LEARNING COORDINATED PREFERENCE FOR MULTI-OBJECTIVE MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative multi-objective multi-agent reinforcement learning (MOMARL) models team decision making under multiple, potentially conflicting objectives. In this setting, conflicts arise not only across objectives but also across agents with different observations, roles, and contributions. We propose Preference Coordinated Multi-agent Policy Optimization (PCMA), which learns coordinated agent-specific preferences to enable complementary trade-offs among agents. Theoretically, we formulate cooperative MOMARL as a team-optimal equilibrium problem, and show that, under suitable conditions, preference diversity yields a first-order improvement in the team objective. Experiments on multiple cooperative MOMA environments and a practical traffic-control scenario show that PCMA improves both performance and trade-off coordination.

1 INTRODUCTION

Multi-objective Multi-agent reinforcement learning (MOMARL) provides a natural framework for modeling complex decision making systems where multiple agents must coordinate under competing objectives. Existing applications on domains such as traffic signal control, dynamic traffic management, and natural resource allocation show that such systems often require agents to balance competing objectives such as efficiency, safety, energy consumption, and fairness (Duan et al., 2010; Chung et al., 2019; Bone & Dragićević, 2009). Therefore, modeling such problems with a single scalar reward is often inadequate (Vamplew et al., 2022). At the same time, conflict structure is extremely intricate in MOMA problem: conflicts may arise not only between objectives within each agent, but also between agents under the same objective and across agents with different objective priorities. Therefore, MOMARL is both practically important and theoretically challenging, calling for learning algorithms that model and coordinate such intertwined objectives in Multi-agent systems.

Despite growing interest, research on MOMARL remains at an early stage. Several notable efforts have begun to structure the field, including the utility based analysis of Multi-objective Multi-agent decision making (Rădulescu et al., 2020) and the introduction of MOMALand (Felten et al., 2024), as the first benchmark environments in this field. A straight-forward approach adopted by existing works is to optimize scalarized objective using the same preference vector across all agents. However, enforcing identical objective emphasis for every agent can limit the effectiveness of coordination. When all agents prioritize objectives in the same way, they may compete along the same dimensions of the reward. For example, consider two autonomous vehicles passing through an unsignalized intersection. If both vehicles strongly prefer efficiency, they may aggressively enter the intersection and cause a collision. If both vehicles strongly prefer safety, they may both wait and waste time. A better joint behavior requires coordinated preferences: one vehicle may emphasize yielding and safety, while the other emphasizes passing efficiency.

This motivates our central idea: rather than forcing all agents to follow the same trade-off, we allow agents to adapt their preferences through a team-level coordination mechanism. From a multi-objective perspective (Roijers et al., 2013; Hayes et al., 2022), each agent's policy can be viewed as inducing a point in a shared objective space. By conditioning policies on different preferences, agents can approach different regions of the Pareto front. Coordinating these preferences enables the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

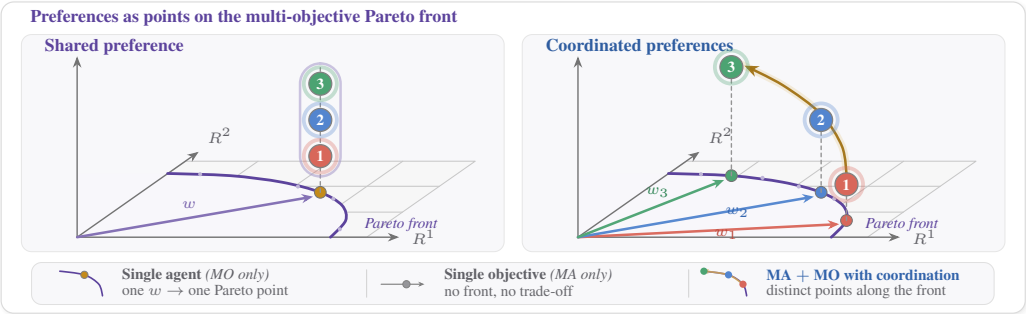


Figure 1: Illustration of preference coordination. Agents' policies are projected on to the multi-objective return space (R^1, R^2). A preference w from the origin selects one point on the front. **Left:** all agents use the same preference vector w, which could cause conflicts. **Right:** distinct w_1, w_2, w_3 project to three different points along the front and the team jointly covers the front. A single agent can only occupy one point, a single objective produces no front at all, and only multi-agent + multi-objective with coordination spreads agents along the front.

multi-agent system to assign diverse but complementary roles, thereby reducing behavioral conflicts and improving team-level performance, as illustrated in Figure. 1.

To realize this idea, in this paper, we propose Preference Coordinated Multi-agent Policy Optimization (PCMA) under the centralized training with decentralized execution (CTDE) framework (Lowe et al., 2017). We model preference as a latent coordination variable and train a stochastic planner for each agent to adapt its preference based on team improvement and its local observation, and then each agent executes a policy conditioned on its sampled preference. To encourage meaningful specialization, we further regularize the learned preference distributions so that agents avoid collapsing to the same preference direction. Our contributions are as follows:

- We formulate cooperative MOMARL as a multi-objective partially observable stochastic game (mo-POSG) with partially aligned interests, and adopt team-optimal equilibrium as the solution concept, seeking a preference profile and an induced equilibrium that maximize the shared team objective.
- Theoretically, we characterize first-order team improvement through a decomposition that isolates the contribution of preference diversity, and establish a local equilibrium-tracking result for preference-conditioned games under suitable conditions.
- We propose PCMA, a practical policy-based method with stochastic preference planners and preference-conditioned actors, and provide empirical evaluations on selected cooperative tasks from MPE, SMAC, and MOMAland, as well as a traffic-control validation scenario.

2 Related Work

Current work on cooperative Multi-objective Multi-agent reinforcement learning (MOMARL) remains at an early stage of exploration. As a pioneer, work on Multi-objective games has put forward solution concepts such as Pareto Nash equilibrium, which extends Nash equilibrium to vector valued utilities by requiring Pareto optimality under unilateral deviations (Borm et al., 1988; Voorneveld et al., 1999; Lozovanu et al., 2005). However, Pareto Nash equilibrium can be ill-posed as an optimization criterion for cooperative MOMARL. Since a policy profile is ruled out only if some unilateral deviation improves at least one objective without decreasing any other objective, many policies that are suboptimal under meaningful trade offs may still satisfy the Pareto Nash condition. We provide an illustrative example in Appendix C.1, where all joint policies in the specified fixed-variance Gaussian class satisfy the Pareto Nash condition, while only a small subset corresponds to meaningful team level trade offs.

More recently, MOMALand (Felten et al., 2024) considers linear scalarization based approaches to convert the vector valued team reward into a scalar reward, and then train policies by standard

MARL algorithms such as MAPPO (Yu et al., 2022). However, they need re-train a separate policy for new preference vector, making them impractical for desired trade off during execution. Another work, MoMix (Hu et al., 2023), extends the value decomposition principle of QMIX (Rashid et al., 2018) to vector valued rewards. While this provides an important step toward Multi-objective credit assignment, it relies on strong individual global max (IGM) assumptions and can't be applied to continous action space.

3 Preliminaries and Problem Formulation

We introduce the problem setting in Subsection 3.1 and formulate the preference coordination objective in Subsection 3.2.

3.1 Multi-objective Partially Observable Stochastic Game

We consider a multi-objective partially observable stochastic game (mo-POSG) with partially aligned interests, following the formulation of Felten et al. (2024, Definition 1). Our setting involves N agents and K objectives and is specified by

$$\mathcal{G} = \left\langle N, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \Omega, O, \{\mathbf{R}_i\}_{i=1}^N, \gamma \right\rangle.$$

Here, $\mathcal{S}$ is the state space, $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$ is the joint action space, $\Omega = \prod_{i=1}^N \Omega_i$ is the joint observation space, and γ is the discount factor. The environment evolves according to $s_{t+1} \sim P(\cdot \mid s_t, \mathbf{a}_t)$, where $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$. Each agent receives only its local observation o_i. We use a state-dependent observation kernel $O(\mathbf{o} \mid s)$, a special case of the general mo-POSG observation model. Each agent has a vector reward function $\mathbf{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^K$.

All agents share the same team objective, such as winning a game, represented by a common scalar reward r_{team}. Agent-specific guiding vector rewards $\mathbf{r}_i$ and preferences may introduce conflicting incentives, resulting in partially aligned interests. We instantiate the per-agent vector rewards as

$$\mathbf{R}_i(s, \mathbf{a}, s') = r_{\text{team}}(s, \mathbf{a})\mathbf{1}_K + \mathbf{r}_i(s, \mathbf{a}),$$

where $\mathbf{1}_K$ is the all-ones vector, $r_{\text{team}}(s, \mathbf{a}) \in \mathbb{R}$ is a sparse team reward, and $\mathbf{r}_i(s, \mathbf{a}) \in \mathbb{R}^K$ provides auxiliary feedback on objectives such as efficiency, cost, or risk.

