

Supplementary Information

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

Contents

S1	Experimental Details	2
S1.1	Environments	2
S1.2	Peer Policy Pool Construction	2
S1.3	Training Details	3
S1.4	Evaluation Metrics for Mechanism Analysis	3
S2	Additional Results	6
S2.1	Impact of Coefficient a	6
S2.2	Peer and Policy Representations	6
S2.3	Observation-to-Representation Sensitivity Analysis	8
S2.4	Policy-Level Behavioral Differences	9
	References	10

S1 Experimental Details

S1.1 Environments

S1.1.1 Partially-Observable Overcooked

Partially-Observable Overcooked (POO)¹ is a partially observable cooperative cooking environment adopted from PACE², which extends the original Overcooked benchmark with a more challenging multi-recipe setting. The environment contains six ingredients (tomato, onion, carrot, lettuce, potato, and broccoli) and nine recipes, requiring agents to coordinate diverse cooking behaviors. Two agents work together to collect ingredients, chop them on cutting boards, assemble dishes according to recipe requirements, and finally deliver completed meals.

To encourage cooperation, the kitchen is divided by a counter that separates the workspace into two areas. Many objects, including ingredients and plates, must therefore be passed between agents to complete recipes efficiently. Partial observability is introduced by splitting the kitchen horizontally into an upper and a lower room. Each agent can observe only entities located in the same room, making successful collaboration dependent on effective coordination under incomplete information.

Observation space. The observation for each agent is represented by a 105-dimensional feature vector, including its position, orientation, held object, nearby interactive objects, and other environment states. Visibility indicators are provided for relevant entities so that invisible objects can be distinguished from observable ones.

Action space. Each agent selects actions from a discrete action space consisting of six actions: moving in four cardinal directions, interacting with nearby objects, and remaining idle.

Reward. The environment uses a shared team reward. Agents receive an interaction reward when an object is manipulated for the first time, a progress reward whenever a recipe advances to the next preparation stage, and a completion reward after a valid dish is successfully delivered.

S1.1.2 Predator-Prey with Watchtowers

Predator-Prey with Watchtowers (PPW)² is a partially observable mixed cooperative-competitive environment introduced by PACE based on the Multi-Agent Particle Environment (MPE)³. The environment consists of two predator agents, two prey agents, multiple static landmarks, and four watchtowers. The predators cooperate to capture all prey, whereas the prey attempt to evade capture. Each episode lasts for at most 40 time steps and terminates immediately once every prey has been touched by a predator.

To increase task difficulty, each predator has only local observations within a fixed observation radius of 0.2. Four watchtowers are placed at the corners of the map. When the ego agent reaches a watchtower, it temporarily gains full observability of all agents and landmarks in the environment.

Observation space. Each observation is represented by a 37-dimensional vector containing relative positions and velocities of agents together with landmark locations. Visibility indicators are included for every observable entity. For entities outside the observation range, the corresponding visibility flag is set to zero, and their state information is masked accordingly. All positions are expressed relative to the ego agent.

Action space. The action space is discrete and consists of five actions: moving left, right, up, or down, and remaining stationary.

Reward. Predators share a common cooperative reward designed to encourage coordinated coverage of all prey. At each time step, the reward is computed as the negative weighted sum of the minimum Euclidean distance between each prey and its nearest predator, promoting strategies in which different predators simultaneously track different prey.

$$r = -c \sum_{b \in \mathcal{B}} \min_{a \in \mathcal{A}} d(a, b), \quad (\text{S1})$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the sets of predators and prey, respectively, $d(\cdot, \cdot)$ is the Euclidean distance, and $c = 0.1$. By minimizing the distance between each prey and its nearest predator, this reward encourages predators to distribute themselves across different prey rather than pursuing the same target, thereby promoting coordinated coverage.

S1.2 Peer Policy Pool Construction

Following PACE², we construct the training and evaluation peer pools using rule-based policies with diverse behavioral preferences. The peer pools are identical to those used in PACE to ensure a fair comparison. Below, we briefly summarize the peer generation process for each environment.

In POO, each peer policy is associated with a preferred recipe among the nine available recipes. The peer prioritizes subtasks related to its preferred recipe, including collecting ingredients, preparing dishes, and delivering completed meals. During execution, the peer repeatedly selects a preferred subtask and follows the shortest path toward its completion. Two stochastic parameters further diversify behaviors by controlling navigation preferences among multiple shortest paths and the probability of taking random actions. Training and evaluation peer pools contain 18 and 9 policies, respectively.

Table S1. Hyperparameters for all algorithms in the Partially-Observable Overcooked (POO) environment.

