arXiv:2305.05909v1 [cs.MA] 10 May 2023

通过进化生成辅助对抗攻击者实现鲁棒的多智能体协调

预印本

袁磊[1,2], **张子谦**[1],**李力禾**[1],**薛科**[1],**尹浩**[1],**陈锋**[1],**关聪**[1],**李力禾**[1],**钱超**[1],**Yang Yu**[1,2,*] 中国南京南京大学新型软件技术国家重点实验室[2] Polixir.ai{yuanl, zhangzq, xuek, yinh, chenf, guanc, lilh}@lamda.nju.edu.cn, {qianc,yuy}@nju.edu.cn

摘要

合作式多智能体强化学习（CMARL）已显示出在许多现实应用中的广阔前景。以往研究主要致力于通过解决多智能体强化学习特有的挑战（如非平稳性、信用分配和可扩展性）来提升协调能力，却忽略了在不同环境中测试时的策略扰动问题。问题建模和高效算法设计均未考虑这一问题。为此，我们首先将问题建模为有限策略对抗者去中心化部分可观测马尔可夫决策过程（LPA-Dec-POMDP）：团队中的部分协调者可能意外且不可预测地遭遇数量有限的恶意动作攻击，而常规协调者仍努力实现预定目标。随后，我们提出**鲁棒多**智**体**协**调**的**进**化**生**成辅助对抗攻击者方法（ROMANCE），使训练后的策略在训练过程中遭遇多样且强大的辅助对抗攻击，从而在各种策略扰动下实现高鲁棒性。具体而言，为避免自我系统对某个特定攻击者过拟合，我们维护一组攻击者，并对其进行优化，以确保攻击者具有较强的攻击质量和行为多样性。攻击质量旨在最大限度地削弱自我系统的协调效果；此外，我们采用一种基于稀疏动作的新型多样性正则项，以增加攻击者之间的行为差异。随后，将自我系统与从所维护的攻击者集合中选出的攻击者种群配对，并与持续进化的攻击者交替训练。在SMAC的多个场景上开展的大量实验表明，与其他基线方法相比，我们提出的ROMANCE具有相当或更强的鲁棒性和泛化能力。

1 引言

近年来，合作式多智能体强化学习（CMARL）受到广泛关注 [Hernandez-Leal 等，2019，Gronauer 和 Diepold，2022] 并展现出在众多领域中的应用潜力，例如自动驾驶车辆团队[Peng 等人， 2021], 多智能体路径规划 [Greshler 等人， 2021], 多无人机控制 [Yun 等人， 2022], 以及动态算法配置 [Xue 等人，2022b]。现有的合作式多智能体强化学习方法主要致力于解决特定挑战，例如非平稳性 [Papoudakis 等人， 2019], 信用分配 [Wang 等，2021a], 和可扩展性 [Christianos 等人， 2021] 以提升复杂场景中的协调能力。无论是基于价值的方法 [Sunehag 等，2018，Rashid 等，2018，Wang 等，2021b] ，还是基于策略梯度的方法 [Foerster 等，2018，Lowe 等，2017，Yu 等人， 2022] 都已在广泛的任务中展现出显著的协调能力（例如 SMAC [Samvelyan 等人，2019] 和花火 [Yu 等人， 2022]）。尽管取得了这些成功，主流的合作式多智能体强化学习方法仍难以应用于现实世界，因为它们主要是在无差异的环境中考虑策略的训练和测试。因此，当训练环境与测试环境存在任何差异时，这些方法学习到的策略可能会出现性能下降 [Guo 等人，2022]。

在部署前训练鲁棒策略对于解决上述问题具有重要潜力，并且在单智能体强化学习（SARL）领域已取得显著进展 [Moos 等人，2022，Xu 等人， 2022]。以往的研究通常

