t^i and R_t not included in $X_t^i = (O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$. It means that once we condition on (S_t, X_t^i), the residual dependence between $R_t(a)$ and A_t^i is removed.
- Assumption C.3 requires that every action remains possible within each stratum (S_t, X_t^i), so that the counterfactual action a is statistically estimated from observed data.
- Assumption C.4 formalizes the validity of proxy variables. It states that T_{t-1}^i must not directly affect the outcome R_t other than through S_t, A_t^i, and X_t^i. Similarly, O_t^{-i} must not affect A_t^i or T_{t-1}^i other than through S_t and X_t^i.

Specifically, Assumption C.4 is justified by the rule of *d-separation* on the Dec-POMDP graph in Figure 2. For the first condition, the path $\mathrm{T}_{t-1}^i \to A_t^i \to R_t$ is blocked by conditioning on A_t^i, and any residual path from T_{t-1}^i to R_t through shared latent causes S_t or other agents' past information is blocked by conditioning on S_t and $(\mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$. For the second condition, the path $O_t^{-i} \leftarrow S_t \to O_t^i \to A_t^i$ is blocked by conditioning on S_t, and paths from O_t^{-i} to T_{t-1}^i through S_{t-1} are blocked once S_t and X_t^i are conditioned on. Hence,

$$\mathrm{T}_{t-1}^i \perp\!\!\!\perp R_t \mid S_t, A_t^i, O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i}, \quad O_t^{-i} \perp\!\!\!\perp (A_t^i, \mathrm{T}_{t-1}^i) \mid S_t, O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i}.$$

At $t = 1$, T_0^i is constant, so the conditions involving T_0^i hold trivially, and the remaining condition $O_1^{-i} \perp\!\!\!\perp A_1^i \mid S_1, X_1^i$ follows from the same argument.

C.2 Discussion on Completeness

Assumption C.5. *(Completeness). For any square-integrable measurable function g, for all $i \in \mathcal{N}$ and $t \in \{2, \dots, T\}$, and for all $a \in \mathcal{A}^i$ and $x \in \mathrm{supp}(X_t^i)$,*

$$\mathbb{E}\left[g(S_t) \mid \mathrm{T}_{t-1}^i, A_t^i = a, X_t^i = x\right] = 0\ a.s. \quad \Longleftrightarrow \quad g(S_t) = 0\ a.s.$$

At $t = 1$, Assumption C.5 is not generally guaranteed, since the constant T_0^i provides no additional state-related variation beyond (A_1^i, X_1^i). Hence, Proposition 1 and Theorem 1 are stated for $t \geq 2$, and the bridge output at the first step of each episode serves as an approximation. As mentioned in Section 5, the above completeness condition reduces to familiar full-rank conditions (actually, full row rank) on conditional probability operators in finite/discrete settings. It should be noted that $Z_t^i = \mathrm{T}_{t-1}^i,\ W_t^i = O_t^{-i},\ X_t^i = (O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$, and $Y_t = R_t$. Here, we assume all variables are

in a discrete and finite domain. Fix $(a, x) \in \mathcal{A}^i \times \text{supp}(X_t^i)$. Define the conditional expectation operator

$$(T_{a,x}g)(z) \triangleq \mathbb{E}\big[g(S_t) \mid Z_t^i = z,\ A_t^i = a,\ X_t^i = x\big],$$

mapping functions of S_t to functions of Z_t^i. Assumption C.5 states that $T_{a,x}$ is injective (one-to-one):

$$T_{a,x}g = 0\ a.s. \quad \Longrightarrow \quad g(S_t) = 0\ a.s.$$

Now assume S_t and Z_t^i take values on $\mathcal{S} = \{s_1, \ldots, s_m\}$ and $\mathcal{Z} = \{z_1, \ldots, z_n\}$, respectively. Identify g with a vector $g \in \mathbb{R}^m$, whose k-th element is $g(s_k)$, and define $P(S_t \mid Z_t^i, a, x)$ as an $m \times n$ matrix whose $(i, j)th$ element is $P(S_t = s_i \mid Z_t^i = z_j, a, x)$. Then the completeness (injectivity) is equivalent to the null space of $P(S_t \mid Z_t^i, a, x)^\top$ being trivial, i.e.,

$$P(S_t \mid Z_t^i, a, x)^\top g = 0 \implies g = 0 \qquad \Longleftrightarrow \qquad \text{rank}(P(S_t \mid Z_t^i, a, x)) = m = |\mathcal{S}|.$$

In particular, a necessary condition is $|\mathcal{Z}| \geq |\mathcal{S}|$ (otherwise full row rank is impossible). Therefore, we note that the following is a sufficient condition to guarantee the completeness condition in this case,

$$\text{rank}\left(P\left(S_t \mid \mathrm{T}_{t-1}^i, a, x\right)\right) = |\mathcal{S}| \quad \forall (a, x).$$

In our finite/discrete Dec-POMDP instantiation, S_t is the unobserved state, and we set $Z_t^i = \mathrm{T}_{t-1}^i$. The full-rank condition means the following: for each fixed (a, x), distinct states S_t induce linearly independent conditional distributions over the proxy Z_t^i. Equivalently, the proxy must be rich enough to separate (or distinguish) the unobserved states that matter for reward generation, beyond what is already contained in (A_t^i, X_t^i). This full-rank condition has a standard interpretation in proximal causal inference studies. In the case of categorical variables, it can be interpreted as requiring the proxy to have at least as many categories as the unobserved confounder [35]. Also, in single-agent partially observed models (POMDPs), it implies that distinct unobserved states induce linearly independent distributions over the proxy. In tabular settings, such rank conditions are routinely used as sufficient identifiability assumptions; for example, Hong et al. [13] provide an explicit tabular identification (see Appendix B.6 of their paper).

Although verifying this condition in arbitrary Dec-POMDPs is difficult, we provide operational intuition for why it is plausible in our setting. First, Z_t^i is a high-dimensional encoded history depending on the implementation. Second, the flexibility of choosing Z_t^i helps avoid degeneracy with X_t^i: completeness requires that Z_t^i contain variation not already included in $X_t^i = (O_t^i, \mathrm{T}_{t-1}^{-i}, A_{t-1}^{-i})$. Our choice $Z_t^i = \mathrm{T}_{t-1}^i$ retains such extra information relative to X_t^i for $t \geq 2$. Our discussion emphasizes that enough information is essential to overcome partial observability in MARL, so it may provide a causal-inference interpretation of CTDE.

