面向开放环境的协作多智能体强化学习

袁雷

南京大学
机器学习与数据挖掘研究所(LAMDA)

协作多智能体强化学习进展

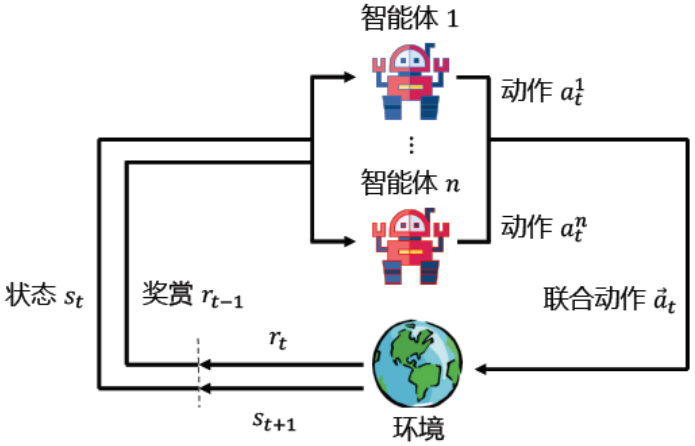


Fig. 5 MARL示意图。

Shoham Y, Powers R, Grenager T. Multi-agent reinforcement learning: a critical survey[R]. Technical report, Stanford University, 2003.

Panait L, Luke S. Cooperative multi-agent learning: The state of the art[J]. Autonomous agents and multi-agent systems, 2005, 11: 387-434.

Shoham Y, Powers R, Grenager T. If multi-agent learning is the answer, what is the question?[J]. Artificial intelligence, 2007, 171(7): 365-377.

Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *IEEE Access, 6:28573–28593, 2018*

Hernandez-Leal P, Kartal B, Taylor M E. A survey and critique of multiagent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *Preprint arXiv:2011.00583, 2020.*

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

Doran J E, Franklin S, Jennings N R, et al. On cooperation in multi-agent systems[J]. The Knowledge Engineering Review, 1997, 12(3): 309-314.

Du W, Ding S. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications[J]. Artificial Intelligence Review, 2021, 54: 3215-3238.

Stefano V. Albrecht, Filippos Christianos, and Lukas SchÅNafer. Multi-Agent Reinforcement Learning: Foundations and Modern Approaches. *MIT Press, 2023.*

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

Zhou Z, Liu G, Tang Y. Multi-Agent Reinforcement Learning: Methods, Applications, Visionary Prospects, and Challenges[J]. arXiv preprint arXiv:2305.10091, 2023.

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

协作多智能体强化学习进展

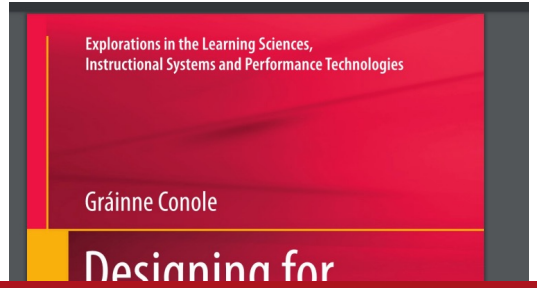
Table 1 封闭环境下的协作多智能体强化学习研究内容。

研究方向	核心内容	代表算法	应用与取得成果
算法框架设计	利用多智能体协作理论或设计神经网络提升协作能力	VDN [30], QMIX [31], QPLEX [85], MADDPG [28], MAPPO [29], HAPPO [92], DOP [86], MAT [32]	在多种典型任务场景如SMAC [33], GRF [32]等环境上取得不错协作效果，展现出巨大潜力
协作探索	设计机制以高效探索环境获得最优队友协作模式，与此同时收集高效的经验轨迹以训练策略找到最优解	MAVEN [93], EITI(EDTI) [94], EMC [95], CMAE [96], Uneven [97], SMMAE [98]	在复杂任务场景下显著提升协作效果，在稀疏奖赏等场景下解决协作能力过低的问题
多智能体通信	设计方法促进智能体间的信息共享，解决局部可观测等问题，专注于**何时**与**那个（些）**队友交换**何种**信息	DIAL [99], VBC [100], I2C [101], TarMAC [102], MAIC [103], MASIA [104]	在局部可观测任务场景或需要强协作场景，可以有效提升协作能力
智能体建模	开发技术以赋能智能体以推断环境中其他智能体（实体）的动作、目标和信念的能力，借此促进系统的协作能力	ToMnet [105], OMDDPG [106], LIAM [107], LILI [108], MBOM [109], MACC [110]	可以显著改善由于其他智能体的存在带来的环境非稳态问题，可以交互强与需要强协作的场景下改善协作性能
策略模仿	智能体通过从给定的轨迹或者示例样本中学习协作策略以完成任务	MAGAIL [111], MA-AIRL [112], CoDAIL [113], DM^2 [114]	实现仅从示例数据进行策略学习的目标
基于模型类方法	从数据中学习世界模型，多智能体在所学的模型中学习数据以避免与环境直接交互，提升样本效率	MAMBPO [115], AORPO [116], MBVD [117], MAMBA [118], VDFD [119]	借助成功的模型学习方法或开发针对多智能体开发方法，可以显著提升系统的样本效率与复杂场景下的协作效能
动作分层学习	将复杂问题分解成多个子问题，分别解决子问题进而实现对原始复杂问题求解	FHM [120], HSD [121], RODE [122], ALMA [123], HAVEN [124], ODIS [34]	在多类任务场景下显著提升多智能体系统的协作效率
拓扑结构学习	建模多智能体间的交互关系，利用如协作图及其他方式刻画智能体间的交互关系	CG [125], DCG [126], DICG [127], MAGIC [128], ATOC [129], CASEC [130]	显（隐）式刻画智能体间的关系，在复杂场景下可以降低系统联合动作空间，提升协作性能
其他方面	包括如可解释、理论分析、社会困境、大规模、延时奖励等方面开展研究	Na2q [131], ACE [132], CM3 [133], MAHHQN [134], 文献 [135, 136, 137, 138]	从其他方面完备协作多智能体强化学习研究

Table 2 典型多智能体测试环境介绍。

环境名称	是否异质	场景类型	观测空间	动作空间	典型数量	是否通信	问题领域
Matrix Games [81] (1998)	是	混合	离散	离散	2	否	矩阵博弈
MPE [28] (2017)	是	混合	连续	离散	2-6	允许	粒子游戏
MACO [130] (2022)	否	混合	离散	离散	5-15	允许	粒子游戏
GoBigger [238] (2022)	否	混合	连续	连续或离散	4-24	否	粒子游戏
MAgent [232] (2018)	是	混合	连续+图像	离散	1000	否	大规模粒子对抗
MARLÖ [239] (2018)	否	混合	连续+图像	离散	2-8	否	对抗游戏
DCA [227] (2022)	否	混合	连续	离散	100-300	否	对抗游戏
Pommerman [240] (2018)	否	混合	离散	离散	4	是	炸弹人游戏
SMAC [33] (2019)	是	协作	连续	离散	2-27	否	星际争霸游戏
Hanabi [241] (2019)	否	协作	离散	离散	2-5	是	卡牌游戏
Overcooked [242] (2019)	是	协作	离散	离散	2	否	烹饪游戏
Neural MMO [243] (2019)	否	混合	连续	离散	1-1024	否	多人游戏
Hide-and-Seek [244] (2019)	是	混合	连续	离散	2-6	否	捉迷藏游戏
LBF [235] (2020)	否	协作	离散	离散	2-4	否	食物搜寻游戏
Hallway [156] (2020)	否	协作	离散	离散	2	是	通信走廊游戏
GRF [231] (2019)	否	协作	连续	离散	1-3	否	足球对抗
Fever Basketball [245] (2020)	是	混合	连续	离散	2-6	否	篮球对抗
SUMO [246] (2010)	否	混合	连续	离散	2-6	否	交通控制
Traffic Junction[151] (2016)	否	协作	离散	离散	2-10	是	通信交通调度
CityFlow [247] (2019)	否	协作	连续	离散	1-50+	否	交通控制
MAPF [248] (2019)	是	协作	离散	离散	2-118	否	路径导航
Flatland [249] (2020)	否	协作	连续	离散	>100	否	列车调度
SMARTS [250] (2020)	是	混合	连续+图像	连续或离散	3-5	否	无人驾驶
MetaDrive [251] (2021)	否	混合	连续	连续	20-40	否	无人驾驶
MATE [252] (2022)	是	混合	连续	连续或离散	2-100+	是	目标追踪
MARBLER [253] (2023)	是	混合	连续	离散	4-6	允许	交通控制
RWARE [235] (2020)	否	协作	离散	离散	2-4	否	仓库物流
MABIM [254] (2023)	否	混合	连续	连续或离散	500-2000	否	库存管理
MaMo [27] (2022)	是	协作	连续	连续	2-4	否	参数调优
Active Voltage Control [26] (2021)	是	协作	连续	连续	6-38	否	电力控制
MAMuJoCo [89] (2020)	是	协作	连续	连续	2-6	否	机器人控制
Light Aircraft Game [255] (2022)	否	混合	连续	离散	1-2	否	智能空战
MaCa [256] (2020)	是	混合	图像	离散	2	否	智能空战
Gathering [226] (2020)	否	协作	图像	离散	2	否	社会困境
Harvest [257] (2017)	否	混合	图像	离散	3-6	否	社会困境
Safe MAMuJoCo [258] (2023)	是	协作	连续	连续	2-8	否	安全多智能体
Safe MARobosuite [258] (2023)	是	协作	连续	连续	2-8	否	安全多智能体
Safe MAIG [258] (2023)	是	协作	连续	连续	2-12	否	安全多智能体
OG-MARL [233] (2023)	是	混合	连续	连续或离散	2-27	否	离线数据集
MASIA [104] (2023)	是	协作	离散或连续	离散	2-11	否	离线通信数据集

开放环境下的机器学习



Recent Advances in Open Set Recognition: A Survey.

Geng C, Hu

Author infor

IEEE Transacti

https://doi.org

Share this arti

nature

Explore content | About the journal | Publish with us | Subscribe

nature > outlook > article

OUTLOOK | 20 July 2022 | Correction 01 September 2022

Learning over a lifetime

Artificial-intelligence researchers turn to lifelong learning in the hopes of making machine intelligence more adaptable.

Neil Savage

Issues | Subject | More Content | Publish | Alerts | About | National Science Review

Volume 9, Issue 8
August 2022

Home

Article Contents

- Abstract
- INTRODUCTION
- EMERGING NEW CLASSES
- DECREMENTAL/INCREMENTAL FEATURES
- CHANGING DATA DISTRIBUTIONS
- VARIED LEARNING OBJECTIVES
- THEORETICAL ISSUES

JOURNAL ARTICLE

Open-environment machine learning

Zhi-Hua Zhou

National Science Review, Volume 9, Issue 8, August 2022, nwac123,
https://doi.org/10.1093/nsr/nwac123
Published: 01 July 2022 **Article history**

PDF | Split View | Annotate | Cite | Permissions | Share

Abstract

Conventional machine learning studies generally assume *close-environment* scenarios where important factors of the learning process hold invariant. With the great success of machine learning, nowadays, more and more practical tasks, particularly those involving *open-environment* scenarios where important factors are subject to change, called *open-environment machine learning* in this article, are p mmunity. Evidently, it is a grand challenge for machine learn close environment to open environment. It becomes even more challenging since, in various big data tasks, data are usually accumulated with time, like *streams*, while it is hard to train the machine learning model after collecting all data as in conventional studies. This article briefly introduces some advances in this line of research, focusing on

Annotate Highlight

Article activity al
Advance article al
New issue aler
In progress issue
Subject alert

Receive exclusive offers a
from Oxford Acade

Recommended

Published at the Workshop on Agent Learning in Open-Endedness (ALOE) at ICLR 2022

A LITTLE TAXONOMY OF OPEN-ENDEDNESS

Asiiah Song
Department of Computational Media
University of California, Santa Cruz
Santa Cruz, CA 95064, USA
julinas@ucsc.edu

开放环境下的机器学习

开放环境中服务机器人的任务规划和解释

Open-World Machine Learning and Classification

(Open-world Recognition, Open Set Recognition, Open-world AI)
A form of Lifelong Learning

Second Edition: "Lifelong Machine Learning." by Z. Chen and B. Liu, Morgan & Claypool, August 2018 (1st edition, 2016)

- Three new chapters have been added and others have been updated and/or reorganized.
- One Chapter is dedicated to Open World Learning
- **Any AI system (e.g., chatbot and self-driving car) that cannot learn in deployment (e.g., chatting and driving) in the real-world open envoronment is not truly intelligent.**

An interview in Nature Outlook, July 20, 2022.

Learning on the Job in the Open World. Invited talk given at the *Continual Learning Workshop* @ ICML-2020, July 17, 2020.

Motivation: Sooner or later, AI agents will need to explore and learn by themsleves in the real world. They cannot forever depend on manually labeled data. The real world is open and dynamic, and full of unknowns. AI agents must be able to detect the unknowns and learn them in a self-supervised manner. They should not make the *closed-world assumption* any more.

Open world learning (OWL) (a.k.a. **open world recognition or classification,** or **open-world AI)** is getting increasingly important as the learning agent is increasingly **working in or facing the real-world open and dynamic environment**, e.g., chatbot and self-driving car, where the agent **cannot assume or expect** what it will see in the real-world contains only what it has learned previously. For example, a chatbot cannot assume that it knows everything that a user may say. A self-driving car cannot assume that the real-world has only things that it has seen and learned before. **The core of open-world learning or open-world AI is about recognizing unknowns and learning them so that the AI agent will become more and more knowledgeable**.

Classic machine learning makes the ***closed world assumption***, i.e., the classes that the agent sees in training are what it will see in testing (no new objects or classes can appear in testing) (Fei and Liu 2016). A more realistic scenario is to expect **unseen classes** during testing (open world). In this case, the goal is to design a learning algorithm that can classify data of the known/seen classes into their respective classes and also to reject/detect instances from unknown/unseen classes. This problem is called **open-world learning (or open-world classification)**. Apart from detecting the unseen classes, open-world learning should also incrementally or continually learn the new classes.

Tasks of open-world learning (OWL)

离度量与表示空间极
义。而实际应用场景
,提出针对或利用度
的泛化能力,并提出策略以降低其样本复杂度。
。本文从目标函数性质以及度量重用两个角度
文通过大量实验进行验证,说明满足理论假...

