Learning Pre-Trained Tacit Behavior for Efficient Multi-Agent Adversarial Coordination

Extended Abstract

Shiqing Yao
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
yaosq23@mails.tsinghua.edu.cn

Jiajun Chai
Institute of Automation, Chinese Academy of Sciences
Beijing, China
chaijiajun2020@ia.ac.cn

Haixin Yu
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
yuhx21@tsinghua.org.cn

Yongzhe Chang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
changyongzhe@sz.tsinghua.edu.cn

Yuanheng Zhu
Institute of Automation, Chinese Academy of Sciences
Beijing, China
yuanheng.zhu@ia.ac.cn

Xueqian Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wang.xq@sz.tsinghua.edu.cn

ABSTRACT

In addressing the multi-agent adversarial coordination problem, existing multi-agent reinforcement learning algorithms primarily rely on team-based rewards to guide agent policy updates, often neglecting the utilization of inter-agent relationships, which limits their performance. Drawing inspiration from human tactics, we introduce the concept of tacit behavior to improve the efficiency of multi-agent reinforcement learning by refining the learning process. This paper presents a novel two-phase framework for learning Pre-trained Tacit Behavior for efficient multi-agent adversarial Coordination (PTBC), comprising a tacit pre-training phase and a centralized adversarial training phase. We demonstrate the superiority of our method through comparisons with several algorithms, each of which possesses distinct strengths.

KEYWORDS

Reinforcement Learning; Multi-Agent Adversarial Coordination; Tacit Pre-Training

ACM Reference Format:
Shiqing Yao, Jiajun Chai, Haixin Yu, Yongzhe Chang, Yuanheng Zhu, and Xueqian Wang. 2025. Learning Pre-Trained Tacit Behavior for Efficient Multi-Agent Adversarial Coordination: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

The multi-agent adversarial coordination problem presents significant challenges, including complex behaviors, a non-stationary environment, and imperfect communication[10, 13]. Multi-agent reinforcement learning (MARL)[1, 6] presents a promising solution by uncovering latent cooperative abilities among agents. Several studies have explored integrating real-world problem-solving insights into the MARL framework to improve policy optimization[14, 16, 17]. However, the expansive policy space in multi-agent settings leads to inefficient training when relying solely on team-based rewards for policy updates[7, 12]. MARL requires the incorporation of additional information to enhance performance[4, 15].

To address the above challenges, we introduce the spatial relationships among agents and their variant trend, as these serve as intuitive and essential embodiment of strategy in most multi-agent adversarial coordination problems. Based on this insight, we expect our agents collectively grasp cognition during the adversarial coordination tasks, which we define as **"tacit"**. In this paper, cognition is represented by guiding formulas that integrate the spatial relationships among our agents and their variant trends. Moreover, we propose that agents rely on the tacit to guide each agent in generating individual actions, enabling joint actions to form advantageous spatial relationships that contribute to team task. We define each agent's corresponding behavior as **"tacit behavior"** and an agent's spatial position relative to the global state as its **"pattern"**. In order to enable multi-agent tacit behavior and refine coordination learning, we propose a novel framework called **P**re-trained **T**acit **B**ehavior for efficient multi-agent **C**oordination **(PTBC)**.

2 METHODS

The PTBC framework emphasizes the formation of advantageous spatial relationships to assist agents defeat their enemies[3, 9, 11]. The PTBC enables the multi-agent system to learn these behaviors, facilitating the local aggregation of more agents than the opponents. As shown in Figure 1, the framework utilizes a two-phase training process: tacit pre-training based on decentralized learning[18], followed by adversarial training using the centralized training and decentralized execution (CTDE) paradigm[2, 5, 8]. The tacit pre-training phase includes two mechanisms: the pattern mechanism and the tacit mechanism.

2.1 Pattern Mechanism

Pattern mechanism is divided into two parts: pattern classification and pattern membership calculation. Pattern classification part

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

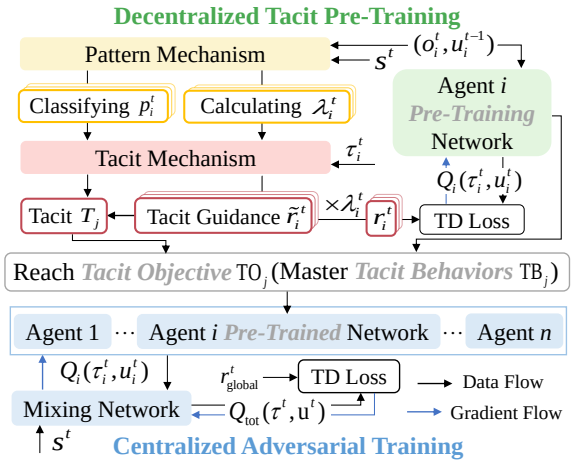

Figure 1: The overview of the PTBC framework.

takes the global state s^t and local observation o_i^t as inputs to classify the agent's **"pattern"**, which is defined as the relative position of the agent with respect to allied agents at a given time. Each pattern corresponds to a specific tacit behavior, but agents within the same pattern require different levels of guidance to learn the corresponding tacit behavior. To capture these variations, we define **"pattern membership"**, which is calculated based on s^t and o_i^t.

At each timestep, we classify the pattern to which agent i belongs, denoted as p_i. These patterns are classified along two dimensions: the spatial relationships among all agents in the global state and the spatial relationship between agent i and other agents within its local observation. The pattern membership of agent i, denoted as λ_i, is calculated based on this classification and quantifies its spatial relationship with other agents.

Pattern membership quantifies spatial relationships and serves as a parameter to regulate reward magnitude, introducing the tacit mechanism for dynamic adjustment of tacit guidance, thereby enabling more targeted coordination.