C.3 Stabilizer for the Minimax Game of the Outcome Bridge Function

The minimax approach for reformulating the conditional moment restriction has been established via various studies in recent years [9, 16]. The rationale behind the connection between conditional moment restriction and minimax reformulation lies in the law of iterated expectations. For the outcome bridge function h, the conditional moment restriction is given by,

$$\mathbb{E}[Y - h(W, A, X)|Z, A, X] = 0.$$

Here, we suppose all variables and functions are scalar for simplicity. By definition, this conditional expectation is zero if and only if the unconditional expectation is zero for any square-integrable test function (or critic function) $\nu \in \mathcal{H}$. This equivalence allows us to reformulate the problem as finding a function h that minimizes the maximum violation of these unconditional moments across a chosen critic class $\mathcal{H}$,

$$\mathbb{E}[Y - h(W, A, X)|Z, A, X] = 0 \iff \sup_{\nu \in \mathcal{H}} \left(\mathbb{E}[\nu(Z, A, X)(Y - h(W, A, X))]\right)^2 = 0.$$

While this forms the basis of the adversarial generalized method of moments (AGMM), directly optimizing this squared objective can be highly unstable in practice. To address this, a stabilizer term is introduced. For any constant $\lambda > 0$, the squared conditional expectation can be reformulated as a penalized supremum problem,

$$0 = \sup_{\nu \in \mathcal{H}} \left(\mathbb{E}[\nu(Z, A, X)(Y - h(W, A, X))] - \lambda \mathbb{E}[\nu^2(Z, A, X)]\right).$$

This mathematical identity motivates the empirical minimax estimator with a stabilizer. Unlike standard regularizers that shrink the target parameters and inherently introduce estimation bias, the stabilizer $-\lambda\mathbb{E}[\nu^2]$ is merely a structural reformulation of the conditional moment. It effectively bounds the norm of the critic function, thereby controlling the critical radius of the function class. Consequently, this stabilization yields a strongly concave inner maximization problem, which improves the finite-sample convergence rates for the bridge function estimation while enabling the use of flexible function approximators. For more information, we provide variants of Eq. (9) and (10) without stabilizer,

$$\hat{h} \in \arg\min_{h} \max_{\nu \in \mathcal{H}} \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \left(\mathbb{E}_k \left[\nu(\tau^i_{t_k-1}, a^i_{t_k}, x^i_{t_k}) \left(h(o^{-i}_{t_k}, a^i_{t_k}, x^i_{t_k}) - r_{t_k} \right) \right] \right)^2 .$$

The counterpart of Eq. (10) is $\hat{h} \in \arg\min_h \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \mathbf{e}_i^\top \boldsymbol{K}_i \mathbf{e}_i$.

D Proofs

Proof of Proposition 1. Fix an arbitrary agent $i \in \mathcal{N}$ and timestep $t \in \{2, \dots, T\}$ and omit (i, t) subscripts for readability. Let $h \in \mathbb{H}^{obs}$, i.e.,

$$\mathbb{E}[R - h(W, A, X) \mid Z, A, X] = 0 \quad a.s. \tag{16}$$

Define the function $g_{a,x} : \mathcal{S} \to \mathbb{R}$ as

$$g_{a,x}(s) \triangleq \mathbb{E}[R - h(W, A, X) \mid S = s, A = a, X = x].$$

Equivalently, define the random variable $g_{A,X}(S)$ by evaluating $g_{a,x}(s)$ at $(a, x, s) = (A, X, S)$. By Assumption C.4, we have

$$\mathbb{E}[R \mid S, Z, A, X] = \mathbb{E}[R \mid S, A, X],$$
$$\mathbb{E}[h(W, A, X) \mid S, Z, A, X] = \mathbb{E}[h(W, A, X) \mid S, A, X].$$

Subtracting the two yields

$$\mathbb{E}[R - h(W, A, X) \mid S, Z, A, X] = \mathbb{E}[R - h(W, A, X) \mid S, A, X] = g_{A,X}(S). \tag{17}$$

Taking conditional expectation of Eq. (17) given (Z, A, X) and using the law of total expectation,

$$\mathbb{E}[R - h(W, A, X) \mid Z, A, X] = \mathbb{E}[g_{A,X}(S) \mid Z, A, X].$$

Combining with Eq. (16), we obtain

$$\mathbb{E}[g_{A,X}(S) \mid Z, A, X] = 0 \quad a.s. \tag{18}$$

For each fixed (a, x), Eq. (18) implies

$$\mathbb{E}[g_{a,x}(S) \mid Z, A = a, X = x] = 0 \quad a.s.$$

By the completeness condition (7), this yields $g_{a,x}(S) = 0$ almost surely for all (a, x) in the support of (A, X). Then by the definition of $g_{A,X}(S)$,

$$\mathbb{E}[R - h(W, A, X) \mid S, A, X] = 0 \quad a.s.$$

This shows $h \in \mathbb{H}$, hence $\mathbb{H}^{obs} \subseteq \mathbb{H}$. Conversely, if $h \in \mathbb{H}$, then

$$\mathbb{E}[R - h(W, A, X) \mid Z, A, X] = \mathbb{E}[\mathbb{E}\left[R - h(W, A, X) \mid S, A, X\right] \mid Z, A, X] = 0,$$

so $h \in \mathbb{H}^{obs}$. Therefore, $\mathbb{H}^{obs} = \mathbb{H}$, and in particular every observed bridge is a valid outcome bridge under the completeness condition.

□

Proof of Theorem 1. Fix an arbitrary agent $i \in \mathcal{N}$ and timestep $t \in \{2, \dots, T\}$. Recall the target

$$\bar{r}_t^i(a, x) \triangleq \mathbb{E}[R_t(a) \mid X_t^i = x].$$

By the law of total expectation,

$$\bar{r}_t^i(a, x) = \mathbb{E}\big[\mathbb{E}\left[R_t(a) \mid S_t, X_t^i\right] \mid X_t^i = x\big]. \tag{19}$$

Under Assumptions C.1 and C.2, $\mathbb{E}\left[R_t(a) \mid S_t, X_t^i = x\right] = \mathbb{E}\left[R_t \mid S_t, A_t^i = a, X_t^i = x\right]$. Here, Assumption C.3 guarantees the relevant conditional expectation is well-defined.