3.2 MOMARL as Team Optimal Equilibrium.

Let θ_i parameterize agent i's policy, and let π_θ denote the joint policy with $\theta = (\theta_1, \ldots, \theta_N)$. We define the shared team objective and the agent-specific vector objective as

$$J_{\text{team}}(\theta) := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_{\text{team}}(s_t, \mathbf{a}_t)\right], \qquad \mathbf{J}_i(\theta) := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t \mathbf{r}_i(s_t, \mathbf{a}_t)\right].$$

Given a preference profile $\mathbf{p} = (p_1, \ldots, p_N)$, with $p_i \in \Delta^{K-1} := \{p \in \mathbb{R}^K_{\geq 0} : \mathbf{1}_K^\top p = 1\}$, agent i evaluates its vector return using linear scalarization:

$$\begin{aligned} U_i(\theta; p_i) &:= p_i^\top \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t \mathbf{R}_i(s_t, \mathbf{a}_t, s_{t+1})\right] \\ &= J_{\text{team}}(\theta) + p_i^\top \mathbf{J}_i(\theta). \end{aligned}$$

For a fixed $\mathbf{p}$, this scalarization induces a preference-conditioned stochastic game $\mathcal{G}(\mathbf{p})$, whose agents generally have different payoffs. We use Nash equilibrium (Nash, 1951; Fink, 1964) as the solution concept: at an equilibrium $\theta^*(\mathbf{p})$, no agent can improve its own payoff U_i by unilaterally changing its policy within the considered policy class while the other agents' policies remain fixed.

However, an equilibrium under a fixed preference profile does not necessarily maximize the final team objective. The key problem is therefore to coordinate the preference profile so that the induced equilibrium achieves high team performance. Motivated by the social-welfare view of evaluating collective outcomes (Harsanyi, 1955) and the optimal-equilibrium perspective in team Markov games (Wang & Sandholm, 2002), we formulate the goal as finding a preference profile $\mathbf{p}$ and an induced equilibrium θ that maximize the team objective J_{team}.

4 THEORETICAL ANALYSIS

We now provide a theoretical explanation for coordinated preference learning. First, we characterize how a policy update affects the team objective and show that preference diversity can contribute a positive first-order improvement. Second, under suitable regularity and second-order conditions, we show that strict local Nash equilibria under nearby preference profiles are connected through a locally smooth path, which implies that preference optimization can proceed gradually from an initial profile instead of treating each preference-conditioned game as an isolated problem.

4.1 FIRST-ORDER TEAM IMPROVEMENT AND PREFERENCE DIVERSITY

In this subsection, we ask how the choice of preference profile $(p_1, \cdots, p_N)$ affects the resulting team improvement. Starting from the current policy parameters $\theta = (\theta_1, \ldots, \theta_N)$, each agent performs one preference-conditioned gradient update

$$\theta_{i,\text{new}} = \theta_i + \eta \nabla_{\theta_i} U_i(\theta; p_i), \qquad i = 1, \ldots, N,$$

We can define the following team-improvement matrix $B \in \mathbb{R}^{N \times K}$ by

$$B_{i,k} := \left(\nabla_{\theta_i} J_{\text{team}}(\theta)\right)^\top \nabla_{\theta_i} J_{i,k}(\theta), \qquad i = 1, \ldots, N,\ k = 1, \ldots, K.$$

Here, $B_{i,k}$ measures the first-order contribution of agent i's k-th individual objective to the team objective. Let $b_i = (B_{i,1}, \ldots, B_{i,K})^\top \in \mathbb{R}^K$ denote the i-th row of the team-improvement matrix B, represented as a column vector. Let $\bar{p} = N^{-1} \sum_{i=1}^N p_i$ and $\bar{b} = N^{-1} \sum_{i=1}^N b_i$ be the averages across agents, and let $\tilde{p}_i = p_i - \bar{p}$ and $\tilde{b}_i = b_i - \bar{b}$ be their centered counterparts. Inspired by gradient-alignment principles in multi-objective optimization (Désidéri, 2012; Liu et al., 2021), we assume the following sufficient condition, which requires agents' deviations from the mean preference to align with their relative contributions to team improvement.

Assumption 4.1 (Preference-improvement alignment)**.** There exists $\kappa > 0$ such that, for every agent i with $\tilde{p}_i \neq 0$, the projection of $\tilde{b}_i$ onto $\tilde{p}_i$ is lower bounded by κ, i.e.,

$$\frac{\tilde{p}_i^\top \tilde{b}_i}{\|\tilde{p}_i\|_2^2} \geq \kappa, \qquad i = 1, \ldots, N.$$

By first-order Taylor expansion, we can derive the following decomposition

Theorem 4.2 (Team Improvement Decomposition)**.** *Under Assumption 4.1, the team improvement as $\eta \to 0^+$ satisfies*

$$J_{\text{team}}(\theta_{\text{new}}) - J_{\text{team}}(\theta) \geq \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta N \left(\bar{p}^\top \bar{b} + \kappa \mathcal{D}_p\right) + o(\eta)$$

where

$$\mathcal{D}_p := \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N \|p_i - p_j\|_2^2.$$

is the pairwise preference distance across agents.

This theorem decomposes team improvement into three first-order terms and a higher-order remainder. The first term captures direct team-gradient improvement. It is non-negative but can be noisy under sparse team rewards. The second term, $\eta N \bar{p}^\top \bar{b}$, captures average preference alignment. The third term, $\eta N \kappa \mathcal{D}_p$, captures the additional improvement from aligned preference diversity.

4.2 EQUILIBRIUM TRACKING

For each preference profile $\mathbf{p}$, there is a corresponding preference-conditioned game $\mathcal{G}(\mathbf{p})$. A natural question is how the solutions of these games are related as $\mathbf{p}$ changes. In this subsection, we first show that, under regularity and second-order conditions, a strict local Nash equilibrium changes continuously with respect to $\mathbf{p}$. Then, when the preference profile varies slowly, gradient updates can track the corresponding moving equilibrium path.

By applying the implicit function theorem (Jittorntrum, 1978) and the second-order sufficient conditions for a strict local Nash equilibrium (Ratliff et al., 2014), we obtain the following local continuity result. The full proof is provided in Appendix A.3.

Lemma 4.3 (Continuity of preference-conditioned local Nash equilibria)**.** *Suppose that $\theta^* \in \operatorname{int}\Theta$ is a strict local Nash equilibrium of $\mathcal{G}(\mathbf{p})$ satisfying*

$$\nabla_{\theta_i} U_i(\theta^*; p_i) = 0, \qquad \nabla^2_{\theta_i\theta_i} U_i(\theta^*; p_i) \prec 0, \qquad i = 1, \dots, N.$$

Assume that the joint gradient $(\nabla_{\theta_1} U_1, \dots, \nabla_{\theta_N} U_N)$ is continuously differentiable near $(\theta^*, \mathbf{p})$, and its Jacobian with respect to θ is nonsingular at this point. Then there exist neighborhoods $\mathcal{U}$ of $\mathbf{p}$ and $\mathcal{V} \subset \Theta$ of θ^*, and a unique continuously differentiable mapping $\theta(\cdot) : \mathcal{U} \to \mathcal{V}$ with $\theta(\mathbf{p}) = \theta^*$, such that, for every $\tilde{\mathbf{p}} \in \mathcal{U}$, $\theta(\tilde{\mathbf{p}})$ is the unique stationary point in $\mathcal{V}$ and a strict local Nash equilibrium of $\mathcal{G}(\tilde{\mathbf{p}})$.

The previous lemma guarantees a locally smooth branch of strict local Nash equilibria. For learning dynamic analysis, we further impose the following two regularity assumptions.

Assumption 4.4 (Attraction property near equilibrium)**.** Let $\theta(\mathbf{p})$ denote a strict local Nash equilibrium under the preference profile $\mathbf{p}$. Assume that there exist $\rho \in (0, 1)$ and a neighborhood of $\theta(\mathbf{p})$ such that

$$\|\theta_{\text{new}} - \theta(\mathbf{p})\| \le \rho \|\theta_{\text{old}} - \theta(\mathbf{p})\|,$$

for all θ_{old} in this neighborhood.

Assumption 4.5 (Local Lipschitz continuity)**.** Let $\theta(\cdot) : \mathcal{U} \to \Theta$ be the strict local Nash equilibrium mapping induced by Lemma 4.3. Assume that this mapping is Lipschitz continuous in $\mathcal{U}$: there exists $C > 0$ such that for any two preference profiles $\mathbf{p}, \mathbf{p}' \in \mathcal{U}$,

$$\|\theta(\mathbf{p}') - \theta(\mathbf{p})\| \le C \|\mathbf{p}' - \mathbf{p}\|.$$

The following theorem shows that, when the preference profile changes slowly, the policy update remains close to the equilibrium associated with the new preference profile. See proof in Appendix A.4.

Theorem 4.6 (Equilibrium tracking)**.** *Consider the iterative policy update*

$$\theta_i^{t+1} = \theta_i^t + \eta \nabla_{\theta_i} U_i(\theta^t; \mathbf{p}^t), \qquad i = 1, \dots, N.$$

Under Assumption 4.4 and Assumption 4.5, the tracking error $e_t := \|\theta^t - \theta(\mathbf{p}^t)\|$ satisfies

$$e_{t+1} \le \rho e_t + C \|\mathbf{p}^{t+1} - \mathbf{p}^t\|.$$

In particular, if $\|\mathbf{p}^{t+1} - \mathbf{p}^t\| \le \delta$, then

$$\limsup_{t \to \infty} e_t \le \frac{C}{1 - \rho} \delta.$$

5 Multi-agent Policy Optimization with Preference Coordinated Learning

Motivated by the preceding analysis, we now implement the preference-coordination idea in a practical PPO-based algorithm, called Preference Coordinated Multi-agent Policy Optimization (PCMA). The key design is to treat preference p_i as a latent coordination variable: it determines the utility optimized by agent i, while the planner adjusts the preference profile to encourage team-beneficial updates. PCMA follows the CTDE (Lowe et al., 2017) paradigm, using centralized critics for training and decentralized policies for execution.

5.1 Preference-conditioned Policy Optimization

Following preference-conditioned policy learning in MORL (Abels et al., 2019; Yang et al., 2019; Basaklar et al., 2023), each agent conditions its actor on its local observation o_i and preference p_i, selecting $a_i \sim \pi_\theta(\cdot \mid o_i, p_i)$.