Martin Mundt (he/him)

OWLL Group Leader

Email: martin.mundt
_at_ tu-darmstadt.de

Personal Webpage:
martin-mundt.com

Martin is the OWLL research group leader at Hessian.AI and TU Darmstadt, where the vision is to create robust systems that can learn continually in an open-ended world. He is a board member of directors at the non-profit organization ContinualAI in the 2022-2024 election term and also a postdoctoral researcher at the Artificial Intelligence and Machine Learning (AIML) lab at TU Darmstadt.

He holds a PhD in computer science from Goethe University Frankfurt in 2021, with a focus on continual

开放环境下的强化学习

Fig. 1. Overview diagram for trustworthy RL against intrinsic vulnerabilities: robustness, safety, generalization

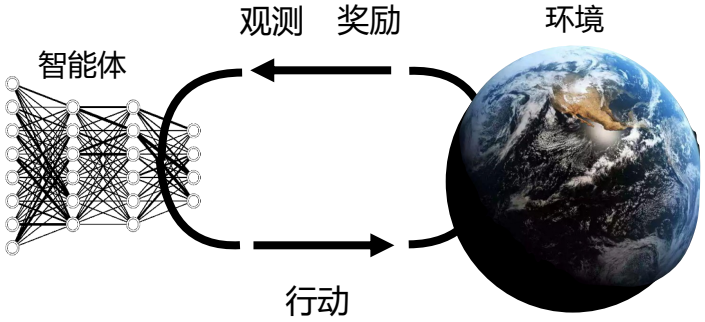
Xu, M., Liu, Z., Huang, P., Ding, W., Cen, Z., Li, B., & Zhao, D. (2022). **Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability**. *arXiv preprint arXiv:2209.08025*.

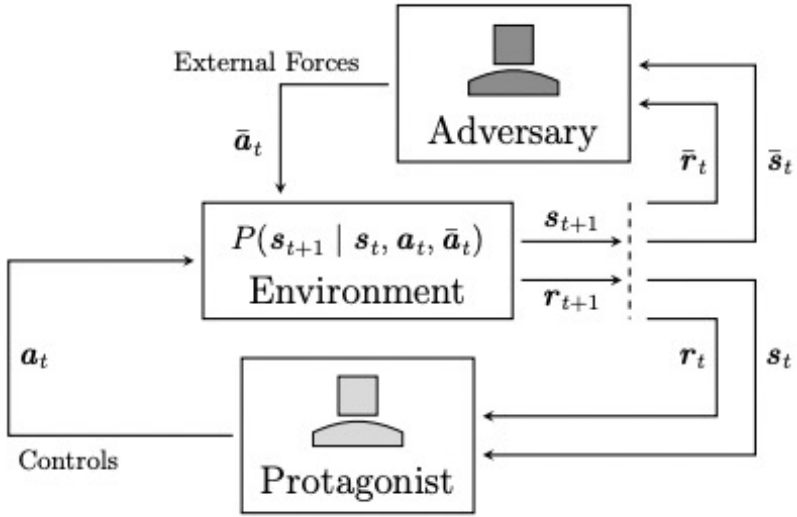
Beck J, Vuorio R, Liu E Z, et al. **A survey of meta-reinforcement learning**[J]. arXiv preprint arXiv:2301.08028, 2023.

Moos J, Hansel K, Abdulsamad H, et al. **Robust reinforcement learning: A review of foundations and recent advances**[J]. Machine Learning and Knowledge Extraction, 2022, 4(1): 276-315.

Kirk R, Zhang A, Grefenstette E, et al. **A survey of zero-shot generalisation in deep reinforcement learning**[J]. Journal of Artificial Intelligence Research, 2023, 76: 201-264.

开放环境下的强化学习





在预定环境中训练 vs 在开放环境中执行

Wang R, Lehman J, Rawal A, et al. **Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions**[C]//International Conference on Machine Learning. PMLR, 2020: 9940-9951.

Meier R, Mujika A. **Open-ended reinforcement learning with neural reward functions**[J]. Advances in Neural Information Processing Systems, 2022, 35: 2465-2479.

Samvelyan M, Khan A, Dennis M D, et al. **MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning**[C]//The Eleventh International Conference on Learning Representations. 20

Team A A, Bauer J, Baumli K, et al. **Human-timescale adaptation in an open-ended task space**[J]. arXiv preprint arXiv:2301.07608, 2023.

Li Y, Zhang S, Sun J, et al. **Cooperative Open-ended Learning** Framework for Zero-shot Coordination[J]. arXiv preprint arXiv:2302.04831, 2023.

Zhang W, Lu Z. RLADAPTER: BRIDGING LARGE LANGUAGE MODELS **TO REINFORCEMENT LEARNING IN OPEN WORLDS**[J].

Eck A, Soh L K, Doshi P. **Decision making in open agent systems**[J]. AI Magazine, 2023.

Prior Work: Policy Reuse

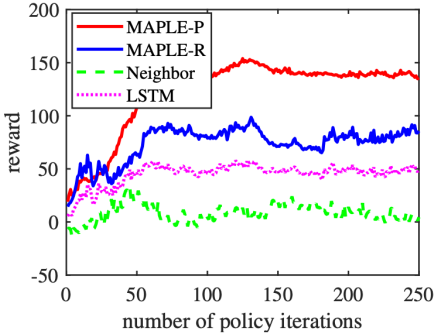
shallow trails: policies trained for a few iterations $\{\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, ...\tilde{\pi}_m\}$

environment feature = the rewards running these policies

$$v_i = (R(\tau_i|\tilde{\pi}_1), R(\tau_i|\tilde{\pi}_2), R(\tau_i|\tilde{\pi}_3), ...R(\tau_i|\tilde{\pi}_m))$$

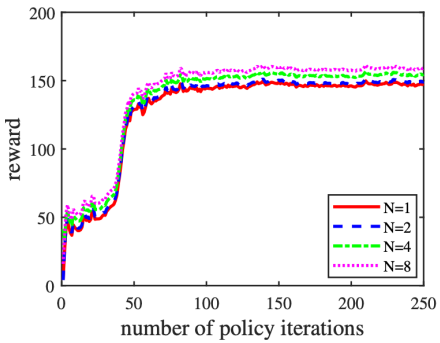
it works, and

feature quality matters



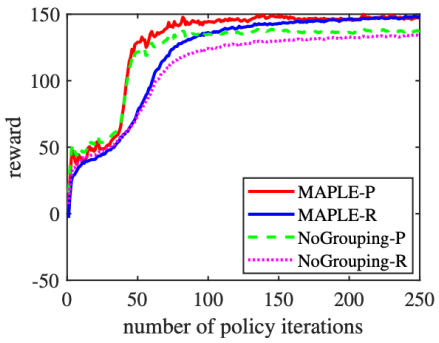
(a) Swimmer

feature discriminativeness



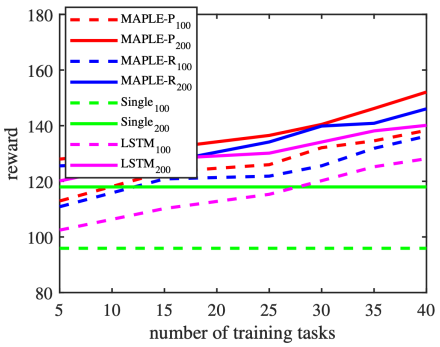
(a) Swimmer

model capacity matters



(a) Swimmer

training tasks matter



(a) Swimmer

we need: interactions, discriminative outcomes, large coverage tasks, and large model

Yang Yu, Shi-Yong Chen, Qing Da, Zhi-Hua Zhou. Reusable reinforcement learning via shallow trails. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2204-2215.

Prior Work: Context Capture

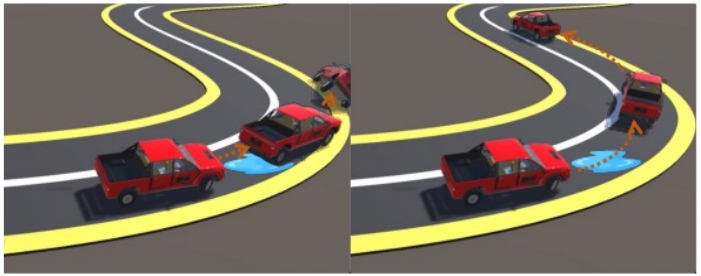
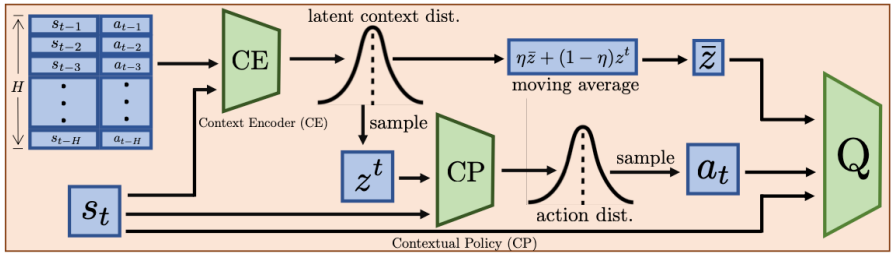


Figure 1: An illustration of sudden changes in an environment. When a car is driving through a water puddle, the water reduces the friction of the road, and the dynamics of the environment suddenly changes. Without the ability of fast adaptation for the new dynamics, an RL trained driving policy will lose control (left). In contrast, we expect the policy can adapt to the environment change rapidly and handle this emergency (right).

有限历史交互轨迹

环境编码平滑



隐变量与环境特征一致：

$$\mathcal{L}^i_{\text{distance}} = \mathbb{E}[\|z^t_i - u_i\|^2_2],$$

分解目标：

$$\mathcal{L}^i_{\text{distance}} = \mathbb{E}[\|z^t_i - \mathbb{E}[z^t_i]\|^2_2] + \|\mathbb{E}[z^t_i] - u_i\|^2_2.$$

改写目标：

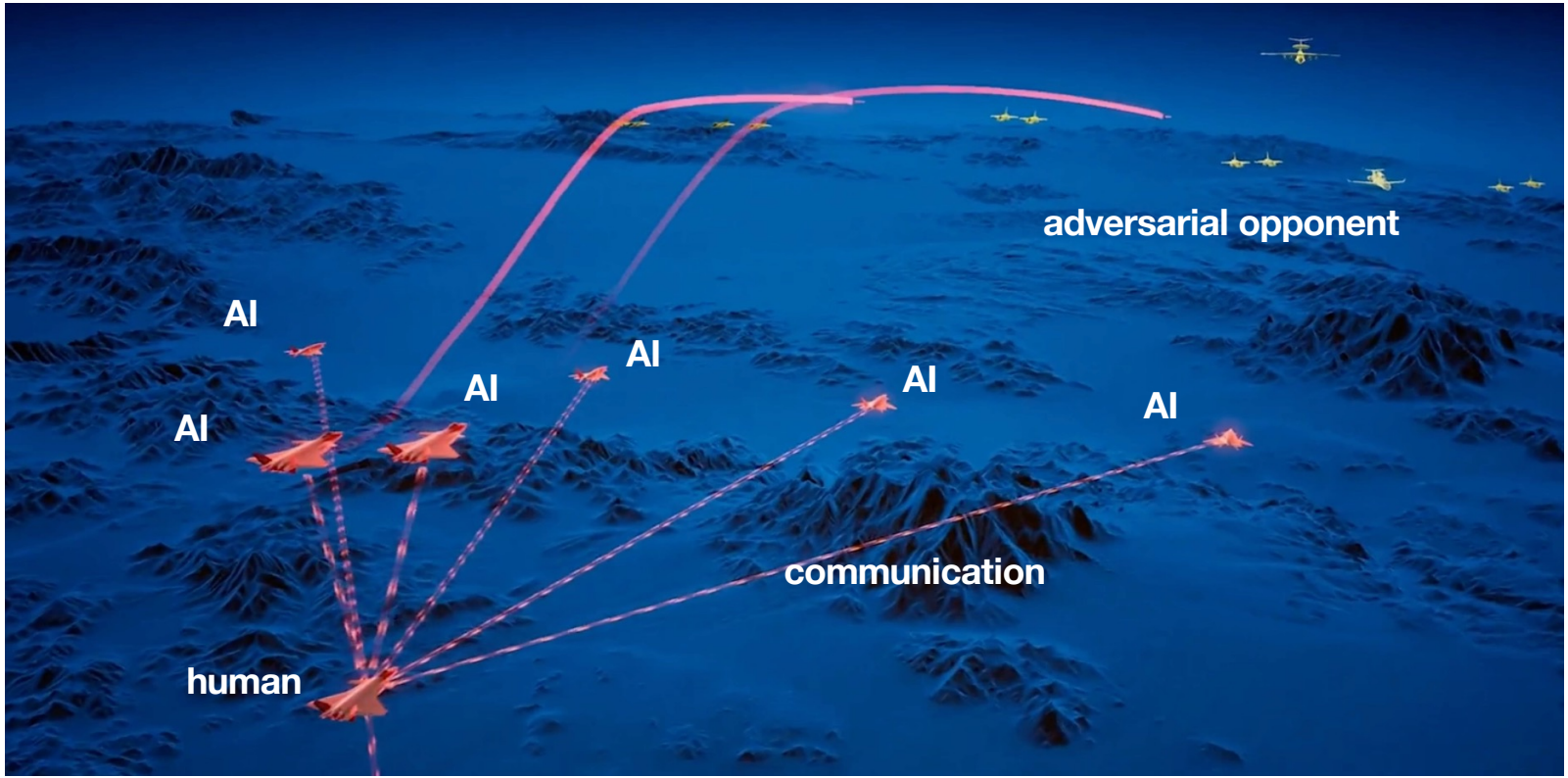
$$\mathcal{L}_{\text{CE}} = \lambda \sum_{i=0}^{M-1} \mathbb{E}\left[\|z^t_i - \mathbb{E}[z^t_i]\|^2_2\right] - \log\det(R_{\{\mathbb{E}[z^t_i]\}}),$$

Fan-Ming Luo, Shengyi Jiang, Yang Yu, Zongzhang Zhang, Yi-Feng Zhang. Adapting environment sudden changes by learning context sensitive policy. AAAI 2022.

