Guiding Team Objectives with Individual Policies: A Two-Stage Model Aggregation Framework for Partially Cooperative MARL

Wanting Liu[1] , **Baoxi Wang**[2] , **Pengyi Li**[3] , **Lu Jiang**[1] and **Chengwei Zhang**[1,*]
[1]Dalian Maritime University, Dalian, China
[2]Dalian University of Technology, Dalian, China
[3]Tianjin University, Tianjin, China
{wantyluu, jiangl761, chenvy}@dlmu.edu.cn, wangbaoxi@mail.dlut.edu.cn, lipengyi@tju.edu.cn

Abstract

In Partially Cooperative Markov Games (PCMG), agents need to learn effective cooperation under individual rewards, with the key challenge lies in leveraging these rewards to enhance team benefits optimally. Model aggregation (MA) is a promising solution due to its simplicity and efficiency, yet it suffers from instability caused by the volatility of individual policies trained with individual rewards and the non-asymptotic nature of aggregation updates. To address these challenges, we propose **T**wo-**S**tage **M**odel **A**ggregation (TSMA), a novel multi-policy MA framework for multi-agent reinforcement learning (MARL) that leverages individual policies to enhance team cooperation. Specifically, to enhance the stability and expressiveness of individual policies, each TSMA agent trains two individual policies (trained using the same individual reward) with parameter-level variation and combines them via adaptive weighted aggregation. The aggregated individual policy is then integrated with the team policy (trained using team reward), with integration weights dynamically adjusted over training. Additionally, L2 regularization is applied during the data-driven training to align their parameter distributions across policies, mitigating instability from non-asymptotic aggregation. As a general framework, TSMA is compatible with standard actor-critic MARL algorithms. We integrate TSMA with MAPPO and MADDPG, and evaluate them on challenging PettingZoo benchmarks, demonstrating significant improvements in PCMG. The code is available at: https://github.com/RL-DLMU/TSMA.

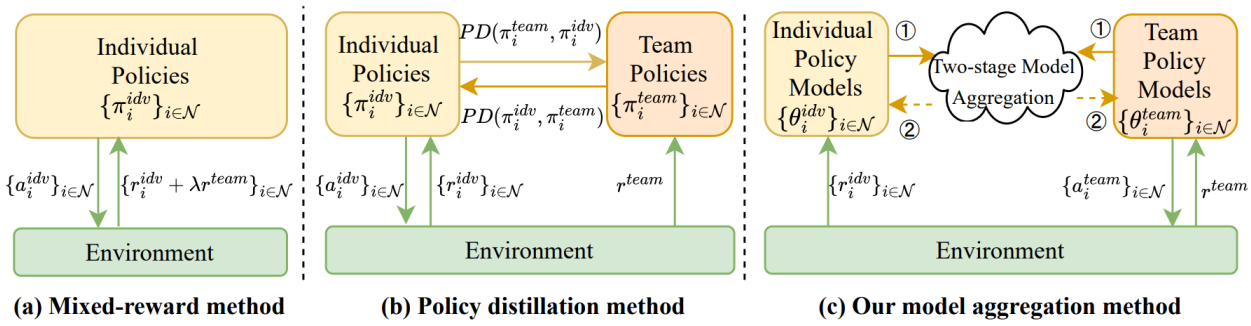
1 Introduction

Multi-agent reinforcement learning (MARL) has succeeded in fully cooperative scenarios [Liu *et al.*, 2025; Li *et al.*, 2022]. However, in real-world settings with partial cooperation, where agents must balance individual and team objectives, such as football players striving to win as a team while showcasing individual performance, MARL cooperation can be undermined by conflicts between these objectives.

Theoretically, using team rewards exclusively to train MARL agents [Foerster *et al.*, 2018; Rashid *et al.*, 2020] can enable cooperation in Partially Cooperative Markov Games (PCMG). However, this fully cooperative setting faces endogenous challenges in credit assignment and exploration, especially when team rewards are sparse [Wang *et al.*, 2022; Xu *et al.*, 2023; Li *et al.*, 2024], making it nearly infeasible to train a cooperative policy. To mitigate these issues, the effective use of individual rewards to guide the training of cooperative policies emerges as a crucial and promising direction in partially cooperative settings. Early works [Durugkar *et al.*, 2020; Ma *et al.*, 2024; Chen *et al.*, 2022] have explored combining individual and team rewards to directly incentivize cooperation, defining the PCMG as an independent learning game (Fig. 1(a)). However, combining multiple rewards often necessitates a detailed understanding of the specific research scenario, limiting the generalizability of such methods. As a result, recent effort [Wang *et al.*, 2022] has sought to combine the strengths of individual and team policies through policy distillation (PD) [Rusu *et al.*, 2016; Lai *et al.*, 2020], as illustrated in Fig. 1(b), to facilitate training of team policies based on individual policies or vice versa. However, PD is data driven and suffers from incomplete knowledge transfer, as it is constrained by its inability to capture policy information outside the distribution of the distillation data, limiting the efficiency of policy mutual learning. These limitations highlight the need for more direct, data-independent methods.

Unlike PD, model aggregation (MA) [Bourdais and Owhadi, 2025; Li *et al.*, 2025], offers a direct and effective approach to integrating multiple policies, and has gained widespread application in fields such as federated learning [Wang *et al.*, 2020; Lee *et al.*, 2023] and ensemble learning [Chan and van der Schaar, 2022; Kotary *et al.*, 2023]. In reinforcement learning (RL), a notable example is A3C [Mnih *et al.*, 2016], which trains multiple policy models in parallel and uses MA to improve training efficiency and policy generalization. However, MA still suffers from instability, mainly due to the volatility of individual policy models as aggregation objects and the discontinuous parameter updates during aggregation, both of which can cause significant performance fluctuations or even severe degradation.

*Corresponding author.



