

Peer-Conditioned Observation Representation for Peer Adaptation in Multi-Agent Reinforcement Learning

Daipeng Li

University of Chinese Academy of Sciences

Pei Xu

Institute of Automation

Shiyue Cao

University of Chinese Academy of Sciences

Likun Yang

University of Chinese Academy of Sciences

Shihang Zhang

University of Chinese Academy of Sciences

Tong Wang

University of Chinese Academy of Sciences

Kaiqi Huang

`kaiqi.huang@nlpr.ia.ac.cn`

Institute of Automation

Article

Keywords:

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10918805/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Peer-Conditioned Observation Representation for Peer Adaptation in Multi-Agent Reinforcement Learning

Daipeng Li^{1,2}, Pei Xu², Shiyue Cao^{3,2}, Likun Yang^{3,2}, Shihang Zhang^{1,2}, Tong Wang^{4,2}, and Kaiqi Huang^{2,*}

¹School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, 101408, China

²National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

³School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China

⁴School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, Beijing, 100049, China

*kaiqi.huang@nlpr.ia.ac.cn

ABSTRACT

Effective adaptation to heterogeneous peers, including cooperative partners and adversarial opponents, remains challenging in multi-agent reinforcement learning, particularly under partial observability. Existing methods focus on what representations of other agents should be learned, while typically concatenating or jointly encoding the learned peer representation with the current observation as a policy input. These joint encoders leave the relationship between these sources to be discovered implicitly during policy learning. Instead, we focus on how peer representations are utilized and propose Peer-Conditioned Observation Representation (PCOR). PCOR imposes an asymmetric inductive bias: observation features provide the base representation of the current interaction state, and peer representations condition how these features are interpreted, rather than entering as parallel inputs to an unconstrained joint encoder. Experiments on two partially observable benchmarks show that PCOR attains higher final Eval performance than the reported baselines on unseen peers. Mechanistic analyses further show that PCOR constructs increasingly peer-discriminative policy representations as interaction history accumulates and induces more consistent peer-dependent observation-to-representation mappings. On the reported unseen-peer splits, explicitly structuring how peer information enters policy representations provides an effective inductive bias for adaptation.

Keywords: Multi-agent reinforcement learning, Opponent modeling, Peer adaptation

Introduction

Learning adaptive policies for heterogeneous peers remains a fundamental challenge in multi-agent reinforcement learning. Unlike single-agent settings, agents operate in environments where dynamics and rewards are influenced by other learning or adaptive entities, introducing strong non-stationarity. This issue arises in a wide range of applications, including human–robot collaboration, multi-robot coordination, strategic games, and open-ended interactive systems, making peer adaptation an important problem in multi-agent reinforcement learning. Throughout this paper, we use “peer” to denote the interacting partner and retain “opponent modeling” as the conventional name of the research area; the interacting partner may be cooperative, competitive, or mixed-motive.

Prior work on opponent modeling has focused on extracting and utilizing peer information from interaction histories^{1–13}. Common approaches infer a peer representation of the interaction partner, capturing behavioral patterns, latent intentions, or contextual signals^{1,2,5–8,10–13}. This peer representation is then incorporated into the decision-making process together with the current observation. A common realization of this paradigm is to jointly encode observations and peer representations as policy inputs.

Existing research has extensively investigated what peer representations should be learned^{1–18}. Comparatively less attention has been given to how learned peer representations should participate in policy learning. Observations and peer representations arise from fundamentally different sources: observations constitute environment-provided information describing the current interaction state, whereas peer representations are learned abstractions inferred from historical interactions. Existing methods typically treat both as input features and jointly encode them as policy inputs. While this design has proven effective in

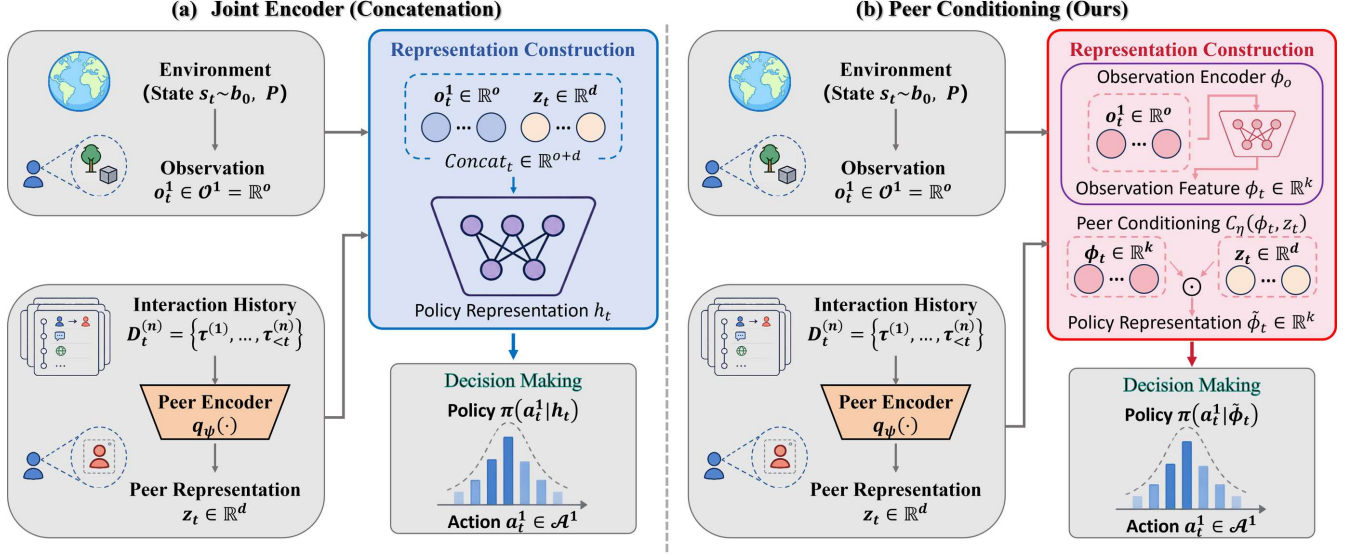


Figure 1. Architectural contrast between concatenation-based and conditioning-based agents. Shared modules that are not modified in this work (environment, interaction History / peer encoder, and decision-making) are shown in gray on both sides. Given the current observation o_t^1 and interaction history $D_t^{(n)}$, a peer representation z_t is inferred by a peer encoder q_ψ . (a) Concatenation-based agent: o_t^1 and z_t are concatenated and jointly encoded without an explicit distinction between observation and peer features, yielding an unconstrained fusion h_t for the policy $\pi(a_t^1 | h_t)$. (b) PCOR: o_t^1 is first encoded into an observation feature $\phi_t = \phi_o(o_t^1)$, which is then conditioned on z_t via $C_\eta(\phi_t, z_t)$ to construct the peer-conditioned policy representation $\tilde{\phi}_t$ for the policy $\pi(a_t^1 | \tilde{\phi}_t)$, thereby imposing an asymmetric inductive bias.

many opponent modeling approaches, it implicitly assumes that information originating from these distinct sources should be incorporated in the same manner. This raises a natural question: *should observation information and peer information necessarily participate in policy learning in the same way?*

To this end, we propose Peer-Conditioned Observation Representation (PCOR), a policy representation construction mechanism for peer adaptation. Rather than treating observations and peer representations as parallel inputs to an unconstrained joint encoder, PCOR introduces an explicit asymmetric inductive bias: observation features provide the base representation of the current interaction state, whereas peer representations modulate how these features are interpreted for policy learning. Figure 1 contrasts the conventional concatenation-based design with the proposed conditioning-based construction.