Let $h_t^i \in \mathbb{H}_t^{obs,i}$ be an observed outcome bridge function. Under Assumption C.5, Proposition 1 implies $h_t^i \in \mathbb{H}_t^i$, i.e.,

$$\mathbb{E}[R_t - h_t^i(W_t^i, A_t^i, X_t^i) \mid S_t, A_t^i, X_t^i] = 0.$$

Conditioning on $\{A_t^i = a, X_t^i = x\}$,

$$\mathbb{E}\big[R_t \mid S_t, A_t^i = a, X_t^i = x\big] = \mathbb{E}\big[h_t^i(W_t^i, a, x) \mid S_t, A_t^i = a, X_t^i = x\big]. \tag{20}$$

Because $h_t^i(W_t^i, a, x)$ depends on W_t^i only through the first argument, Assumption C.4 implies

$$\mathbb{E}[h_t^i(W_t^i, a, x) \mid S_t, A_t^i = a, X_t^i = x] = \mathbb{E}[h_t^i(W_t^i, a, x) \mid S_t, X_t^i = x].$$

Substituting into (19) and applying the law of total expectation,

$$\bar{r}_t^i(a, x) = \mathbb{E}\big[\mathbb{E}[h_t^i(W_t^i, a, x) \mid S_t, X_t^i = x] \mid X_t^i = x\big] = \mathbb{E}[h_t^i(W_t^i, a, x) \mid X_t^i = x].$$

This equation involves only observable variables and the trainable outcome bridge function h_t^i, which completes the identification.

□

E Theoretical analysis on the counterfactual baseline

Unlike COMA [11], which uses a value-level counterfactual baseline, and potential-based reward shaping [8, 24], whose temporal structure preserves policy invariance, our counterfactual baseline is subtracted directly from the global reward without such an invariance constraint. The renormalization further makes the baseline depend on the focal action. Consequently, its effect generally accumulates over time and changes the policy-gradient contribution associated with the original global reward. In this section, we characterize this policy-gradient bias and show how it appears as a change in the magnitude and direction of the policy gradient.

Frozen counterfactual reward shaping. As introduced in Section 6 and Algorithm 1, an auxiliary value function $\tilde{V}_t^i$ is estimated from the credit-aligned reward. Although the counterfactual baseline depends on the individual policy, its computation is detached during the policy update. We therefore distinguish the fixed baseline parameter $\bar{\theta}$ from the policy parameter θ being differentiated, such that

$$\nabla_\theta b_{\bar{\theta},t}^i = 0.$$

The credit-aligned reward can be written as $\tilde{r}_t^i = r_t - b_{\bar{\theta},t}^i$. For notational simplicity, we suppress the frozen parameter $\bar{\theta}$ from the baseline and its derived value quantities below. Next, we define the discounted return of the counterfactual baseline as

$$B_t^i \triangleq \sum_{k=0}^{T-t} \gamma^k b_{t+k}^i.$$

Value-level effect. We first extend the reward-level counterfactual baseline to the corresponding return, value, and advantage quantities that directly enter the policy-gradient update.

Proposition 2 (Value-level decomposition). *Let Q^i and A^i denote the action value and advantage induced by the team reward, and let Q_b^i and A_b^i denote the corresponding quantities induced by the discounted counterfactual-baseline return. Then*

$$\tilde{Q}^i(\tau, a) = Q^i(\tau, a) - Q_b^i(\tau, a), \tag{21}$$

$$\tilde{A}^i(\tau, a) = A^i(\tau, a) - A_b^i(\tau, a). \tag{22}$$

Proof. Define the standard and credit-aligned returns from timestep t as

$$G_t \triangleq \sum_{k=0}^{T-t} \gamma^k r_{t+k}, \quad \tilde{G}_t^i \triangleq \sum_{k=0}^{T-t} \gamma^k \tilde{r}_{t+k}^i.$$

Using $\tilde{r}_{t+k}^i = r_{t+k} - b_{t+k}^i$,

$$\tilde{G}_t^i = \sum_{k=0}^{T-t} \gamma^k \left(r_{t+k} - b_{t+k}^i\right) = G_t - B_t^i.$$

Then,

$$\tilde{Q}^i(\tau, a) = Q^i(\tau, a) - Q_b^i(\tau, a),$$

where

$$Q^i(\tau, a) \triangleq \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[G_t \mid \tau_t^i = \tau,\, a_t^i = a\right],\; Q_b^i(\tau, a) \triangleq \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[B_t^i \mid \tau_t^i = \tau,\, a_t^i = a\right].$$

Taking the expectation over the current policy yields

$$\tilde{V}^i(\tau) = V^i(\tau) - V_b^i(\tau), \quad V_b^i(\tau) \triangleq \mathbb{E}_{a\sim\pi_\theta^i(\cdot|\tau)}\left[Q_b^i(\tau, a)\right].$$

Therefore,

$$\tilde{A}^i(\tau, a) = \tilde{Q}^i(\tau, a) - \tilde{V}^i(\tau) = A^i(\tau, a) - A_b^i(\tau, a),$$

where $A_b^i(\tau, a) \triangleq Q_b^i(\tau, a) - V_b^i(\tau)$. □

Proposition 2 shows that subtracting the counterfactual baseline at the reward level induces a corresponding baseline advantage A_b^i after temporal accumulation. Hence, the effect on the policy gradient is determined by the discounted baseline return rather than by the baseline at a single timestep.

Policy-gradient effect. We next propagate the value-level decomposition to the policy gradient and characterize the resulting bias and displacement.

Proposition 3 (Policy-gradient bias and displacement). *Let*

$$\boldsymbol{g}_r^i \triangleq \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\sum_{t=1}^{T} \gamma^{t-1} \nabla_\theta \log \pi_\theta^i(a_t^i \mid \tau_t^i) A^i(\tau_t^i, a_t^i)\right]$$

denote the standard agent-wise policy-gradient contribution, and define the baseline-return contribution as

$$\boldsymbol{g}_b^i \triangleq \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\sum_{t=1}^{T} \gamma^{t-1} \nabla_\theta \log \pi_\theta^i(a_t^i \mid \tau_t^i) A_b^i(\tau_t^i, a_t^i)\right].$$

Then the policy-gradient contribution induced by the credit-aligned reward satisfies

$$\boldsymbol{g}_{\tilde{r}}^i = \boldsymbol{g}_r^i - \boldsymbol{g}_b^i.$$

Consequently, the combined contribution of the standard and credit-aligned branches is

$$\boldsymbol{g}^i \triangleq \boldsymbol{g}_r^i + \lambda_c \boldsymbol{g}_{\tilde{r}}^i = (1 + \lambda_c)\boldsymbol{g}_r^i - \lambda_c \boldsymbol{g}_b^i.$$

Proof. Using Eq. (22),

$$\boldsymbol{g}_{\tilde{r}}^i = \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\sum_{t=1}^{T} \gamma^{t-1} \nabla_\theta \log \pi_\theta^i(a_t^i \mid \tau_t^i)\big(A^i(\tau_t^i, a_t^i) - A_b^i(\tau_t^i, a_t^i)\big)\right] = \boldsymbol{g}_r^i - \boldsymbol{g}_b^i.$$