Parameter	PPO	Generalist	LIAM(X)	LILI	GSCU	PACE(TF)	PCOR
Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	7×10^{-4}	1×10^{-3}	1×10^{-3}
PPO Clip ϵ	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Entropy Coefficient	0.03	0.03	0.03	0.03	0.01	0.03	0.03
γ	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GAE λ	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Batch Size	72,000	72,000	72,000	72,000	2,500	72,000	72,000
# Update Epochs	4	4	4	4	8	4	4
# Mini Batches	18	18	18	18	2	18	18
Gradient Clipping	15.0	15.0	15.0	15.0	0.5	15.0	15.0
Activation Function	ReLU	ReLU	ReLU	ReLU	Tanh	ReLU	ReLU
Actor/Critic Hidden Dims	[128,128]	[128,128]	[128,128]	[128,128]	[64,64]	[128,128]	[128,128]
Loss Weight λ_p	–	–	1.0	1.0	–	1.0	1.0

Table S2. Hyperparameters for all algorithms in the Predator-Prey with Watchtowers (PPW) environment.

Parameter	PPO	Generalist	LIAM(X)	LILI	GSCU	PACE(TF)	PCOR
Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	5×10^{-4}	1×10^{-3}	1×10^{-3}
PPO Clip ϵ	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Entropy Coefficient	0.03	0.03	0.03	0.03	0.01	0.03	0.03
γ	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GAE λ	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Batch Size	64,000	64,000	64,000	64,000	2,500	64,000	64,000
# Update Epochs	2	2	2	2	8	2	2
# Mini Batches	150	150	150	150	2	150	150
Gradient Clipping	15.0	15.0	15.0	15.0	0.5	15.0	15.0
Activation Function	ReLU	ReLU	ReLU	ReLU	Tanh	ReLU	ReLU
Actor/Critic Hidden Dims	[128,128]	[128,128]	[128,128]	[128,128]	[64,64]	[128,128]	[128,128]
Loss Weight λ_p	–	–	1.0	1.0	–	1.0	1.0

In PPW, predator peers are parameterized by their preferred prey, which they consistently pursue under full observability. Prey peers follow predefined movement trajectories, resulting in diverse escape behaviors. Four trajectories are used for training, and another four unseen trajectories are reserved for evaluation. The final peer pools are obtained by sampling different combinations of predator and prey policies, yielding 16 training combinations and 24 testing combinations.

S1.3 Training Details

All methods are trained using PPO⁴ under identical training budgets and environment interaction steps whenever applicable. Supplementary Table S1 summarizes the hyperparameters used in the POO experiments, whereas Supplementary Table S2 reports those used in the PPW experiments. Unless otherwise specified, all PPO-based methods use the same optimization settings within each environment. GSCU⁵ follows its original three-stage training pipeline and therefore adopts different optimization hyperparameters during the conditional RL stage. We adopt the implementations and experimental settings of PACE² for all baselines, including LIAM/LIAMX⁶ and LILI⁷. Unless otherwise stated, all hyperparameters remain identical to those used in PACE. PCOR is implemented by replacing only the policy representation module, while the remaining training and evaluation pipeline is kept unchanged. The total training budget is fixed to 3×10^7 and 1.5×10^7 environment interaction steps for POO and PPW, respectively.

S1.4 Evaluation Metrics for Mechanism Analysis

This subsection specifies the evaluation metrics used in the *Mechanism Analysis* section of the main paper. Unless otherwise stated, analyses are conducted on POO with unseen evaluation peers under a shared fixed initial configuration across sessions. The results are reported as means over five independently trained models (with standard deviation across models where shown). Same-peer and cross-peer comparisons follow the protocols below.

Notation aligned with the *Instantiation and Optimization* section of the main paper. We use o_t for the current observation, $z_t = q_\psi(D_t)$ for the peer representation, $\phi_t = \phi_o(o_t)$ for the observation feature, and $\tilde{\phi}_t = C_\eta(\phi_t, z_t)$ for the peer-conditioned policy representation of PCOR. The multiplicative instantiation used throughout is $\tilde{\phi}_t = (1 + a \tanh(\hat{\alpha}_t)) \odot \phi_t$ with $\hat{\alpha}_t = g_\eta(z_t)$ (Eq. (16) in the *Instantiation and Optimization* section of the main paper). For PACE, h_t denotes the jointly encoded policy representation. When a generic actor-facing policy representation is needed, we write $v \in \{\tilde{\phi}_t, h_t\}$.

S1.4.1 Peer Representation Metrics

Let $z_t = q_\psi(D_t) \in \mathbb{R}^{d_z}$ denote the peer representation (encoder latent). We collect *step-wise* z_t over all timesteps of g1–g5 for each unseen peer, with labels y given by peer identity, and evaluate peer discriminability with two complementary metrics.