The main contributions of this paper are summarized as follows:

1. We revisit a widely adopted design paradigm in opponent modeling and argue that adaptive decision making depends not only on learning informative peer representations but also on how such representations participate in policy learning.
2. We propose Peer-Conditioned Observation Representation (PCOR), which conditions observation features on peer representations and thereby imposes an asymmetric inductive bias, instead of unconstrained joint encoding of heterogeneous information sources.
3. Experiments on two partially observable benchmarks show improved final adaptation performance over the reported opponent-modeling baselines on unseen peers. Mechanistic analyses further reveal distinct empirical patterns of PCOR across representation and policy levels, clarifying how policy representation construction relates to adaptive decision making under the accumulation of interaction history.

Related Work

Peer Adaptation

Peer adaptation concerns how agents adjust decision-making when interacting partners change. Existing studies address this problem from two main perspectives. One treats interaction variation as latent task or context shift inferred from limited experience, whereas another explicitly models such variation through peer behavior, strategy, or intent and constructs peer representations to support decision-making.

Task-oriented adaptation. Representative approaches include recurrent adaptation frameworks such as RL² that encode task inference implicitly through sequential interaction experience¹⁹; meta-reinforcement learning methods that infer latent task context from trajectories and perform context-conditioned policy adaptation, such as PEARL²⁰; and Bayesian or probabilistic formulations that maintain uncertainty-aware beliefs over latent task variables or dynamics, such as VariBAD²¹. have further extended task adaptation through offline meta-reinforcement learning²², transformer-based in-context reinforcement learning²³, and in-context planning methods that exploit large-scale trajectory data and sequence modeling capabilities^{24,25}. These methods establish generic adaptation mechanisms based on latent context inference, belief updating, and context-conditioned policy optimization. In peer adaptation settings, the dominant source of interaction variation often lies in different peers. Consequently, purely task-oriented formulations provide limited explicit characterization of peer information and how such inferred information should participate in policy learning.

Peer-oriented adaptation via opponent modeling. Existing approaches construct peer representations through diverse forms of supervision, including opponent action prediction^{1,2,10}, hidden-goal inference⁴, opponent-learning awareness³, policy representation learning^{6,8,12}, behavior modeling^{7,11}, and approximating the learning dynamics or mental states of other agents^{5,9}. The resulting representations capture complementary aspects of interacting peers, such as behavioral regularities, latent intentions, policy identities, and strategic preferences, enabling agents to adapt their decisions according to different interaction partners. More have further extended this paradigm toward increasingly challenging adaptation scenarios, including subgoal inference for long-horizon strategic reasoning¹⁴, offline opponent modeling via in-context learning from pre-collected trajectories¹⁵, hierarchical opponent modeling and planning in mixed-motive environments¹⁶, context-aware exploration for rapid peer identification¹³, uncertainty-aware opponent modeling under stochastic behaviors²⁶, contextual representation learning for adaptive multi-agent reinforcement learning¹⁷, and in-context coordination with previously unseen teammates¹⁸. These advances substantially improve adaptation efficiency and generalization across diverse interaction partners.

Despite their different adaptation mechanisms, both task-oriented and peer-oriented approaches ultimately rely on a common representation learning paradigm, where the current observation is integrated with an inferred context representation to construct the policy representation. Existing studies have substantially advanced adaptive decision-making by investigating what information should be inferred, but comparatively less attention has been given to how these learned representations should participate in policy representation construction. Although observations and inferred context jointly contribute to decision making, they play different functional roles: observations describe the current interaction state provided by the environment, whereas inferred context summarizes historical experience and provides additional information for interpreting the current observation. Nevertheless, existing methods generally integrate them through a unified representation learning process before policy optimization, without explicitly considering their distinct roles during policy representation construction.

Conditional Representation Learning

Conditional representation learning suggests a complementary perspective. In addition to treating conditioning variables as additional inputs to joint encoders, conditioning can also act on intermediate features and modulate the way learned features are transformed. In reinforcement learning, universal value function approximators condition value estimation on task or goal descriptors, demonstrating that contextual variables can shape internal computation rather than merely augment raw inputs²⁷. Hypernetworks generate network parameters conditioned on auxiliary information, enabling context-dependent function realization²⁸. Conditional normalization and feature-wise transformation mechanisms further separate base feature extraction from the utilization of contextual information, allowing contextual variables to modulate learned representations after observation features have already been extracted^{29,30}. In particular, Feature-wise Linear Modulation (FiLM) applies feature-wise affine transformations to intermediate activations according to conditioning variables³⁰.

This perspective is particularly relevant to partially observable multi-agent interaction. Observations provide environment-derived descriptions of the current state, whereas peer representations summarize interaction histories and encode contextual information about the current partner. Since these two sources serve fundamentally different purposes, there is no need for them to participate identically in policy representation construction. Instead of jointly encoding heterogeneous information sources, PCOR uses peer representations to condition observation features, allowing peer information to regulate how environment-derived representations are transformed before policy learning. Our concrete instantiation uses a multiplicative feature-wise scaling of observation features, which is closely related to the FiLM family of conditional transforms³⁰ but restricted to a gated scaling form rather than a full affine transform. Consequently, PCOR complements existing opponent modeling methods from a different perspective: rather than proposing another objective for learning more informative peer representations, it revisits how learned peer representations should participate in constructing policy representations.

Method

Problem Formulation

We consider an n -agent partially observable stochastic game (POSG),

$$G = \langle I, S, b_0, \{A_i\}_{i=1}^n, \{O_i\}_{i=1}^n, P, \{R_i\}_{i=1}^n \rangle, \quad (1)$$

where $I = \{1, \dots, n\}$ denotes the set of agents, S is the state space, b_0 is the initial state distribution, A_i and O_i are the action and observation spaces of agent i , P is the state transition function, and R_i is the reward function of agent i .

We focus on adapting a controlled agent, referred to as the ego agent, in the above POSG. The observation, action, and reward of the ego agent are denoted by o_t^1 , a_t^1 , and r_t^1 , respectively, while the remaining agents are regarded as peers.

In peer adaptation, the ego agent repeatedly interacts with the same peer over multiple consecutive episodes and exploits the accumulated interaction experience to adapt its behavior. Since peer policies are not directly observable, the ego agent must infer peer-related information from interaction histories. We refer to the N_{eps} consecutive episodes spent interacting with a fixed peer as an *adaptation horizon*:

$$\mathcal{H} = \{1, \dots, N_{\text{eps}}\}, \quad (2)$$

where each $n \in \mathcal{H}$ indexes an adaptation episode within the horizon. The n -th adaptation episode is a trajectory

$$\tau^{(n)} = \{(o_t^1, a_t^1, r_t^1, \xi_t)\}_{t=0}^{T-1}, \quad (3)$$

where T is the episode horizon and ξ_t denotes peer-related observable information, such as peer actions, public states, or communication messages.