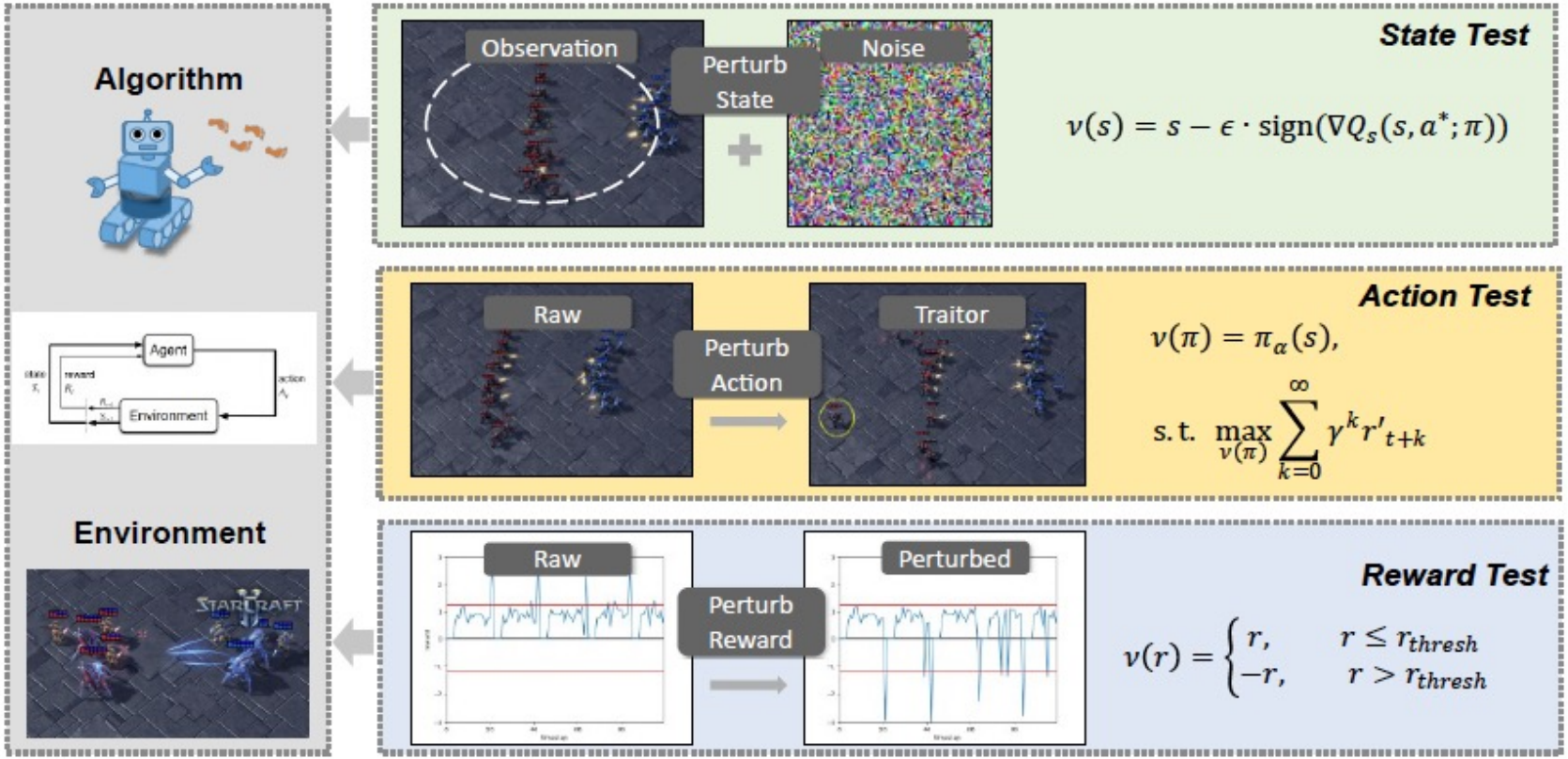
From SARL to MARL

A higher level complexity of the environment

Using similar ideas: large coverage training tasks, interactions to acquire features



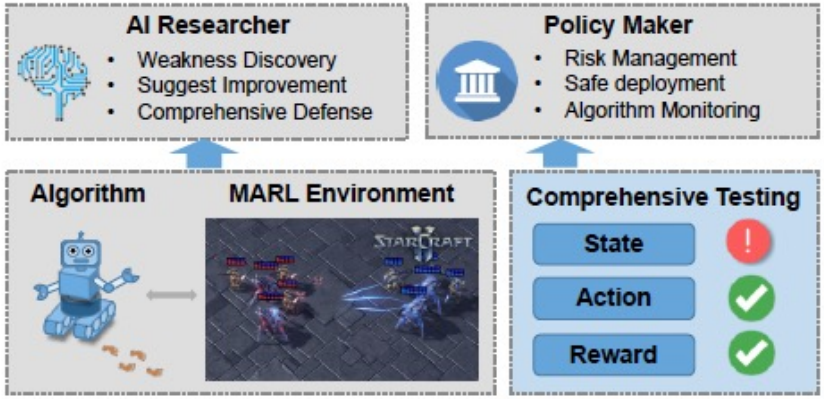
The Robustness of MARL



Guo, J., Chen, Y., Hao, Y., Yin, Z., Yu, Y., & Li, S. (2022). Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 115-122).

The Robustness of MARL

Win rate (WR), team reward (TR), mean number of dead allies (mDA), and mean number of dead enemies (mDE).



Map	Algorithm	WR	TR	mDA	mDE
2s3z	QMIX	0.00%	5.32	4.41	0.06
	MAPPO	0.00%	0.00	3.72	0.00
11m	QMIX	0.00%	0.00	2.09	0.00
	MAPPO	0.00%	5.89	11.00	0.16

Table 3. Performance of QMIX and MAPPO under reward test. None of the algorithms are robust under reward-based attack.

Map	Algorithm	WR	TR	mDA	mDE
2s3z	QMIX	0.00%	10.07	4.97	1.34
	MAPPO	0.00%	11.59	5.00	2.13
11m	QMIX	6.25%	9.88	10.69	4.81
	MAPPO	0.00%	10.53	11.00	6.31

Table 4. Performance of QMIX and MAPPO under action test. None of the algorithms are robust under action-based attack.

Map	Algorithm	WR	TR	mDA	mDE
2s3z	QMIX	9.38%	11.85	4.72	1.69
	MAPPO	65.62%	17.91	3.34	4.25
11m	QMIX	0.00%	9.76	10.94	5.53
	MAPPO	31.25%	14.62	9.69	8.66

Table 2. Performance of QMIX and MAPPO under state test. Both algorithms shows weak robustness, while MAPPO is relatively robust.

Guo, J., Chen, Y., Hao, Y., Yin, Z., Yu, Y., & Li, S. (2022). Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 115-122).

Openness in MARL

Table 3　开放环境下的多智能体研究内容。

研究方向	核心内容	代表算法	应用与取得成果
离线学习	将单智能体强化学习中成功应用的技术扩展到多智能体场景或针对性设计多智能体离线方法	ICQ [317], MABCQ [318], SIT [319], ODIS [34], MADT [320], OMAC [321], CFCQL [322]	从收集的静态离线数据中学习策略，避免与环境交互带来的问题，实现从大规模及多样性的数据中进行策略的学习目标
策略迁移与泛化	跨任务间学得多智能体策略的迁移与直接泛化，实现知识重用	LeCTR [323], MAPTF [324], EPC [325], 文献 [326], MATTAR [327]	实现任务任务间的知识重用，加快在新任务上的学习速度与能力
持续协作	在面对任务或样本以顺序的方式出现的情况下的协作任务学习	文献 [328], MACPro [329], Macop [330]	扩展单智能体已有技术，在多智能体中处理协作任务流出现情况
演化多智能体强化学习	模拟生物自然进化过程的启发式随机优化算法，包括遗传算法、演化策略、粒子群算法等，赋能多智能体协作	MERL [331], BEHT [332], MCAA [333], EPC [325], ROMANCE [334], MA3C [159]	通过演化手段模拟多智能体策略或生成辅助训练对手帮助多智能体策略训练，在很多项目场景中得到广泛应用
稳健性研究	考虑系统环境发生变化条件下的策略学习与执行，学习可以应对环境噪声、队友变化等情况出现下的稳健策略	R-MADDPG [47], 文献 [46], RAMAO [335], ROMANCE [334], MA3C [158], CroMAC [159]	在环境中状态、观测、动作与通信信道遭受噪声甚至恶意攻击条件下具备稳健的协作能力
多目标（约束）协作	优化问题中存在多个目标函数，需要同时考虑多个目标函数的最优解	MACPO(MAPPO-Lagrangian) [258], CAMA [336], MDBC [337], 文献 [192, 338, 339]	考虑环境中存在的多个约束目标，在有约束或安全领域等取得进展，为多智能体协作落地提供基础
风险敏感多智能体协作	使用值分布等手段将环境中的变量数值（奖赏）建模为分布，运用风险函数等评估系统的风险等	DFAC [340], RMIX [341], ROE [149], DRE-MARL [342], DRIMA [343], 文献 [344, 345]	可以在复杂场景下提升协作性能，在风险敏感场景可以有效感知风险并评估性能
自组织协作	创建自主智能体，使其能够有效、稳健地与之前未知的队友在需要协作的任务上合作	文献 [346, 347], ODITS [348], OSBG [349], BRDiv [350], L-BRDiv [351], TEAMSTER [352]	赋予智能体与未训练过的智能体高效协作的能力，在多种任务场景下可以实现与未见队友高效协作目的

所学的策略应该可以应对环境因素的变化，考虑真实环境下的约束等，至少具备以下能力：

- 包括**离线策略学习**、策略具备**迁移与泛化能力**、策略支持**持续学习**以及系统系统应该具备**演化与演进能力**；
- 策略在部署过程中具备应对环境因素发生变化的能力，具体而言，在多智能体环境如状态、观测、动作与通信等发生变化下的时具备**稳健协作**能力；
- 真实环境部署应该考虑的**多目标（约束）**策略优化、面对真实搞高动态任务场景时具备**风险感知与评估**能力；
- 训练好的策略，在部署时，应该具备**自组织协作**能力，同时应该具有**零（少）样本策略**适应能力；另一方面，应该支持**人智协同**，赋予多智能体系统为人类服务的能力；
- 最后，考虑各种多智能体协作任务的差别与相似度，为每一类任务学习一个策略模型往往代价大且浪费资源，策略应该具备诸如ChatGPT一样的**大模型**能力。

Table 4　开放环境下的多智能体研究内容（续）。

研究方向	核心内容	代表算法	应用与取得成果
零（少）样本协作	设计训练范式以使得多智能体系统在使用少量样本甚至零交互样本的条件下具备与未见队友协作的能力	FCP [353], TrajeDi [354], MAZE [355], CSP [356], LIPO [357], HSP [358], Macop [329], 文献 [52, 359]	在当前多类环境如Overcook上的结果表明，当前的部分算法可以有效实现与未见队友协作的目标
人智协同	为人智交互或人机交互提供支持，使得人类参与者与智能体之间更好地协作以完成特定任务	FCP [353], 文献 [360], HSP [358], Latent Offline RL [361], RILI [362], PECAN [363]	在给定的仿真环境或者真实机器人场景等实现一定程度上的人机协作目标
协作大模型	借助通用大模型思想开发决策协作大模型，或者借助当前现有大模型技术促进多智能体协作	MADT [320], MAT [32], MAGENTA [364], MADiff [365], ProAgent [217], SAMA [366]	部分工作学习针对部分任务的大模型，在部分场景具备一定的通用能力，一些工作借助大语言模型促进协作

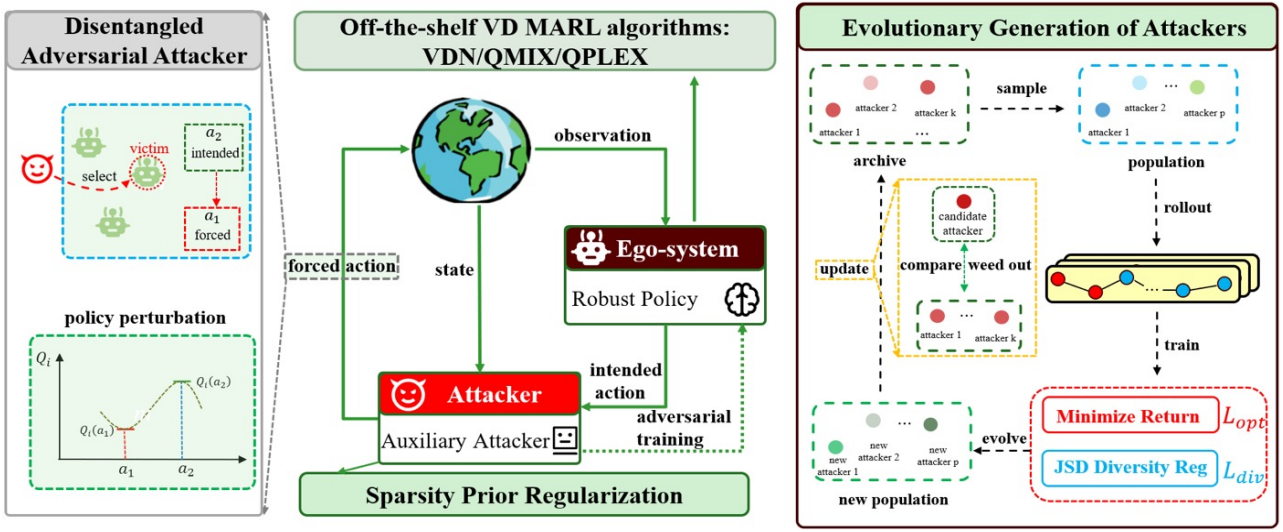
Openness in MARL: Action Perturbation

Adversarial attacker
$\pi_{adv} : \mathcal{S} \times \mathcal{A} \times \mathbb{N} \to \mathcal{A}$

forcing to execute
$\hat{\boldsymbol{a}} \sim \pi_{adv}(\cdot|s, \boldsymbol{a}, k)$ with $s' \sim P(\cdot|s, \hat{\boldsymbol{a}}),\ r = R(s, \hat{\boldsymbol{a}}, s')$

Attacker learning:
- minimize the reward of the ego-system
- sparsity prior regularization
- JSD diversity regularization

Adversarial training:
- Maintain attacker population
- Quality-Diversity algorithm
- customized selection and update mechanism



Lei Yuan, Zi-Qian Zhang, Ke Xue, Hao Yin, Feng Chen, Cong Guan, Li-He Li, Chao Qian, Yang Yu. Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. In: AAAI'23.