(a) Mixed-reward method **(b) Policy distillation method** **(c) Our model aggregation method**

Figure 1: (a) The policy of each agent is trained using a mixed reward, *i.e.*, a weighted sum of the individual and team rewards. (b) Individual and team policies are trained through individual and team rewards, respectively, and then aligned using policy distillation (PD). (c-ours) Align the team policies with the assistance of individual policies by model aggregation.

To this end, this paper introduces **T**wo-**S**tage **M**odel **A**ggregation (TSMA), as depicted in Fig. 1(c), a multi-policy MA framework for MARL. TSMA leverages individual policies to guide optimal team policy learning in partially cooperative scenarios. It achieves objective consistency through model aggregation, avoids the data dependency of knowledge distillation, and improves learning efficiency by integrating intermediate-layer knowledge. TSMA consists of two adaptive aggregation stages. In the first stage, each agent aggregates two individual policies trained on the same individual reward using loss-based adaptive weights to enhance expressiveness and stability. In the second stage, the aggregated individual policy is further combined with the team policy trained on team rewards, where the aggregation weights gradually decrease over training to enable smooth knowledge transfer and effective team learning. To enhance individual policy diversity, TSMA introduces parameter-level variation in individual policies after aggregation, providing a broader range of policy options for subsequent training. Additionally, TSMA applies L2 regularization across all policy models to mitigate instability caused by aggregation-induced abrupt parameter updates. The TSMA agent interacts with the environment via the team policy, making it a natural extension compatible with all Actor-Critic (AC)-based MARL algorithms. To demonstrate its generality, we integrate TSMA into two representative AC algorithms: MAPPO [Yu *et al.*, 2022] and MADDPG [Lowe *et al.*, 2017]. These methods are selected because they exemplify the on-policy and off-policy learning paradigms, respectively, providing complementary characteristics in terms of sample efficiency and training stability. We then evaluated TSMA-MAPPO and TSMA-MADDPG on PCMG benchmarks, including the Multi-Agent Particle Environment (MPE) [Mordatch and Abbeel, 2018] and the Pursuit scenario of the Stanford Intelligent Systems Laboratory (SISL) [Gupta *et al.*, 2017] at PettingZoo. The results show that TSMA-MAPPO and TSMA-MADDPG outperform the state-of-the-art methods in cooperative performance within PCMG while ensuring stable aggregation.

The contributions of this paper are summarized as follows: (1) We introduce TSMA, a multipolicy MA framework for MARL, designed to address cooperation challenges in partially cooperative scenarios by guiding team policy learning through individual policies. (2) TSMA is a versatile aggregation framework compatible with all AC-based MARL algorithms. We demonstrate its effectiveness by integrating it into both on-policy (MAPPO) and off-policy (MADDPG) based algorithms. (3)We provide a theoretical analysis of the convergence of TSMA, and through extensive experiments in partially cooperative scenarios, confirm its higher cooperative efficiency over state-of-the-art baselines.

The paper is structured as follows. Sect. 2 reviews related work. Sect. 3 introduces the PCMG model. Sect. 4 details the TSMA framework and algorithm with qualitative analysis. Sect. 5 presents experimental results and analysis. Finally, Sect. 6 provides the conclusion and future work.

2 Related Work

Mixed-reward cooperation methods. The mixed-reward method incorporates individual rewards alongside team rewards, enabling it to train policies using a combination of both in PCMG. For instance, Durugkar [Durugkar *et al.*, 2020] proposed a weighted combination of individual rewards and team rewards to promote effective learning of cooperative policies. Some works directly incentivize cooperative behavior through tailored rewards. ReLara [Ma *et al.*, 2024] introduces an auxiliary reward agent to automatically learn dense reward functions, allowing the principal policy agent to benefit from auxiliary and sparse team rewards. Similarly, DCIR [Lin *et al.*, 2023] rewards or penalizes agents based on their behavioral consistency with others and employs a dynamic scale network to dynamically adjust reward strength, allowing agents to flexibly modify their level of cooperation based on environmental changes. Mixing multiple rewards often relies heavily on expert experience to fine-tune the weighting coefficients of multiple rewards. Additionally, dimensional inconsistencies among rewards can complicate direct weighting. Furthermore, the independent reinforcement learning-based training paradigm used in these methods may struggle with the non-stationarity problem.

Policy mutual learning methods. Mutual learning methods for policies focus primarily on designing loss functions that ensure consistency between individual and team goals, quantifying policy differences, and utilizing these to guide policy optimization or achieve goal consistency between policies through shared loss functions. For example, IRAT [Wang

et al., 2022] employs knowledge distillation to efficiently align individual and team policies by applying differentiated constraints. Similarly, TRP [Wang *et al.*, 2024] predicts truncated returns of historical trajectories and minimizes the prediction error between current and historical policies, thus achieving consistency in the value function between different policies. D3C [Kwon *et al.*, 2023], on the other hand, adjusts each agent's loss function by learning a set of shared weights that combine individual losses into a modified objective, encouraging alignment with the global objective. Although data-driven methods effectively ensure that cooperative policies remain consistent with global goals, they are inherently limited in extracting policy information beyond the data distribution. This limitation restricts their ability to achieve more efficient policy mutual learning.

Model aggregation methods. Model aggregation, a core technique in federated learning [Wang *et al.*, 2020; Karimireddy *et al.*, 2020], combines multiple client models trained on their respective local datasets to produce a more accurate and generalized global model while preserving data privacy and security. This technique has shown effectiveness in RL, *e.g.*, A3C [Mnih *et al.*, 2016], which improves learning efficiency by training multiple agents in parallel environments and aggregating their models. Beyond efficiency, MA is crucial in achieving objective consistency. AgA [Li *et al.*, 2025] aligns individual and team objectives by performing a weighted aggregation of gradients from different policies to generate a global gradient. Similarly, SPAHM [Yurochkin *et al.*, 2019] maps local model parameters to a unified parameter space through parameter matching and leverages MA to minimize parameter distribution differences, ensuring stability during learning. While some of these methods address the stability issues in aggregation, they often suffer from high training complexity. Additionally, they fail to account for the prioritization of multiple models, which is critical in PCMG scenarios, as policies trained for different objectives may conflict. In this work, we focus on leveraging individual policies to guide team policy learning while maintaining stability throughout the aggregation process.