Before timestep t in the n -th adaptation episode, let

$$D_t^{(n)} = \{\tau^{(1)}, \dots, \tau^{(n-1)}, \tau_{<t}^{(n)}\}, \quad (4)$$

denote the interaction history, where $\tau_{<t}^{(n)} = \{(o_k^1, a_k^1, r_k^1, \xi_k)\}_{k<t}$ denotes the partial trajectory collected in the current episode. $D_t^{(n)}$ contains all interaction information accumulated before the current decision, including previous adaptation episodes and the partial trajectory of the current episode. A peer representation is inferred from the accumulated interaction history as

$$z_t = q_\psi(D_t^{(n)}), \quad (5)$$

which summarizes the behavioral characteristics of the current peer based on accumulated interaction history. For notational simplicity, we write D_t in place of $D_t^{(n)}$ in the remainder of the paper when the adaptation episode index is clear from context. The ego policy is defined as

$$\pi_\theta(a_t^1 | o_t^1, z_t). \quad (6)$$

The return of the n -th adaptation episode is

$$G^{(n)} = \sum_{t=0}^{T-1} \gamma^t r_t^{1,(n)}, \quad (7)$$

where $\gamma \in [0, 1)$ is the discount factor. The objective is to maximize the cumulative discounted return over the adaptation horizon:

$$\max_{\pi} \mathbb{E} \left[\sum_{n \in \mathcal{H}} G^{(n)} \right]. \quad (8)$$

The formulation above specifies *what* information is available for adaptive decision making. The remaining question is *how* the inferred peer representation should participate in policy representation construction, which is the focus of this work.

Peer-Conditioned Observation Representation

Existing peer adaptation methods typically incorporate peer representations through joint encoding before policy learning:

$$h_t = f_\theta(o_t^1, z_t), \quad (9)$$

where h_t denotes the policy representation. Under this paradigm, research has focused primarily on learning more informative peer representations, including behavior patterns, latent goals, policy identities, and strategic contexts. Comparatively less attention has been given to how peer information participates in policy representation construction, and most methods ultimately rely on the same joint encoding mechanism regardless of how the peer representation is learned.

We revisit this design from the perspective of the functional roles of the two information sources. The current observation describes the interaction state at the present timestep, whereas the peer representation summarizes historical interaction experience and provides contextual information for interpreting the current observation. Rather than treating them symmetrically through joint encoding, we argue that peer information should regulate the construction of policy representations.

Motivated by this perspective, we propose **Peer-Conditioned Observation Representation (PCOR)**, which replaces conventional joint encoding with feature-wise conditioning. Specifically, the current observation is first encoded into an observation feature:

$$\phi_t = \phi_o(o_t^1), \quad (10)$$

which is subsequently conditioned by the inferred peer representation:

$$\tilde{\phi}_t = C_\eta(\phi_t, z_t), \quad (11)$$

where C_η denotes a peer-conditioned feature transformation. The resulting policy representation is directly consumed by the policy network as

$$\pi_\theta(a_t^1 | \tilde{\phi}_t). \quad (12)$$

Unlike conventional joint encoding, PCOR explicitly separates the functional roles of observations and peer representations. Concatenation (or other unconstrained joint encoders) does not explicitly distinguish observation features from peer features: both enter as parallel inputs, and any interaction between them must be discovered implicitly during policy learning. In contrast, PCOR imposes an asymmetric inductive bias—observation features provide the base representation of the current interaction state, while the peer representation modulates their interpretation for downstream decision making (Figure 1).

Instantiation and Optimization

To instantiate the proposed PCOR framework, we implement the abstract peer encoder q_ψ using a transformer-based peer-history encoder:

$$z_t = W_z E_\psi(D_t), \quad (13)$$

where E_ψ denotes the peer-history encoder and W_z is a learnable projection. During training, a peer-identification objective is introduced to encourage policy-discriminative peer representations:

$$L_{\text{peer}} = \text{CE}(F_{\text{cls}}(z_t), y_t). \quad (14)$$

Following the PCOR formulation, the learned peer representation is mapped to feature-wise conditioning parameters:

$$\hat{\alpha}_t = g_\eta(z_t), \quad (15)$$

which construct the peer-conditioned policy representation as

$$\tilde{\phi}_t = (1 + a \tanh(\hat{\alpha}_t)) \odot \phi_t, \quad (16)$$

where a controls the strength of peer-dependent feature scaling. Unless otherwise stated, we use $a = 0.9$, which was selected from the sensitivity study in Supplementary Section S2.1 (Supplementary Fig. S1). The resulting representation $\tilde{\phi}_t$ is shared by the policy and value networks for action selection and value estimation. Gradients from both the reinforcement learning objective and the peer-identification loss update the peer encoder q_ψ .

The policy is optimized using PPO, while the peer encoder is jointly trained with the peer-identification objective. The overall optimization objective is as follows:

$$L = L_{\text{RL}} + \lambda_p L_{\text{peer}}. \quad (17)$$

Experimentation

Experimental Setup

Environments

We evaluate PCOR on two representative partially observable multi-agent environments that require adaptation to diverse unseen peers: Partially-Observable Overcooked (POO) and Predator-Prey with Watchtowers (PPW). All the compared methods use identical environment configurations, observation settings, and evaluation protocols. Detailed environment specifications, including observation and action spaces and reward designs, are provided in Supplementary Section S1.1.

Partially-Observable Overcooked (POO)³¹ is a cooperative cooking environment where two agents must coordinate to complete a sequence of recipe-related tasks. Due to partial observability, agents only observe a local region of the kitchen and therefore cannot directly access the full environment state. Successful coordination requires agents to infer both the global task status and the behavioral preferences of their partners from interaction history. Different partners exhibit distinct preferences over ingredients and recipes, making peer adaptation essential for achieving high task success rates.

Predator-Prey with Watchtowers (PPW)¹³, built on the Multi-Agent Particle Environment³², models a mixed cooperative-competitive scenario involving two predators and two prey agents. The predators seek to maximize successful capture through cooperation, whereas the prey attempt to avoid being caught. Information about the global environment is not continuously available and can only be obtained intermittently, resulting in long periods of decision-making under incomplete observations. Consequently, effective behavior depends not only on coordinating with teammates but also on recognizing and responding to the movement patterns of adversarial agents. The coexistence of cooperation and competition makes PPW a challenging benchmark for evaluating adaptive interaction strategies.

Peer Policy Pool Construction

Following the benchmark protocol of prior work¹³, we employ diverse rule-based peer policy pools in both POO and PPW. Further details on peer-pool construction are provided in Supplementary Section S1.2.

For POO, peer agents are implemented using rule-based policies with distinct ingredient and recipe preferences. Each peer consistently prioritizes a particular subset of ingredients or dishes, resulting in diverse collaboration styles that require different coordination strategies from the ego agent.

For PPW, peer agents are designed to exhibit diverse movement, exploration, and coordination behaviors. Different peers induce substantially different interaction patterns, requiring the ego agent to continuously infer peer characteristics and adjust its policy accordingly.

The complete peer pool is partitioned into disjoint training and evaluation sets. During training, agents interact only with peers from the training pool. During evaluation, all methods are tested exclusively against unseen peers sampled from the evaluation pool. This protocol evaluates whether learned peer representations z_i and peer-conditioned policy representations $\tilde{\phi}_i$ generalize beyond previously observed peer behaviors.

Baselines

We compare PCOR against the following baselines. Hyperparameters and implementation settings for all methods are reported in Supplementary Section S1.3 and Supplementary Tables S1 and S2.