Openness in MARL: Action Perturbation

Different attackers

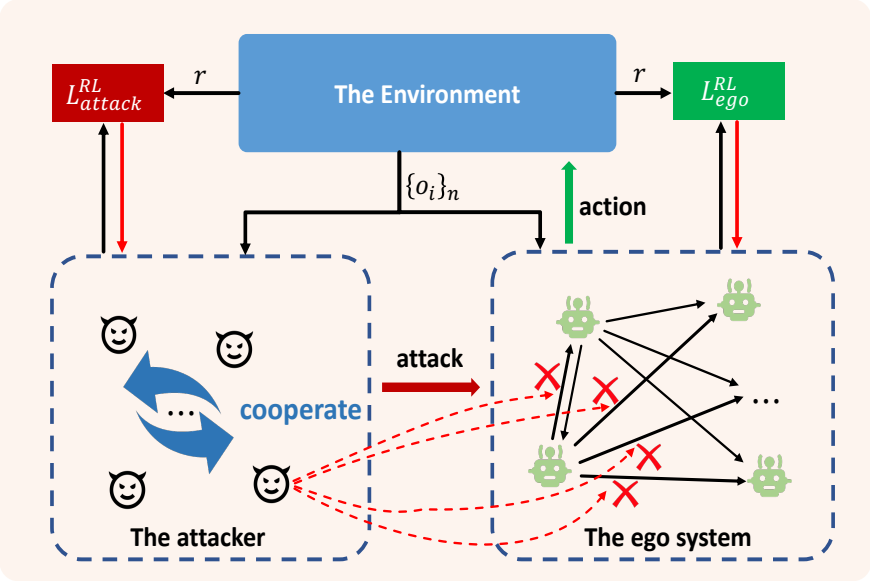
Method	Map_Name	2s3z $K=8$	3m $K=4$	3s_vs_3z $K=8$	8m $K=5$	MMM $K=8$	1c3s5z $K=6$	+/ − / ≈
Natural	vanilla QMIX	92.8 ± 1.62	$\mathbf{97.9 \pm 1.02}$	98.3 ± 0.78	98.2 ± 0.45	95.8 ± 1.59	88.8 ± 2.13	1/1/4
	RARL	96.4 ± 1.19	86.0 ± 5.38	80.6 ± 27.5	95.3 ± 3.31	89.3 ± 7.01	76.9 ± 9.85	0/4/2
	RAP	$\mathbf{98.1 \pm 0.76}$	91.3 ± 4.93	$\mathbf{99.3 \pm 0.51}$	91.7 ± 7.96	95.3 ± 4.98	86.7 ± 10.5	0/1/5
	RANDOM	98.0 ± 0.60	95.3 ± 2.07	99.6 ± 0.35	$\mathbf{98.6 \pm 0.90}$	93.8 ± 7.56	93.1 ± 4.41	1/0/5
	ROMANCE	97.9 ± 1.34	96.0 ± 1.83	97.8 ± 1.78	94.3 ± 3.94	$\mathbf{97.1 \pm 1.49}$	$\mathbf{93.9 \pm 1.24}$	
Random Attack	vanilla QMIX	78.8 ± 1.28	$\mathbf{78.7 \pm 1.49}$	87.0 ± 0.36	66.2 ± 2.08	70.0 ± 3.97	66.6 ± 2.03	0/5/1
	RARL	84.3 ± 2.40	67.6 ± 5.01	70.1 ± 29.1	75.7 ± 7.00	62.2 ± 10.2	56.5 ± 10.8	0/5/1
	RAP	87.3 ± 1.87	73.5 ± 3.49	89.8 ± 4.81	$\mathbf{78.4 \pm 8.22}$	84.2 ± 9.05	66.8 ± 9.66	0/1/5
	RANDOM	83.9 ± 6.38	76.4 ± 2.27	91.9 ± 1.32	72.0 ± 3.46	72.9 ± 7.09	60.5 ± 21.3	0/2/4
	ROMANCE	$\mathbf{89.1 \pm 1.97}$	78.1 ± 5.13	$\mathbf{93.0 \pm 1.82}$	76.2 ± 5.36	$\mathbf{85.8 \pm 8.66}$	$\mathbf{77.9 \pm 1.96}$	
EGA	vanilla QMIX	26.7 ± 4.28	20.7 ± 2.13	30.9 ± 1.52	42.7 ± 9.79	37.9 ± 3.13	35.2 ± 8.66	0/6/0
	RARL	56.1 ± 11.8	86.1 ± 0.98	60.9 ± 14.2	66.3 ± 7.25	41.5 ± 11.6	35.3 ± 4.00	0/6/0
	RAP	64.1 ± 11.9	84.0 ± 4.27	65.1 ± 4.41	84.4 ± 8.88	74.9 ± 15.5	45.4 ± 6.83	0/4/2
	RANDOM	48.3 ± 17.3	66.2 ± 16.6	54.4 ± 7.83	55.6 ± 12.5	53.1 ± 6.09	43.3 ± 10.3	0/6/0
	ROMANCE	$\mathbf{81.6 \pm 0.84}$	$\mathbf{89.7 \pm 1.52}$	$\mathbf{90.5 \pm 1.97}$	$\mathbf{86.2 \pm 5.11}$	$\mathbf{84.0 \pm 11.5}$	$\mathbf{66.5 \pm 3.24}$	

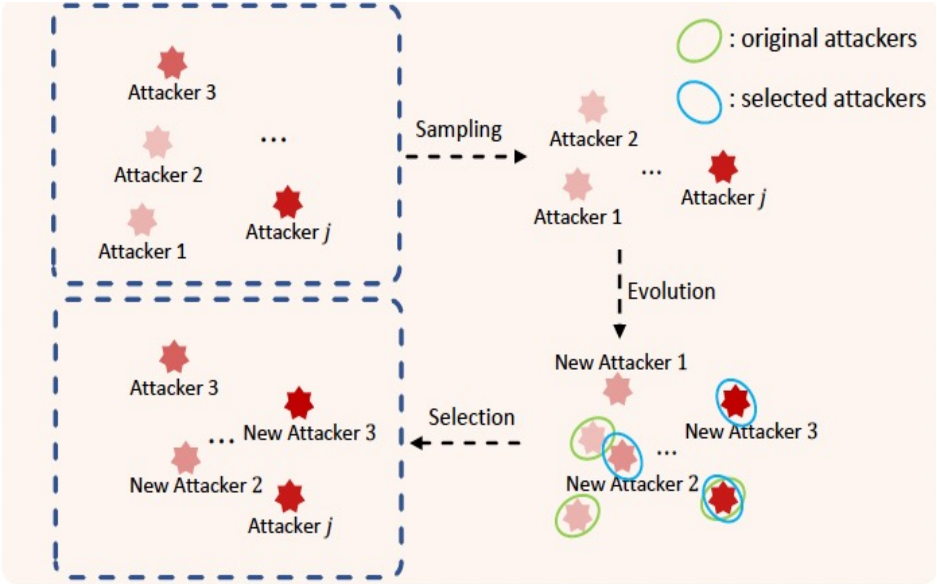
Different attack strengths

Method	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	$K=11$	$K=12$	$K=14$
vanilla QMIX	59.2 ± 2.66	42.1 ± 0.81	26.7 ± 4.28	17.3 ± 0.62	12.2 ± 0.33	8.74 ± 0.14	6.42 ± 0.84	2.82 ± 0.70
RARL	72.7 ± 4.22	65.2 ± 9.11	56.1 ± 11.8	46.3 ± 12.1	38.0 ± 13.5	31.8 ± 13.1	25.9 ± 12.3	18.6 ± 10.9
RAP	81.7 ± 7.37	73.6 ± 7.46	64.1 ± 11.9	53.5 ± 11.7	42.5 ± 11.6	33.9 ± 11.4	25.8 ± 10.7	14.0 ± 7.50
RANDOM	69.3 ± 10.9	56.8 ± 12.8	48.3 ± 17.3	34.7 ± 17.3	25.5 ± 15.8	19.8 ± 14.3	14.9 ± 12.6	10.0 ± 9.69
ROMANCE	$\mathbf{89.9 \pm 1.19}$	$\mathbf{86.4 \pm 1.87}$	$\mathbf{81.6 \pm 0.84}$	$\mathbf{75.1 \pm 0.58}$	$\mathbf{66.7 \pm 1.56}$	$\mathbf{57.4 \pm 1.61}$	$\mathbf{48.6 \pm 2.60}$	$\mathbf{41.5 \pm 2.17}$

Lei Yuan, Zi-Qian Zhang, Ke Xue, Hao Yin, Feng Chen, Cong Guan, Li-He Li, Chao Qian, Yang Yu. Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. In: AAAI'23.

Openness in MARL: Message Perturbation





Lei Yuan Feng Chen, Zongzhang Zhang, Yang Yu. Communication-Robust Multi-Agent Learning by Adaptable Auxiliary Multi-Agent Adversary Generation. https://arxiv.org/pdf/2305.05116.pdf.

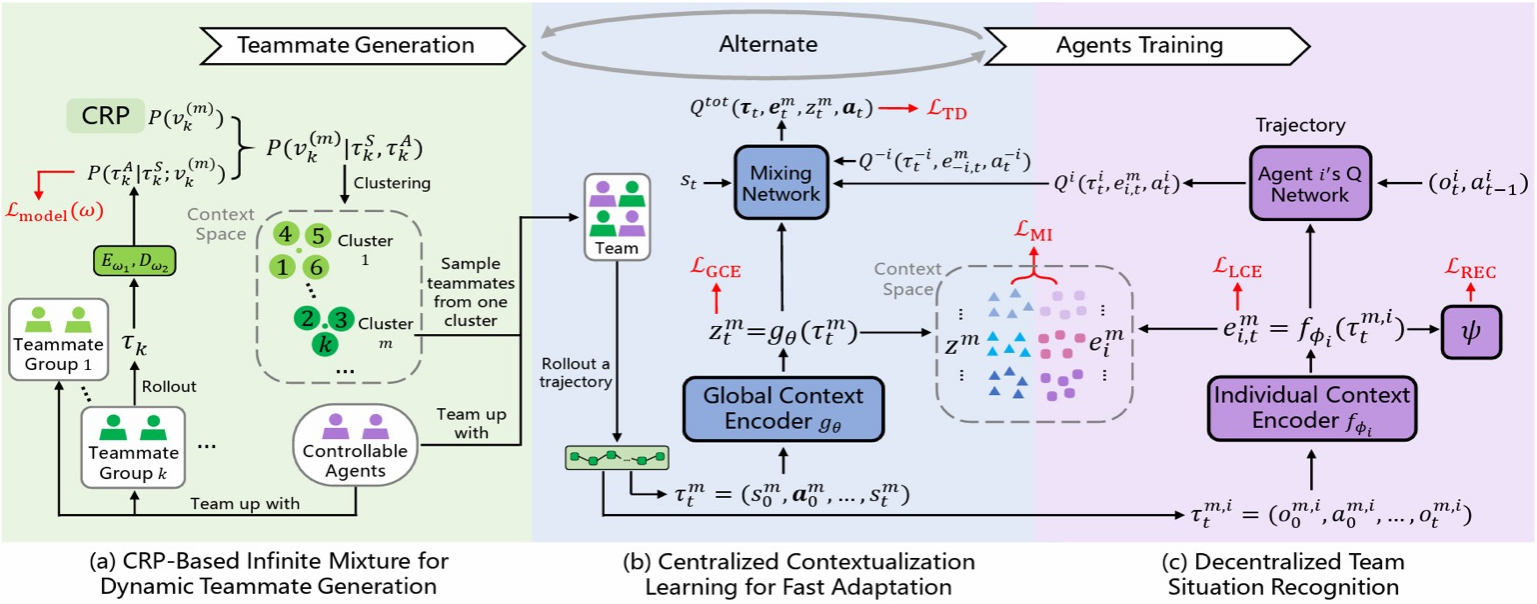
Openness in MARL: Message Perturbation

Table 1 Performance comparison under different attack modes.

		Hallway-6x6	Hallway-4x5x9	SMAC-1o_2r_vs_4r	SMAC-1o_10b_vs_1r	GP-4r	GP-9r
Normal	MA3C	0.94±0.05	0.97±0.05	0.86±0.02	0.62±0.01	0.87±0.02	0.82±0.01
	Vanilla	**1.00**±0.00	**1.00**±0.00	0.81±0.06	**0.63**±0.04	**0.88**±0.03	0.82±0.02
	Noise Adv.	**1.00**±0.00	0.99±0.01	**0.88**±0.04	0.6±0.05	**0.88**±0.03	**0.85**±0.02
	MA3C w/o div.	0.98±0.02	0.66±0.46	0.86±0.02	0.62±0.03	0.86±0.09	0.81±0.03
	Instance Adv.	0.52±0.48	0.67±0.47	0.84±0.02	0.57±0.04	0.86±0.03	0.82±0.03
Random Noise	MA3C	0.91±0.07	0.79±0.18	**0.87**±0.01	**0.67**±0.03	0.88±0.01	0.80±0.07
	Vanilla	0.58±0.03	0.53±0.06	0.73±0.07	0.60±0.02	0.86±0.03	0.79±0.02
	Noise Adv.	**0.97**±0.02	**1.00**±0.00	0.82±0.02	0.56±0.02	0.88±0.01	**0.82**±0.01
	MA3C w/o div.	0.68±0.07	0.68±0.29	0.73±0.07	0.53±0.01	0.82±0.06	0.80±0.07
	Instance Adv.	0.56±0.34	0.67±0.47	0.79±0.07	0.60±0.08	**0.90**±0.03	0.81±0.02
Aggressive Attackers	MA3C	**0.91**±0.22	**0.98**±0.01	**0.67**±0.03	**0.62**±0.03	**0.81**±0.02	**0.76**±0.03
	Vanilla	0.09±0.19	0.00±0.00	0.26±0.12	0.57±0.03	0.38±0.02	0.30±0.05
	Noise Adv.	0.61±0.37	0.13±0.14	0.51±0.02	0.54±0.03	0.41±0.13	0.48±0.11
	MA3C w/o div.	0.57±0.39	0.96±0.03	0.54±0.05	0.61±0.02	0.68±0.06	0.71±0.01
	Instance Adv.	0.63±0.42	0.88±0.14	0.28±0.01	0.61±0.04	**0.81**±0.02	**0.76**±0.03

Lei Yuan, Feng Chen, Zongzhang Zhang, Yang Yu. Communication-Robust Multi-Agent Learning by Adaptable Auxiliary Multi-Agent Adversary Generation. https://arxiv.org/pdf/2305.05116.pdf.

Openness in MARL: Policy Sudden Change



(a) CRP-Based Infinite Mixture for Dynamic Teammate Generation

(b) Centralized Contextualization Learning for Fast Adaptation

(c) Decentralized Team Situation Recognition

CRP-based Infinite Mixture

- How to deal with infinite groups of teammates?
- Instantiate a DPMM with Chinese Restaurant Process.