Linear probe accuracy. Following linear classifier probes for frozen representations⁸, let $X \in \mathbb{R}^{N \times d_z}$ stack the pooled step-wise representations and $y \in \{1, \dots, P\}^N$ the peer labels. Features are first standardized feature-wise (zero mean, unit variance across samples),

$$X'_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{N} \sum_{i=1}^N X_{ij}, \quad \sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_{ij} - \mu_j)^2}, \quad (\text{S2})$$

and a multinomial logistic regression probe is fit on X' ⁹:

$$p(y | x') = \text{softmax}(Wx' + b), \quad \hat{y} = \arg \max_k p(y=k | x'). \quad (\text{S3})$$

The probe accuracy is the classification accuracy

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i]. \quad (\text{S4})$$

We estimate Acc by stratified K -fold cross-validation⁹ ($K=5$ when feasible, otherwise $K \leq \min_c n_c$), with shuffling within folds, and report the mean accuracy over folds. The encoder producing z_t is frozen; only (W, b) are trained. Higher accuracy indicates that peer identity is more linearly decodable from z_t .

Silhouette score. We compute the silhouette coefficient¹⁰ of the *raw* (unstandardized) z_t under Euclidean distance. For each sample i ,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{silhouette} = \frac{1}{N} \sum_{i=1}^N s(i), \quad (\text{S5})$$

where $a(i)$ is the mean distance to other samples with the same peer label and $b(i)$ is the mean distance to samples in the nearest other peer. Unlike the linear probe, silhouette does *not* apply feature standardization. Values closer to 1 indicate tighter within-peer clusters and clearer separation across peers.

S1.4.2 Policy Representation Similarity

Let v denote the actor-facing policy representation used for action selection: for PCOR this is the peer-conditioned policy representation $\tilde{\phi}_t$; for PACE this is the jointly encoded policy representation h_t . Mechanism comparisons use the *opening* timestep (step 0) of each adaptation episode.

Raw and centered cosine similarity. For two vectors v, v' ,

$$S_{\text{raw}}(v, v') = \frac{v^\top v'}{\|v\|_2 \|v'\|_2}. \quad (\text{S6})$$

Centered similarity removes the shared component over the analysis set. Let $\{v_k\}$ be all peer $\times$ session opening vectors under consideration and $\mu = \text{mean}_k v_k$. Define

$$\tilde{v}_k = \frac{v_k - \mu}{\|v_k - \mu\|_2}, \quad S_{\text{cen}}(\tilde{v}_i, \tilde{v}_j) = \tilde{v}_i^\top \tilde{v}_j. \quad (\text{S7})$$

Centering follows the spirit of RSA / CKA mean-centering^{11–13} and is closely related to Pearson correlation after mean removal¹⁴.

Same / Cross / Gap (protocol in the *Mechanism Analysis* section of the main paper). For pairwise comparisons we collect opening-step vectors under a matched peer–replay design: the number of independent replays equals the number

of evaluation peers, with the interaction history cleared between replays. Let $v_{p,r}$ denote the opening vector for peer p on replay r . Same-peer similarity averages $S(v_{p,r}, v_{p,r'})$ over all pairs $r \neq r'$ within each peer (and then over peers). Cross-peer similarity averages $S(v_{p,r}, v_{p',r})$ over $p \neq p'$ under the same replay index r (and then over r), restricted to peer pairs with different preferred-ingredient keys. We report

$$\text{Gap} = S_{\text{same}} - S_{\text{cross}}. \quad (\text{S8})$$

The *Peer and Policy Representations* section of the main paper focuses on centered similarities at g5; Supplementary Section S2.2 additionally reports raw trajectories over g1–g5 where informative.

S1.4.3 Observation-to-Representation Jacobian Sensitivity

We measure local observation-to-representation sensitivity by differentiating the actor-facing policy representation v (as above: $\tilde{\phi}_t$ for PCOR; h_t for PACE) w.r.t. the current flat observation o_t , with interaction-history encodings held fixed (no backpropagation into the current observation through the peer encoder). The Jacobian

$$J = \frac{\partial v}{\partial o} \in \mathbb{R}^{D \times N} \quad (\text{S9})$$

is obtained by automatic differentiation¹⁵, where D is the dimension of the policy representation and N is the observation dimension. Absolute entries yield two descriptors used in the analyses:

$$u_{\text{full}} = \text{vec}(|J|) \in \mathbb{R}^{DN}, \quad I_j = \frac{1}{D} \sum_{i=1}^D |J_{ij}| \Rightarrow u_{\text{obs}} = I \in \mathbb{R}^N. \quad (\text{S10})$$

Here u_{full} retains the full entry-wise sensitivity profile, whereas u_{obs} aggregates over units of the policy representation to an observation-coordinate importance vector. Jacobians are evaluated at episode openings. Same-peer / cross-peer cosine similarities and Gap of either descriptor follow the matched peer–replay protocol in Supplementary Section S1.4.2, under both raw and centered variants. The *Observation-to-Representation Sensitivity Analysis* section of the main paper reports centered similarities of u_{full} ; Supplementary Section S2.3 reports complementary Gap / raw trajectories for u_{full} and a centered view of u_{obs} . Higher same-peer and lower cross-peer similarity indicate more consistent peer-conditioned observation-to-representation mappings.