1. **PACE**¹³ encourages informative exploration by maximizing the mutual information between the peer policy and the ego agent’s cross-episode trajectories, thereby facilitating peer identification and adaptation.
2. **PACE-TF**¹³ replaces the recurrent peer encoder in PACE with a transformer-based encoder while keeping the remaining components unchanged.
3. **GSCU**¹² learns a conditional response policy based on offline policy embeddings and opponent-conditioned reinforcement learning. Importantly, GSCU assumes access to peer observations and actions after each episode for opponent inference, which is a stronger information assumption than the partially observable setting used by the other methods; we therefore report it mainly as a reference under its original protocol rather than as a strictly matched baseline.
4. **LIAM**¹⁰ introduces an auxiliary opponent reconstruction objective that predicts the peer’s observations and actions from the ego agent’s experience, yielding peer representations for adaptive decision-making.
5. **LIAMX**¹⁰ extends LIAM by incorporating cross-episode contextual information, enabling opponent inference over longer temporal horizons.
6. **LILI**¹¹ leverages transition dynamics collected in previous episodes to capture peer-dependent environmental changes, treating peer behaviors as latent environment dynamics during policy learning.

(a) Partially Observable Overcooked (POO)				(b) Predator–Prey with Watchtowers (PPW)			
Method	Train	Eval	All	Method	Train	Eval	All
PPO	0.076 ± 0.019	0.062 ± 0.017	0.069 ± 0.014	PPO	−3.93 ± 0.07	−4.66 ± 0.02	−4.36 ± 0.06
Generalist	0.139 ± 0.020	0.140 ± 0.034	0.133 ± 0.018	Generalist	−4.47 ± 0.26	−4.67 ± 0.08	−4.66 ± 0.15
LIAM	0.079 ± 0.012	0.071 ± 0.015	0.075 ± 0.015	LIAM	−3.91 ± 0.12	−4.65 ± 0.08	−4.39 ± 0.06
LIAMX	0.146 ± 0.019	0.152 ± 0.019	0.141 ± 0.023	LIAMX	−2.83 ± 0.07	−4.14 ± 0.13	<u>−3.65 ± 0.18</u>
LILI	0.117 ± 0.009	0.134 ± 0.024	0.116 ± 0.009	LILI	−4.45 ± 0.11	−4.81 ± 0.21	−4.72 ± 0.11
GSCU	0.019 ± 0.019	0.022 ± 0.022	0.020 ± 0.022	GSCU	−5.77 ± 0.08	−5.93 ± 0.12	−5.86 ± 0.12
PACE	<u>0.452 ± 0.149</u>	<u>0.406 ± 0.117</u>	<u>0.438 ± 0.142</u>	PACE	−3.19 ± 0.15	<u>−4.00 ± 0.14</u>	−3.69 ± 0.09
PACE-TF	0.114 ± 0.028	0.146 ± 0.031	0.121 ± 0.018	PACE-TF	−4.44 ± 0.11	−4.83 ± 0.08	−4.70 ± 0.02
PCOR (ours)	0.571 ± 0.095	0.511 ± 0.080	0.554 ± 0.061	PCOR (ours)	<u>−2.91 ± 0.07</u>	−3.72 ± 0.07	−3.35 ± 0.05

Table 1. Performance comparison of peer adaptation methods. (a) Success rates ($\uparrow$) on POO. (b) Average episodic returns ($\uparrow$) on PPW. Train, Eval, and All denote the seen peer set, unseen peer set, and their union, respectively. The values are mean $\pm$ s.d. over five random seeds.

- Generalist** is a recurrent policy equipped with cross-episode memory, allowing it to exploit historical interaction trajectories without employing explicit peer modeling mechanisms.
- PPO**³³ serves as a non-adaptive baseline that relies solely on the current observation and performs no cross-episode reasoning or peer-specific adaptation.

Training and Evaluation Protocol

All methods are trained using PPO under identical training budgets and environment interaction steps. Training budgets and optimization hyperparameters are summarized in Supplementary Section S1.3.

During training, a peer is sampled from the training peer pool at the beginning of an adaptation horizon and remains fixed for $N_{\text{eps}} = 5$ consecutive episodes. This setting enables the accumulation of interaction history with the same peer across episodes. After the adaptation horizon terminates, a new peer is sampled and the adaptation process is reset.

The evaluation follows the same protocol using peers drawn from the training peer pool (Train), the unseen peer pool (Eval), or their union (All). For each evaluation peer, we perform 20 independent adaptation trials. In each trial, interaction proceeds over $N_{\text{eps}} = 5$ consecutive episodes, denoted g1–g5, with the interaction history initialized at the beginning of the trial and reset after the fifth episode. The first episode corresponds to the initial interaction stage with limited peer-specific information, whereas later episodes provide progressively richer interaction histories. We report the mean performance averaged over g1–g5 and the 20 adaptation trials (Eval20 protocol). All the results are further averaged over peers and five independently trained models with different random seeds.

Main Results

Overall Performance

Figure 2 reports the training and evaluation performance of all methods on peers from the training pool and previously unseen peers. Tables 1 and 2 summarize the corresponding final performance.

On PPW, PACE achieves stronger performance during the early stage of training and remains superior to PCOR before approximately 0.3×10^7 environment steps. As training progresses, PCOR steadily improves and eventually surpasses PACE. This result suggests that while mutual-information-based peer identification provides useful signals during early learning, later-training returns favor PCOR over PACE. On the training peer set of PPW, LIAMX attains the highest final return (Table 1), whereas PCOR achieves the strongest returns on the unseen evaluation set and on the combined All set.

A similar trend is observed on POO. LILI exhibits stronger performance during the initial stage of training and remains competitive before approximately 0.6×10^7 environment steps. After this point, PCOR consistently outperforms all competing methods and converges to the highest final performance. Compared with existing approaches that primarily focus on modeling peer identities or interaction dynamics, this pattern is consistent with more effective use of z_t when constructing $\hat{\phi}_t$ under partial observability.

The evaluation results on unseen peers closely mirror the trends observed during training. PCOR attains the highest final Eval performance on both environments. The persistence of this advantage on held-out peers from the same environments indicates that the learned peer-conditioned policy representations $\hat{\phi}_t$ are not limited to the training partners. We also note that PACE-TF, which uses a transformer peer encoder within the PACE pipeline but retains concatenation-style fusion, underperforms both PACE and PCOR; together with the Concat ablation below, this suggests that encoder choice alone does not explain the gains of PCOR.