Centralized Contextualization Learning

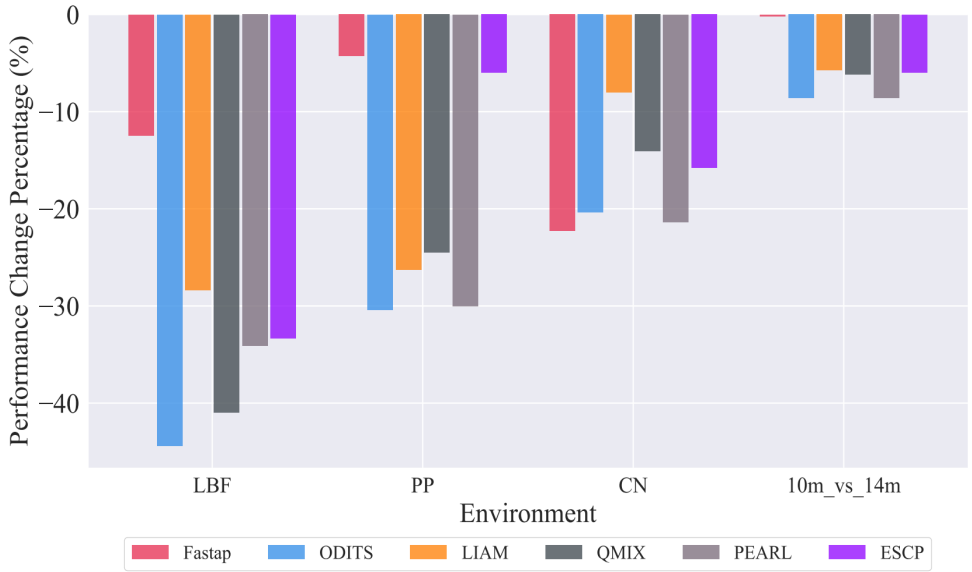
- Learn a global context encoder which is able to identify and track the sudden change of teammates.

Decentralized Team Situation Recognition

- Learn informatively consistent local embeddings based on mutual information objective and auxiliary objectives.

Ziqian Zhang, Lei Yuan, Lihe Li, Ke Xue, Chengxing Jia, Cong Guan, Chao Qian, Yang Yu. Fast teammate adaptation in the presence of sudden policy change. In: UAI 2023.

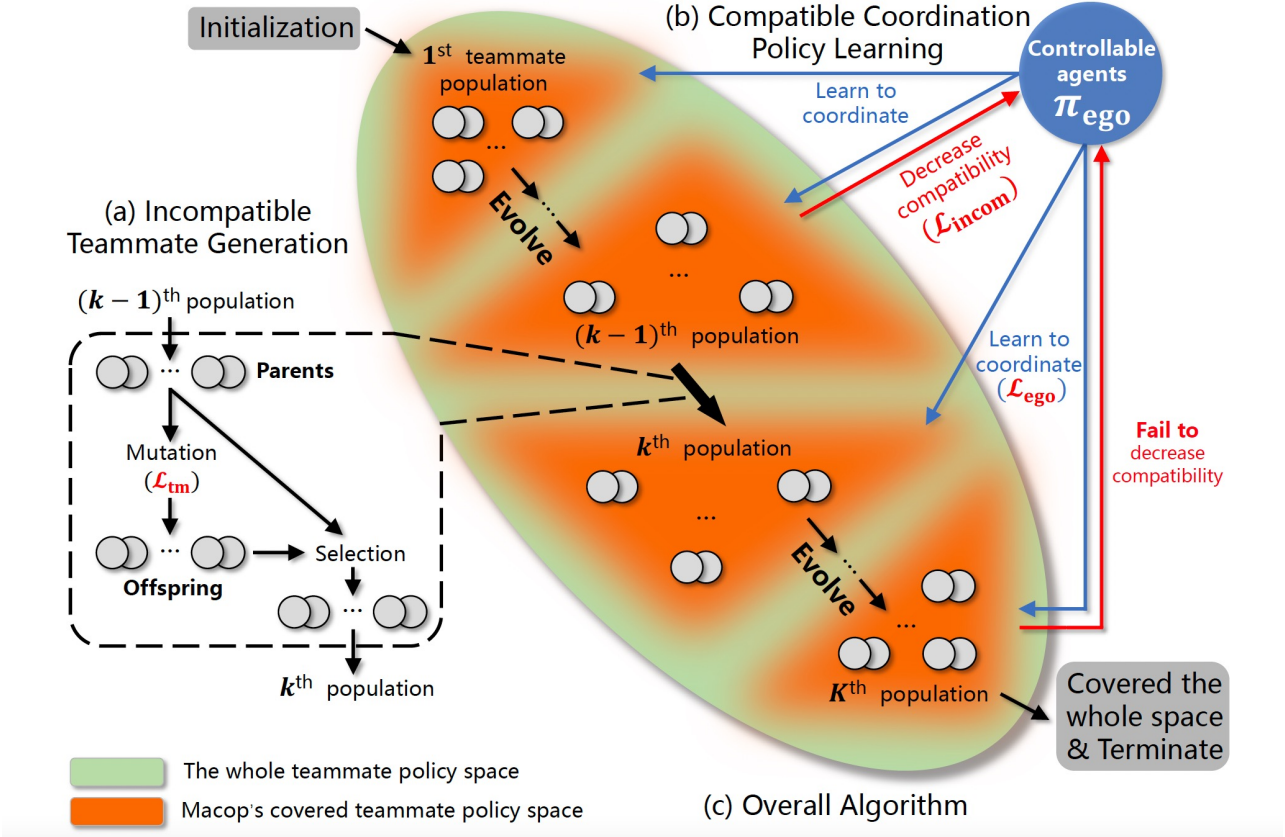
Openness in MARL: Policy Sudden Change



https://www.bilibili.com/video/BV1hm4y1E7y4/?buvid=XX38F533E5D4584B88F79166E23FBE6E0051D&from_spmid=main.space-contribution.0.0&is_story_h5=false&mid=EU%2BHoPaulypzlHXNVjg1YA%3D%3D&p=1&plat_id=114&share_from=ugc&share_medium=android&share_plat=android&share_session_id=d334e409-ce14-46e1-aaf2-e657ea81bbc4&share_source=COPY&share_tag=s_i&spmid=united.player-video-detail.0.0&timestamp=1697191225&unique_k=OKRohtI&up_id=604515161

Ziqian Zhang, Lei Yuan, Lihe Li, Ke Xue, Chengxing Jia, Cong Guan, Chao Qian, Yang Yu. Fast teammate adaptation in the presence of sudden policy change. In: UAI 2023.

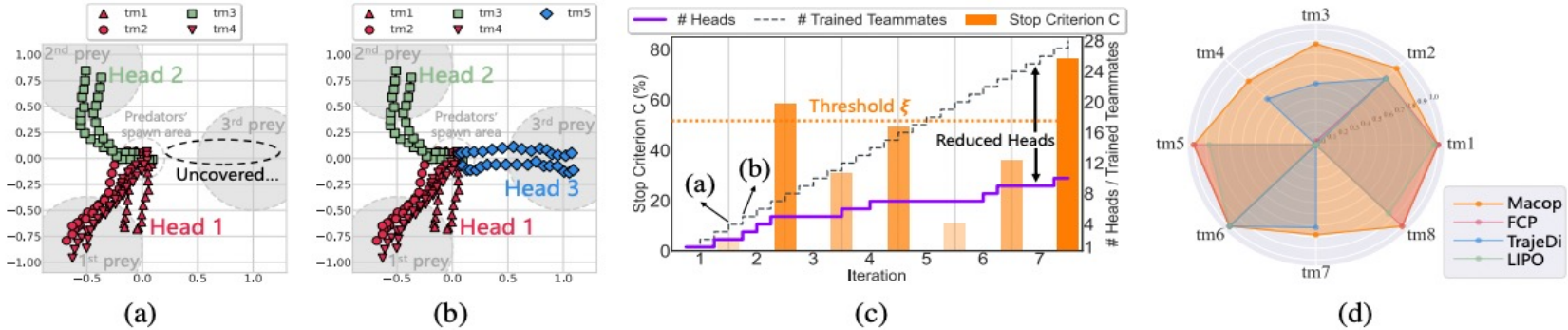
Openness in MARL: High Coordination Generalization Ability



Yuan, Lei and Li, Lihe and Zhang, Ziqian and Chen, Feng and Zhang, Tianyi and Guan, Cong and Yu, Yang and Zhou, Zhi-Hua. “Learning to Coordinate with Anyone.” DAI 2023.

Openness in MARL: High Coordination Generalization Ability

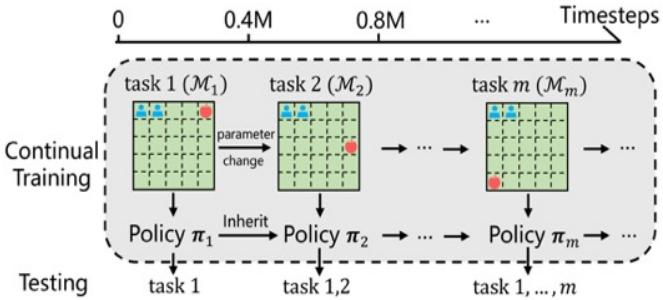
Method \ Env	LBF		PP		CN		SMAC		Avg. Performance
	LBF1	LBF4	PP1	PP2	CN2	CN3	SMAC1	SMAC2	Improvement (%)
Macop (ours)	1.14 ± 0.02	$\mathbf{1.64 \pm 0.03}$	$\mathbf{1.73 \pm 0.11}$	$\mathbf{2.14 \pm 0.53}$	$\mathbf{1.66 \pm 0.03}$	$\mathbf{1.70 \pm 0.06}$	$\mathbf{1.26 \pm 0.42}$	1.56 ± 0.17	**60.44**
Single Head	0.98 ± 0.07	1.10 ± 0.32	0.87 ± 0.58	1.44 ± 0.52	1.01 ± 0.49	0.99 ± 0.24	1.06 ± 0.14	1.25 ± 0.40	8.92
Random Head	0.92 ± 0.05	0.85 ± 0.10	0.88 ± 0.17	1.18 ± 0.39	0.98 ± 0.23	0.92 ± 0.11	0.97 ± 0.14	1.28 ± 0.21	−0.25
LIPO [Charakorn et al.(2023)]	1.07 ± 0.09	1.53 ± 0.14	1.64 ± 0.21	1.93 ± 0.52	1.13 ± 0.41	1.33 ± 0.25	1.19 ± 0.18	1.08 ± 0.21	36.27
FCP [Strouse et al.(2021)]	$\mathbf{1.16 \pm 0.02}$	1.33 ± 0.06	1.17 ± 0.85	1.34 ± 0.12	0.90 ± 0.48	1.41 ± 0.23	0.97 ± 0.19	1.54 ± 0.10	25.82
TrajeDi [Lupu et al.(2021)]	1.16 ± 0.06	1.34 ± 0.11	1.68 ± 0.33	1.56 ± 0.52	1.29 ± 0.23	1.53 ± 0.11	1.25 ± 0.12	$\mathbf{1.57 \pm 0.16}$	42.26
EWC [Kirkpatrick et al.(2017)]	0.97 ± 0.08	0.99 ± 0.16	0.83 ± 0.48	0.77 ± 0.43	0.57 ± 0.37	0.71 ± 0.27	1.03 ± 0.13	0.61 ± 0.09	−18.82
Finetune	1.00 ± 0.16	1.00 ± 0.27	1.00 ± 0.58	1.00 ± 0.68	1.00 ± 0.31	1.00 ± 0.24	1.00 ± 0.17	1.00 ± 0.23	/
$+/\approx/-$	3/4/0	6/1/0	2/5/0	6/1/0	7/0/0	7/0/0	5/2/0	4/3/0	7/0/0

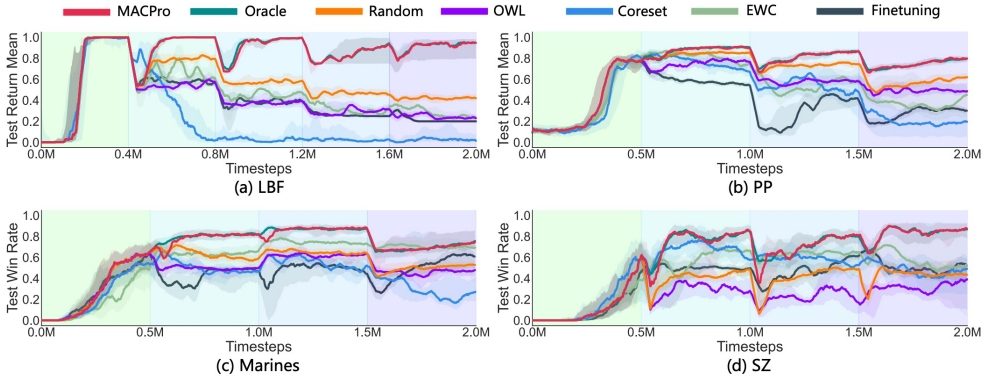


Yuan, Lei and Li, Lihe and Zhang, Ziqian and Chen, Feng and Zhang, Tianyi and Guan, Cong and Yu, Yang and Zhou, Zhi-Hua. “Learning to Coordinate with Anyone.” DAI 2023.

Openness in MARL: Continual Coordination

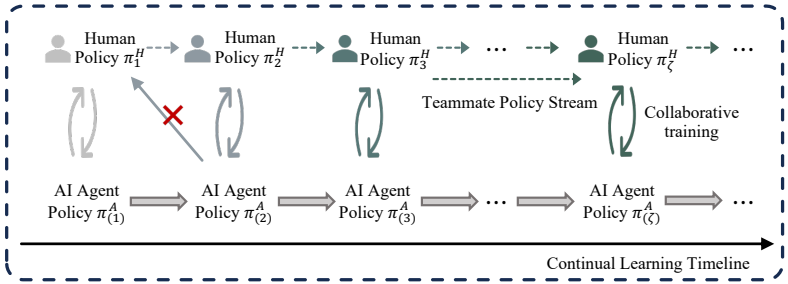
Continual coordination

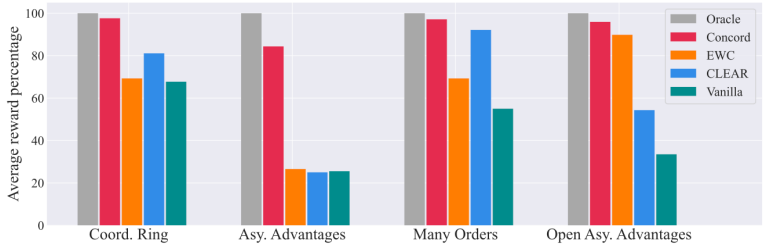




Lei Yuan, et al. “Multi-agent Continual Coordination via Progressive Task Contextualization.” arXiv:2305.13937 (2023).