Thus, relative to the standard policy gradient, the counterfactual baseline introduces the bias

$$\boldsymbol{g}_{\tilde{r}}^i - \boldsymbol{g}_r^i = -\boldsymbol{g}_b^i.$$

The combined gradient therefore becomes

$$\boldsymbol{g}^i = (1 + \lambda_c)\boldsymbol{g}_r^i - \lambda_c \boldsymbol{g}_b^i.$$

If the counterfactual baseline induced no policy-gradient bias, i.e., $\boldsymbol{g}_b^i = \mathbf{0}$, the corresponding two-branch reference would be

$$\boldsymbol{g}_0^i \triangleq (1 + \lambda_c)\boldsymbol{g}_r^i.$$

Hence, the displacement induced by the counterfactual baseline is $\boldsymbol{g}^i - \boldsymbol{g}_0^i = -\lambda_c \boldsymbol{g}_b^i$. □

Table 3: Actor-gradient deformation across selected evaluation settings. The gain $s_i = \|\boldsymbol{g}^i\|/\|\boldsymbol{g}_0^i\|$ measures the gradient norm relative to the bias-free two-branch reference, while ϕ_i measures the steering angle from the standard actor-gradient direction. Entries report mean $\pm$ 95% Student-t intervals across five training seeds.

Domain	Model	λ_c	s_i	ϕ_i
Stag-Hunt:vanilla	PCI-IPPO	1.0	0.84 ± 0.11	$28.6^\circ \pm 7.9^\circ$
Stag-Hunt:vanilla	PCI-MAPPO	1.0	0.91 ± 0.08	$29.0^\circ \pm 6.0^\circ$
SMAX 3m:hard	PCI-IPPO	1.0	0.85 ± 0.17	$39.3^\circ \pm 17.0^\circ$
SMAX 3m:hard	PCI-MAPPO	1.0	0.74 ± 0.24	$51.8^\circ \pm 14.8^\circ$
SMAX smacv2_5_units:hard	PCI-IPPO	0.1	0.96 ± 0.02	$7.7^\circ \pm 2.0^\circ$
SMAX smacv2_5_units:hard	PCI-MAPPO	0.1	0.95 ± 0.02	$6.9^\circ \pm 1.1^\circ$

The bias therefore acts as a systematic displacement from the gradient that would be obtained if the credit-aligned branch preserved the standard policy gradient. We characterize this deformation by the relative gain and steering angle

$$s_i \triangleq \frac{\|\boldsymbol{g}^i\|}{\|\boldsymbol{g}_0^i\|}, \quad \phi_i \triangleq \angle(\boldsymbol{g}_0^i, \boldsymbol{g}^i).$$

Since $\boldsymbol{g}_0^i = (1 + \lambda_c)\boldsymbol{g}_r^i$ with $\lambda_c \geq 0$, $\phi_i = \angle(\boldsymbol{g}_r^i, \boldsymbol{g}^i)$. Thus, without counterfactual-baseline-induced bias, there is neither magnitude deformation nor directional steering relative to the bias-free reference,

$$\boldsymbol{g}_b^i = \mathbf{0} \quad \Longrightarrow \quad s_i = 1, \qquad \phi_i = 0.$$

Empirical gradient steering. In practice, the exact advantages above are replaced by the critic and GAE estimates used by the implementation. We evaluate the corresponding actor-gradient vectors at the pre-update policy, before any optimizer step on the collected rollout. At this point, the PPO likelihood ratio equals one and clipping is inactive, allowing the two gradient branches to be compared directly using the same gain and steering definitions.

Table 3 shows that the counterfactual baseline produces a measurable displacement of the actor gradient in both magnitude and direction. Across the settings with $\lambda_c = 1$, the resulting gradient has approximately 0.74–0.91 times the norm of the bias-free reference and is rotated by about 29°–52°. In `smacv2_5_units:hard`, where $\lambda_c = 0.1$, the deformation is more modest, with $s_i \approx 0.95$–0.96 and a steering angle of about 7°. These results show that the policy-gradient bias induced by the counterfactual baseline appears in the implemented update as both a change in gradient magnitude and a measurable steering away from the standard policy-gradient direction.

F Implementation Details

We reuse and modify the MLP and RNN modules from the `marlbenchmark/on-policy` codebase (MIT License), following Yu et al. [42].

F.1 RBF Kernels

To construct the RKHS critic kernel, we use a product kernel over the three input blocks $(\tau_{t-1}^i, a_t^i, x_t^i)$. Specifically, the history feature τ_{t-1}^i and the context feature x_t^i are standardized block-wise, while the action a_t^i is represented by a one-hot vector. We then define block-specific RBF kernels,

$$k_\tau(\tau, \tau') = \exp\left(-\gamma_\tau \|\tilde{\tau} - \tilde{\tau}'\|^2\right), \quad k_a(a, a') = \exp\left(-\gamma_a \|\tilde{a} - \tilde{a}'\|^2\right),$$
$$k_x(x, x') = \exp\left(-\gamma_x \|\tilde{x} - \tilde{x}'\|^2\right).$$

Here, $\tilde{\tau}$ and $\tilde{x}$ denote standardized inputs and $\tilde{a}$ denotes the one-hot encoding of the action. For simplicity, super/subscripts are omitted. The context variable is itself heterogeneous: in our setting, $x_t^i = (o_t^i, \tau_{t-1}^{-i}, a_{t-1}^{-i})$ consists of the focal agent's current observation, the other agents' previous histories, and the other agents' previous actions. Accordingly, $k_x(x, x')$ is not implemented on x_t^i as a single raw vector. Instead, its constituent feature blocks are preprocessed separately—o_t^i and τ_{t-1}^{-i} are standardized block-wise, while a_{t-1}^{-i} is represented by a one-hot vector—and the resulting

block-wise similarities are combined multiplicatively before forming the context-level Gram matrix. Thus, the final kernel still takes the product form

$$k\big((\tau, a, x), (\tau', a', x')\big) = k_\tau(\tau, \tau')\, k_a(a, a')\, k_x(x, x').$$

Accordingly, the final Gram matrix is constructed as the Hadamard product in Eq. (11). For each preprocessed block, the bandwidth is set as $\gamma_b = 1/d_b$, where d_b is the dimension of the corresponding input block after preprocessing, except for one-hot encoded action components with $\gamma_b = 1$. This yields separate bandwidths for the focal history, local observation, and other-agent history, reflecting their different scales and dimensions.