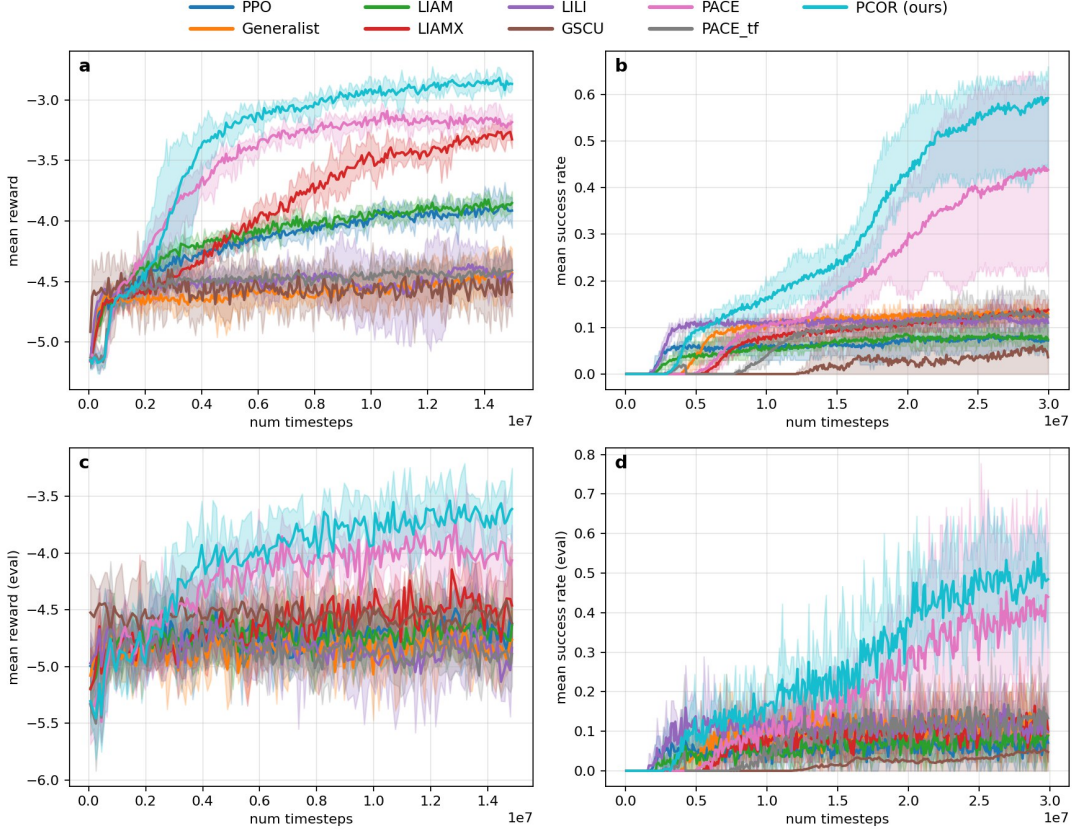


Figure 2. Baseline comparison learning curves. (a) Mean reward on PPW (training peers). (b) Mean success rate on POO (training peers). (c) Mean reward on PPW (unseen evaluation peers). (d) Mean success rate on POO (unseen evaluation peers). The shaded regions denote standard deviation across five random seeds.

(a) Partially Observable Overcooked (POO)				(b) Predator–Prey with Watchtowers (PPW)			
Method	Train	Eval	All	Method	Train	Eval	All
PCOR	0.571 ± 0.095	0.511 ± 0.080	0.554 ± 0.061	PCOR	-2.91 ± 0.07	-3.72 ± 0.07	-3.35 ± 0.05
Concat	0.289 ± 0.133	0.302 ± 0.059	0.273 ± 0.119	Concat	-3.10 ± 0.35	-3.87 ± 0.27	-3.56 ± 0.31
Pure Obs	0.076 ± 0.019	0.062 ± 0.017	0.069 ± 0.014	Pure Obs	-3.93 ± 0.07	-4.66 ± 0.02	-4.36 ± 0.06
Pure Z	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	Pure Z	-5.00 ± 0.03	-5.25 ± 0.14	-5.19 ± 0.03
Weak-Z	0.116 ± 0.019	0.120 ± 0.023	0.127 ± 0.034	Weak-Z	-4.42 ± 0.09	-4.76 ± 0.08	-4.68 ± 0.07

Table 2. Ablation study results for PCOR. (a) Success rates ($\uparrow$) on POO. (b) Average episodic returns ($\uparrow$) on PPW. Train, Eval, and All denote the seen peer set, unseen peer set, and their union, respectively. Variants are defined as in Fig. 3. The values are mean $\pm$ s.d. over five random seeds.

Ablation Study

To investigate the respective contributions of observation information, peer representations, and their integration into the policy representation, we perform an ablation study by systematically removing individual information sources or altering how peer representations are utilized. Figure 3 shows the learning curves on training peers and unseen evaluation peers. PCOR achieves the strongest mean performance among the ablation variants on both environments and on both training and unseen evaluation peers.

Among all ablated variants, Concat is the most competitive. Since Concat employs the same peer encoder q_ψ and observation encoder ϕ_o as PCOR while differing only in how peer information is incorporated—through unconstrained concatenation rather than peer-conditioned feature transformation—the performance gap highlights the benefit of the asymmetric inductive bias over unconstrained fusion. On POO, PCOR improves the evaluation success rate from 30.2% to 51.1%, a gain of 20.9 percentage points. On PPW, Concat remains the strongest ablation variant; PCOR attains higher mean returns on Train and All, while the

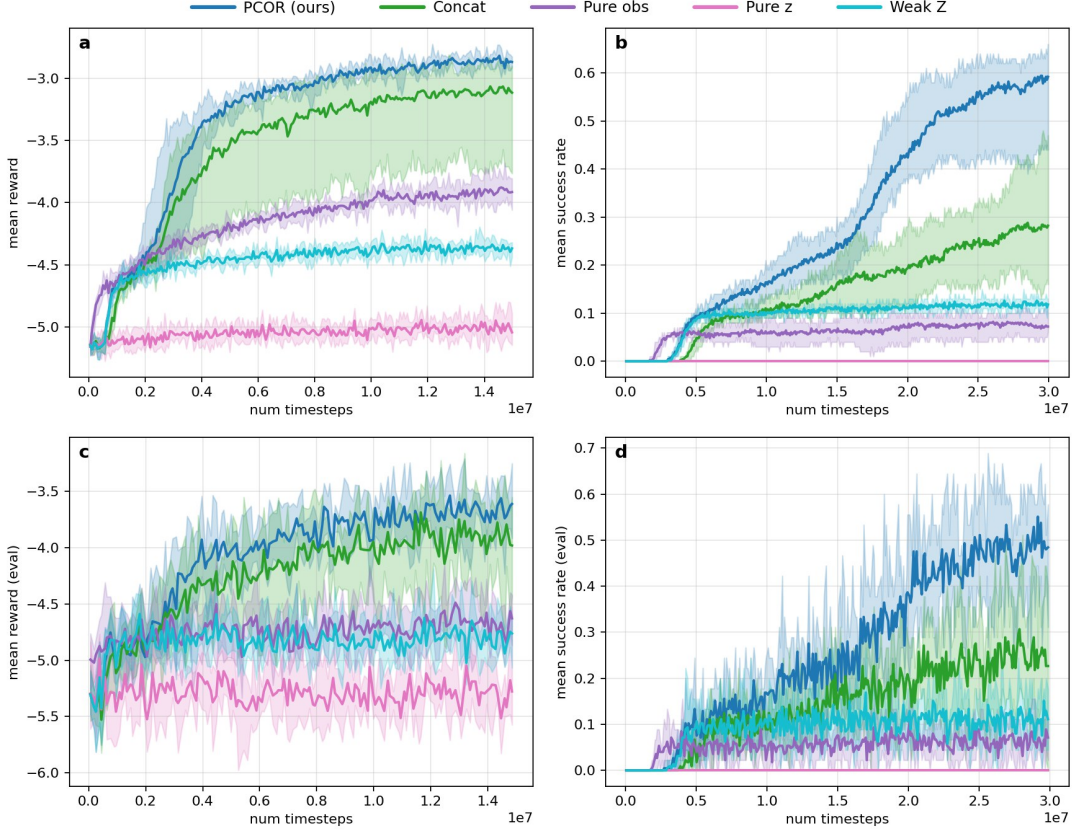


Figure 3. Ablation study on the contributions of observation information, peer representations, and their integration into the policy representation. (a) Mean episodic return on PPW (training peers). (b) Mean success rate on POO (training peers). (c) Mean episodic return on PPW (unseen evaluation peers). (d) Mean success rate on POO (unseen evaluation peers). The shaded regions denote standard deviation across five random seeds. Concat replaces peer-conditioned policy representation construction with direct concatenation of observation and peer representations; Pure Obs removes peer representations; Pure Z removes observation information; Weak-Z preserves the PCOR architecture while using a near-zero conditioning coefficient, rendering peer representations nearly ineffective.