3 Preliminaries

The Partially Cooperative Markov Game (PCMG) can be formalized as a seven-tuple: $< \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{O}, \gamma, \mathcal{R} >$, where $\mathcal{N} = \{1, ..., n\}$ is the set of agents, $\mathcal{S}$ is the global state space of the environment and $\mathcal{A}$ is the action space for each agent. The state transition function $\mathcal{P}$ describes the probability distribution of the environment changing from state s to next state s' given the joint action $\boldsymbol{a}_t = (a_{1,t}, ..., a_{n,t}), \forall a_i \in \mathcal{A}$. $\mathcal{O}$ represents the partially observable state space and γ is the discount factor. $\mathcal{R}$ is the reward function that comprises each agent's individual reward and the shared team reward based on the current state and the joint action. At the time step t, all agents execute a joint action $\boldsymbol{a}_t$. The environment provides each agent with two rewards for $\boldsymbol{a}_t$, *i.e.*, individual reward $r_{i,t}^{idv}$ and shared team reward r_t^{team}.

In PCMG, each agent has two objectives, *i.e.*, individual and team objectives. The individual objective is to maximize their discounted cumulative individual return, $J_i^{idv}(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_{i,t}^{idv}]$. All agents cooperate to learn a joint policy simultaneously, to maximize the cumulative discount return for the entire team, $J_i^{team}(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t^{team}\right]$. In PCMG, agents pursue individual objectives while also cooperating with others to achieve team objectives. However, discrepancies in reward functions among agents create only partial alignment of individual rewards, often leading to incomplete cooperation. This misalignment can cause the team objectives of the agents to converge to the local optimum rather than the global optimum. In this paper, we address the challenge of balancing the conflict between individual and team objectives in PCMG, aiming to maximize team benefit.

4 Method

To address the challenges of PCMG, we propose **T**wo-**S**tage **M**odel **A**ggregation (TSMA), a multi-policy MA framework for MARL that uses individual policies to guide the learning of optimal team policy. Fig. 2 shows the learning process of a TSMA agent, including data-based training and a two-stage model aggregation. Within this framework, each agent is equipped with a team actor for interaction with the environment and two individual actors to guide the team policy. Individual and team critics separately evaluate individual and team returns, guiding the learning of the actors. TSMA, as a universal framework, seamlessly integrates with all standard AC algorithms. In this paper, we implement TSMA under the CTDE architecture by integrating it with two representative MARL backbones: the on-policy MAPPO and the off-policy MADDPG, forming TSMA-MAPPO and TSMA-MADDPG respectively. This section first details the TSMA aggregation process, followed by the training details of TSMA-MARL and the convergence analysis of the TSMA.

4.1 TSMA

In this section, we detail TSMA, which focuses on enabling individual policies to promote effective learning of team policies through MA. The core challenge is addressing MA's instability, which arises from two main sources: the inherent volatility of individual policy models and the non-asymptotic nature of aggregation updates. Specifically, TSMA first performs a weighted aggregation of two individual policies (trained on the same individual reward). It then integrates the aggregated individual policy with the team policy (trained in team reward) using a structure-level MA approach, bypassing data dependency in conventional policy distillation. To further enhance individual policy diversity, TSMA introduces parameter-level variation. Additionally, TSMA incorporates L2 regularization to guide all policy parameters toward a consistent norm space, mitigating the instability caused by the non-asymptotic nature of the aggregation updates. Note that reference [Dohare *et al.*, 2024] provides precedents for the use of variation injection and L2 regularization, demonstrating their theoretical foundation and effectiveness.

Aggregation of individual policies: Model aggregation can efficiently transfer knowledge from individual to team policies, but the instability of individual policies can lead to fluctuations or degradation in team policy performance. This issue arises because these individual policies, as foundational