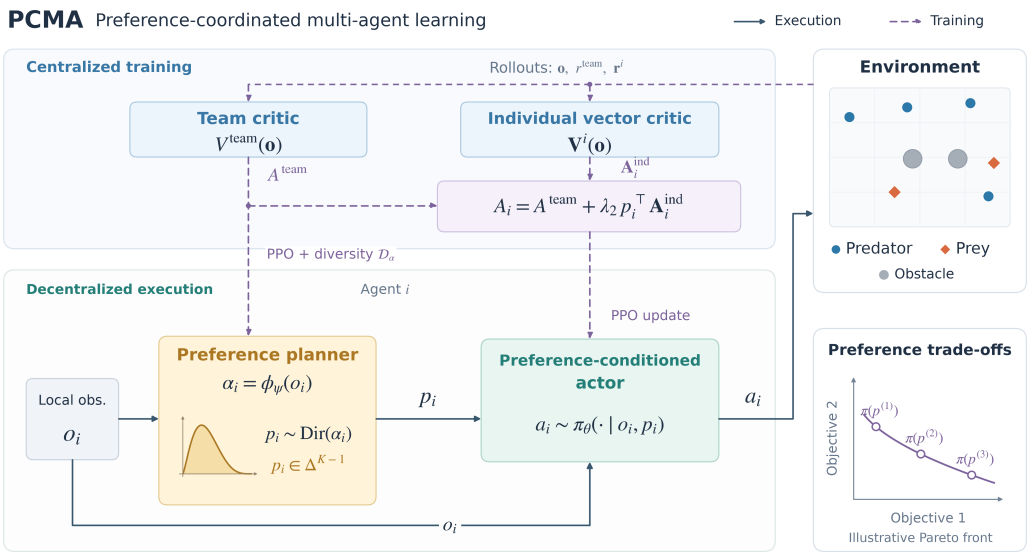


Figure 2: Overview of PCMA. Each agent uses a stochastic preference planner and a preference-conditioned actor. The planner samples preferences from a Dirichlet distribution, and the actor selects actions conditioned on the sampled preference. During training, the team critic provides coordination feedback, while individual vector critics provide dense preference-aligned learning signals.

The team critic $V^{\text{team}}(\mathbf{o})$, with concatenated observations $\mathbf{o} = (o_1, \ldots, o_N)$, estimates the expected discounted team return. Each agent i has its own vector critic $\mathbf{V}^i(\mathbf{o}) \in \mathbb{R}^K$, which estimates the returns for its objectives. We use these value estimates and rewards to compute the scalar team advantage A^{team} and individual advantage vector $\mathbf{A}_i^{\text{ind}}$ via generalized advantage estimation (GAE) (Schulman et al., 2016), applied componentwise to individual objectives. Agent i combines the team and preference-weighted individual advantages for its PPO update:

$$\mathcal{L}_{\text{actor}}(\theta) = \mathcal{L}_{\text{PPO}}\left(\pi_\theta(\cdot \mid o_i, p_i), A_i\right), \qquad A_i = A^{\text{team}} + \lambda_2 p_i^\top \mathbf{A}_i^{\text{ind}}. \tag{1}$$

Here, λ_2 weights individual guidance relative to the team signal. We train the critics on discounted return targets R^{team} and $\mathbf{R}^i$ by minimizing

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_\tau\left[\left(V^{\text{team}}(\mathbf{o}) - R^{\text{team}}\right)^2 + \sum_{i=1}^{N} \left\|\mathbf{V}^i(\mathbf{o}) - \mathbf{R}^i\right\|_2^2\right], \tag{2}$$

where the expectation is over collected rollouts τ. We next learn the preferences themselves to improve team coordination.

5.2 Coordinated Preference Planning

We then equip each agent with a stochastic preference planner that adapts its own preference based on its local observation. Since preferences lie on the simplex, a natural choice is the Dirichlet distribution. Specifically, the planner outputs $\alpha_i = \phi_\psi(o_i)$, and samples $p_i \sim \text{Dir}(\alpha_i)$, where $p_i \in \Delta^{K-1}$. Motivated by Theorem 4.2, we encourage the planner to maintain diverse preference distributions across agents. We regularize the expected pairwise diversity of the sampled preferences $\mathcal{D}_\alpha := \mathbb{E}[\mathcal{D}_p]$. This regularizer promotes coordinated specialization: different agents can focus on different objective trade-offs, rather than collapsing to the same preference direction. The planner is then trained with the team advantage:

$$\mathcal{L}_{\text{plan}}(\psi) = \mathcal{L}_{\text{PPO}}\left(\phi_\psi(\cdot \mid o_i), A^{\text{team}}\right) - \lambda_1 \mathcal{D}_\alpha. \tag{3}$$

The PPO term encourages preference distributions that yield higher team advantage. Therefore, our algorithm 1 not only let agent learn from local utilities, but are guide them towards team-level optimality.

Algorithm 1 Preference Coordinated Multi-agent Policy Optimization(PCMA)

Initialize preference planner ϕ_ψ, actor π_θ, team critic V^{team}, and vector critics $\{\mathbf{V}^i\}_{i=1}^N$
for each iteration **do**
  for each environment step **do**
    for each agent $i = 1, \ldots, N$ **do**
      sample local preference $p_i \sim \text{Dir}(\phi_\psi(o_i))$ and Sample action $a_i \sim \pi_\theta(\cdot \mid o_i, p_i)$
    end for
    Execute joint action $\mathbf{a} = (a_1, \ldots, a_N)$ and store $(\mathbf{o}, \mathbf{p}, \mathbf{a}, r^{\text{team}}, \{\mathbf{r}^i\}_{i=1}^N, \mathbf{o}')$
  end for
  Estimate the team advantage A^{team} and individual vector advantages $\{\mathbf{A}_i^{\text{ind}}\}_{i=1}^N$
  Update V^{team} and $\{\mathbf{V}^i\}_{i=1}^N$ by minimizing $\mathcal{L}_{\text{critic}}$ defined in Eq. (2)
  Compute A_i and Update π_θ by minimizing the $\mathcal{L}_{\text{actor}}$ defined in Eq. (1)
  Compute preference diversity $\mathcal{D}_\alpha$, and update ϕ_ψ by minimizing the $\mathcal{L}_{\text{plan}}$ defined in Eq. (3)
end for

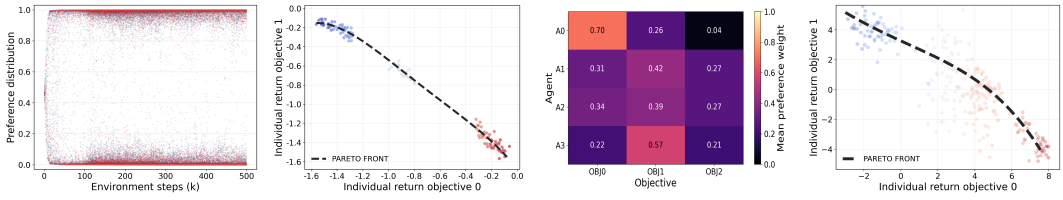
6 Experiments

In this section, we present empirical results to validate the effectiveness of our method. We evaluate on a diverse set of cooperative multi-agent environments, including particle-world coordination (Lowe et al., 2017), drone control (Felten, 2024), walker locomotion (Terry et al., 2021), and StarCraft combat (Samvelyan et al., 2019). To better study coordinated preference learning, we modify these environments by separating the sparse team reward from vector-valued individual rewards. This design allows us to test whether our method can organize vectorized local learning signals toward improved team performance. The numbers of agents and objectives for each environment, together with detailed environment settings, are provided in Appendix B.1 (Table 2). We report team success rate and mean reward across agents and objectives in Table 1. Additional PCMA-SAC results on OpenCDA-MARL/CARLA are reported in Appendix B.3.

6.1 How Does Preference Coordination Improve Cooperation?

Preference specialization in particle-world tasks. We first analyze preference specialization in MOMPE tasks. In Cooperative Spread, we use objective-specific distance rewards, where each objective encourages agents to approach a different landmark. In Predator-Prey, the objectives correspond to approaching different prey while maintaining safety.

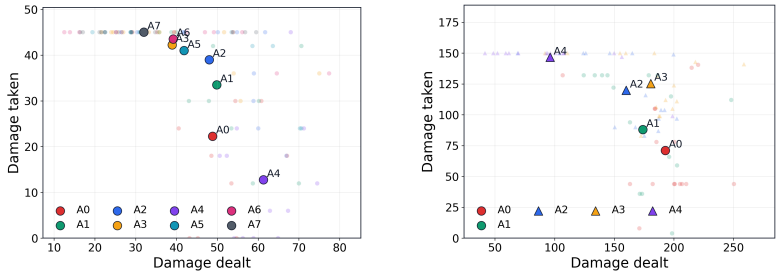
Cooperative Spread has two agents and two objectives: the negative distances to landmarks 1 and 2 (OBJ0 and OBJ1, respectively). Figure 3a shows that, as training progresses, the two agents learn preferences for different landmarks. This specialization encourages them to cover distinct landmarks rather than crowd around the same one. Safe Predator-Prey has four predators and three objectives: progress toward prey 1 (OBJ0), progress toward prey 2 (OBJ1), and safety (OBJ2). Figure 3c similarly shows that the four predators learn different preferences. A0 favors prey 1, while A3 favors prey 2. A1 and A2 balance pursuit and safety more evenly. Figures 3b and 3d show that agents' returns are distributed along the corresponding return fronts.



(a) Preferences (Spread). (b) Return front (Spread). (c) Preferences (Predator). (d) Return front (Predator).