S1.4.4 Policy Divergence and Per-Episode Success Rate

Symmetric policy KL divergence. Let $p, q \in \Delta^{|A|-1}$ be discrete action distributions (softmax policy probabilities) at a matched timestep. After clipping probabilities by $\epsilon=10^{-8}$ and renormalizing, we use the symmetric Kullback–Leibler (Jeffreys) divergence^{16,17}

$$D_{\text{sym}}(p, q) = \frac{1}{2} \text{KL}(p||q) + \frac{1}{2} \text{KL}(q||p). \quad (\text{S11})$$

We consider two temporal aggregations: *opening* uses only the first action distribution after each episode reset; *episode-mean* averages D_{sym} over all valid timesteps within the episode before same/cross aggregation. Same-peer pairs compare a reference session against other sessions of the same peer at matched (game, timestep); cross-peer pairs compare the reference session against peers with different preferred-ingredient keys at the same (game, timestep). The *Policy-Level Behavioral Differences* section of the main paper reports opening-step same/cross D_{sym} over g1–g5; Supplementary Section S2.4 additionally reports episode-mean curves.

Per-episode success rate (Eval20, g1–g5). For each unseen evaluation peer, we run 20 independent adaptation sessions under a shared fixed initial configuration. In each session, the ego agent interacts with the fixed peer for $N_{\text{eps}} = 5$ consecutive episodes (g1–g5), clearing the interaction history only between sessions. An episode is counted as successful if the environment termination indicates recipe completion (valid dish delivery). For adaptation episode $n \in \{1, \dots, 5\}$,

$$\text{Success}(n) = \frac{\#\{\text{successful episodes at gn}\}}{\#\{\text{peer-session pairs}\}} = \frac{\#\{\text{successful episodes at gn}\}}{\#\{\text{peers}\} \times 20}. \quad (\text{S12})$$

Thus g1–g5 measures completion rate as interaction history of length $n-1$ accumulates within a session. The overall Eval20 score in the main tables averages performance over g1–g5 and the 20 sessions (further averaged over peers and training seeds).

S2 Additional Results

S2.1 Impact of Coefficient a

We further investigate the influence of the scaling coefficient a in Eq. (16) of the *Instantiation and Optimization* section of the main paper,

$$\tilde{\phi}_t = (1 + a \tanh(\hat{\alpha}_t)) \odot \phi_t, \quad (\text{S13})$$

where $\hat{\alpha}_t = g_\eta(z_t)$ and a controls the strength of peer-dependent feature scaling. We evaluate five values of a , namely 0.1, 0.5, 0.9, 1.3, and 1.7, while keeping all other hyperparameters unchanged.

As shown in Supplementary Fig. S1, the sensitivity to a differs across environments. In PPW, PCOR is relatively insensitive to the choice of a , with comparable performance achieved across a broad range of values. In contrast, POO exhibits a much stronger dependence on the scaling coefficient. Both an insufficient scaling strength ($a = 0.1$) and an excessively large scaling strength ($a = 1.7$) lead to a substantial degradation in evaluation performance, whereas intermediate values achieve substantially better adaptation. Among all tested settings, $a = 0.9$ provides the best overall trade-off between training stability and evaluation performance, and is therefore adopted in all experiments.