F.2 Pseudocode

Algorithm 1 PCI-IPPO/MAPPO with credit-aligned reward using the outcome bridge function

Initialize bridge, policy, critic, and delta-critic parameters: $\phi, \boldsymbol{\theta}, \boldsymbol{\omega}$, and $\tilde{\boldsymbol{\omega}}$.
Collect a batch $\mathcal{D} = \{(o_t^i, a_t^i, \tau_t^i, r_t)\}$ with $\boldsymbol{\pi_\theta} = (\pi^1_{\theta_1}, \ldots, \pi^n_{\theta_n})$
for gradient steps **do**
    Update outcome bridge network h_ϕ with objective Eq. (10).
end for
Compute credit-aligned reward $\tilde{r}_t^i = r_t - b_t^i$ with b_t^i detached during the subsequent updates, Eq. (14).
Compute returns G_t and $\tilde{G}_t^i$ from r_t and $\tilde{r}_t^i$, respectively.
for gradient steps **do**
    Update policies with objective Eq. (15).
    Update critics with objective $\mathcal{L}_\omega = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T (V_\omega - G_t)^2$.
    Update delta-critics with objective $\mathcal{L}_{\tilde{\omega}} = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T (\tilde{V}_{\tilde{\omega}} - \tilde{G}_t^i)^2$.
end for

G Experimental Settings and Extended Results

G.1 Oracle validation

This section gives the full specification of the toy model summarized in Section 7.1. The variable distributions are

$$S \sim \text{Uniform}\{-1, 1\},\;\; X \sim \text{Uniform}\{-1, 1\},\;\; Z \sim P(Z = S) = p_z,\;\; W \sim P(W = S) = p_w,$$
$$A \sim \mathbf{1}(Z = 1) \text{ flipped w.p. } \eta,\;\; \text{and } A^- = \mathbf{1}(W = 1),$$

with S and X independent, so that the proxy conditional independence (C.4) holds by construction. The flip probability $\eta > 0$ ensures that both actions can occur at each value of Z, which is the positivity requirement (C.3) in this discrete model; at $\eta = 0$ the action is a deterministic function of Z and the bridge system is underdetermined. It also controls how strongly S influences A: at $\eta = 0.5$ the action is independent of Z and hence unconfounded, so we fix $\eta = 0.05$ to keep confounding strong while all cells remain reachable. Note that η does not enter $\det \boldsymbol{M}$, so positivity and completeness are governed by separate parameters in this model. The reward is

$$R = \beta \cdot SA + \mathbf{1}(A = A^-)/2 + XA + \epsilon,$$

where β is the confounding coefficient and $\epsilon \sim N(0, \sigma^2)$ is the Gaussian noise. Under $do(A{=}a)$ the confounding term becomes 0 since $\mathbb{E}[S] = 0$, and the coordination term contributes $P(A^-{=}a)/2$, which cancels in the contrast because W is marginally uniform; the true contrast is therefore $\delta(x) = x$. The coefficient matrix of the bridge system is the same for both actions, since $W \perp\!\!\!\perp A \mid Z$, and has the explicit form

$$\boldsymbol{M} = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix}, \qquad \alpha = p_z p_w + (1-p_z)(1-p_w),$$

where α is the probability that the two proxies agree, giving $\det \boldsymbol{M} = 2\alpha - 1 = (2p_z - 1)(2p_w - 1)$. The interventional response is obtained by averaging the solution over the marginal of W,

$$\mathbb{E}[R \mid do(A{=}a), X{=}x] = \frac{1}{2}\sum_w h(w, a, x). \tag{23}$$

We evaluate $\boldsymbol{M}$ and $r_{a,x}$ by exact enumeration over the discrete joint distribution induced by (p_w, p_z, η, β), so both oracles carry no finite-sample error. The observational oracle substitutes $\mathbb{E}[R \mid W, A, X]$ for the interventional response and takes its difference across actions. The two trained estimators share architecture, optimizer, learning rate, and initialization, and differ only in the objective—the bridge objective (10) against a regression (MSE) loss. The bridge objective uses the same kernel construction and the same Tikhonov and stabilizer settings as our MARL experiments. Reported results use 10 seeds at sample size $n = 5000$ with $\beta = 1$, $p_z = p_w = 0.85$, $\eta = 0.05$, and $\sigma^2 = 0.05$; the error metric is $\max_x |\hat{\delta}(x) - \delta(x)|$.

Figure 3 provides two additional validations: (1) sensitivity to confounding strength and (2) sensitivity to W (outcome-side proxy) quality. We vary β and p_w, respectively, while holding all other parameters fixed. Values of $p_w < 0.5$ are omitted without loss of generality, since they would yield symmetric results. The left panel shows that the true credit is invariant to β, since the hidden confounder contributes equally to both arms of the contrast. The bridge class tracks it, whereas the observational oracle diverges linearly in β. In the right panel, as the outcome proxy W becomes noisier (i.e., p_w decreases toward 0.5), simply conditioning on W, i.e., what the regression does, fails to block the backdoor paths. The bridge oracle stays exact, and the trained bridge degrades far more slowly, while the regressor class suffers from rapidly growing bias.

G.2 Environmental Details and Results

Diagnostic Stag-Hunt Benchmark. The Stag-Hunt environment is a 7×7 partially observable grid world with n agents, hares, and stags. Each hare can be captured individually for a small reward (+3). Each stag i can be captured only through successful coordination between a designated pair of agents $(i,\ i+1)$ (with wrap-around), yielding a larger reward (+5). At each timestep, every agent receives a 5×5 local observation: a multi-channel tensor that encodes the positions of entities and boundaries. Each channel indicates the presence of the i-th stag, the i-th agent, or a hare. Hares are not distinguished by their indices. Consequently, the channel vector for each grid cell has a length of $2n + 1$. For our specific scenario with $n = 3$, this results in a channel vector length of 7 per cell. Therefore, each flattened observation $o_t^i \in \mathbb{R}^{175}$. Any grid cell outside the environment boundaries is masked with a value of -1 across all channels.

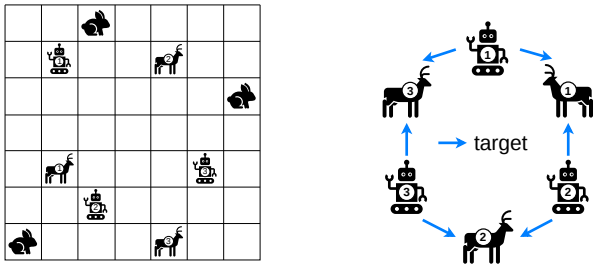


Figure 6: Stag-Hunt environment snapshot and cooperation structure.