Figure 3: Learned preferences and return fronts for PCMA in particle-world tasks. (a) Preference weights during Cooperative Spread training. (b) 2D vector returns along the front for the two landmark objectives. (c) Mean preference weights in Safe Predator-Prey. (d) 2D vector returns along the front for the two prey objectives.

Role differentiation in SMAC. The same effect appears in SMAC. Figure 4 measures each agent by damage dealt and damage taken, which reflect different combat roles such as aggressive attackers or front-line units. With coordinated preferences, agents spread across the behavior space. This indicates that preference coordination induces role differentiation: different agents emphasize different combat trade-offs while still contributing to the same team objective. Thus, coordinated preference diversity provides a simple mechanism for improving cooperation.

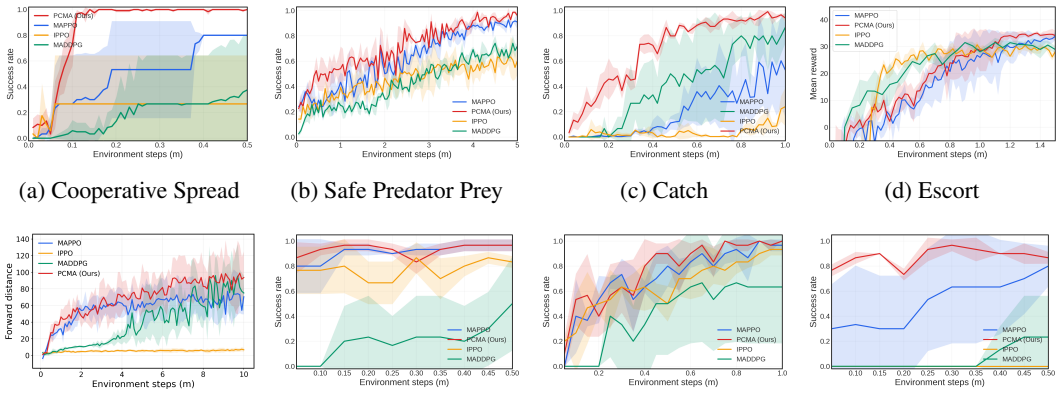


(a) 8m. (b) 2s3z.

Figure 4: Combat behavior under PCMA in (a) 8m and (b) 2s3z. The horizontal and vertical axes show damage dealt and damage taken. Large labeled markers show each agent's episode-averaged values.

6.2 Quantitative Performance

We compare PCMA with MADDPG, IPPO, and MAPPO, covering deterministic actor-critic learning, independent policy optimization, and policy optimization shared critic. All baselines use uniform scalarization weights for the vector-valued guiding reward. More details are shown in Appendix B.4. For most environments, we report success rate and average reward as our main metric in Table 1. For Escort and MOMAwalker, where success rate is not directly defined, we instead report task-specific metrics such as average reward and forward distance. Fig. 5 shows the learning curves across the evaluated tasks. PCMA achieves the best or tied-best performance on most metrics, showing consistent improvements across diverse cooperative multi-agent tasks.



(a) Cooperative Spread (b) Safe Predator Prey (c) Catch (d) Escort

(e) Walker 1 (f) SMAC 3m (g) SMAC 2s3z (h) SMAC 8m

Figure 5: Learning curves of PCMA and baseline methods across the evaluated cooperative multi-agent tasks. Curves are averaged over 3 random seeds, and shaded regions indicate standard deviation.

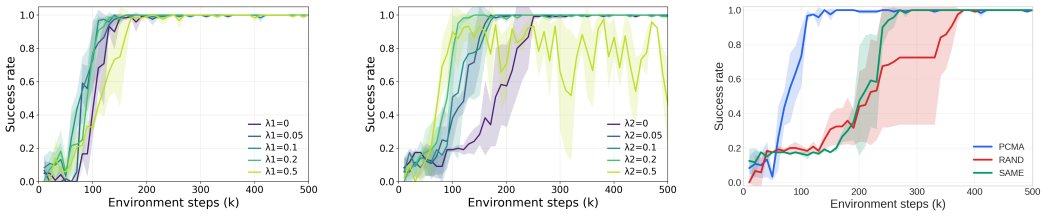
6.3 Ablation Study

As shown in Figure 6, we conduct ablation studies on Cooperative Spread to analyze three key factors in PCMA: the diversity regularization coefficient λ_1, the actor balancing coefficient λ_2, and the learned preference planner. Specifically, λ_1 controls the strength of the pairwise diversity regularizer

Table 1: Results on multi-agent tasks. Values in parentheses denote standard deviation.

Environment	Metric	MADDPG	IPPO	MAPPO	PCMA (Ours)
Cooperative Spread	Success rate	$0.38_{(0.43)}$	$0.27_{(0.38)}$	$0.80_{(0.00)}$	$\mathbf{1.00}_{(0.00)}$
	Average reward	$-0.63_{(1.09)}$	$0.20_{(0.35)}$	$0.69_{(0.00)}$	$\mathbf{0.89}_{(0.00)}$
Safe Predator Prey	Success rate	$0.68_{(0.06)}$	$0.60_{(0.01)}$	$0.91_{(0.05)}$	$\mathbf{0.96}_{(0.02)}$
	Average reward	$2.30_{(0.10)}$	$2.21_{(0.02)}$	$2.39_{(0.04)}$	$\mathbf{2.45}_{(0.05)}$
Catch	Success rate	$0.87_{(0.12)}$	$0.24_{(0.15)}$	$0.53_{(0.37)}$	$\mathbf{0.94}_{(0.03)}$
	Average reward	$4.71_{(7.94)}$	$8.61_{(0.64)}$	$11.33_{(3.83)}$	$\mathbf{14.21}_{(0.83)}$
Escort	Average reward	$29.19_{(0.80)}$	$27.74_{(0.74)}$	$33.72_{(0.63)}$	$\mathbf{34.27}_{(0.99)}$
MOMAwalker	Forward Distance	$75.04_{(21.28)}$	$6.69_{(2.36)}$	$70.52_{(21.08)}$	$\mathbf{93.64}_{(31.47)}$
	Average reward	$18.28_{(16.18)}$	$-0.80_{(0.37)}$	$10.11_{(15.35)}$	$\mathbf{21.62}_{(20.05)}$
SMAC-3m	Success rate	$0.50_{(0.37)}$	$0.83_{(0.05)}$	$\mathbf{0.97}_{(0.05)}$	$\mathbf{0.97}_{(0.05)}$
	Average reward	$0.75_{(0.31)}$	$1.02_{(0.03)}$	$0.29_{(0.02)}$	$\mathbf{1.11}_{(0.03)}$
SMAC-2s3z	Success rate	$0.63_{(0.45)}$	$0.93_{(0.05)}$	$0.97_{(0.05)}$	$\mathbf{1.00}_{(0.00)}$
	Average reward	$0.22_{(0.10)}$	$0.29_{(0.01)}$	$0.30_{(0.01)}$	$\mathbf{0.31}_{(0.01)}$
SMAC-8m	Success rate	$0.23_{(0.33)}$	$0.00_{(0.00)}$	$0.80_{(0.16)}$	$\mathbf{0.87}_{(0.05)}$
	Average reward	$0.08_{(0.12)}$	$0.00_{(0.00)}$	$0.27_{(0.04)}$	$\mathbf{0.28}_{(0.01)}$

in the planner objective, λ_2 balances the team advantage and the scalarized individual advantage in the actor update. Moreover, to validate the effectiveness of our preference coordination, we compare PCMA with RAND and SAME. RAND samples each agent's preference randomly without a learned planner, while SAME forces all agents to use the same preference by averaging their sampled preferences.



(a) Study of λ_1. (b) Study of λ_2. (c) Study of preference planner.

Figure 6: Ablation studies on Cooperative Spread. (a) $\lambda_1 = 0.1$ and $\lambda_1 = 0.2$ improve early learning over $\lambda_1 = 0$, suggesting that preference diversity helps avoid collapse, while $\lambda_1 = 0.5$ slows learning due to excessive diversity. (b) $\lambda_2 = 0.1$ and $\lambda_2 = 0.2$ achieve stable convergence, whereas $\lambda_2 = 0.5$ becomes unstable, showing that individual guidance should not dominate the team signal. (c) The planner learns coordinated, agent-specific preferences rather than merely injecting random diversity.

7 Conclusion and Future Work

We model cooperative MOMARL as a mo-POSG with partially aligned interests and seek a team-optimal equilibrium. Our analysis links preference diversity to first-order team improvement and establishes local equilibrium tracking. Based on this analysis, we propose PCMA to coordinate agent-specific preferences. Experiments show that PCMA can improve team performance and induce interpretable agent specialization.

One limitation of this work is that our study focuses on cooperative benchmarks where the objectives and reward decomposition are explicitly specified. A promising direction for future work is to extend preference coordination to more complex real-world multi-agent systems, such as autonomous driving and open-ended agentic systems, where agents may have different capabilities, contexts, and objectives that naturally emerge from real task demands.

AI USE STATEMENT

Generative AI tools were used to polish language and improve manuscript clarity, assist with algorithm implementation, check citations and bibliographic metadata, and resolve LaTeX formatting and layout issues. AI assistance was also used to refine theoretical formulations and mathematical statements and clarify the interpretation of experimental results. The authors take responsibility for the final content, including AI-assisted text, code, mathematical claims, and references.

REPRODUCIBILITY STATEMENT

Section 4 states the theoretical assumptions and results, with proofs in Appendix A. Section 5 specifies the policy, critic, and preference-planner objectives, while Algorithm 2 details rollout collection, advantage estimation, and parameter updates. Appendix B.1 documents the task setups, action spaces, reward definitions, and success criteria. Appendix B.2 describes the OpenCDA-MARL/CARLA setup and evaluation protocol. Appendix B.4 provides network architectures, training hyperparameters, computing hardware, and measured training times.