Eval margin is smaller and less conclusive given the overlapping seed-level variation (Table 2). The POO Concat comparison indicates that the improvement cannot be attributed solely to learning informative z_t , but also depends on how z_t is integrated into the policy representation for downstream decision making; the corresponding PPW comparison provides weaker support for the same claim. PCOR also exhibits lower seed-to-seed variation than Concat on the POO training and combined sets (0.095 vs. 0.133 on Train; 0.061 vs. 0.119 on All).

Weak-Z substantially underperforms the full model despite preserving the same architectural structure. In this variant, the influence of z_t on constructing $\tilde{\phi}_t$ is reduced to a negligible level by scaling its contribution with a coefficient of 10^{-6} , making $\tilde{\phi}_t$ effectively equivalent to ϕ_t . As a result, performance drops from 51.1% to 12.0% on POO and deteriorates similarly on PPW. This observation suggests that the gains of PCOR do not arise from the architecture itself, but from allowing peer representations to meaningfully influence the construction of $\tilde{\phi}_t$.

Finally, removing either source of information leads to severe degradation. Pure Obs performs substantially worse than PCOR, confirming the importance of z_t for peer-specific adaptation. Conversely, Pure Z completely fails on POO and achieves the worst performance on PPW, indicating that z_t alone cannot support effective decision making without constructing $\tilde{\phi}_t$ from ϕ_t . Together, these results demonstrate that successful adaptation requires both observation features ϕ_t and peer representations z_t , as well as an effective mechanism for integrating them into the policy representation.

Overall, the results indicate that both the availability of peer information and, more critically, its structured integration into the policy representation are essential for effective adaptation.

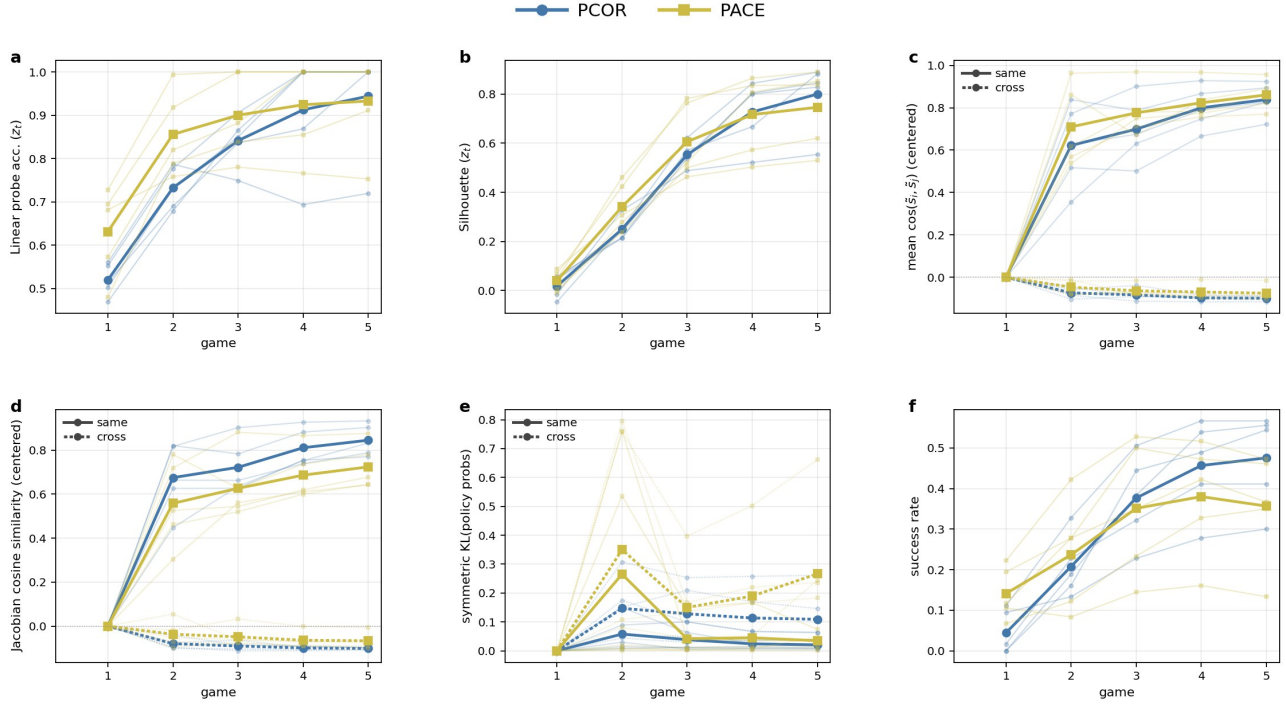


Figure 4. Mechanism analysis of PCOR and PACE on POO across the five-game adaptation horizon (pooled over five training seeds: thin lines, individual seeds; thick lines, means). (a) Peer representation quality of z_t measured by linear probe accuracy for peer identity ($\uparrow$). (b) Peer representation clusterability of z_t measured by silhouette score ($\uparrow$). (c) Discriminability of policy representations (ϕ_t for PCOR, h_t for PACE) measured by centered pairwise cosine similarity (same-peer $\uparrow$; cross-peer $\downarrow$). (d) Consistency of the observation-to-representation mapping measured by centered Jacobian cosine similarity (same-peer $\uparrow$; cross-peer $\downarrow$). (e) Policy-level behavioral difference measured by symmetric KL divergence between opening-step action distributions (same-peer vs. cross-peer). (f) Adaptation performance measured by Eval20 success rate on unseen peers ($\uparrow$). In (c)–(e), solid lines denote same-peer comparisons and dashed lines denote cross-peer comparisons.

Mechanism Analysis

To further investigate the performance difference between PCOR and PACE, we conduct additional analyses on POO. Since POO involves interaction with a single peer and exhibits interpretable coordination patterns, it provides a suitable setting for analyzing how peer information is represented and used during adaptation. The analyses are presented in the same order as peer information is incorporated into policy learning, beginning with learned peer representations, followed by peer-conditioned policy representations, and finally observation sensitivity and policy behavior. Results are aggregated over five training seeds unless otherwise specified.

For all analyses in this section, PCOR and PACE are evaluated using the same set of unseen peers, initial configurations, and adaptation sessions to ensure fair comparison. Each session consists of five consecutive games, with the interaction history reset at the beginning of every session. For analyses in Fig. 4c–e, same-peer comparisons are computed across independent replays of the same peer, whereas cross-peer comparisons are computed across different peers under the same replay index. Therefore, same-peer metrics characterize the consistency of representations or behaviors for a given peer, while cross-peer metrics measure the degree of separation induced by different peer identities. Formal definitions of all mechanism-analysis metrics are provided in Supplementary Section S1.4.