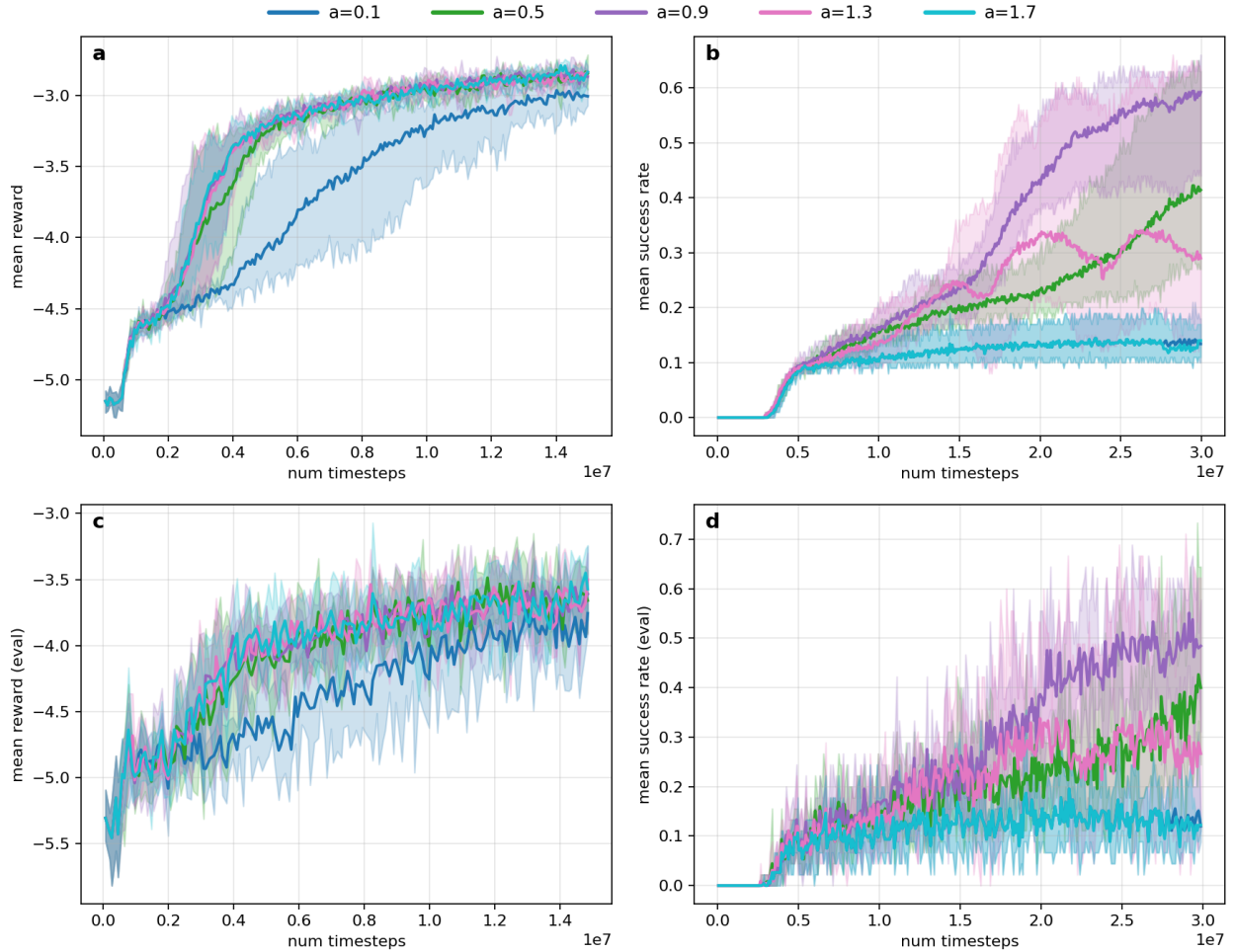


Figure S1. Sensitivity analysis of scaling coefficient a . Top: learning curves on training peers. Bottom: learning curves on unseen evaluation peers. Left: mean reward on PPW. Right: mean success rate on POO. The shaded regions denote standard deviation across five random seeds.

S2.2 Peer and Policy Representations

The *Peer and Policy Representations* section of the main paper shows that both PCOR and PACE construct increasingly peer-discriminative peer representations z_t and policy representations ($\tilde{\phi}_t$ for PCOR, h_t for PACE) over the adaptation horizon,

yet these representation-level trends alone do not fully account for the eventual performance gap. We provide complementary pooled statistics, a centered-versus-raw comparison of policy representations, and a qualitative visualization. Unless otherwise stated, the results are means over five training seeds. Same-peer and cross-peer comparisons follow the protocol in the *Mechanism Analysis* section of the main paper.

Supplementary Fig. S2a reports linear probe accuracy (stratified 5-fold logistic regression predicting peer identity) and Euclidean silhouette score of step-wise z_t pooled over all timesteps of g1–g5 for 9 unseen evaluation peers. Averaged over the horizon, PACE exceeds PCOR (probe ≈ 0.85 vs. 0.79; silhouette ≈ 0.24 vs. 0.19), consistent with the observation in the *Peer and Policy Representations* section of the main paper that PACE yields more discriminative z_t during early adaptation.

Supplementary Fig. S2b complements the centered analysis of policy representations in the *Peer and Policy Representations* section of the main paper by reporting same-peer similarity, cross-peer similarity, and Gap = $S_{\text{same}} - S_{\text{cross}}$ for opening-step $\tilde{\phi}_t/h_t$ at g5, computed over 9 peers and 9 independent replays. After centering, both methods attain high same-peer similarity (≈ 0.84 –0.86), near-zero or negative cross-peer similarity, and large Gaps (≈ 0.92 –0.94), in line with the finding there that peer-specific information is preserved after policy representation construction; PACE’s Gap is slightly larger.

Supplementary Fig. S2c reports the corresponding *raw* (uncentered) similarities over g1–g5 under the same protocol. Same-peer similarity remains high for both methods, whereas cross-peer similarity decreases more for PCOR (from ≈ 1.0 at g1 to ≈ 0.86 at g5) than for PACE (remaining near ≈ 0.98). Thus the stronger raw cross-peer separation of PCOR is largely masked under centering, which emphasizes peer-specific deviations relative to a shared component.