REFERENCES

Axel Abels, Diederik M. Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. arXiv preprint arXiv:1809.07803, 2019.

Toygun Basaklar, Suat Gumussoy, and Umit Y. Ogras. Pd-morl: Preference-driven multi-objective reinforcement learning algorithm. arXiv preprint arXiv:2208.07914, 2023.

Christopher Bone and Suzana Dragićević. Gis and intelligent agents for multiobjective natural resource allocation: A reinforcement learning approach. *Transactions in GIS*, 13(3):253–272, 2009.

P. E. M. Borm, S. H. Tijs, and J. C. M. van den Aarssen. Pareto equilibria in multiobjective games. *Methods of Operations Research*, 60:303–312, 1988. URL `https://research.tilburguniversity.edu/en/publications/pareto-equilibria-in-multiobjective-games/`.

Jen Jen Chung, Carrie Rebhuhn, Connor Yates, Geoffrey A. Hollinger, and Kagan Tumer. A multi-agent framework for learning dynamic traffic management strategies. *Autonomous Robots*, 43(6): 1375–1391, 2019. doi: 10.1007/s10514-018-9800-z.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathématique*, 350(5–6):313–318, 2012. doi: 10.1016/j.crma.2012.03.014. URL `https://www.numdam.org/item/10.1016/j.crma.2012.03.014/`.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 13–15 Nov 2017.

Houli Duan, Zhiheng Li, and Yi Zhang. Multiobjective reinforcement learning for traffic signal control using vehicular ad hoc network. *EURASIP Journal on Advances in Signal Processing*, 2010:724035, 2010. doi: 10.1155/2010/724035.

Florian Felten. *Multi-Objective Reinforcement Learning*. PhD Thesis, Unilu - Université du Luxembourg [FSTM], Luxembourg, jun 2024.

Florian Felten, Umut Ucak, Hicham Azmani, Gao Peng, Willem Röpke, Hendrik Baier, Patrick Mannion, Diederik M. Roijers, Jordan K. Terry, El-Ghazali Talbi, Grégoire Danoy, Ann Nowé, and Roxana Rădulescu. Momaland: A set of benchmarks for multi-objective multi-agent reinforcement learning. arXiv preprint arXiv:2407.16312, 2024.

Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I*, 28:89–93, 1964.

Lihao Guo, Louis Liu, Jiahao Tang, Bo Liu, and Siyang Cao. Opencda-marl: A unified benchmarking framework for cooperative autonomous intersection management with multi-agent reinforcement learning. *IEEE Robotics and Automation Letters*, 11(4):4291–4298, 2026.

John C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321, 1955.

Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. Mo-mix: Multi-objective multi-agent cooperative decision-making with deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12098–12112, 2023.

Krisorn Jittorntrum. An implicit function theorem. *Journal of Optimization Theory and Applications*, 25(4):575–577, 1978.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Information Processing Systems (NIPS)*, 2017.

D. Lozovanu, D. Solomon, and A. Zelikovsky. Multiobjective games and determining Pareto-Nash equilibria. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica*, (3(49)): 115–122, 2005. URL https://www.math.md/en/publications/basm/issues/y2005-n3/9536/.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34(1):10, 2020. doi: 10.1007/s10458-019-09433-x.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. arXiv preprint arXiv:1803.11485, 2018.

Lillian J. Ratliff, Samuel A. Burden, and S. Shankar Sastry. On the characterization of local nash equilibria in continuous games. arXiv preprint arXiv:1411.2168, 2014.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

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. *arXiv preprint arXiv:1902.04043*, 2019.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Justin K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S. Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15032–15043, 2021.

Peter Vamplew, Benjamin J. Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M. Roijers, Conor F. Hayes, Fredrik Heintz, Patrick Mannion, Pieter J. K. Libin, Richard Dazeley, and Cameron Foale. Scalar reward is not enough: a response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, 2022. doi: 10.1007/s10458-022-09575-5.

Mark Voorneveld, Dries Vermeulen, and Peter Borm. Axiomatizations of pareto equilibria in multicriteria games. *Games and Economic Behavior*, 28(1):146–154, 1999.

Xiaofeng Wang and Tuomas Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *Advances in Neural Information Processing Systems*, volume 15, 2002.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. arXiv preprint arXiv:1908.08342, 2019.

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. arXiv preprint arXiv:2103.01955, 2022.

A Detailed Proofs for Theoretical Results

A.1 Assumptions

We begin by restating the assumptions required for the theoretical analysis.

Assumption A.1 (Preference-improvement alignment)**.** There exists $\kappa > 0$ such that, for every agent i with $\tilde{p}_i \neq 0$, the projection of $\tilde{b}_i$ onto $\tilde{p}_i$ is lower bounded by κ, i.e.,

$$\frac{\tilde{p}_i^\top \tilde{b}_i}{\|\tilde{p}_i\|_2^2} \geq \kappa, \qquad i = 1, \ldots, N.$$

where $\tilde{p}_i := p_i - \bar{p}$ and $\tilde{b}_i := b_i - \bar{b}$.

Assumption A.2 (Attraction property near equilibrium)**.** Let $\theta(\mathbf{p})$ denote a strict local Nash equilibrium under the preference profile $\mathbf{p}$. Assume that there exist $\rho \in (0, 1)$ and a neighborhood of $\theta(\mathbf{p})$ such that

$$\|\theta_{\text{new}} - \theta(\mathbf{p})\| \leq \rho \|\theta_{\text{old}} - \theta(\mathbf{p})\|,$$

for all θ_{old} in this neighborhood.

Assumption A.3 (Local Lipschitz continuity)**.** Let $\theta(\cdot) : \mathcal{U} \to \Theta$ be the strict local Nash equilibrium mapping induced by Lemma 4.3. Assume that this mapping is Lipschitz continuous in $\mathcal{U}$: there exists $C > 0$ such that for any two preference profiles $\mathbf{p}, \mathbf{p}' \in \mathcal{U}$,

$$\|\theta(\mathbf{p}') - \theta(\mathbf{p})\| \leq C \|\mathbf{p}' - \mathbf{p}\|.$$

A.2 Proof of Theorem 4.2 in the main text

Theorem A.4 (Team Improvement Decomposition)**.** *Under Assumption A.1, the team improvement as $\eta \to 0^+$ satisfies*

$$J_{\text{team}}(\theta_{\text{new}}) - J_{\text{team}}(\theta) \geq \eta \sum_{i=1}^{N} \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta N \left(\bar{p}^\top \bar{b} + \kappa \mathcal{D}_p\right) + o(\eta)$$

where

$$\mathcal{D}_p := \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|p_i - p_j\|_2^2.$$

is the pairwise preference distance across agents.

Proof. Given preferences $p_1, \ldots, p_N$, each agent takes one gradient update

$$\theta_{i,\text{new}} = \theta_i + \eta \nabla_{\theta_i} U_i(\theta; p_i), \qquad i = 1, \ldots, N.$$

where

$$U_i(\theta; p_i) = J_{\text{team}}(\theta) + \langle p_i, J_i(\theta) \rangle,$$

We write $\theta_{\text{new}} = (\theta_{1,\text{new}}, \ldots, \theta_{N,\text{new}})$. Let $\Delta J := J_{\text{team}}(\theta_{\text{new}}) - J_{\text{team}}(\theta)$. By the first-order Taylor expansion of J_{team} at θ,

$$\begin{aligned}
\Delta J &= \eta \sum_{i=1}^{N} \langle \nabla_{\theta_i} J_{\text{team}}(\theta), \nabla_{\theta_i} U_i(\theta; p_i) \rangle + o(\eta) \\
&= \eta \sum_{i=1}^{N} \langle \nabla_{\theta_i} J_{\text{team}}(\theta), \nabla_{\theta_i} J_{\text{team}}(\theta) + \nabla_{\theta_i} \langle p_i, J_i(\theta) \rangle \rangle + o(\eta) \\
&= \eta \sum_{i=1}^{N} \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \sum_{i=1}^{N} \left\langle \nabla_{\theta_i} J_{\text{team}}(\theta), \sum_{k=1}^{K} p_{i,k} \nabla_{\theta_i} J_{i,k}(\theta) \right\rangle + o(\eta) \\
&= \eta \sum_{i=1}^{N} \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \sum_{i=1}^{N} \sum_{k=1}^{K} p_{i,k} \langle \nabla_{\theta_i} J_{\text{team}}(\theta), \nabla_{\theta_i} J_{i,k}(\theta) \rangle + o(\eta) \\
&= \eta \sum_{i=1}^{N} \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \sum_{i=1}^{N} \langle p_i, b_i \rangle + o(\eta),
\end{aligned}$$

Let $\tilde{p}_i = p_i - \bar{p}$ and $\tilde{b}_i = b_i - \bar{b}$. Then $\sum_{i=1}^N \tilde{p}_i = 0$ and $\sum_{i=1}^N \tilde{b}_i = 0$. Let $P = (\tilde{p}_1, \ldots, \tilde{p}_N)^\top \in \mathbb{R}^{N\times K}$ be the centered preference matrix, and then crossing term can be decomposed as follows.

$$\begin{aligned}
\Delta J &= \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \sum_{i=1}^N \left\langle \bar{p} + \tilde{p}_i, \bar{b} + \tilde{b}_i \right\rangle + o(\eta) \\
&= \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \left(N\langle \bar{p}, \bar{b}\rangle + \sum_{i=1}^N \langle \tilde{p}_i, \tilde{b}_i\rangle \right) + o(\eta) \\
&\geq \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta \left(N\langle \bar{p}, \bar{b}\rangle + \kappa \sum_{i=1}^N \|\tilde{p}_i\|_2^2 \right) + o(\eta) \\
&= \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta N \bar{p}^\top \bar{b} + \eta\kappa \|P\|_F^2 + o(\eta)
\end{aligned}$$