SMAX. SMAX from JaxMARL benchmark [29] serves as a lightweight, faster alternative to the popular StarCraft Multi-Agent Challenge (SMAC; [30]). SMAX provides discrete-action tasks based on SMAC and SMACv2 [10]. Here we add target-allocation constraints under which each agent i is assigned a fixed admissible target set $\mathcal{T}_i$ over training, and an attack on enemy e_j is effective only if $e_j \in \mathcal{T}_i$; otherwise, it nullifies all attacks on that target. For evaluation on SMAX, we select 6 maps (`3m`, `2s3z`, `3s5z`, `6h_vs_8z`, `smacv2_5_units`, and `smacv2_10_units`) that allow for a perfectly symmetric target allocation. Also, the *Medium* constraint is used for `2s3z`, `smacv2_5_units`, and `smacv2_10_units`.

Level-Based Foraging (LBF; [6]). LBF emphasizes the coordination between agents. Agents are placed within a grid world and assigned different levels. To collect food, the cumulative level of participating agents must meet or exceed the food's designated level. Agents receive points equivalent to the level of the collected food and their own level. Each task specification is encoded by a map name of the form `{sight}s-{grid_size}x{grid_size}-{num_agents}p-{num_foods}f-{force_cooperation}`. The sight parameter specifies each agent's local observation range. If the last term, coop, is activated, every food level is fixed as $\sum_{i=1}^{n}$ `agent_level`^{i}, which forces strict cooperation.

Table 4: Test success rates (%) on the Stag-Hunt and LBF domains at the final checkpoint. Entries report mean ± 95% Student-t interval across the five training seeds, each evaluated on 64 episodes with distinct seeds; to be read as truncated at 0 and 100. Gray columns represent proposed methods. **Gain** columns give the *paired* per-seed improvement of each PCI variant over its base with a 95% paired-t interval; gains whose interval excludes zero are shown in blue.

Map	IPPO	PCI-IPPO	Gain (pp)	MAPPO	PCI-MAPPO	Gain (pp)	QMIX
SH	70.0 ± 17.8	**92.8 ± 4.0**	+22.8 ± 18.5	84.7 ± 17.6	90.9 ± 7.9	+6.2 ± 20.5	64.1 ± 14.2
SH-noise	39.1 ± 11.1	79.4 ± 5.9	+40.3 ± 15.1	59.4 ± 11.7	**80.6 ± 10.5**	+21.2 ± 14.6	55.9 ± 17.0
2s-8x8-2p-2f-coop	**98.1 ± 1.6**	97.8 ± 2.2	-0.3 ± 0.9	97.8 ± 2.2	97.2 ± 1.6	-0.6 ± 1.7	**98.1 ± 4.2**
2s-10x10-3p-3f	**82.8 ± 7.4**	80.6 ± 9.4	-2.2 ± 9.4	74.7 ± 9.6	76.2 ± 6.6	+1.6 ± 10.4	65.3 ± 10.6
2s-15x15-3p-3f	30.6 ± 5.8	**33.8 ± 5.6**	+3.1 ± 10.2	31.6 ± 10.7	25.3 ± 8.9	-6.2 ± 14.8	21.9 ± 4.9
2s-15x15-4p-3f	60.0 ± 19.2	**67.2 ± 7.1**	+7.2 ± 13.3	65.3 ± 6.6	**67.2 ± 8.3**	+1.9 ± 11.5	46.2 ± 5.6
2s-15x15-4p-5f	31.2 ± 10.2	**36.6 ± 5.8**	+5.3 ± 6.5	24.1 ± 6.7	33.8 ± 7.3	+9.7 ± 10.3	25.9 ± 5.8

Table 5: Test success rates (%) on SMAX at the final checkpoint. Entries and **Gain** columns report the same information as Table 4.

Map	Target Constraint	IPPO	PCI-IPPO	Gain (pp)	MAPPO	PCI-MAPPO	Gain (pp)	QMIX
3m	Vanilla	90.3 ± 8.7	**97.5 ± 1.0**	+7.2 ± 8.7	87.5 ± 6.6	93.1 ± 3.5	+5.6 ± 5.6	85.6 ± 4.4
	Hard	40.9 ± 8.4	85.6 ± 8.6	+44.7 ± 11.5	55.6 ± 30.6	**88.8 ± 4.2**	+33.1 ± 30.8	40.3 ± 8.9
2s3z	Vanilla	98.4 ± 4.3	98.7 ± 2.5	+0.3 ± 5.7	99.7 ± 0.9	**100.0 ± 0.0**	+0.3 ± 0.9	99.4 ± 1.1
	Medium	79.1 ± 16.3	95.3 ± 6.0	+16.3 ± 16.8	98.1 ± 2.5	**99.7 ± 0.9**	+1.6 ± 2.4	93.4 ± 7.6
	Hard	73.4 ± 28.4	84.4 ± 36.9	+10.9 ± 26.9	72.8 ± 7.1	**99.4 ± 1.1**	+26.5 ± 7.3	89.1 ± 7.9
6h_vs_8z	Vanilla	78.7 ± 54.7	80.0 ± 55.5	+1.3 ± 1.6	**100.0 ± 0.0**	**100.0 ± 0.0**	+0.0 ± 0.0	99.4 ± 1.7
	Hard	98.1 ± 3.2	98.7 ± 1.6	+0.6 ± 3.2	98.7 ± 1.6	**100.0 ± 0.0**	+1.3 ± 1.6	99.7 ± 0.9
3s5z	Vanilla	70.9 ± 50.8	73.4 ± 53.0	+2.5 ± 78.2	98.4 ± 1.4	**99.4 ± 1.1**	+0.9 ± 1.7	96.3 ± 5.2
	Hard	29.4 ± 48.8	35.9 ± 54.7	+6.6 ± 11.7	66.6 ± 11.8	**97.2 ± 2.5**	+30.6 ± 12.9	90.9 ± 7.2
smacv2_5_units	Vanilla	67.2 ± 3.1	77.8 ± 4.8	+10.6 ± 7.6	73.8 ± 4.6	**80.9 ± 2.9**	+7.2 ± 5.9	67.8 ± 2.9
	Medium	46.9 ± 2.7	65.9 ± 7.2	+19.1 ± 6.9	57.2 ± 5.6	**67.8 ± 5.0**	+10.6 ± 2.5	45.9 ± 5.9
	Hard	23.8 ± 3.7	40.6 ± 7.3	+16.9 ± 5.2	31.9 ± 7.5	**43.8 ± 4.9**	+11.9 ± 7.3	35.9 ± 5.7
smacv2_10_units	Vanilla	36.6 ± 9.4	46.6 ± 4.2	+10.0 ± 10.2	52.5 ± 5.9	62.8 ± 2.5	+10.4 ± 5.3	**64.0 ± 7.7**
	Medium	4.1 ± 3.5	14.1 ± 6.6	+10.0 ± 8.1	36.9 ± 8.1	**60.3 ± 4.0**	+23.4 ± 8.4	48.7 ± 14.6
	Hard	8.4 ± 11.0	11.2 ± 11.2	+2.8 ± 16.2	18.1 ± 5.6	**29.4 ± 10.8**	+11.2 ± 11.7	20.6 ± 3.7