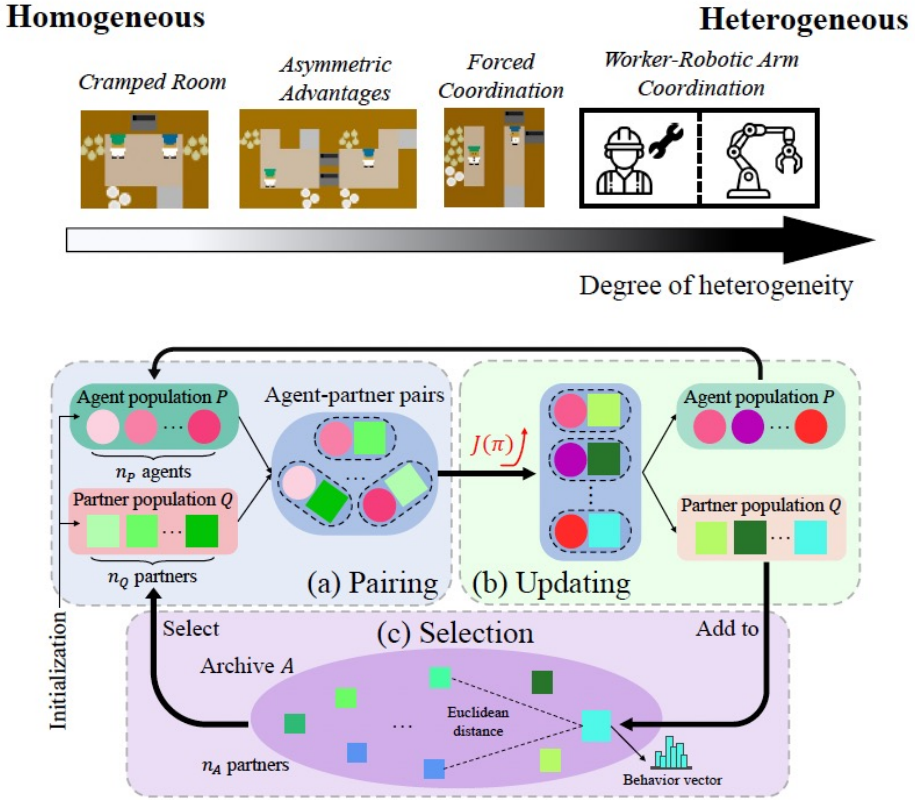
Continual learning for human-AI Coordination

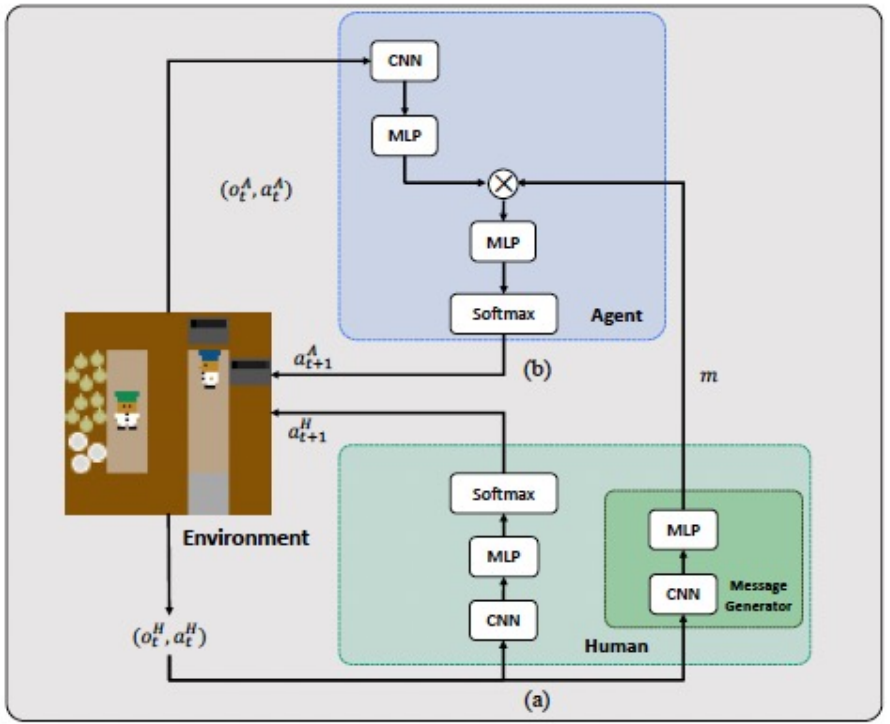




Cong Guan, et al. One by One, Continual Coordinating with Humans via Hyper-Teammate Identification. Submitted.

Openness in MARL: Open and Real-World Human-AI Coordination





Xue, Ke, et al. “Heterogeneous multi-agent zero-shot coordination by coevolution." arXiv preprint arXiv:2208.04957 (2022).

Cong Guan, et al. Open and Real-World Human-AI Coordination by Heterogeneous Training with Communication. Submitted.

Openness in MARL: LLMs for Human-AI Coordination

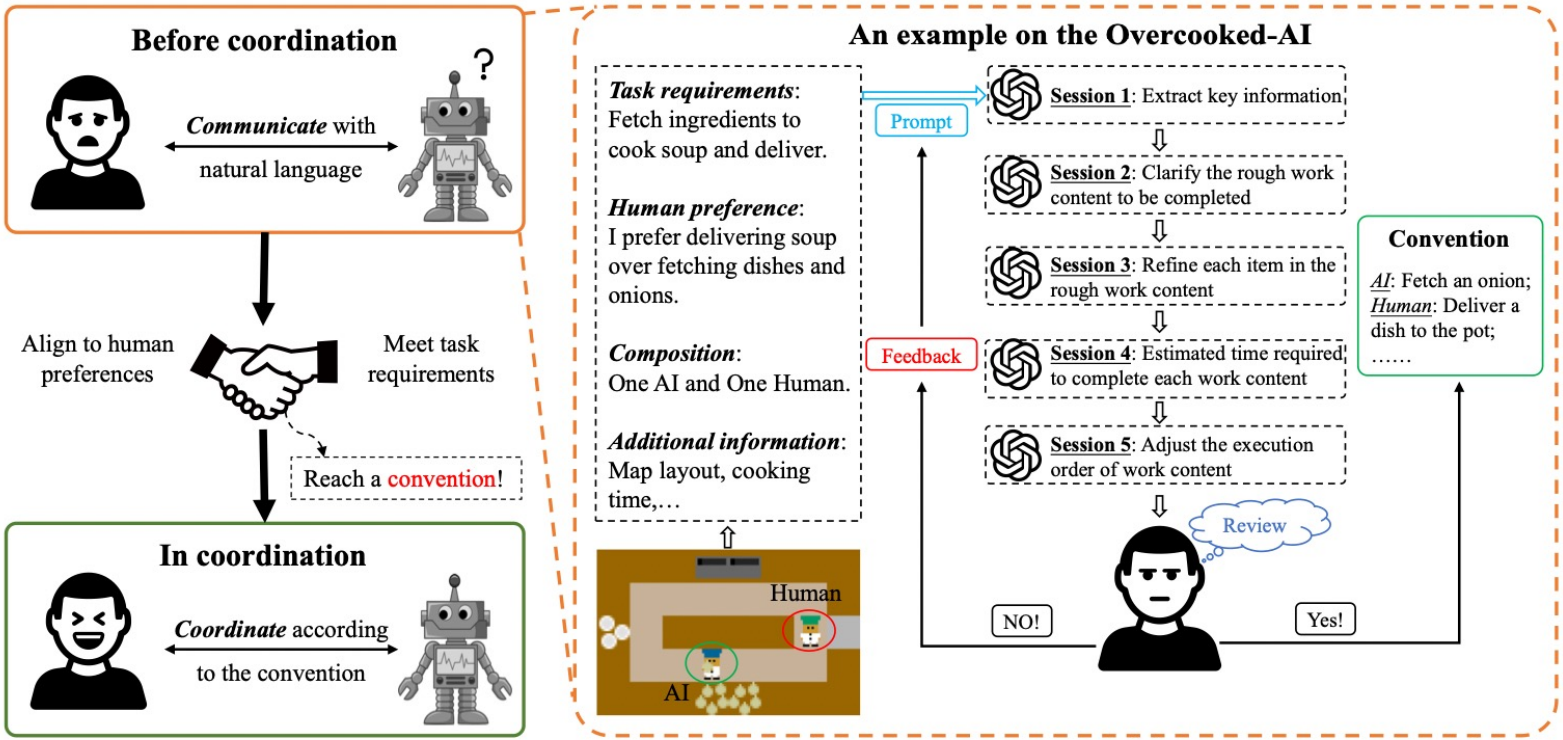
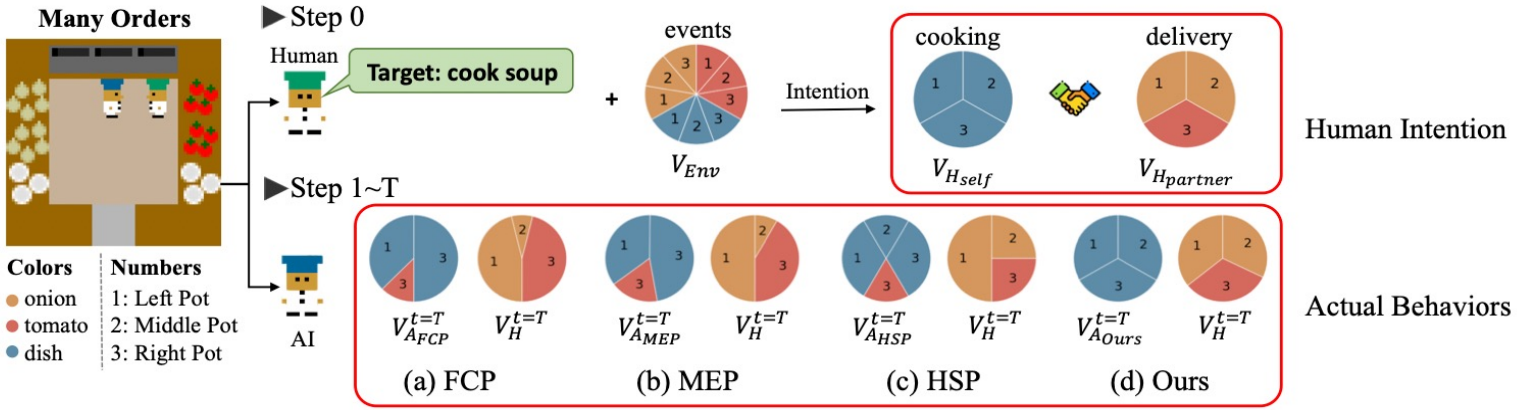


Figure 1: Overview of our proposed HAPLAN on the Overcooked-AI.

Cong Guan, et al. Efficient Human-AI Coordination via Preparatory Language-based Convention. Submitted.

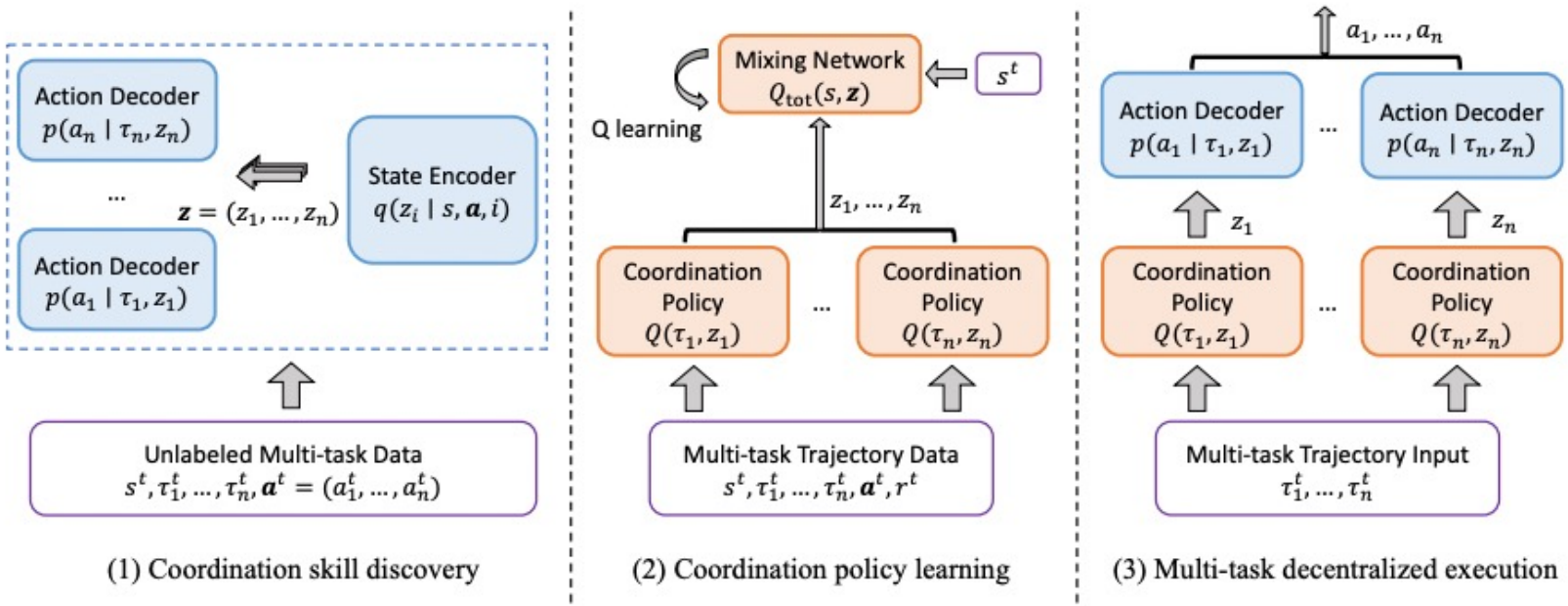
Openness in MARL: LLMs for Human-AI Coordination

Layout	Partner	FCP	MEP	HSP	HAPLAN
Counter Circle	Onion Placement Delivery	104.38±9.66 86.88±9.49	133.75±20.27 83.12±7.26	135.38±15.19 96.25±7.81	**140.00±26.92** **103.75±10.53**
Asymmetric Advantages	Onion Placement & Delivery (Pot1) Delivery (Pot2)	233.13±17.75 215.00±16.58	256.25±18.66 250.00±19.36	**282.88±17.03** 258.13±21.71	260.63±18.36 **268.00±9.79**
Soup Coordination	Onion Placement & Delivery Tomato Place & Delivery	199.38±6.09 44.38±29.04	105.00 ± 32.78 192.50±9.68	198.75±4.84 128.12±30.76	**219.38±3.47** **220.63±3.47**
Distant Tomato	Tomato Placement Tomato Place & Delivery	38.75±30.79 175.62±24.35	27.50±27.27 180.00±22.36	148.75±68.36 198.12±37.20	**210.00±15.00** **251.25±23.41**
Many Orders	Tomato Placement Delivery	140.62±32.59 194.38±12.48	170.00±33.91 175.63±35.61	248.75±29.55 208.13±25.42	**256.36±35.99** **241.21±12.97**



Cong Guan, et al. Efficient Human-AI Coordination via Preparatory Language-based Convention. Submitted.