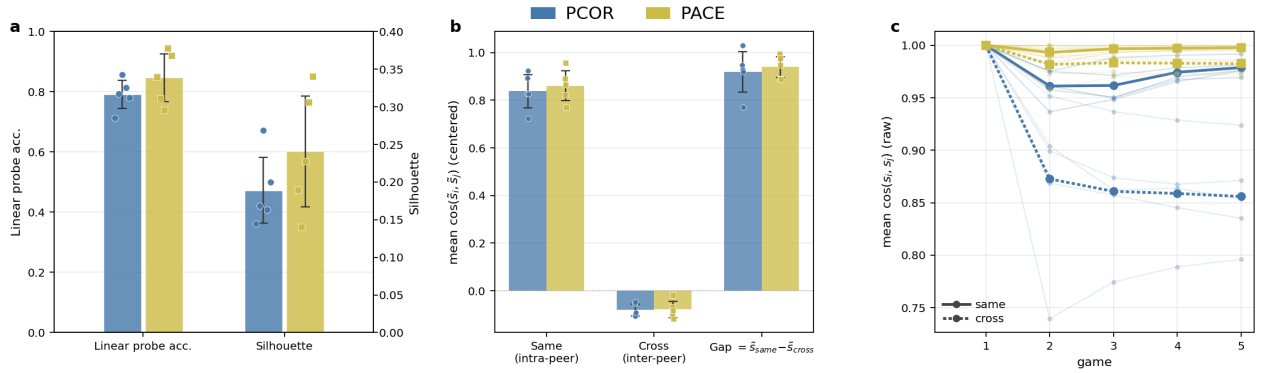


Figure S2. Complementary peer and policy representation analysis of PCOR and PACE on POO (pooled over five training seeds). (a) Peer representation quality of z_t measured by linear probe accuracy for peer identity ($\uparrow$) and silhouette score ($\uparrow$). (b) Discriminability of opening-step policy representations ($\tilde{\phi}_t$ for PCOR, h_t for PACE) at g5 measured by centered pairwise cosine similarity (same-peer $\uparrow$; cross-peer $\downarrow$; Gap = $S_{\text{same}} - S_{\text{cross}}$). (c) Temporal evolution of raw same-peer/cross-peer cosine similarity of opening-step $\tilde{\phi}_t/h_t$ over g1–g5. In (c), solid lines denote same-peer comparisons and dashed lines denote cross-peer comparisons.

Supplementary Fig. S3 provides a qualitative t-SNE visualization of step-wise peer representations z_t for a representative training run, using the same evaluation setting as Supplementary Fig. S2a: all timesteps over g1–g5 for 9 unseen peers. Supplementary Fig. S3a,b correspond to PCOR and PACE (fitted independently); colors index peer identity. The linear probe accuracy and silhouette score in the *Peer and Policy Representations* section of the main paper and in Supplementary Fig. S2a are computed on this overall (g1–g5) collection of z_t . Ellipses highlight the g5 subset for each peer, illustrating where representations concentrate by the end of the adaptation horizon within each peer-colored cloud. The figure is illustrative; quantitative conclusions rely on Supplementary Fig. S2a.

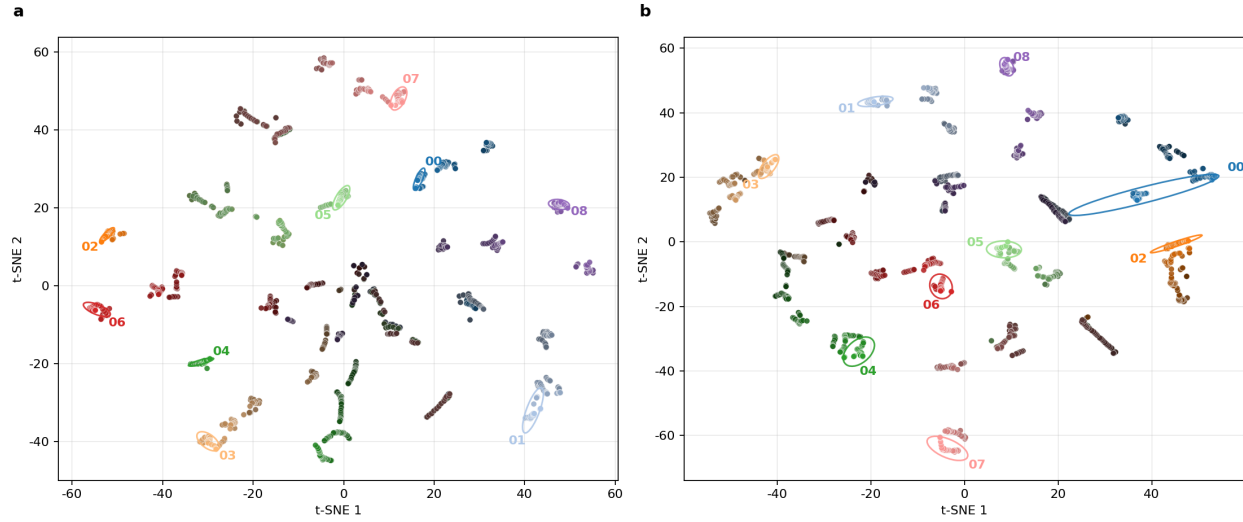


Figure S3. t-SNE visualization of step-wise peer representations z_t over all timesteps of g1–g5 for a representative training run. (a) PCOR. (b) PACE. Colors correspond to nine unseen evaluation peers (00–08). Ellipses mark the g5 subset for each peer. Each embedding is fitted independently.