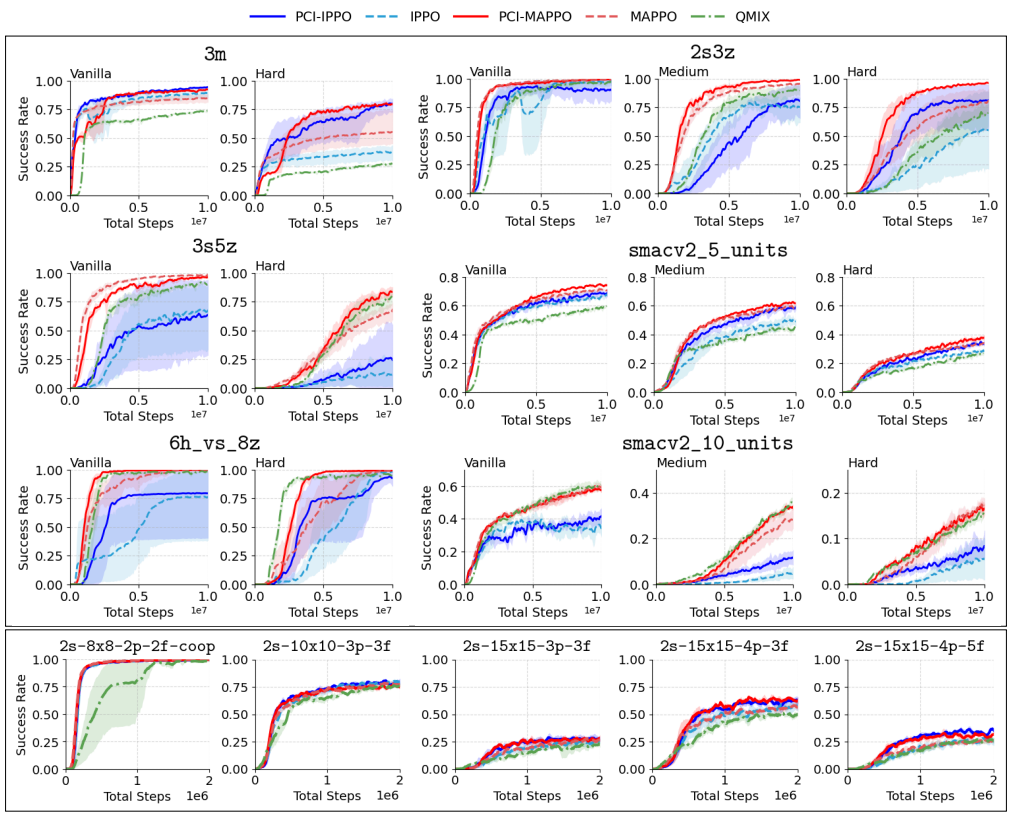


Figure 7: Training curves on SMAX (upper box) and LBF (lower box) show the mean success rates across 5 training seeds. The shaded regions indicate 95% bootstrap CIs across training seeds.

Table 6: Ablation study in `3m:hard` scenario. Each entry reports mean success rate $\pm$ 95% Student-t intervals across five training seeds. The top row represents the proposed methods.

Credit predictor	Counterfactual weighting	IPPO-based	MAPPO-based
outcome bridge	$\bar{\pi}^i$	**85.6** $\pm$ 8.6	**88.8** $\pm$ 4.2
outcome bridge	π^i	46.2 ± 26.9	38.6 ± 1.6
observational regressor	$\bar{\pi}^i$	67.0 ± 22.0	87.0 ± 2.0
observational regressor	π^i	46.0 ± 26.0	41.0 ± 2.0
none (IPPO / MAPPO)	–	40.9 ± 8.4	55.6 ± 30.6

G.3 Ablation Study

On Stag-Hunt and SMAX `3m:hard`, we conduct an ablation study to validate the outcome bridge function and proposed renormalization trick. Table 6 shows the results in SMAX `3m:hard`. It reveals that removing renormalization reduces the IPPO- and MAPPO-based results by 39 and 50 percentage points, bringing both to backbone performance or below. Once it is removed, predictor choice has little effect because $\mathbb{E}_{\pi^i}[\cdot]$ approaches the factual prediction as the policy concentrates, leaving little counterfactual contrast. On Stag-Hunt, the reductions are smaller ($92.8 \to 81.6$ and $90.9 \to 85.3$), so the magnitude is domain-dependent. Conditional on renormalization, the bridge's incremental gain over the observational regressor is not resolved on this map with five seeds: for PCI-IPPO the regressor has a lower mean and about three times the across-seed interval width, whereas for PCI-MAPPO it nearly matches the bridge.

G.4 Hyperparameters

Table 7 and Table 8 summarize hyperparameters for experiments in all experimental domains. We follow the hyperparameter settings in [29, 36] for IPPO and MAPPO. However, we use a lower learning rate for PCI-IPPO/MAPPO in several tasks because our method introduces a more delicate update signal, as $\tilde{r}_t^i$ is influenced by updating policies.

Table 7: Hyperparameters for PPO-type methods in all three domains. The selected hyperparameters and hyperparameters specific to the PCI variants are highlighted in boldface. For SMAX, the learning rates are representative values; their values for each map are shown in Table 10. Note that λ_c is set to 1.0 only for the `3m` map. It is 0.1 for other SMAX maps.

Hyperparameter	**Stag-Hunt**	**SMAX**	**LBF**
optimizer	Adam	Adam	Adam
lr	1e-4, **3e-4**, 1e-3	5e-4	1e-4, **3e-4**, 1e-3
total_timesteps	5×10^6	1×10^7	1×10^6
num_steps	2000	6400	2000
fc_hidden_dim	64	128	64
gru_hidden_dim	64	128	64
clip_eps	0.2	0.2	0.2
update_epochs	4	4	2
gamma	0.99	0.99	0.99
gae_lambda	0.95	0.95	0.95
entropy_coefficient	0.01	0.01 (only for PCI variants)	0.01
value_coefficient	0.5	0.5	0.5
max_grad_norm	0.5	0.25	0.5
activation	ReLU	ReLU	ReLU
seed	42, 1, 2, 31, 32	42, 1, 2, 31, 32	42, 1, 2, 31, 32
bridge_lr	**0.001**	**0.002**	**0.001**
$\boldsymbol{\lambda_c}$	**1**	**1** (`3m`) or **0.1** (others)	**1**
Tikhonov regularization γ_{st}	**0.0001**	**0.0001**	**0.0001**

Table 8: Hyperparameters for QMIX in all three domains.