The inequality uses Assumption A.1, and $\|\cdot\|_F$ denotes the Frobenius norm. Finally, the squared Frobenius norm of the centered preference matrix, divided by N, equals the pairwise preference distance:

$$\begin{aligned}
\frac{1}{N}\|P\|_F^2 &= \frac{1}{N}\sum_{i=1}^N \|p_i - \bar{p}\|_2^2 \\
&= \frac{1}{N}\sum_{i=1}^N \|p_i\|_2^2 - \|\bar{p}\|_2^2 \quad \text{(by the variance identity)} \\
&= \frac{1}{N}\sum_{i=1}^N \|p_i\|_2^2 - \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N p_i^\top p_j \quad \text{(by bilinearity)} \\
&= \frac{1}{2N^2}\sum_{i=1}^N\sum_{j=1}^N \left(\|p_i\|_2^2 + \|p_j\|_2^2 - 2p_i^\top p_j\right) \\
&= \frac{1}{2N^2}\sum_{i=1}^N\sum_{j=1}^N \|p_i - p_j\|_2^2 = \mathcal{D}_p.
\end{aligned}$$

We obtain the following asymptotic lower bound:

$$J_{\text{team}}(\theta_{\text{new}}) - J_{\text{team}}(\theta) \geq \eta \sum_{i=1}^N \|\nabla_{\theta_i} J_{\text{team}}(\theta)\|_2^2 + \eta N \bar{p}^\top \bar{b} + \eta\kappa N \mathcal{D}_p + o(\eta).$$

□

A.3 Proof of Lemma 4.3 in the main text

Lemma A.5 (Continuity of preference-conditioned local Nash equilibria)**.** *Suppose that $\theta^* \in \operatorname{int}\Theta$ is a strict local Nash equilibrium of $\mathcal{G}(\mathbf{p})$ satisfying*

$$\nabla_{\theta_i} U_i(\theta^*; p_i) = 0, \qquad \nabla^2_{\theta_i\theta_i} U_i(\theta^*; p_i) \prec 0, \qquad i = 1, \ldots, N.$$

Assume that the joint gradient $(\nabla_{\theta_1} U_1, \ldots, \nabla_{\theta_N} U_N)$ is continuously differentiable near $(\theta^*, \mathbf{p})$, and its Jacobian with respect to θ is nonsingular at this point. Then there exist neighborhoods $\mathcal{U}$ of $\mathbf{p}$ and $\mathcal{V} \subset \Theta$ of θ^*, and a unique continuously differentiable mapping $\theta(\cdot) : \mathcal{U} \to \mathcal{V}$ with $\theta(\mathbf{p}) = \theta^*$, such that, for every $\tilde{\mathbf{p}} \in \mathcal{U}$, $\theta(\tilde{\mathbf{p}})$ is the unique stationary point in $\mathcal{V}$ and a strict local Nash equilibrium of $\mathcal{G}(\tilde{\mathbf{p}})$.

Proof. Define the joint stationarity mapping

$$F(\theta, \mathbf{p}) := \left(\nabla_{\theta_1} U_1(\theta; p_1), \ldots, \nabla_{\theta_N} U_N(\theta; p_N)\right).$$

The assumed first-order conditions give

$$F(\theta^*, \mathbf{p}) = 0.$$

By assumption, F is continuously differentiable near $(\theta^*, \mathbf{p})$, and its Jacobian $\nabla_\theta F(\theta^*, \mathbf{p})$ is non-singular. Therefore, by the Implicit Function Theorem, there exist sufficiently small neighborhoods $\mathcal{U}$ of $\mathbf{p}$ and $\mathcal{V} \subset \operatorname{int}\Theta$ of θ^*, and a unique continuously differentiable mapping $\theta(\cdot) : \mathcal{U} \to \mathcal{V}$ such that $\theta(\mathbf{p}) = \theta^*$ and

$$F(\theta(\tilde{\mathbf{p}}), \tilde{\mathbf{p}}) = 0, \qquad \forall\, \tilde{\mathbf{p}} \in \mathcal{U}.$$

For each $\tilde{\mathbf{p}} \in \mathcal{U}$, $\theta(\tilde{\mathbf{p}})$ is the unique stationary point in $\mathcal{V}$, satisfying

$$\nabla_{\theta_i} U_i(\theta(\tilde{\mathbf{p}}); \tilde{p}_i) = 0, \qquad i = 1, \dots, N.$$

The Hessians are negative definite at $(\theta^*, \mathbf{p})$. By continuity of the Hessians and $\theta(\cdot)$, this property also holds for all sufficiently nearby preference profiles:

$$\nabla^2_{\theta_i \theta_i} U_i(\theta(\tilde{\mathbf{p}}); \tilde{p}_i) \prec 0, \qquad i = 1, \dots, N, \quad \tilde{\mathbf{p}} \in \mathcal{U}.$$

For each agent, hold the other agents' policies fixed. Stationarity and the second-order conditions then imply that its policy is a strict local maximizer of its payoff. Hence every $\theta(\tilde{\mathbf{p}})$ is a strict local Nash equilibrium of $\mathcal{G}(\tilde{\mathbf{p}})$, which proves the result. □

A.4 Proof of Theorem 4.6 in the main text

Theorem A.6 (Equilibrium tracking)**.** *Consider the iterative policy update*

$$\theta_i^{t+1} = \theta_i^t + \eta \nabla_{\theta_i} U_i(\theta^t; \mathbf{p}^t), \qquad i = 1, \dots, N.$$

Under Assumption 4.4 and Assumption 4.5 on page 5 of the main text, the tracking error $e_t := \|\theta^t - \theta(\mathbf{p}^t)\|$ *satisfies*

$$e_{t+1} \le \rho e_t + C\|\mathbf{p}^{t+1} - \mathbf{p}^t\|.$$

In particular, if $\|\mathbf{p}^{t+1} - \mathbf{p}^t\| \le \delta$*, then*

$$\limsup_{t \to \infty} e_t \le \frac{C}{1-\rho}\delta.$$

Proof. For any $t \ge 0$, by adding and subtracting $\theta(\mathbf{p}^t)$, we have

$$\begin{aligned} e_{t+1} &= \|\theta^{t+1} - \theta(\mathbf{p}^{t+1})\| \\ &\le \|\theta^{t+1} - \theta(\mathbf{p}^t)\| + \|\theta(\mathbf{p}^t) - \theta(\mathbf{p}^{t+1})\|. \end{aligned}$$

First, by Assumption 4.4, the policy update is locally contractive around $\theta(\mathbf{p}^t)$. Therefore,

$$\|\theta^{t+1} - \theta(\mathbf{p}^t)\| \le \rho\|\theta^t - \theta(\mathbf{p}^t)\| = \rho e_t.$$

Second, by Assumption 4.5, the equilibrium path changes Lipschitz continuously with the preference profile, so

$$\|\theta(\mathbf{p}^t) - \theta(\mathbf{p}^{t+1})\| \le C\|\mathbf{p}^{t+1} - \mathbf{p}^t\|.$$

Combining the two bounds gives

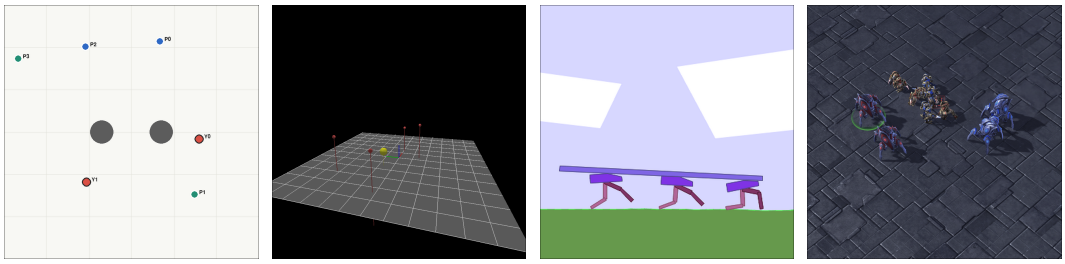
$$e_{t+1} \le \rho e_t + C\|\mathbf{p}^{t+1} - \mathbf{p}^t\|.$$

If $\|\mathbf{p}^{t+1} - \mathbf{p}^t\| \le \delta$, then

$$e_{t+1} \le \rho e_t + C\delta.$$

Applying this inequality recursively gives

$$\begin{aligned} e_t &\le \rho e_{t-1} + C\delta \\ &\le \rho^2 e_{t-2} + C\delta(1+\rho) \\ &\le \cdots \\ &\le \rho^t e_0 + C\delta \sum_{k=0}^{t-1} \rho^k \\ &= \rho^t e_0 + C\delta \frac{1-\rho^t}{1-\rho}. \end{aligned}$$



(a) MOMPE (b) Catch/Escort (c) MOMAWalker (d) SMAC

Figure 7: Environments used in our experiments.

Since $\rho \in (0, 1)$, take $\limsup_{t\to\infty}$, and we obtain

$$\limsup_{t\to\infty} e_t \leq \frac{C}{1-\rho}\delta.$$

□

B Implementation Details

To support reproducibility, we provide detailed environment specifications, training hyperparameters, and pseudocode in the appendix. The source code, including environment modifications and training scripts, will be released upon acceptance.

B.1 Environment Setting

This subsection provides additional details of the environments used in our experiments. We describe the task setup, action space, reward design, and success criterion for each environment. The detailed reward design and success criterion are summarized in Table 2.