S2.3 Observation-to-Representation Sensitivity Analysis

The *Observation-to-Representation Sensitivity Analysis* section of the main paper reports that PCOR maintains higher same-peer Jacobian similarity than PACE while keeping comparable cross-peer similarity, yielding a larger same–cross separation in the observation-to-representation mapping. We supplement that analysis with an aggregated Gap summary and raw trajectories for the full descriptor $u_{\text{full}} = \text{vec}(|J|)$, together with a centered observation-level view $u_{\text{obs}} = I$. Opening-step Jacobians are collected under the matched peer–replay protocol of Supplementary Section S1.4.2 over g1–g5, with history encodings held fixed; the results are pooled over five training seeds.

Supplementary Fig. S4a summarizes centered u_{full} same-peer/cross-peer similarities and Gap, averaged over g1–g5. PCOR attains a larger Gap (≈ 0.68 vs. ≈ 0.56), with higher same-peer similarity (≈ 0.61 vs. ≈ 0.52) and more negative cross-peer similarity, quantifying the stronger peer-specific sensitivity structure already suggested by the curves in the *Observation-to-Representation Sensitivity Analysis* section of the main paper.

Supplementary Fig. S4b shows raw u_{full} trajectories. At g1, same-peer and cross-peer similarities coincide near 1 for both methods; thereafter PCOR cross-peer similarity decreases to ≈ 0.80 , while PACE remains around 0.88–0.90, with same-peer similarity staying high for both. Supplementary Fig. S4c instead uses centered u_{obs} . At g1, same-peer and cross-peer similarities are near zero; subsequently same-peer similarity rises (PCOR to ≈ 0.84 , PACE to ≈ 0.74) while cross-peer similarity remains near zero. Thus PCOR forms a clearer peer-specific observation-sensitivity structure under both the full and observation-aggregated descriptors.

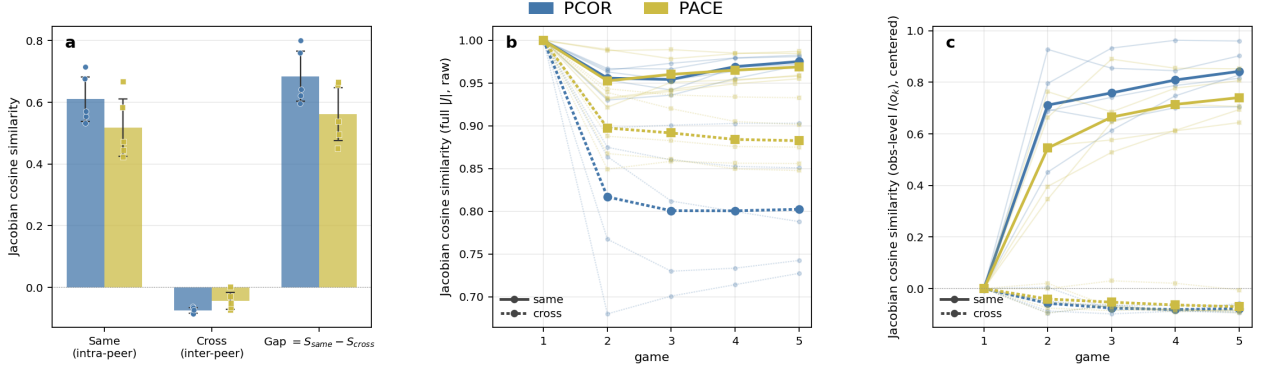


Figure S4. Complementary observation-to-representation sensitivity analysis of PCOR and PACE on POO (pooled over five training seeds). The Jacobian is $J = \partial v / \partial o$ for the actor-facing policy representation v ($\hat{q}_t$ for PCOR; h_t for PACE) with history encodings held fixed. (a) Aggregate same-peer/cross-peer Gap of centered $u_{\text{full}} = \text{vec}(|J|)$, averaged over g1–g5 (same-peer $\uparrow$; cross-peer $\downarrow$). (b) Evolution of raw u_{full} cosine similarity over g1–g5 (same-peer $\uparrow$; cross-peer $\downarrow$). (c) Evolution of centered observation-level $u_{\text{obs}} = I$ same-peer/cross-peer similarity over g1–g5. In (b)–(c), solid lines denote same-peer comparisons and dashed lines denote cross-peer comparisons.