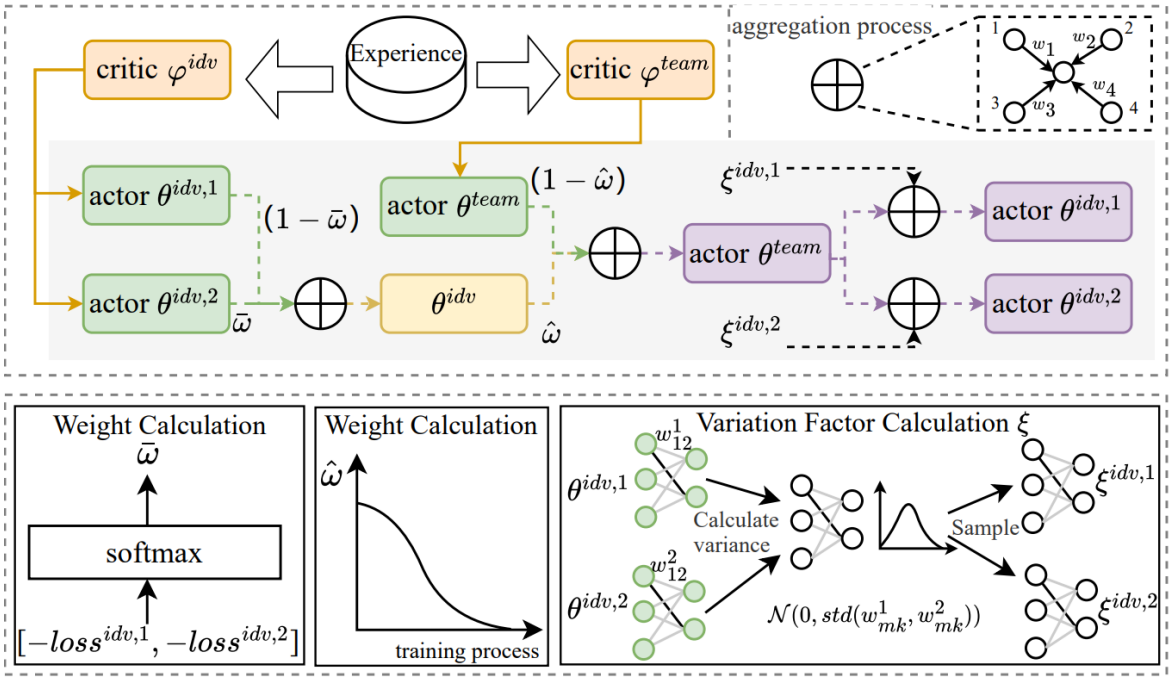


Figure 2: The upper panel illustrates the training and aggregation process of a TSMA agent, which includes two individual actors and one team actor. After the data-driven training phase, TSMA's two-stage aggregation first weights and combines individual policies, then performs an adaptive merge between the resulting individual and team policies with injected variation. The lower panel shows how the aggregation weights and variation factors are derived in TSMA.

guidance signals, directly influence team policy learning outcomes. To mitigate this problem, we train two individual policies for each agent and use their aggregated results to guide the team policy, enhancing the expressiveness and stability of individual policies. Furthermore, to reduce the interference that noise from low-quality policies might cause in the aggregation results, we employ a weighted aggregation that dynamically assigns weights based on the loss of each individual policy, highlighting the contributions of high-quality policies and suppressing the interference from low-quality ones. This 'majority voting, retaining the best' ensemble strategy effectively improves the stability of policy aggregation.

$$\theta^{idv} = \bar{w} \cdot \theta^{idv,1} + (1 - \bar{w}) \cdot \theta^{idv,2}, \tag{1}$$

where $\bar{w} = \exp(-L_1)/(\exp(-L_1) + \exp(-L_2))$ signifies the aggregated weight between individual policies and L_j denotes the average loss value of the individual policy $\theta^{idv,j}$ based on data-driven training.

Aggregation of individual and team policies: The core of this study is to integrate individual and team policies for each agent to help learn the more challenging team policy, ultimately improving the efficiency of multi-agent cooperation in PCMG. The approach of using individual policies to guide the learning of team policies has been proven to be effective in IRAT [Wang *et al.*, 2022], but the policy distillation method used in IRAT is limited by data distribution and struggles to achieve comprehensive policy transfer. To overcome this limitation, we propose replacing the data-driven distillation method with model aggregation. Model aggregation not only offers higher efficiency, but also bypasses the dependence on data. Integrates individual and team policies at the structural level, thus supporting the learning of more complex cooperative policies. To further improve the stability of the aggregation process, we enable individual and team policies to leverage their respective strengths at different training stages. Specifically, we introduce a weight annealing function $\hat{w}^{(e)}$ from MAPTF [Yang *et al.*, 2021] to adjust the aggregation ratio between individual and team policies, ensuring that the importance of the team policy progressively increases throughout the training process. This design retains the rich exploratory signals provided by individual policies during the early stages of training, while also ensuring that team policies can effectively lead to cooperative behaviors in later stages. The process of aggregating individual and team policies can be formalized as:

$$\theta^{team} = \hat{w}^{(e)} \cdot \theta^{idv} + (1 - \hat{w}^{(e)}) \cdot \theta^{team}, \tag{2}$$

Variation injection: Although different individual policies can effectively help to learn team policy models and help escape local optima, applying the aggregated individual and team policy model θ^{team} directly to all individual policies can lead to homogenization. To prevent this, we have introduced a variation mechanism to ensure the diversity of individual policies. Specifically, the aggregated model is updated through two paths: one directly updates the team policy, and the other updates the individual policies after variation injection. To enhance policy diversity while maintaining model stability, the TSMA performs fine-grained variation control

at the parameter level. For individual policy models with the same structure, TSMA adjusts the variation magnitude based on the variance of corresponding parameters in each agent's individual policies. Parameters with higher variance, indicating greater behavioral differences between individual policies, will undergo larger variations to explore a wider solution space. In contrast, parameters with lower variance, which exhibit stability across different policies, will receive smaller variations to maintain model stability. Ultimately, the parameters of each individual policy model follow the distribution:

$$\{\theta^{idv,j}\}_{j\in\{1,2\}} \sim \mathcal{N}(\theta^{team}, \beta \cdot std(\theta^{idv,1}, \theta^{idv,2})). \quad (3)$$

L2 regularization: The non-asymptotic nature of aggregation updates means that significant discrepancies between the parameter distributions of the policies to be aggregated can cause the resulting parameters to deviate from the original distribution characteristics when directly applying aggregation. Such deviations may cause the parameters to fall into ineffective regions of the parameter space, ultimately causing a sharp decline in policy performance[Dohare *et al.*, 2024]. This phenomenon fundamentally arises from the lack of a unified metric in the parameter space, where different policies may reside in entirely different scales or distributional forms. To effectively address this issue, we apply L2 regularization to force all aggregated policy parameters to contract toward a common canonical space. This alleviates inconsistencies and ensures that the aggregated policy retains the core features of the original. We incorporate L2 regularization penalties into the loss functions of both individual and team policies.