Peer and Policy Representations

To investigate how peer information is represented before and after policy representation construction, we analyze both the learned peer representations z_t and the policy representations produced by the two methods, namely the peer-conditioned policy representation ϕ_t in PCOR and the jointly encoded representation h_t in PACE. Linear probe accuracy and silhouette score are used to evaluate the peer discriminability of z_t throughout the adaptation horizon (formal definitions in Supplementary Section S1.4.1). As shown in Fig. 4a,b, both metrics increase steadily from g1 to g5, indicating that peer representations become progressively more discriminative as interaction history accumulates. PACE exhibits higher probe accuracy and silhouette

scores during the early stages of adaptation, whereas PCOR gradually catches up and achieves comparable or slightly higher discriminability by the end of the adaptation horizon.

The policy representations are further analyzed using centered pairwise cosine similarity (definitions of raw and centered cosine similarity in Supplementary Section S1.4.2). Figure 4c reports the same-peer and cross-peer similarities of $\tilde{\phi}_t$ for PCOR and h_t for PACE throughout the adaptation horizon. Both methods maintain high same-peer similarity and low cross-peer similarity, indicating that peer-specific information is preserved after policy representation construction. Compared with PCOR, PACE consistently exhibits slightly higher same-peer similarity, while the overall same–cross separation remains comparable between the two methods.

These results suggest that both methods progressively construct increasingly peer-discriminative policy representations as interaction history accumulates. Although differences exist in the evolution of z_t and the resulting policy representations, the representation-level analyses alone do not fully explain the eventual differences in adaptation performance. Complementary pooled statistics, raw similarity trajectories, and a qualitative t-SNE visualization are provided in Supplementary Section S2.2. This motivates further investigation into how these representations are reflected in subsequent policy behavior.

Observation-to-Representation Sensitivity Analysis

To investigate how peer information influences the construction of policy representations, we analyze the observation-to-representation mapping using Jacobian-based sensitivity analysis. Specifically, the Jacobian characterizes how changes in the current observation o_t affect the resulting policy representation, namely $\tilde{\phi}_t$ for PCOR and h_t for PACE, while keeping the learned peer representation fixed. This provides a local characterization of how observations are transformed into policy representations under different peer conditions. If interactions with the same peer induce similar policy representation construction mechanisms, comparable observation-to-representation transformation patterns are expected across repeated interactions. In contrast, different peers may induce distinct transformation behaviors. Based on this intuition, Jacobian similarities are compared between same-peer and cross-peer trajectories. More detailed definitions of the observation–representation Jacobian descriptors and associated similarity metrics are provided in Supplementary Section S1.4.3, with complementary analyses in Supplementary Section S2.3.

Figure 4d shows centered Jacobian cosine similarities throughout the adaptation horizon. Both methods exhibit increasingly consistent observation-to-representation transformations for repeated interactions with the same peer as interaction history accumulates. However, PCOR consistently achieves higher same-peer similarity than PACE while maintaining comparable cross-peer similarity, resulting in a larger separation between same-peer and cross-peer Jacobian patterns across the adaptation process.

These results indicate that PCOR constructs more consistent peer-conditioned observation-to-representation mappings than PACE under the reported similarity metrics. Although both methods preserve peer-specific information in their policy representations, PCOR exhibits more coherent local transformation patterns for repeated interactions with the same peer. This is consistent with the hypothesis that conditioning observation features yields a more stable policy representation construction mechanism under peer contexts; we emphasize that the Jacobian analysis is correlational and does not by itself establish causality.

Policy-Level Behavioral Differences

To examine whether peer-conditioned policy representations are reflected in policy behavior, we analyze the symmetric KL divergence between policy action distributions at the opening timestep of each adaptation episode (formal definition in Supplementary Section S1.4.4). We focus on the episode opening because all evaluations use the same initial environment configuration, resulting in identical initial observations across trajectories. Under this controlled setting, differences in the initial policy arise primarily from the accumulated interaction history and the inferred peer representation rather than variations in the current observation.

Figure 4e reports same-peer and cross-peer policy divergences across the adaptation horizon. Regardless of the method, the cross-peer KL divergence remains consistently higher than the same-peer KL divergence after g2. This indicates that both methods learn to produce different initial policies for different peers while maintaining relatively consistent behaviors for repeated interactions with the same peer. As interaction history accumulates, the policy therefore becomes increasingly conditioned on the inferred peer identity rather than remaining peer-agnostic.

Compared with PACE, PCOR consistently exhibits lower policy divergence under both same-peer and cross-peer comparisons. Notably, although PCOR still preserves a clear separation between same-peer and cross-peer policies, the overall magnitude of policy variation is substantially smaller. Combined with the preceding Jacobian analysis, this pattern is consistent with PCOR adapting through comparatively smoother policy adjustments under peer-conditioned representations, while still preserving peer-specific behaviors. We do not claim that smaller policy KL is the causal cause of higher success; rather, it is an accompanying behavioral signature observed alongside the performance gap. Episode-mean policy divergences and a longer ten-episode adaptation session are reported in Supplementary Section S2.4.

Connecting Mechanism Analysis to Adaptation Performance

To relate the mechanism analyses to adaptation performance, we compare the adaptation trajectories of PCOR and PACE throughout the Eval20 protocol (per-episode success-rate definition in Supplementary Section S1.4.4). Figure 4f reports the success rates on unseen evaluation peers from g1 to g5, illustrating how performance evolves as interaction history accumulates. PACE achieves higher success rates during the early stages of adaptation, particularly in g1 and g2, when limited interaction history is available. As adaptation proceeds, PCOR improves more steadily and surpasses PACE in the later stages. On unseen peers, PCOR achieves success rates of 45.7% and 47.6% in g4 and g5, compared with 38.0% and 35.7% for PACE.

These trajectories reveal an early-late trade-off rather than uniform dominance: PACE is preferable when only one or two episodes of interaction are available, whereas PCOR yields higher success once richer histories have been collected. Averaged Eval20 scores therefore summarize the full horizon and can mask stage-dependent differences. Combined with the preceding analyses, the results reveal different empirical characteristics of the two methods across representation, policy representation construction, and policy behavior. While PACE learns more discriminative peer representations during the early stages of adaptation, PCOR constructs more consistent peer-conditioned observation-to-representation mappings and exhibits smaller policy divergences while still preserving clear peer-specific policy behaviors. These findings suggest that adaptation performance is empirically associated not only with how peer information is represented, but also with how the learned representations are incorporated into policy learning. Establishing causal links among these quantities remains an open question, as discussed below.

Discussion

Adapting to previously unseen peers is a fundamental challenge in partially observable multi-agent environments. Existing opponent modeling methods learn peer representations from interaction history and concatenate or jointly encode them with observations as policy inputs. In this work, we revisited this paradigm from the perspective of policy representation construction, using peer representations to condition observation features before policy learning. Across both cooperative and mixed cooperative-competitive environments, PCOR achieves competitive or superior final performance relative to existing baselines on unseen peers. The performance gains are particularly evident during the later stages of the adaptation horizon, as richer interaction history becomes available; early-horizon episodes can still favor methods with stronger immediate peer identification, such as PACE. On the PPW training-peer set, LIAMX attains a higher final return than PCOR.