Hyperparameter	Stag-Hunt	SMAX	LBF
optimizer	Adam	Adam	Adam
lr	1e-5, **5e-5**, 1e-4	5e-5	1e-5, **5e-5**, 1e-4
total_timesteps	5×10^6	1×10^7	1×10^6
num_steps	2048	2048	2048
fc_hidden_dim	64	512	64
gru_hidden_dim	64	512	64
gamma	0.99	0.99	0.99
max_grad_norm	10.0	10.0	10.0
batch-size	32	32	32
replay_buffer_size	5000	5000	5000
epsilon_annealing_time	8×10^5	8×10^5	8×10^5
epsilon-start, -finish	1.0, 0.05	1.0, 0.05	1.0, 0.05
target_update_interval	200	200	200
double_q	true	true	true
mixing_embed_dim	32	32	32
hypernet_layers	2	2	2
hypernet_embed_dim	64	64	64
activation	ReLU	ReLU	ReLU
seed	42, 1, 2, 31, 32	42, 1, 2, 31, 32	42, 1, 2, 31, 32

Table 9: Neural Network Parameters. We refer to the codebase of Yu et al. [42] for the network architecture.

Parameter	policy / Value Network	Outcome Bridge Network
hidden_sizes	64	64, 64
initialization_method	orthogonal	orthogonal
use_feature_normalization	false	true
use_layer_norm	true	true
recurrent_n	1	–

Table 10: Learning Rate Sweeps and Final Values in SMAX. P- refers to the proposed PCI variants. For each method and scenario, we selected the learning rate from $\{0.0001, 0.0005, 0.001, 0.002, 0.004\}$ for PPO-based methods and $\{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$ for QMIX.

Map	Target Size	P-IPPO	P-MAPPO	IPPO	MAPPO	QMIX
3m	2 (Hard)	0.0005	0.0005	0.004	0.002	0.00005
	3 (Vanilla)	0.0005	0.0005	0.004	0.002	0.00005
2s3z	3 (Hard)	0.002	0.001	0.004	0.002	0.00005
	4 (Medium)	0.0005	0.002	0.004	0.002	0.00005
	5 (Vanilla)	0.001	0.002	0.004	0.002	0.00005
3s5z	6 (Hard)	0.004	0.002	0.004	0.002	0.00005
	8 (Vanilla)	0.002	0.002	0.0005	0.002	0.00005
6h_vs_8z	6 (Hard)	0.001	0.002	0.002	0.002	0.00005
	8 (Vanilla)	0.001	0.001	0.001	0.002	0.00005
smacv2_5_units	3 (Hard)	0.002	0.001	0.004	0.002	0.00005
	4 (Medium)	0.002	0.001	0.004	0.002	0.00005
	5 (Vanilla)	0.002	0.001	0.002	0.002	0.00005
smacv2_10_units	6 (Hard)	0.002	0.001	0.001	0.002	0.00005
	8 (Medium)	0.002	0.001	0.001	0.002	0.00005
	10 (Vanilla)	0.002	0.001	0.001	0.002	0.00005

G.5 Computational Resources

Table 11 summarizes the computational resources for the benchmark experiments in Section 7. We conducted experiments on multi-core CPU workstations and GPU-equipped servers. Stag-Hunt, SMAX, and LBF experiments required approximately 125, 600, and 50 total compute hours, respectively. In Section 8, we note that the Gram matrix computation can be a computational burden. Here, we profile PCI variants against their corresponding backbones on GPU across three domains: Stag-Hunt, SMAX, and LBF, using device-synchronized timing. A bridge network update involves Gram matrix construction, which scales quadratically with the batch size $|\mathcal{B}|$, and a regularized kernel-ridge solve, which scales cubically with $|\mathcal{B}|$. Gram construction accounts for 0.3–2.3% of a single update; construction plus solving accounts for about 5% on Stag-Hunt/LBF and 27% on SMAX, where the solve acts on a larger batch. Bridge forward and backward passes account for 5–20%, and batch preprocessing for 2–6%. Overall, a PCI update takes approximately 1.8–2.7 times as long as the corresponding backbone update.

Table 11: Computational Resources by Experimental Types

Experimental Domain	**Hardware Configuration**	**Execution Time**	**Total Hours**
Stag-Hunt Experiments	10 cores on Intel Core i9-10850K (3.60GHz) 10-Core Processor	2–2.5 hours per run (algorithm dependent)	125 hours
SMAX Experiments	**Server A:** 20 cores on Intel Xeon Silver 4210 (2.20GHz) + NVIDIA GeForce RTX 2080 Ti: 11GB **Server B:** 16 cores on Intel Xeon E5-2620 v4 (2.10GHz) + NVIDIA GeForce RTX 3090 Ti: 24GB	1.5–3 hours per run (scenario & algorithm dependent)	600 hours
LBF Experiments	**Server A:** 20 cores on Intel Xeon Silver 4210 (2.20GHz) + NVIDIA GeForce RTX 2080 Ti: 11GB **Server B:** 16 cores on Intel Xeon E5-2620 v4 (2.10GHz) + NVIDIA GeForce RTX 3090 Ti: 24GB	0.25–0.5 hours per run (scenario & algorithm dependent)	50 hours

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to Sections 4, 5, 6, and 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see Appendix C for the identification assumptions. Appendix D provides the proofs of Proposition 1 and Theorem 1, and Appendix E provides the policy-gradient analysis and proofs of Propositions 2 and 3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section 6, Algorithm 1, and Appendix F, G. We also provide all the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide all the code and script files to reproduce results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 7 and Appendix F and G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean $\pm$ 95% confidence intervals or standard deviations for all experiments. Also, we report a cross-domain aggregate analysis (Table 2) and a performance profile (Figure 5) using IQM and the probability of improvement, with stratified bootstrap CIs, following Agarwal et al. [1].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Appendix G.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appendix B. The paper discusses that the work is foundational and does not involve human subjects, personal data, surveillance data, or protected attributes. It also discusses potential downstream risks in real-world multi-agent control or allocation systems, including safety, security, robustness, fairness, and environmental considerations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such a risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the creators of the existing assets used in this work, including the `marlbenchmark/on-policy` codebase used as a reference implementation for MAPPO-style training. Its third-party attribution and MIT License are documented in Appendix F and in the supplementary code package.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented, and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a supplementary code package for the proposed method, including implementation, training scripts, configuration files, README documentation, and license information. The released code is accompanied by an MIT LICENSE file and third-party attribution notices for derived components.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.