4.2 Implement TSMA into MARL

As mentioned above, TSMA is a universal framework compatible with all standard AC algorithms. To comprehensively evaluate its compatibility across different learning paradigms, this paper selects MAPPO and MADDPG as backbones, which are representative on-policy and off-policy methods, respectively, producing the TSMA-MAPPO and TSMA-MADDPG algorithms. Within the CTDE framework, both TSMA-MAPPO and TSMA-MADDPG employ individual critics $\{\varphi_i^{idv}\}_{i\in\mathcal{N}}$ and a team critic φ^{team} to estimate individual and team objectives based on individual rewards $\{r_i^{idv}\}_{i\in\mathcal{N}}$ and the shared team reward r^{team}. In practice, we share the parameters of the individual critics across all agents to enhance training efficiency. Each agent is equipped with three actors, including two individual actors parameterized by $\{\theta_i^{idv,1}, \theta_i^{idv,2}\}$ and a team actor θ_i^{team}, where the team actor is responsible for interacting with the environment and sampling data. The individual and team actors' policies are evaluated by their corresponding critics. In TSMA-MADDPG, both critics and actors are additionally associated with their respective target networks to stabilize training.

The TSMA training process for MARL algorithms consists of two stages: the data-based training stage and the model aggregation stage, as described in Sect. 4.1. During the data-based training phase, the individual and team critics, together with the actors of each agent, are updated following the standard training paradigms of the underlying algorithms (*e.g.*, MAPPO or MADDPG), utilizing trajectory data collected through the interaction of the team policies with the environment. As detailed in Sect. 4.1, to ensure stability during the model aggregation stage, L2 regularization is applied to the loss functions of individual and team actors. This regularization effectively constrains the parameters of the aggregation model within a consistent distribution space. The actor's loss function for each agent is defined as follows:

$$L(\boldsymbol{\theta}_i) = L_{\text{MARL}}(\boldsymbol{\theta}_i) + \lambda \left\|\boldsymbol{\theta}_i\right\|_2^2, \quad (4)$$

where $\boldsymbol{\theta}_i \in \{\theta_i^{idv,1}, \theta_i^{idv,2}, \theta_i^{team}\}, i \in \mathcal{N}$, λ is the hyperparameter that controls regularization intensity, and $\|\theta\|_2^2$ is the L2 norm square of the parameter.

4.3 Convergence Analysis

This study adopts the FedAvg convergence analysis framework [Li *et al.*, 2020], as both methods share the fundamental mechanics of multiple local updates, then central aggregation.

Let $J(\theta)$ denote the objective function, and assume it is L-smooth. Under stochastic gradient descent with step size η_t, the expected improvement per iteration satisfies

$$\begin{aligned}\mathbb{E}[J(\theta_{t+1})] \leq \mathbb{E}[J(\theta_t)] - \eta_t \mathbb{E}\|\nabla J(\theta_t)\|^2 \\ + \frac{L\eta_t^2}{2}\mathbb{E}\|g_t - \nabla J(\theta_t)\|^2,\end{aligned} \quad (5)$$

where g_t is the stochastic gradient. The term $\mathbb{E}\|g_t - \nabla J(\theta_t)\|^2$ captures the variance due to client sampling and local drift. By invoking bounded variance assumptions and gradient deviation bounds from FedAvg, it can be shown that

$$\mathbb{E}\|g_t - \nabla J(\theta_t)\|^2 \leq C_1\sigma^2 + C_2\mathcal{O}(\eta_t^2). \quad (6)$$

Substituting this bound leads to

$$\mathbb{E}[J(\theta_{t+1})] \leq \mathbb{E}[J(\theta_t)] - \eta\mathbb{E}\|\nabla J(\theta_t)\|^2 + \eta^2 \cdot O(\eta^2). \quad (7)$$

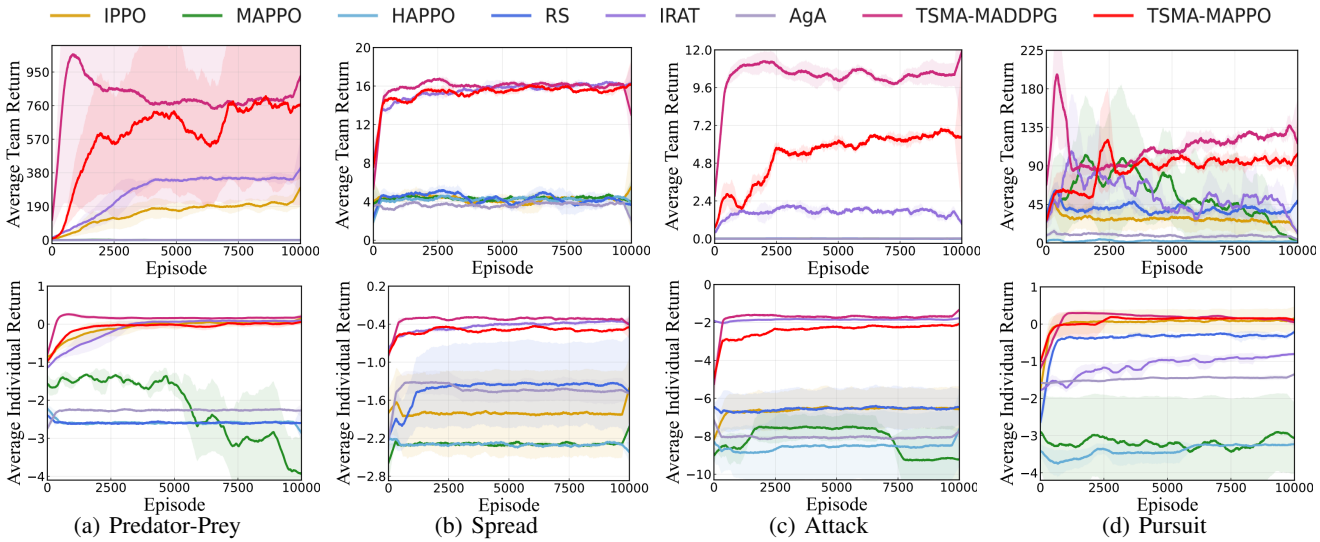
Since $J(\theta)$ is bounded below, $\mathbb{E}[J(\theta_0) - J(\theta_T)]$ is finite. Therefore, for a fixed learning rate η:

$$\lim_{T\to\infty} sup \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2 \leq O(\eta), \quad (8)$$

implying convergence to an $O(\eta)$-neighborhood of a stationary point. By decaying η appropriately, the algorithm can asymptotically approach an exact stationary point.