Cooperative Spread and Safe Predator-Prey. We use two particle-world tasks with discrete actions. Spread is a two-agent, two-landmark variant of MPE simple spread, where agents must coordinate to cover distinct landmarks. Predator-Prey uses four predators, two moving prey, and obstacles in a bounded 2D arena. Compared with Spread, Predator-Prey introduces moving targets, obstacle avoidance, and a three-objective reward vector that separates progress toward the two prey from safety.

Catch and Escort. Catch and Escort are continuous-control drone tasks adapted from CrazyRL/MOMALand. In both environments, four drones move in a bounded 3D space with 3D velocity actions. Catch focuses on approaching and capturing a target while avoiding unsafe proximity to other drones. Escort instead requires drones to maintain an escort formation around a moving target as it travels toward a goal. Thus, Catch emphasizes interception, while Escort emphasizes sustained formation control.

MOMAWalker. MOMAWalker is adapted from the Multiwalker domain. Three walkers jointly transport a package across terrain using continuous joint-torque actions. We use this environment to evaluate coordinated continuous control under a shared physical task, where agents must trade off forward progress with package and body stability.

SMAC. SMAC comprises cooperative StarCraft Multi-Agent Challenge maps, including 3m, 2s3z, and 8m, and the difficulty is set to 5. In these tasks, agents control decentralized units using discrete actions with action masks. We use a two-dimensional combat reward vector based on damage dealt and health loss, together with a sparse team-success signal indicating whether all enemy units are defeated within the episode limit.

Table 2: Reward and success details

Environment	Dimension	Team reward	Individual reward	Success criterion
Cooperative Spread	2A2O	Sparse task reward or punishment: $+1$ when both landmark reached, -1 on collision.	[- distance to landmark 1, - distance to landmark 2].	Two agents occupy two distinct landmarks.
Safe Predator-Prey	4A3O	Sparse task reward or punishment: $+1$ when catch prey, -1 on collision or bump into obstacle.	[progress to prey 1, progress to prey 2, - close distance to obstacle/other predator].	All prey are captured.
Catch	4A2O	Sparse task reward or punishment: $+1$ when catch target, -1 on collision.	[progress to target, - close distance to other drone]	catch the target
Escort	4A2O	only collision penalty	[- distance to target, distance to other drones]	Not Applicable
MOMAWalker	3A2O	Package progress distance, fall/drop penalties.	[Walker progress, stability penalty]	Not Applicable
SMAC	xA2O	+1 on success.	[damage dealt, -health loss]	Defeat all enemies within episode time
OpenCDA-MARL	$\leq$10A2O	Route success, progress, step penalty, and collision penalty.	[efficiency, safety/ interaction risk]	Route completion without collision.

OpenCDA-MARL. OpenCDA-MARL extends OpenCDA with a MARL controller for CARLA intersection control. The environment is an unsignalized four-way intersection on a local OpenDRIVE map, with traffic entering from north, south, east, and west, and each episode lasts at most 2400 simulation steps. During training, we use live traffic with at most 10 simultaneously active vehicles, all of them MARL-controlled CAVs (there is no human-driven background traffic); for checkpoint evaluation, we replay a fixed traffic file so that all algorithms face the same arrival pattern.

B.2 OpenCDA-MARL/CARLA Experimental Details

CAVs spawn on the four approaches and drive through the intersection to their destinations. Each CAV follows a fixed route generated by the OpenCDA planner. The learned action is a one-dimensional target speed command in $\mathrm{km} \cdot \mathrm{h}^{-1}$, clipped to $[0, 45]$, and the low-level controller tracks this speed along the assigned route. The learned speed is applied directly: CAVs have no rule-based right-of-way or collision-avoidance layer, so the collision rate measures the learned policies themselves, which explains the high absolute collision rates of all methods. The experiment therefore tests learned yielding and crossing-order behavior under speed control, not route selection. The base observation includes relative position to the intersection, absolute position, lane position, heading angle, distance to the intersection, distance to the front vehicle, and waypoint-buffer information. Preference-conditioned variants append a two-dimensional preference vector to this state.

PCMA-SAC. In this benchmark, PCMA is instantiated on a SAC backbone. Each CAV is assigned a preference vector when it spawns; the vector is appended to its state and weights its two objectives. During training, a learned, observation-dependent shift adapts the preference that conditions each CAV's actor. The Dirichlet preference planner of Section 5 is not used here, so Table 3 validates preference-conditioned coordination in the off-policy setting rather than the full on-policy algorithm.

Rewards and objectives. We use two reward profiles, `cooperative` and `competitive`, with the same learned terms: collision penalty -400, success reward 300, step penalty -0.8, and progress shaping scaled by 0.4. Both group the reward components into two objectives: efficiency (progress, success, and step penalty) and a collision-driven second objective, called safety in the cooperative profile and interaction risk in the competitive one. Both settings also share the traffic and the preference assignment: CAVs entering from the north and east receive the preference $[0.6, 0.4]$ over efficiency and safety, and those entering from the south and west receive $[0.4, 0.6]$. The competitive setting differs in two respects: it additionally records time-to-collision, yielding, clearance, and per-group outcomes as metric-only signals, and PCMA-SAC additionally uses a separation term that encourages the adapted preferences of the two groups to stay at least 0.25 apart (weight 0.02). SAC and MAPPO use the same configuration in both settings, so their two rows are independent repetitions.

Metrics and protocol. We report success rate (percentage of vehicles that complete their route), collision rate (percentage of vehicles that collide), and throughput (successful completions per simulated hour). Each run is evaluated at the checkpoint of its highest-return training episode, on the fixed evaluation replay; we report the mean and standard deviation over three seeds.

B.3 OpenCDA-MARL/CARLA Results

We evaluate PCMA-SAC on OpenCDA-MARL (Guo et al., 2026) in CARLA (Dosovitskiy et al., 2017); Table 3 reports the results under the protocol in Appendix B.2.

Table 3: OpenCDA-MARL/CARLA best-checkpoint evaluation. Mean $\pm$ std over three seeds; bold marks the best mean per setting. Without preference separation, PCMA-SAC performs comparably to SAC; with preference separation, it favors a more conservative efficiency–safety trade-off.

Setting	Method	Success (%) ↑	Collision (%) ↓	Throughput (veh/h) ↑
Coop.	SAC	68.6 ± 1.9	31.4 ± 1.9	1692 ± 46
Coop.	MAPPO	55.4 ± 4.7	42.5 ± 3.5	1417 ± 100
Coop.	PCMA-SAC	$\mathbf{69.6 \pm 1.2}$	$\mathbf{30.4 \pm 1.2}$	$\mathbf{1716 \pm 29}$
Comp.	SAC	67.9 ± 0.7	32.1 ± 0.7	$\mathbf{1674 \pm 17}$
Comp.	MAPPO	42.1 ± 30.3	$\mathbf{23.3 \pm 17.6}$	1066 ± 759
Comp.	PCMA-SAC	$\mathbf{68.0 \pm 2.2}$	$\mathbf{23.3 \pm 8.3}$	1192 ± 291

B.4 Training Details

Algorithm 2 gives the full training procedure of PCMA.

Training setup. Unless otherwise stated, all on-policy methods use the same PPO-style training setup. We use Adam for both actor and critic optimization, and all MLPs use Tanh activations. For PCMA and MAPPO-style baselines, the actor is a shared preference-conditioned MLP with two hidden layers of size 64. The actor separately encodes the local observation and the preference, then fuses the two embeddings before the action head. The preference embedding dimension is 32. The centralized team critic takes the concatenated observations $\mathbf{o}$ and outputs a scalar team value, with two hidden layers of size 128. Each individual critic has independent parameters, the same input and hidden-layer architecture, and a K-dimensional output for its agent. Table 5 lists the additional hyperparameters used by PCMA.

Training time. All PCMA experiments were run on an internal Linux workstation with NVIDIA RTX PRO 6000 Blackwell GPUs (98GB GPU memory each). Each training run used one GPU and one training process. Non-SMAC experiments used 4 parallel environment workers, while SMAC used 1 environment worker due to the single-environment action-mask path. The measured wall-clock training time for one PCMA seed was 11.6 minutes for Cooperative Spread, 21.2 minutes for Catch, 1.12 hours for Safe Predator-Prey, 4.49 hours for MultiWalker, 1.12 hours for SMAC-3m, 2.31 hours for SMAC-2s3z, and 1.54 hours for SMAC-8m. Main comparisons use three seeds, so

the reported PCMA training compute is approximately three times these per-seed costs for each environment. Execution-time evaluation and rendering use fixed policy rollouts without gradient updates and were run on the same machine; these jobs are lightweight compared with training and typically finish within minutes per checkpoint.