Openness in MARL: Learning From Offline Data



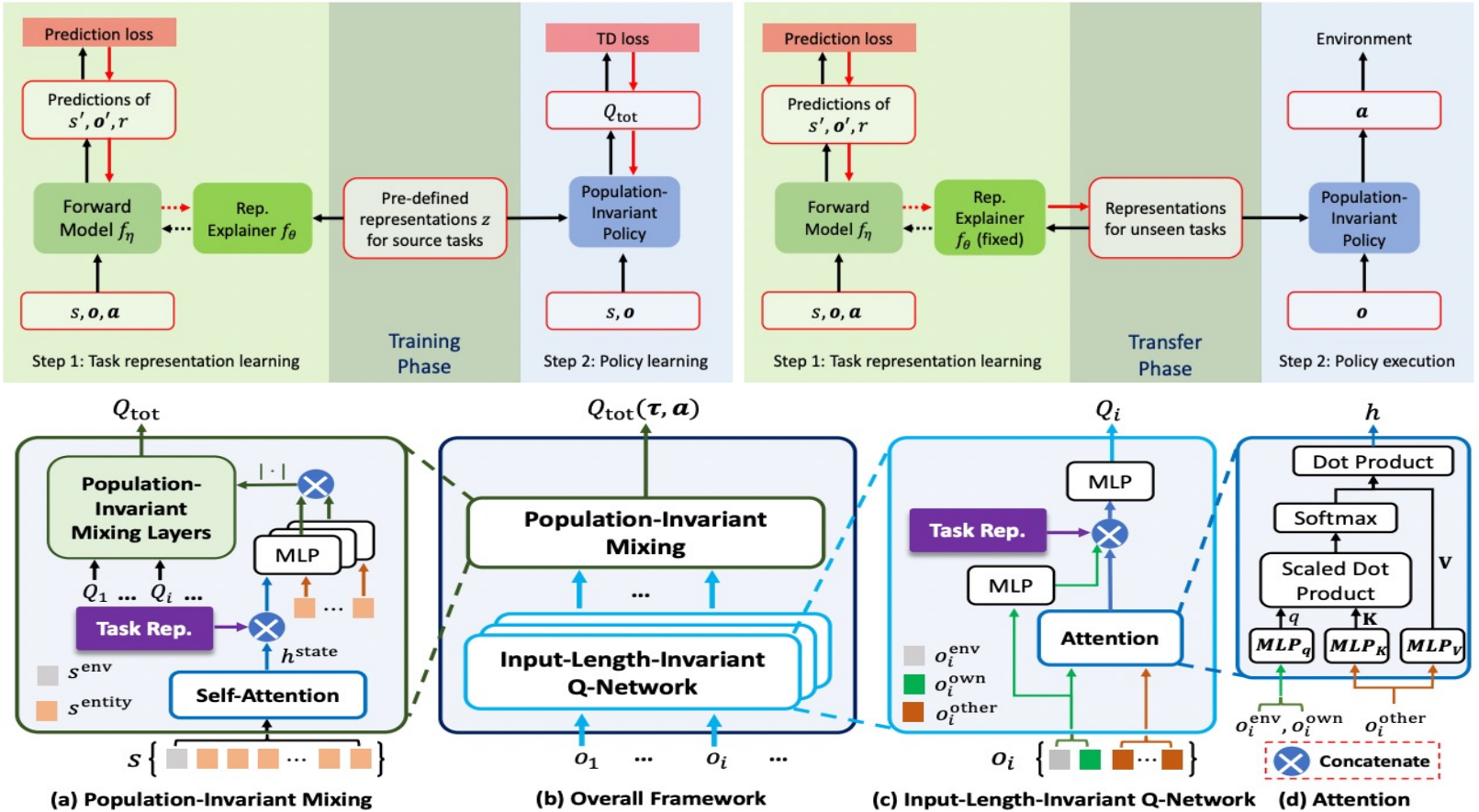
Fuxiang Zhang, Chengxing Jia, Yi-Chen Li, Lei Yuan, Yang Yu, Zongzhang Zhang. "Discovering generalizable multi-agent coordination skills from multi-task offline data." The Eleventh International Conference on Learning Representations. 2022.

Openness in MARL: Learning From Offline Data

Task	Expert				Medium			
	BC-best	UPDeT-l	UPDeT-m	ODIS (ours)	BC-best	UPDeT-l	UPDeT-m	ODIS (ours)
Source tasks								
3m	97.7 ± 2.6	71.0 ± 16.6	82.8 ± 16.0	**98.4 ± 2.7**	65.4 ± 14.7	56.6 ± 14.2	51.2 ± 3.4	**85.9 ± 10.5**
5m6m	50.4 ± 2.3	12.1 ± 12.6	17.2 ± 28.0	**53.9 ± 5.1**	21.9 ± 3.4	5.6 ± 4.8	6.3 ± 4.9	**22.7 ± 7.1**
9m10m	**95.3 ± 1.6**	26.6 ± 12.0	3.1 ± 5.4	80.4 ± 8.7	63.8 ± 10.9	34.4 ± 13.9	28.5 ± 10.2	**78.1 ± 3.8**
Unseen Tasks								
4m	92.1 ± 3.5	28.6 ± 21.6	33.0 ± 27.1	**95.3 ± 3.5**	48.8 ± 21.1	21.6 ± 17.2	14.1 ± 5.2	**61.7 ± 17.7**
5m	87.1 ± 10.5	40.1 ± 25.9	33.6 ± 40.2	**89.1 ± 10.0**	76.6 ± 14.1	77.4 ± 16.0	67.2 ± 21.3	**85.9 ± 11.8**
10m	90.5 ± 3.8	33.9 ± 25.2	54.7 ± 44.4	**93.8 ± 2.2**	56.2 ± 20.6	36.8 ± 20.7	32.9 ± 11.3	**61.3 ± 11.3**
12m	**70.8 ± 15.2**	10.9 ± 18.9	17.2 ± 28.0	58.6 ± 11.8	24.0 ± 10.5	4.0 ± 5.3	3.2 ± 3.8	**35.9 ± 8.1**
7m8m	18.8 ± 3.1	0.8 ± 1.4	0.0 ± 0.0	**25.0 ± 15.1**	1.6 ± 1.6	2.4 ± 2.6	0.0 ± 0.0	**28.1 ± 22.0**
8m9m	15.8 ± 3.3	1.6 ± 1.6	0.0 ± 0.0	**19.6 ± 6.0**	3.1 ± 3.8	3.1 ± 3.1	2.3 ± 2.6	**4.7 ± 2.7**
10m11m	**45.3 ± 11.1**	0.8 ± 1.4	0.0 ± 0.0	42.2 ± 7.2	19.7 ± 8.9	2.4 ± 1.4	4.0 ± 3.4	**29.7 ± 15.4**
10m12m	1.0 ± 1.5	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**
13m15m	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 2.6**	0.6 ± 1.3	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**
	Medium-Expert				Medium-Replay			
Source Tasks								
3m	67.7 ± 23.7	50.1 ± 23.9	**85.2 ± 17.9**	73.6 ± 22.0	81.1 ± 8.8	27.3 ± 25.9	41.4 ± 20.1	**83.6 ± 14.0**
5m6m	**31.3 ± 6.3**	2.3 ± 2.6	1.6 ± 1.6	9.4 ± 2.2	**25.0 ± 3.1**	0.8 ± 1.4	0.8 ± 1.4	16.6 ± 4.7
9m10m	26.0 ± 13.9	27.7 ± 24.1	24.3 ± 18.7	**31.3 ± 14.5**	33.4 ± 13.1	2.3 ± 4.1	0.8 ± 1.4	**34.4 ± 8.0**
Unseen Tasks								
4m	81.3 ± 18.9	41.0 ± 8.0	43.9 ± 39.0	**82.8 ± 13.5**	**61.5 ± 9.0**	23.4 ± 15.5	35.9 ± 12.6	55.6 ± 14.5
5m	74.0 ± 2.9	65.7 ± 10.1	33.6 ± 40.2	**82.8 ± 17.7**	75.0 ± 24.2	54.7 ± 23.5	61.7 ± 20.3	**96.1 ± 4.1**
10m	78.1 ± 6.7	39.8 ± 20.1	32.8 ± 38.1	**82.8 ± 16.8**	82.4 ± 8.2	8.6 ± 8.7	11.0 ± 7.8	**84.4 ± 15.1**
12m	64.8 ± 24.3	9.4 ± 7.9	9.4 ± 8.6	**81.3 ± 20.6**	83.4 ± 4.5	2.3 ± 4.1	2.3 ± 2.6	**84.4 ± 6.6**
7m8m	13.3 ± 4.5	4.0 ± 4.2	2.3 ± 4.1	**15.6 ± 4.4**	7.3 ± 6.4	2.3 ± 2.6	1.6 ± 2.7	**9.4 ± 2.2**
8m9m	10.2 ± 4.6	5.6 ± 4.8	9.5 ± 8.6	**10.9 ± 4.7**	11.5 ± 3.9	0.8 ± 1.4	0.8 ± 1.4	**11.7 ± 8.7**
10m11m	26.6 ± 4.7	8.0 ± 12.2	11.8 ± 8.1	**33.6 ± 8.9**	**46.8 ± 6.6**	2.3 ± 4.1	0.8 ± 1.4	35.9 ± 5.2
10m12m	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**1.6 ± 1.6**	1.6 ± 2.7	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 1.4**
13m15m	0.8 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	**2.3 ± 2.6**	1.6 ± 1.6	0.0 ± 0.0	0.0 ± 0.0	**2.4 ± 1.4**

Fuxiang Zhang, Chengxing Jia, Yi-Chen Li, Lei Yuan, Yang Yu, Zongzhang Zhang. "Discovering generalizable multi-agent coordination skills from multi-task offline data." The Eleventh International Conference on Learning Representations. 2022.

Openness in MARL: Policy Transfer



Rong-Jun Qin, Feng Chen, Tonghan Wang, Lei Yuan, Xiaoran Wu, Yipeng Kang, Zongzhang Zhang, Chongjie Zhang, Yang Yu . (2022). Multi-agent policy transfer via task relationship modeling. arXiv preprint arXiv:2203.04482.

Openness in MARL: Policy Transfer

Table 2 Transfer performance (mean win rates with variance) on the second series of SMAC maps.

	Source Tasks			Unseen Tasks				
	MMM	MMM2	MMM4	MMM0	MMM1	MMM3	MMM5	MMM6
MATTAR	**1.00**±0.00	**0.92**±0.20	**0.93**±0.12	**0.98**±0.02	**0.97**±0.04	**0.86**±0.10	**0.47**±0.15	**0.09**±0.02
w/o task rep.	0.94±0.05	0.23±0.39	0.33±0.25	0.81±0.15	0.37±0.36	0.07±0.05	0.22±0.30	**0.09**±0.17
0 task rep.	0.61±0.07	0.07±0.06	0.21±0.22	0.28±0.19	0.11±0.13	0.08±0.10	0.08±0.12	0.02±0.04
UPDeT-b	**1.00**±0.00	0.78±0.04	0.41±0.14	0.73±0.21	0.84±0.07	0.57±0.15	0.00±0.00	0.00±0.00
UPDeT-m	0.48±0.03	0.15±0.19	0.20±0.07	0.30±0.16	0.27±0.13	0.28±0.08	0.00±0.00	0.00±0.00
REFIL	0.97±0.01	0.04±0.02	0.06±0.03	0.93±0.02	0.38±0.06	0.12±0.04	0.00±0.00	0.00±0.00

Table 3 Transfer performance (mean win rates with variance) on the third series of SMAC maps.