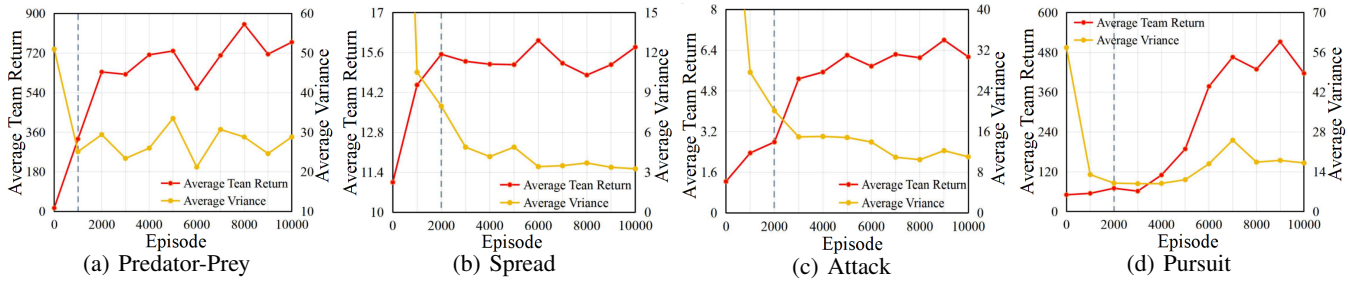
	Parameter	Value
TSMA	L2 weight w	0.01
	Variation weight coefficients β	0.0001
	Weight annealing coefficient μ	-0.0005
	Weight annealing start episode κ	3500
	Threshold T_{th}	0.5
MAPPO	GAE λ	0.95
	Actor/Critic learning rates $\alpha_{\boldsymbol{\theta}}, \alpha_{\boldsymbol{\varphi}}$	1e-3, 1e-3
	PPO clip ε	0.2
MADDPG	Actor/Critic learning rates $\alpha_{\boldsymbol{\theta}}, \alpha_{\boldsymbol{\varphi}}$	1e-4, 1e-3
	Buffer size	1e5
	Batch size	256

Table 1: Main hyperparameters of TSMA-MARL



(a) Predator-Prey (b) Spread (c) Attack (d) Pursuit

Figure 3: Learning curves of comparison methods in the four partial cooperative scenarios, plotted based on five training runs using different random seeds, with smoothing applied every 600 episodes.



(a) Predator-Prey (b) Spread (c) Attack (d) Pursuit

Figure 4: Analysis of the impact from variation injection degree on TSMA-MAPPO in four scenarios. The curves are plotted based on five training runs with different random seeds, where 200 episodes are averaged after every 800 episodes interval.

5 Experiment

TSMA is a general framework applicable to different MARL algorithms, as evidenced by its integration with both MAPPO and MADDPG in the main comparative experiments. For ablation and sensitivity studies, we use MAPPO as the representative backbone to focus on analyzing TSMA's internal mechanisms. In this section, we conduct experiments to answer the following questions: **RQ1**: Can TSMA-MARL enable efficient teamwork in partial cooperation scenarios? **RQ2**: How does the degree of variation injection affect the performance of TSMA-MAPPO? **RQ3**: How do different components (*i.e.*, L2 regularization, variation injection, multiple individual policies, and weight annealing function) affect TSMA? **RQ4**: How do individual policy counts and coefficient weighting of variation affect TSMA? **RQ5**: How does TSMA-MARL perform as the number of agents scales up?

5.1 Experiment Settings

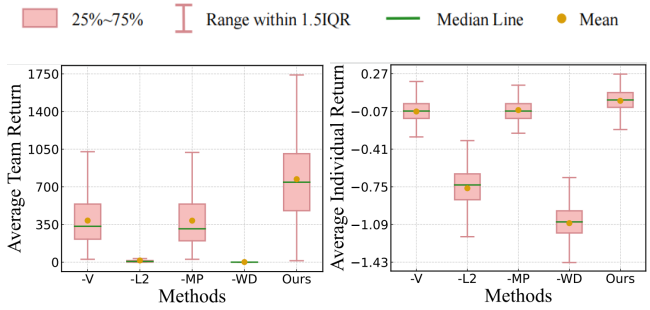
This section will introduce the experiment's scenario setting and comparison methods.

Scenario Setting: To evaluate the team cooperation capability of TSMA-MAPPO and TSMA-MADDPG in partially cooperative scenarios, we conduct experiments across four challenging environments: Predator-Prey, Spread, and Attack from the MPE, and the Pursuit task from the SISL suite. These environments feature a hybrid reward structure, requiring agents to balance individual gains (*e.g.*, distance-based rewards) with sparse, high-value team rewards that are only triggered upon successful team coordination.