Algorithm 2 Preference Coordinated Multi-Agent Policy Optimization (PCMA)

Require: number of agents N, objective dimension K, rollout length T, PPO epochs E
Require: actor π_θ, preference planner ϕ_ψ, team critic V_η^{team}, individual critics $\{\mathbf{V}_{\xi_i}^i\}_{i=1}^N$

for iteration $= 1, 2, \ldots$ **do**
  for $t = 0, \ldots, T-1$ **do**
    for agent $i = 1, \ldots, N$ **do**
      Observe local observation $o_{i,t}$
      Generate inherent preference

$$p_{i,t} \sim \text{Dirichlet}\big(\phi_\psi(o_{i,t})\big)$$

      Sample action

$$a_{i,t} \sim \pi_\theta(\cdot \mid o_{i,t}, p_{i,t})$$

    end for
    Execute joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$
    Store

$$(\mathbf{o}_t, \{p_{i,t}\}_{i=1}^N, \mathbf{a}_t, \log \pi_{\theta_{\text{old}}}, r_t^{\text{team}}, \{\mathbf{r}_{i,t}^{\text{ind}}\}_{i=1}^N, \mathbf{o}_{t+1})$$

  end for
  Estimate team returns $\hat{R}^{\text{team}}$ and team advantages A^{team} with GAE
  Estimate individual vector returns $\hat{\mathbf{R}}_i^{\text{ind}}$ and advantages $\mathbf{A}_i^{\text{ind}}$ with GAE
  Form each agent's actor advantage:

$$A_i = A^{\text{team}} + \lambda_2 p_i^\top \mathbf{A}_i^{\text{ind}}$$

  for PPO epoch $= 1, \ldots, E$ **do**
    for minibatch $\mathcal{B}$ **do**
      Update critics by minimizing

$$\mathcal{L}_{\text{critic}} = \mathcal{L}_{\text{team}} + \mathcal{L}_{\text{ind}}.$$

      Compute PPO ratio

$$\rho_i = \frac{\pi_\theta(a_i \mid o_i, p_i)}{\pi_{\theta_{\text{old}}}(a_i \mid o_i, p_i)}.$$

      Update actor with clipped PPO objective:

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_{\mathcal{B}}\left[\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right)\right] - \beta_{\text{ent}} \mathcal{H}(\pi_\theta).$$

      Compute preference diversity regularizer $\mathcal{D}_\alpha(\{p_i\}_{i=1}^N)$
      Update planner with $\mathcal{L}_{\text{plan}}$.
    end for
  end for
end for

Table 4: Common training setup.

Hyperparameter	Value
Optimizer	Adam
Actor learning rate	3×10^{-4}
Critic learning rate	3×10^{-4}
Discount factor γ	0.99
GAE parameter λ	0.95
Rollout length	256
PPO epochs E	4
Minibatches per update	4
PPO clipping coefficient ϵ	0.2
Value loss coefficient	0.5
Entropy coefficient	0.01
Max gradient norm	0.5
Actor hidden layers	$[64, 64]$
Preference encoder dimension	32
Critic hidden layers	$[128, 128]$
Actor parameter sharing	Fully shared across agents
Agent-ID embedding	Dimension 8

Table 5: PCMA-specific hyperparameters.

Environment	λ_1	λ_2
Simple Spread	0.10	0.20
Catch	0.05	0.10
Escort	0.05	0.10
Predator-Prey	0.05	0.20
MultiWalker	0.02	0.20
SMAC	0.02	0.10

C Illustrative Examples

C.1 Example for Pareto-Nash Equilibrium

For example, consider a two agent two objective game. Each agent follows a Gaussian policy with fixed variance, $a_1 \sim \mathcal{N}(\mu_1, 1)$ and $a_2 \sim \mathcal{N}(\mu_2, 1)$, where $\mu_1, \mu_2 \in [-1, 1]$. Define team reward as

$$(r_1, r_2) = \left(\exp\left[-\tfrac{1}{2}\left((a_1-1)^2 + \tfrac{1}{2}(a_2-1)^2\right)\right],\ \exp\left[-\tfrac{1}{2}\left(\tfrac{1}{2}(a_1+1)^2 + 2(a_2+1)^2\right)\right]\right).$$

Under independent action sampling, the expected returns are

$$R_1(\mu) = \frac{1}{\sqrt{3}} \exp\left[-\frac{(\mu_1-1)^2}{4} - \frac{(\mu_2-1)^2}{6}\right],$$
$$R_2(\mu) = \frac{\sqrt{2}}{3} \exp\left[-\frac{(\mu_1+1)^2}{6} - \frac{(\mu_2+1)^2}{3}\right].$$

Within the specified fixed-variance Gaussian policy class, R_1 is strictly increasing and R_2 is strictly decreasing in either mean on $[-1, 1]$. Consequently, every $(\mu_1, \mu_2) \in [-1, 1]^2$ is a Pareto–Nash equilibrium: no unilateral change can improve one objective without worsening the other. This does not imply team Pareto optimality. For example, a coordinated change from $(0, 0)$ to $(1/4, -1/4)$ strictly improves both returns, since

$$\frac{R_1(1/4, -1/4)}{R_1(0, 0)} = e^{1/64} > 1, \qquad \frac{R_2(1/4, -1/4)}{R_2(0, 0)} = e^{5/96} > 1.$$

C.2 A Toy Example of Better Team Improvement with Agent-specific Preferences

We give a one-step, two-agent game with a coupled team objective and two individual guiding objectives, following the utility definition in Section 3.2. The example uses dense quadratic rewards to make both the learning dynamics and their limits exact. Actions are sampled independently as $a_i \sim \mathcal{N}(\mu_i, 1)$, with $\mu_i \in \mathbb{R}$ the only trainable parameters. Define

$$r_{\text{team}}(a_1, a_2) = -\frac{1}{4}\left[(a_1 - a_2 - 1)^2 + 2(a_1 + a_2)^2\right], \qquad \mathbf{r}_i(a_i) = -\frac{1}{2}\begin{pmatrix}(a_i-1)^2\\(a_i+1)^2\end{pmatrix}.$$

The team reward favors a prescribed separation and a zero joint center, whereas the guiding objectives favor actions near $+1$ and -1. Let $s = \mu_1 + \mu_2$ and $d = \mu_1 - \mu_2$. Taking expectations gives

$$J_{\text{team}}(\mu) = -\frac{3}{2} - \frac{1}{4}(d-1)^2 - \frac{1}{2}s^2, \qquad \mathbf{J}_i(\mu_i) = -\frac{1}{2}\begin{pmatrix}(\mu_i-1)^2+1\\(\mu_i+1)^2+1\end{pmatrix}.$$

The constant $-3/2$ is the contribution of the fixed action variances. The unique team optimum in this policy class is $\mu^{\text{opt}} = (1/2, -1/2)$, with $J_{\text{team}} = -3/2$.

Exact dynamics and equilibria. Write $p_i = (\alpha_i, 1-\alpha_i)$ and $q_i = 2\alpha_i - 1$. Each agent maximizes $U_i = J_{\text{team}} + p_i^\top \mathbf{J}_i$ using simultaneous exact gradient ascent with the same step size η:

$$\mu^{t+1} = \mu^t + \eta g(\mu^t), \qquad g(\mu) = \begin{pmatrix}\frac{1}{2} + q_1 - \frac{5}{2}\mu_1 - \frac{1}{2}\mu_2\\ -\frac{1}{2} + q_2 - \frac{1}{2}\mu_1 - \frac{5}{2}\mu_2\end{pmatrix}.$$

Equivalently,

$$s^{t+1} = (1-3\eta)s^t + \eta(q_1+q_2), \qquad d^{t+1} = (1-2\eta)d^t + \eta(1+q_1-q_2).$$

Thus, for any fixed preference profile and $0 < \eta < 2/3$, the iterates converge from any initial means to the unique stationary point

$$s^* = \frac{q_1+q_2}{3}, \qquad d^* = \frac{1+q_1-q_2}{2}.$$

This point is a Nash equilibrium because each U_i is strictly concave in its own mean, with $\partial^2 U_i/\partial\mu_i^2 = -5/2$.

Comparison with every fixed shared preference. For any shared preference $p_1 = p_2 = (\alpha, 1-\alpha)$, let $q = 2\alpha - 1$. The induced equilibrium and its team return are

$$\mu^*_{\text{shared}} = \left(\frac{q}{3} + \frac{1}{4}, \frac{q}{3} - \frac{1}{4}\right), \qquad J_{\text{team}}(\mu^*_{\text{shared}}) = -\frac{25}{16} - \frac{2q^2}{9} \leq -\frac{25}{16}.$$

The best shared preference is $(1/2, 1/2)$. Instead, choose the interior, agent-specific preferences

$$p_1 = (3/4, 1/4), \qquad p_2 = (1/4, 3/4).$$

Their average is still $(1/2, 1/2)$, but now $q_1 = 1/2$ and $q_2 = -1/2$, so the induced equilibrium is exactly $\mu^{\text{opt}} = (1/2, -1/2)$. Its team return is $-3/2$, a strict improvement of $1/16$ over the best shared-preference equilibrium. Both runs can start from $\mu^0 = (0, 0)$ and use any common step size $0 < \eta < 2/3$.

First-order alignment and the role of coordination. At $\mu^0 = (0, 0)$, the team gradient is $(1/2, -1/2)$ and the team-improvement matrix defined in Section 4.1 is

$$B = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

For the coordinated preferences above, $\bar{b} = 0$, $\tilde{b}_i = 2\tilde{p}_i$, and $\mathcal{D}_p = 1/8$. Hence Assumption 4.1 holds at this initial point with $\kappa = 2$, giving an additional first-order term $\eta N \kappa \mathcal{D}_p = \eta/2$ relative to the shared mean preference. Indeed, writing $\Delta J = J_{\text{team}}(\mu^1) - J_{\text{team}}(\mu^0)$, the exact one-step improvements are

$$\Delta J_{\text{shared}} = \frac{\eta}{2} - \frac{\eta^2}{4}, \qquad \Delta J_{\text{coordinated}} = \eta - \eta^2.$$

Their difference, $\eta/2 - 3\eta^2/4$, is positive for $0 < \eta < 2/3$; the quadratic terms also show why the Taylor remainder in Theorem 4.2 must be retained.

Finally, swapping the two coordinated preferences preserves both their mean and diversity but gives $g(0, 0) = 0$: learning stays at $(0, 0)$ with team return $-7/4$, worse than the best shared preference. Thus, the benefit comes from aligning preferences with agents' contributions, not diversity alone. The example establishes the existence of beneficial preference coordination under the stated utilities.