S2.4 Policy-Level Behavioral Differences

The *Policy-Level Behavioral Differences* section of the main paper focuses on opening-step policy KL divergence under a shared initial observation, and the *Connecting Mechanism Analysis to Adaptation Performance* section of the main paper reports Eval20 success rates over g1–g5. We complement these results with episode-mean policy divergence and a longer adaptation session.

Supplementary Fig. S5a reports symmetric KL divergence between step-wise action distributions, averaged within each adaptation episode, for 9 peers and 9 independent replays over g1–g5 (five training seeds). Same-peer divergence remains roughly stable for both methods (≈ 2.8 – 3.2). Cross-peer divergence for PCOR rises from ≈ 1.62 at g1 to ≈ 3.4 – 3.5 at g4–g5, exceeding the same-peer level, whereas PACE cross-peer divergence rises earlier but later approaches same-peer. Relative to the opening-step analysis in the *Policy-Level Behavioral Differences* section of the main paper—where PCOR exhibits smaller overall KL magnitudes while preserving same–cross separation—the episode-mean view indicates that PCOR’s peer-dependent behavioral differentiation is sustained over the full trajectory as interaction history accumulates.

Supplementary Fig. S5b extends the Eval20 protocol to 10 consecutive adaptation episodes per session (9 unseen peers $\times$ 20 sessions; five training seeds). PACE is slightly higher early (g1 ≈ 0.13 vs. 0.05), the two methods are close around g3, and thereafter PCOR continues to rise to ≈ 0.51 by g10 while PACE plateaus near 0.34–0.42. This longer-horizon trajectory reinforces the observation in the *Connecting Mechanism Analysis to Adaptation Performance* section of the main paper that PCOR’s advantage emerges as richer interaction history becomes available.

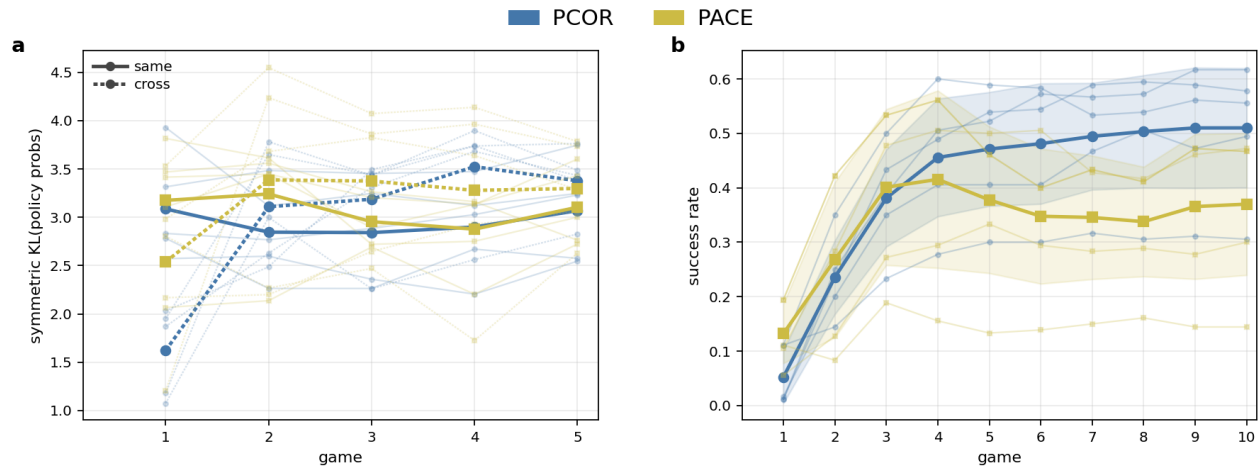


Figure S5. Complementary policy-level behavioral analyses of PCOR and PACE on POO (pooled over five training seeds). (a) Symmetric KL divergence between episode-mean action distributions over g1–g5 (same-peer vs. cross-peer). (b) Eval20 success rate on unseen peers over a 10-episode session (↑). In (a), solid lines denote same-peer comparisons and dashed lines denote cross-peer comparisons.

References

- Charakorn, R., Manoonpong, P. & Dilokthanakul, N. Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations* (2023).
- 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).
- Lowe, R. *et al.* Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- 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).
- 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).
- 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).
- Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop Track* (2017).
- Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
- Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
- Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
- Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 3519–3529 (PMLR, 2019).
- Luo, C., Zhan, J., Wang, L. & Yang, Q. Cosine normalization: Using cosine similarity instead of dot product in neural networks. *Neurocomputing* **319**, 1–11 (2018).

15. Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res.* **18**, 1–43 (2018).
16. Kullback, S. & Leibler, R. A. On information and sufficiency. *The Annals Math. Stat.* **22**, 79–86 (1951).
17. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Royal Soc. London. Ser. A, Math. Phys. Sci.* **186**, 453–461 (1946).