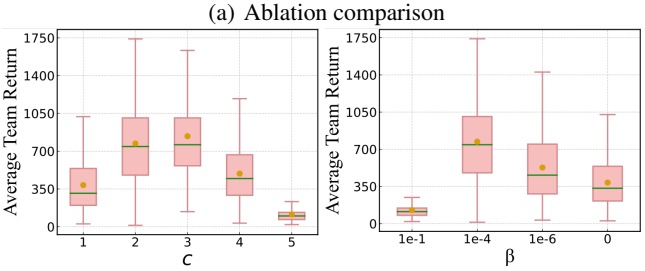
Compared Methods: We evaluated TSMA-MAPPO and TSMA-MADDPG against six RL baselines grouped by reward design: independent learning with individual rewards (IPPO [Schulman *et al.*, 2017]), cooperative learning with team rewards (MAPPO [Yu *et al.*, 2022], HAPPO [Kuba *et al.*, 2022]), and hybrid approaches combining both (RS [Durugkar *et al.*, 2020], IRAT [Wang *et al.*, 2022], AgA [Li *et al.*, 2025]). To fairly compare the algorithms, we use the same network architecture and hyperparameters (summarized in Tab. 1). Notably, $T_{th} = 0$ is optimal in the Pursuit scenario due to individual-team policy conflicts, and this value is used in all experiments except the comparative study.

5.2 Experiment Results

Comparison with SOTA methods (RQ1): Fig. 3 shows that TSMA-MAPPO and TSMA-MADDPG achieve the best cooperative performance across all four partially cooperative tasks, with TSMA-MADDPG exhibiting superior data efficiency in 'Predator-Prey' and 'Attack'. Performance swings



(a) Ablation comparison



(b) Hyperparameter (c and β) comparison

Figure 5: Performance comparisons of TSMA-MAPPO ablations and hyperparameter studies in Predator-Prey, averaged over the last 200 episodes across five seeds.

of TSMA-MAPPO in Predator–Prey, possibly related to variation, prompted further experiments in RQ2. Among baselines, IRAT yields suboptimal cooperation as data limitations restrict individual-team policy mutual learning. IPPO, RS, and AgA deliver strong individual performance but fail in team cooperation: IPPO is limited by individual rewards, RS suffers from fixed reward weighting that causes optimization interference, and AgA's fixed-ratio gradient aggregation lacks task-driven adaptation, hindering effective collective behavior despite strong individual performance. MAPPO and HAPPO fail to learn cooperation due to sparse team rewards.

Impact analysis of variation injection degree (RQ2): We hypothesize that TSMA-MAPPO's performance fluctuations in Predator-Prey are caused by variation injection. To verify this, we analyze the correlation between the degree of variation (measured by individual model parameter variance) and algorithm performance. Fig. 4 shows a clear positive correlation between performance fluctuations and parameter variation magnitude after the initial training phase (dashed line). This confirms variation's key role: insufficient variation limits diversity and traps models in suboptimal solutions, while moderate variation aids in escaping local optima.

Ablation study (RQ3): We conduct ablation studies on TSMA components in the Predator-Prey scenario (Fig. 5(a)). We analyze four variants of TSMA-MAPPO: removing L2 regularization (-L2), variation injection (-V), the multiple individual policies(-MP), or weight annealing (-WD). The results confirm each component's critical role. Both -V and -MP suffer from insufficient policy diversity, leading to suboptimal cooperation. The absence of L2 regularization causes parameter divergence and model collapse, while fixed weight aggregation in -WD impedes the dynamic knowledge transfer required across training, hindering team policy learning.

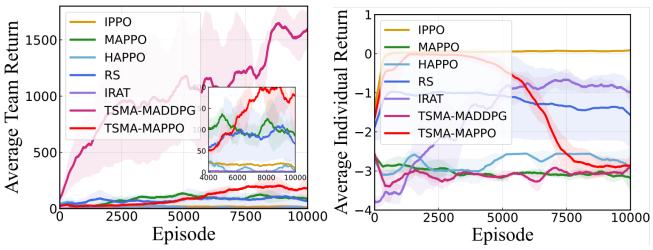


Figure 6: Learning curves of comparison methods in Pursuit-large, plotted based on five training runs using different random seeds, with smoothing applied every 600 episodes.

Hyperparameter analysis study (RQ4): We present a sensitivity analysis of two TSMA hyperparameters, the number of individual policies c and the variation intensity β, using the Predator-Prey scenario (Fig. 5(b)). The left panel reveals that maintaining two or three individual policies yields optimal performance. A single policy provides insufficient information for optimizing team policies, while too many policies degrade performance due to inefficiencies; thus, we set $c = 2$ as the default. The right panel indicates that the optimal variation weight is $\beta = 0.0001$, as lower values have negligible effect while higher values disrupt cooperation.

Scalability study (RQ5): To verify scalability, we test our methods on an extended Pursuit-large environment with 30 pursuers, 50 evaders, a 40x40 map, and a 100-step episode limit. As shown in Fig. 6, TSMA-MADDPG and TSMA-MAPPO maintain optimal performance at scale. TSMA-MADDPG's data efficiency becomes more advantageous with more evaders. Both methods achieve the best team cooperation despite lower individual rewards, showing robustness to individual-team objective conflicts. IRAT performs poorly due to limited exploration under short episodes, while AgA is omitted due to prohibitive computational cost at this scale.

6 Conclusion

To promote cooperation in partial cooperative multiagent scenarios, we introduce TSMA, a multipolicy model aggregation framework for MARL. TSMA aligns individual and team objectives, allowing individual policies to actively support team policy optimization and promoting efficient cooperation. As a versatile framework compatible with standard AC algorithms, we implement it with MAPPO and MADDPG to cover both on-policy and off-policy paradigms, thus developing TSMA-MAPPO and TSMA-MADDPG. Experimental results in various scenarios in the MPE and the SISL Pursuit scenario demonstrate that TSMA-MARL significantly enhances team cooperation in partial cooperative scenarios.

Despite notable progress in enhancing algorithm diversity via variation and multi-policy training, the method still suffers from training fluctuation and high computational cost, leaving room for further optimization. Future work will refine the variation mechanism to improve stability while preserving policy diversity. Another promising direction is to adaptively adjust the number of individual policy models, so as to efficiently utilize limited policies, lower training costs, and build a more effective MARL framework.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the National Natural Science Foundation of China with Grant No. 62376048.