The mechanism analyses suggest that learning peer representations and incorporating them into adaptive decision making are related but distinct aspects of opponent modeling. Although PACE learns more discriminative peer representations during the early stages of adaptation, this advantage does not directly persist throughout the adaptation process. At the representation level, both methods progressively construct peer-discriminative policy representations as interaction history accumulates. However, PCOR exhibits more consistent peer-conditioned observation-to-representation mappings and smaller policy divergences while still maintaining clear peer-specific policy behaviors. These findings suggest that the effectiveness of opponent modeling is associated not only with the information captured by learned peer representations, but also with how such representations are incorporated into policy-relevant representation construction. This perspective complements existing opponent modeling approaches that primarily focus on learning informative peer representations by highlighting the importance of representation utilization in adaptive policy learning.

Several limitations remain. First, the experiments focus on two benchmark environments with predefined rule-based peer pools and finite adaptation horizons ($N_{\text{eps}} = 5$ in the main protocol). It remains unclear whether the observed advantages persist for larger behavioral diversity, longer interaction horizons, learning or non-stationary peers, or additional multi-agent domains. Moreover, the reported gains of PCOR emerge mainly in later episodes of the horizon; when only one or two episodes of interaction are available, methods with stronger early peer identification, such as PACE, can still perform better, and averaged Eval20 scores can mask this trade-off. Second, PCOR is evaluated primarily within the PACE training and evaluation pipeline, including a peer-identification loss that classifies discrete peer identities during training. Although identity labels are not required at execution time, this supervision may not extend to unlabeled or open-ended peer populations. The Concat ablation isolates fusion under matched encoders, but broader comparisons with alternative conditioning operators (for example, full FiLM affine transforms or hypernetwork-generated layers) are left for future work. Third, the mechanism analyses are conducted only on POO. The analyses reveal consistent empirical associations among peer representations, policy representations, observation sensitivity, policy behavior, and adaptation performance, but do not establish causal relationships among them, nor do they show whether the same representation-level patterns hold on PPW. Therefore, developing more principled analytical tools and controlled interventions to understand how peer information influences sequential decision making remains an important direction. In addition, although this study focuses on opponent modeling, the proposed policy representation construction mechanism may extend to broader adaptive decision-making problems in which latent contextual information, rather than peer representations specifically, must be integrated with the agent’s observations for policy learning.

Code and data availability

The experimental environments are built on the publicly available Overcooked environment³¹ and the Multi-Agent Particle Environment (MPE)³². Source code implementing PCOR and the reported baselines is available at <https://huggingface.co/CooCooDaYo/PCOR>.

References

1. He, H., Boyd-Graber, J., Kwok, K. & Daumé III, H. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, 1804–1813 (2016).
2. Hong, Z.-W., Su, S.-Y., Shann, T.-Y., Chang, Y.-H. & Lee, C.-Y. A deep policy inference Q-network for multi-agent systems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1388–1396 (2018).
3. Foerster, J. *et al.* Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 122–130 (2018).
4. Raileanu, R., Denton, E., Szlam, A. & Fergus, R. Modeling others using oneself in multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 4257–4266 (2018).
5. Rabinowitz, N. *et al.* Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 4218–4227 (2018).
6. Grover, A., Al-Shedivat, M., Gupta, J., Burda, Y. & Edwards, H. Learning policy representations in multiagent systems. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 1802–1811 (2018).
7. Papoudakis, G. & Albrecht, S. V. Variational autoencoders for opponent modeling in multi-agent systems. *arXiv preprint arXiv:2001.10829* (2020).
8. Jiang, H., Yu, Y. & Lu, Z. Metric policy representations for opponent modeling. *arXiv preprint arXiv:2106.05802* (2022).
9. Yu, X., Jiang, J., Zhang, W., Jiang, H. & Lu, Z. Model-based opponent modeling. In *Advances in Neural Information Processing Systems*, vol. 35, 28208–28221 (2022).
10. Papoudakis, G., Christianos, F. & Albrecht, S. Agent modelling under partial observability for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, vol. 34, 19210–19222 (2021).
11. Xie, A., Losey, D., Tolsma, R., Finn, C. & Sadigh, D. Learning latent representations to influence multi-agent interaction. In *Proceedings of the 2020 Conference on Robot Learning*, vol. 155, 575–588 (2021).
12. Fu, H. *et al.* Greedy when sure and conservative when uncertain about the opponents. In *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 6829–6848 (2022).
13. Ma, L., Wang, Y., Zhong, F., Zhu, S.-C. & Wang, Y. Fast peer adaptation with context-aware exploration. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 33963–33982 (2024).
14. Yu, X., Jiang, J. & Lu, Z. Opponent modeling based on subgoal inference. In *Advances in Neural Information Processing Systems*, vol. 37, 60531–60555 (2024).
15. Jing, Y. *et al.* Towards offline opponent modeling with in-context learning. In *The Twelfth International Conference on Learning Representations* (2024).
16. Huang, Y. *et al.* Efficient adaptation in mixed-motive environments via hierarchical opponent modeling and planning. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 20004–20022 (2024).
17. Xu, Z. *et al.* Reidentify: Context-aware identity generation for contextual multi-agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267, 69262–69278 (2025).
18. Wang, H.-C., Chuang, H.-C., Cheng, H.-C., Wu, D.-J. & Sun, S.-H. CooT: Learning to coordinate in-context with coordination transformers. *arXiv preprint arXiv:2506.23549* (2025).
19. Duan, Y. *et al.* RL²: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2017).
20. Rakelly, K., Zhou, A., Finn, C., Levine, S. & Quillen, D. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 5331–5340 (2019).
21. Zintgraf, L. *et al.* VariBAD: Variational bayes-adaptive deep RL via meta-learning. *J. Mach. Learn. Res.* **22**, 1–39 (2021).
22. Mitchell, E., Rafailov, R., Peng, X. B., Levine, S. & Finn, C. Offline meta-reinforcement learning with advantage weighting. In *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 7780–7791 (2021).

23. Laskin, M. *et al.* In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations* (2023).
24. Son, J., Lee, S. & Kim, G. Distilling reinforcement learning algorithms for in-context model-based planning. In *The Thirteenth International Conference on Learning Representations* (2025).
25. Wu, W. *et al.* Mixture-of-experts meets in-context reinforcement learning. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems* (2025).
26. Yang, L. *et al.* Uncertainty-aware opponent modeling for deep reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 2217–2225 (2025).
27. Schaul, T., Horgan, D., Gregor, K. & Silver, D. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 1312–1320 (2015).
28. Ha, D., Dai, A. & Le, Q. V. HyperNetworks. *arXiv preprint arXiv:1609.09106* (2016).
29. Dumoulin, V., Shlens, J. & Kudlur, M. A learned representation for artistic style. In *International Conference on Learning Representations* (2017).
30. Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 3942–3951 (2018).
31. Charakorn, R., Manoonpong, P. & Dilokthanakul, N. Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations* (2023).
32. Lowe, R. *et al.* Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Author contributions statement

D.L. conceived the study, designed the methodology, implemented the algorithms, conducted the experiments, analyzed the results, and drafted the manuscript. P.X. contributed substantially to the study design, experimental analysis, interpretation of the results, and manuscript revision. S.C., L.Y., S.Z., and T.W. contributed to methodological discussions, experimental support, and manuscript review and editing. K.H. supervised the project and critically revised the manuscript. All authors reviewed the manuscript.

Funding

This research was funded by the Beijing Natural Science Foundation (Grant No. 4264131).

Additional information

Competing interests: The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [submissionSI.pdf](#)