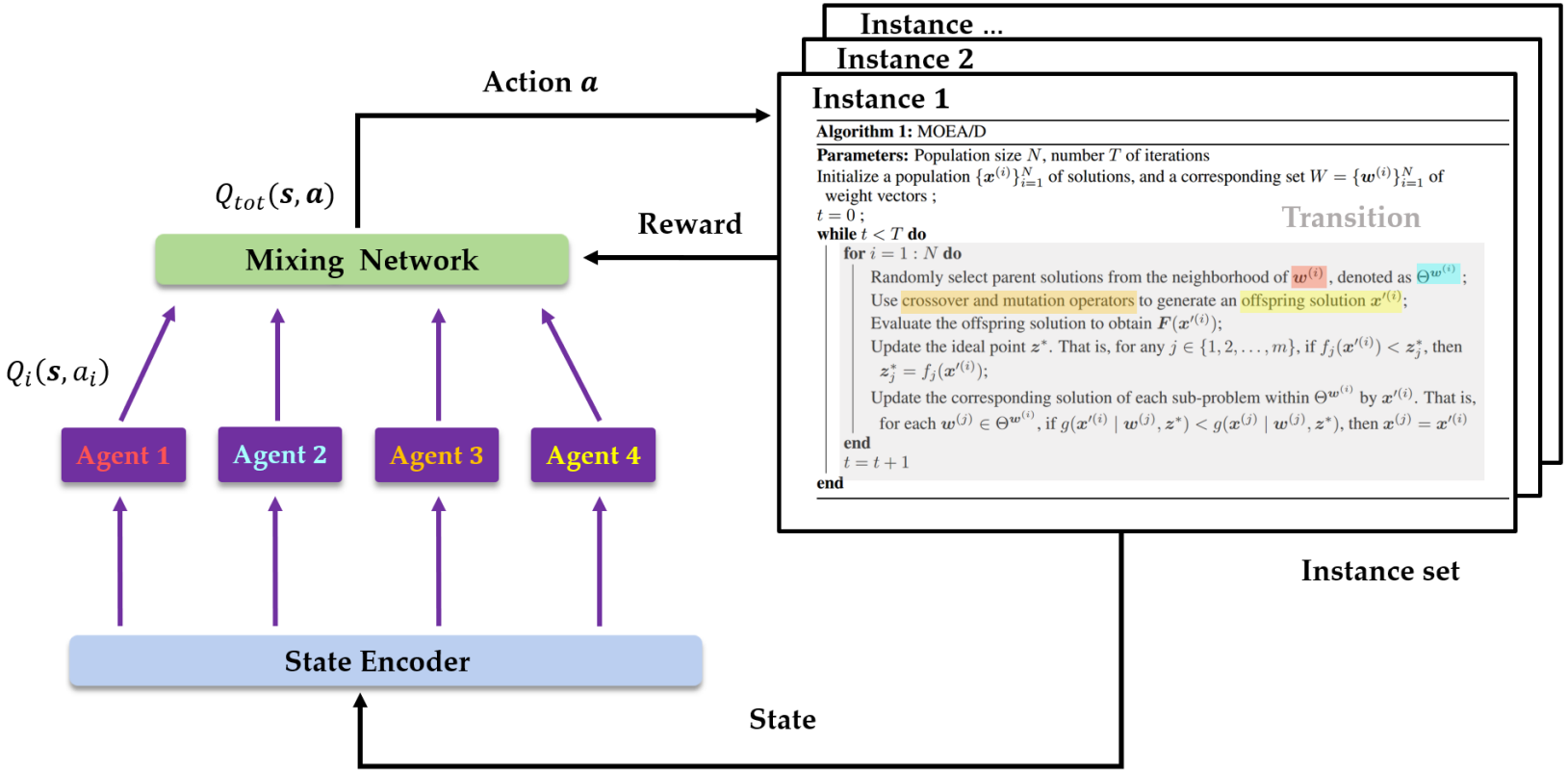
	Source Tasks				Unseen Tasks			
	5m	5m_6m	8m_9m	10m_11m	3m	4m	4m_5m	6m
MATTAR	**1.00**±0.00	0.72±0.05	**0.83**±0.05	0.81±0.09	**0.94**±0.27	**0.97**±0.02	0.04±0.05	**1.00**±0.00
w/o task rep.	0.97±0.01	0.01±0.02	0.01±0.01	0.01±0.03	0.86±0.03	0.88±0.04	0.00±0.00	0.95±0.03
0 task rep.	0.78±0.39	0.16±0.12	0.30±0.24	0.40±0.28	0.00±0.00	0.21±0.15	0.01±0.01	0.67±0.47
UPDeT-b	**1.00**±0.00	**0.93**±0.05	0.81±0.19	**0.94**±0.04	0.81±0.08	0.95±0.06	**0.29**±0.17	**1.00**±0.00
UPDeT-m	0.77±0.09	0.32±0.03	0.35±0.05	0.43±0.02	0.36±0.04	0.57±0.03	0.10±0.06	0.91±0.09
REFIL	0.73±0.03	0.00±0.00	0.01±0.01	0.03±0.02	0.68±0.06	0.74±0.02	0.00±0.00	0.71±0.02

	Unseen Tasks							
	6m_7m	7m	7m_8m	8m	9m	9m_10m	10m	10m_12m
MATTAR	0.74±0.15	**1.00**±0.00	**0.83**±0.04	**1.00**±0.00	**1.00**±0.00	**0.84**±0.09	**1.00**±0.00	**0.07**±0.01
w/o task rep.	0.03±0.02	0.94±0.03	0.08±0.10	0.93±0.04	0.86±0.05	0.04±0.02	0.52±0.22	0.00±0.00
0 task rep.	0.31±0.22	0.67±0.47	0.49±0.35	0.67±0.47	0.66±0.46	0.32±0.24	0.65±0.46	0.00±0.00
UPDeT-b	**0.78**±0.05	0.99±0.01	0.73±0.11	0.99±0.02	0.99±0.01	0.80±0.16	0.99±0.01	**0.07**±0.04
UPDeT-m	0.35±0.10	0.92±0.03	0.38±0.05	0.83±0.05	0.66±0.11	0.33±0.09	0.17±0.08	0.03±0.02
REFIL	0.01±0.00	0.66±0.03	0.01±0.01	0.63±0.05	0.55±0.05	0.01±0.00	0.46±0.02	0.00±0.00

ICLR 2022

Rong-Jun Qin, Feng Chen, Tonghan Wang, Lei Yuan, Xiaoran Wu, Yipeng Kang, Zongzhang Zhang, Chongjie Zhang, Yang Yu . (2022). Multi-agent policy transfer via task relationship modeling. arXiv preprint arXiv:2203.04482.

Openness in MARL: Application



Ke Xue, Jiacheng Xu, Lei Yuan, Miqing Li, Chao Qian, Zongzhang Zhang, Yang Yu. Multi-agent dynamic algorithm configuration[J]. Advances in Neural Information Processing Systems, 2022, 35: 20147-20161.

Openness in MARL: Application

Table 2: IGD values obtained by MOEA/D, DQN, MA-UCB and MA-DAC on different problems. Each result consists of the mean and standard deviation of 30 runs. The best mean value on each problem is highlighted in **bold**. The symbols '+', '−' and '≈' indicate that the result is significantly superior to, inferior to, and almost equivalent to MA-DAC, respectively, according to the Wilcoxon rank-sum test with confidence level 0.05.

Problem	M	MOEA/D	DQN	MA-UCB	MA-DAC
DTLZ2	3	4.605E-02 (3.54E-04) −	4.628E-02 (2.96E-04) −	4.671E-02 (3.70E-04) −	**3.807E-02** (5.05E-04)
	5	3.006E-01 (1.55E-03) −	3.016E-01 (1.34E-03) −	3.041E-01 (1.69E-03) −	**2.442E-01** (1.26E-02)
	7	4.455E-01 (1.41E-02) −	4.671E-01 (1.15E-02) −	4.826E-01 (9.59E-03) −	**3.944E-01** (1.17E-02)
WFG4	3	5.761E-02 (5.41E-04) −	6.920E-02 (1.20E-03) −	7.165E-02 (1.83E-03) −	**5.200E-02** (1.19E-03)
	5	3.442E-01 (1.21E-02) −	2.810E-01 (6.86E-03) −	2.859E-01 (6.77E-03) −	**1.868E-01** (2.81E-03)
	7	4.529E-01 (1.79E-02) −	3.725E-01 (1.14E-02) −	3.868E-01 (1.54E-02) −	**3.033E-01** (3.66E-03)
WFG6	3	6.938E-02 (5.50E-03) −	6.834E-02 (1.78E-02) −	6.601E-02 (1.00E-02) −	**4.831E-02** (8.95E-03)
	5	3.518E-01 (2.82E-03) −	3.160E-01 (2.40E-02) −	3.359E-01 (1.47E-02) −	**1.942E-01** (6.90E-03)
	7	4.869E-01 (3.03E-02) −	4.322E-01 (2.95E-02) −	4.389E-01 (3.41E-02) −	**3.112E-01** (4.93E-03)
Train: +/−/≈		0/9/0	0/9/0	0/9/0	
DTLZ4	3	6.231E-02 (8.85E-02) ≈	**5.590E-02** (5.77E-03) −	6.011E-02 (5.08E-03) −	6.700E-02 (6.14E-02)
	5	3.133E-01 (4.45E-02) ≈	3.457E-01 (1.61E-02) −	3.492E-01 (1.69E-02) −	**2.995E-01** (2.10E-02)
	7	4.374E-01 (2.57E-02) −	4.552E-01 (1.47E-02) −	4.756E-01 (2.01E-02) −	**4.182E-01** (1.21E-02)
WFG5	3	6.327E-02 (1.10E-03) −	6.212E-02 (5.54E-04) −	6.118E-02 (7.03E-04) −	**4.730E-02** (7.89E-04)
	5	3.350E-01 (9.77E-03) −	3.077E-01 (6.36E-03) −	3.036E-01 (8.83E-03) −	**1.811E-01** (3.02E-03)
	7	4.101E-01 (2.08E-02) −	4.996E-01 (1.32E-02) −	5.024E-01 (1.38E-02) −	**3.206E-01** (8.04E-03)
WFG7	3	5.811E-02 (6.31E-04) −	5.930E-02 (7.32E-04) −	6.014E-02 (7.11E-04) −	**4.066E-02** (5.31E-04)
	5	3.572E-01 (5.47E-03) −	2.993E-01 (1.43E-02) −	3.207E-01 (1.71E-02) −	**1.858E-01** (2.12E-03)
	7	5.236E-01 (2.19E-02) −	4.576E-01 (2.38E-02) −	4.879E-01 (2.75E-02) −	**3.258E-01** (1.25E-02)
WFG8	3	8.646E-02 (3.44E-03) −	9.280E-02 (1.06E-03) −	9.612E-02 (1.48E-03) −	**7.901E-02** (1.19E-03)
	5	4.258E-01 (8.42E-03) −	3.969E-01 (1.26E-02) −	3.956E-01 (1.32E-02) −	**2.479E-01** (7.20E-03)
	7	5.816E-01 (1.30E-02) −	5.575E-01 (1.39E-02) −	5.642E-01 (1.38E-02) −	**4.127E-01** (5.93E-03)
WFG9	3	5.817E-02 (1.24E-03) −	5.628E-02 (7.29E-04) −	7.953E-02 (2.45E-02) −	**4.159E-02** (6.10E-04)
	5	3.633E-01 (1.20E-02) −	3.258E-01 (1.61E-02) −	3.396E-01 (1.55E-02) −	**1.832E-01** (7.10E-03)
	7	5.538E-01 (2.63E-02) −	5.115E-01 (2.15E-02) −	5.227E-01 (1.79E-02) −	**3.278E-01** (7.21E-03)
Test: +/−/≈		0/13/2	0/15/0	0/15/0	

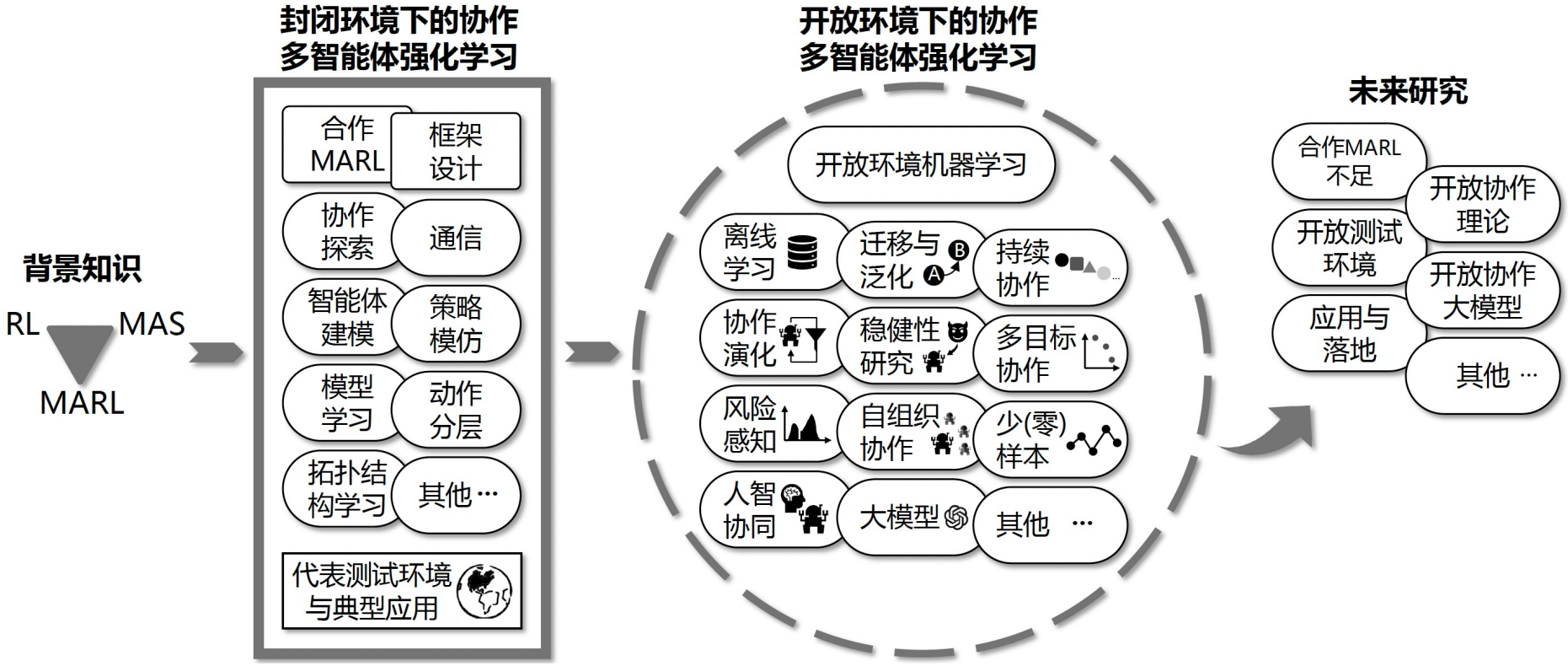
Train on DTLZ2, WFG4, and WFG6 with m objectives, and test on the other problems with m objectives

Significantly better on almost all the 24 problems

Good generalization ability

Ke Xue, Jiacheng Xu, Lei Yuan, Miqing Li, Chao Qian, Zongzhang Zhang, Yang Yu. Multi-agent dynamic algorithm configuration[J]. Advances in Neural Information Processing Systems, 2022, 35: 20147-20161.

总结


背景知识
RL
MAS
MARL
封闭环境下的协作
多智能体强化学习
合作MARL
框架设计
协作探索
通信
智能体建模
策略模仿
模型学习
动作分层
拓扑结构学习
其他…
代表测试环境与典型应用
开放环境下的协作
多智能体强化学习
开放环境机器学习
离线学习
迁移与泛化
持续协作
协作演化
稳健性研究
多目标协作
风险感知
自组织协作
少(零)样本
人智协同
大模型
其他 …
未来研究
合作MARL不足
开放协作理论
开放测试环境
开放协作大模型
应用与落地
其他 …

Thanks !

Q&A