References

[Bourdais and Owhadi, 2025] Theo Bourdais and Houman Owhadi. Minimal variance model aggregation: A principled, non-intrusive, and versatile integration of black box models. In *International Conference on Representation Learning*, volume 2025, pages 82217–82242, 2025.

[Chan and van der Schaar, 2022] Alex Chan and Mihaela van der Schaar. Synthetic model combination: An instance-wise approach to unsupervised ensemble learning. In *Advances in the 36th Neural Information Processing Systems*, pages 27797–27809, 2022.

[Chen *et al.*, 2022] Eric Chen, Zhang-Wei Hong, Joni Pajarinen, and Pulkit Agrawal. Redeeming intrinsic rewards via constrained optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 4996–5008, 2022.

[Dohare *et al.*, 2024] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

[Durugkar *et al.*, 2020] Ishan Durugkar, Elad Liebman, and Peter Stone. Balancing individual preferences and shared objectives in multiagent reinforcement learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2505–2511, 2020.

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

[Gupta *et al.*, 2017] Jayesh K. Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems*, volume 10642, pages 66–83, 2017.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5132–5143, 2020.

[Kotary *et al.*, 2023] James Kotary, Vincenzo Di Vito, and Ferdinando Fioretto. Differentiable model selection for ensemble learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1954–1962, 2023.

[Kuba *et al.*, 2022] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[Kwon *et al.*, 2023] Minae Kwon, John P Agapiou, Edgar A. Duéñez-Guzmán, Romuald Elie, Georgios Piliouras, Kalesha Bullard, and Ian Gemp. Auto-aligning multiagent incentives with global objectives. In *ICML Workshop on Localized Learning (LLW)*, 2023.

[Lai *et al.*, 2020] Kwei-Herng Lai, Daochen Zha, Yuening Li, and Xia Hu. Dual policy distillation. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3146–3152, 2020.

[Lee *et al.*, 2023] Sunwoo Lee, Tuo Zhang, and A. Salman Avestimehr. Layer-wise adaptive model aggregation for scalable federated learning. *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 8491–8499, 2023.

[Li *et al.*, 2020] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[Li *et al.*, 2022] Pengyi Li, Hongyao Tang, Tianpei Yang, Xiaotian Hao, Tong Sang, Yan Zheng, Jianye Hao, Matthew E. Taylor, Wenyuan Tao, and Zhen Wang. PMIC: Improving multi-agent reinforcement learning with progressive mutual information collaboration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12979–12997, 2022.

[Li *et al.*, 2024] Xinran Li, Zifan Liu, Shibo Chen, and Jun Zhang. Individual contributions as intrinsic exploration scaffolds for multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28387–28402, 2024.

[Li *et al.*, 2025] Yang Li, Wenhao Zhang, Jianhong Wang, Shao Zhang, Yali Du, Ying Wen, and Wei Pan. Aligning individual and collective objectives in multi-agent cooperation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 44735 – 44760, 2025.

[Lin *et al.*, 2023] Kunyang Lin, Yufeng Wang, Peihao Chen, Runhao Zeng, Siyuan Zhou, Mingkui Tan, and Chuang Gan. Dcir: Dynamic consistency intrinsic reward for multi-agent reinforcement learning. *arXiv preprint arXiv:2312.05783*, 2023.

[Liu *et al.*, 2025] Wanting Liu, Chengwei Zhang, Wanqing Fang, Kailing Zhou, Yihong Li, Furui Zhan, Qi Wang, Wanli Xue, and Rong Chen. Vehicle-level fairness-oriented constrained multi-agent reinforcement learning for adaptive traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 26(4):4878–4890, 2025.

[Lowe *et al.*, 2017] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pages 6379–6390, 2017.

[Ma *et al.*, 2024] Haozhe Ma, Kuankuan Sima, Thanh Vinh Vo, Di Fu, and Tze-Yun Leong. Reward shaping for reinforcement learning with an assistant reward agent. In *Proceedings of the 41st International Conference on Machine Learning*, pages 33925–33939, 2024.

[Mnih *et al.*, 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1928–1937, 2016.

[Mordatch and Abbeel, 2018] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 1495–1502, 2018.

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

[Rusu *et al.*, 2016] Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation, 2016.

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

[Wang *et al.*, 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 7611–7623, 2020.

[Wang *et al.*, 2022] Li Wang, Yupeng Zhang, Yujing Hu, Weixun Wang, Chongjie Zhang, Yang Gao, Jianye Hao, Tangjie Lv, and Changjie Fan. Individual reward assisted multi-agent reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 23417–23432, 2022.

[Wang *et al.*, 2024] Shuo Wang, Zhihao Wu, Xiaobo Hu, Jinwen Wang, Youfang Lin, and Kai Lv. What effects the generalization in visual reinforcement learning: Policy consistency with truncated return prediction. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, volume 38, pages 5590–5598, 2024.

[Xu *et al.*, 2023] Pei Xu, Junge Zhang, and Kaiqi Huang. Exploration via joint policy diversity for sparse-reward multi-agent tasks. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 326–334, 2023.

[Yang *et al.*, 2021] Tianpei Yang, Weixun Wang, Hongyao Tang, Jianye Hao, Zhaopeng Meng, Hangyu Mao, Dong Li, Wulong Liu, Chengwei Zhang, Yujing Hu, Yingfeng Chen, and Changjie Fan. An efficient transfer learning framework for multiagent reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 17037 – 17048, 2021.

[Yu *et al.*, 2022] 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. In *Proceedings of the 36th Advances in Neural Information Processing Systems*, volume 35, pages 24611–24624, 2022.

[Yurochkin *et al.*, 2019] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, and Nghia Hoang. Statistical model aggregation via parameter matching. In *Advances in the 33rd International Conference on Neural Information Processing Systems*, volume 32, pages 10954–10964, 2019.