2.2 Tacit Mechanism

The tacit mechanism takes each agent's assigned pattern, corresponding membership, and local action-observation history as inputs to construct rewards that guide the agent in learning appropriate tacit behavior. Given the significant differences in observation spaces among agents in distinct patterns, applying uniform tacit behavior guidance leads to inefficient learning and a lack of specificity. To address this, the tacit mechanism designs targeted reward functions tailored to each pattern, offering more effective learning guidance based on agents' varying perceptual and surrounding situations. Tacit guidance is designed from two perspectives: the agent's ability to perceive allied agents and the relationships among agents within the global state. The tacit mechanism enables agent to efficiently master tacit behaviors associated with distinct patterns.

2.3 Overall Training Framework

In the PTBC framework, the effectiveness of tacit pre-training directly influences the achievement of spatial positioning advantages in adversarial learning, resulting in enhanced learning performance. The pre-training phase utilizes both the pattern and tacit mechanisms to generate tacit rewards that guide agents in learning tacit behavior strategies. Meanwhile, the tacit mechanism evaluates the level of tacit mastery to determine when the pre-training objective is achieved, facilitating the transition to centralized adversarial training. Once the tacit pre-trained network is established, global rewards are applied for cooperative adversarial training, building upon the pre-trained network.

3 EXPERIMENTS

We apply our method and five well-established algorithms as baselines to the StarCraft Multi Agent Challenge (SMAC) benchmark. Additionally, we modify the maps by randomizing the initial positions of both teams and introducing scenarios where agents cannot initially perceive each other.

Table 1 and Table 2 present SMAC results on three challenging SMAC maps, featuring both homogeneous and heterogeneous teams in asymmetric battles. Table 1 shows the mean and standard deviation (std) of win rate differences among algorithms at the same training step. Table 2 presents the mean and standard deviation of timesteps required for each algorithm to achieve the target win rate.

Table 1: Mean and std of the winning rates in SMAC

Maps	Steps	PTBC	GoMARL	HAVEN	RODE	QPLEX	QMIX
3s_vs_5z	5M	**96(1)**	89(2)	94(1)	65(9)	60(9)	77(6)
6h_vs_8z	10M	**76(2)**	63(4)	38(5)	72(3)	66(2)	58(4)
MMM2	10M	**72(3)**	70(4)	61(3)	67(6)	67(4)	62(4)

Table 2: Mean and std of the timesteps required to achieve the target win rate in SMAC

Maps (win rate)	PTBC	GoMARL	HAVEN	RODE	QPLEX	QMIX
3s_vs_5z (1.0)	**7.43±0.05**	8.47±0.05	7.85±0.11	8.64±0.05	8.38±0.05	7.96±0.06
6h_vs_8z (0.65)	**7.50±0.05**	8.19±0.11	No step reach	8.18±0.05	8.28±0.07	9.09±0.06
MMM2 (0.7)	**8.15±0.09**	9.24±0.10	8.39±0.03	8.36±0.18	8.46±0.03	8.90±0.13

The results demonstrate that PTBC outperforms other methods, achieving a balance between final win rates (Table 1) and learning efficiency across multiple maps (Table 2). In contrast, baseline methods achieve satisfactory results only on tasks where they excel.

4 CONCLUSION

This study introduces the PTBC framework, which integrates advantageous spatial positioning through tacit coordination. By leveraging pre-trained tacit behaviors, the PTBC enables agents to develop efficient strategies in multi-agent adversarial coordination tasks. We compare PTBC with several algorithms, including group-based, role-based, and hierarchical approaches, and the results demonstrate that PTBC achieves superior performances.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.62103225), Natural Science Foundation of Shenzhen (No.JCYJ20230807111604008), Natural Science Foundation of Guangdong Province (No.2024A1515010003) and National Key Research and Development Program (No.2022YFB4701402).

REFERENCES

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

[2] Landon Kraemer and Bikramjit Banerjee. 2016. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing* 190 (2016), 82–94.

[3] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 4501–4510.

[4] Dapeng Li, Zhiwei Xu, Bin Zhang, Guangchong Zhou, Zeren Zhang, and Guoliang Fan. 2024. From explicit communication to tacit cooperation: A novel paradigm for cooperative MARL. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2360–2362.

[5] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[6] Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence* 53, 11 (2023), 13677–13722.

[7] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4295–4304.

[8] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 5887–5896.

[9] Yan Song, He Jiang, Zheng Tian, Haifeng Zhang, Yingping Zhang, Jiangcheng Zhu, Zonghong Dai, Weinan Zhang, and Jun Wang. 2024. An empirical study on google research football multi-agent scenarios. *Machine Intelligence Research* 21, 3 (2024), 549–570.

[10] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[11] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *nature* 575, 7782 (2019), 350–354.

[12] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*.

[13] Lexing Wang, Tenghai Qiu, Zhiqiang Pu, and Jianqiang Yi. 2024. A cooperation and decision-making framework in dynamic confrontation for multi-agent systems. *Computers and Electrical Engineering* 118 (2024), 109300.

[14] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. 2021. RODE: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*.

[15] Yanan Wang, Tong Xu, Xin Niu, Chang Tan, Enhong Chen, and Hui Xiong. 2020. STMARL: A spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control. *IEEE Transactions on Mobile Computing* 21, 6 (2020), 2228–2242.

[16] Jiachen Yang, Igor Borovikov, and Hongyuan Zha. 2020. Hierarchical cooperative multi-agent reinforcement learning with skill discovery. In *International Conference on Autonomous Agents and Multi-Agent Systems*.

[17] Yifan Zang, Jinmin He, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. 2024. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

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