*通讯作者

采用对抗训练范式来获得鲁棒策略。这些方法通常从博弈论的角度将策略学习过程建模为一个极小极大问题 [Yu 等人， 2021] 并在最坏情况下对策略进行优化 [Pinto 等，2017，Zhang 等，2020a， 2022]。然而，多智能体问题要复杂得多 [Zhang 等， 2021], 因为多个智能体会在环境中同时做出决策。此外，近期研究表明，多智能体强化学习系统通常容易受到各种攻击 [Guo 等人， 2022]。一些多智能体强化学习研究从多个方面探讨了鲁棒性, 包括局部观测的不确定性 [Lin 等人， 2020], 模型函数 [Zhang 等，2020a], 以及消息传递 [Xue 等，2022c]。上述方法要么侧重于从不同方面研究鲁棒性，要么采用单智能体强化学习中使用的启发式规则和正则化器等技术来训练鲁棒协调策略。然而，在合作式多智能体强化学习中，不可预测的恶意动作攻击如何导致策略扰动，尚未得到充分探索。

在本研究中，我们旨在开发一种鲁棒的合作式多智能体强化学习框架，以应对团队中部分协调者遭受恶意动作攻击的情况。具体而言，我们将该问题建模为受限策略对手分散式部分可观测马尔可夫决策过程（LPA-Dec-POMDP）：部分协调者可能遭受恶意动作攻击，而常规协调者仍应努力完成既定目标。

为开发此类鲁棒策略，我们提出了 ROMANCE，这是一种基于进化生成辅助攻击者的对抗训练范式。具体而言，在所有已生成的攻击者中，我们保留一组攻击质量高且行为多样的攻击者，以避免自我系统对某个特定攻击者过拟合。其中，高攻击质量要求攻击者最小化自我系统的奖励，而多样性则指在攻击者之间生成不同的行为。我们还引入了稀疏动作正则化器，以促进不同攻击者之间的行为多样性。此外，为防止攻击者对自我系统完成既定任务设置过于棘手的障碍，我们将攻击总次数限制为固定值。对于自我系统的训练，我们从维护的集合中选取一组攻击者与其配对，以完成给定任务；随后，根据定制的质量分数和多样性距离，迭代选择并更新攻击者种群。最终，我们在不同类型和程度的动作扰动下获得了高度鲁棒的协调策略。

为了评估所提出的方法，我们在 SMAC [Samvelyan 等人， 2019] 的多个地图上进行了大量实验，并将 ROMANCE 与多种基线方法进行了比较。实验结果表明，我们提出的对抗训练范式确实能够获得攻击能力强且行为多样的攻击者。此外，针对该种群训练的协调策略，在测试时面对不同数量的攻击时，能够表现出较高的鲁棒性和泛化能力。此外，可视化实验表明，ROMANCE 如何提升策略在恶意动作扰动下的鲁棒性。

2 相关工作

多智能体强化学习（MARL）近年来取得了显著进展 [Hernandez-Leal 等，2019，Gronauer 和 Diepold， 2022]。许多方法已成为促进智能体之间协调的高效手段，其中大多数方法大致可分为基于策略的方法和基于价值的方法。MADDPG [Lowe 等人，2017], COMA [Foerster 等人， 2018], DOP [Wang 等，2021c], 以及 MAPPO [Yu 等人， 2022]是典型的基于策略梯度的方法，旨在探索多智能体策略梯度方法的优化。MADDPG 采用集中式训练分散式执行（CTDE）范式来训练策略，并通过 DDPG [Lillicrap 等人， 2016]对每个策略进行优化。COMA 同样采用集中式评论家来优化策略，但使用反事实模型计算多智能体系统中每个智能体的边际贡献。DOP 进一步采用集中式线性混合网络，在合作系统中分解全局奖励，并显著提升了 MADDPG 和 COMA 的性能。近年来，MAPPO 将近端策略优化技术在单智能体强化学习中得到广泛验证的学习效率应用于多智能体强化学习。另一类多智能体强化学习方法，即基于价值的方法，主要关注价值函数的分解。VDN [Sunehag 等人， 2018]旨在通过简单的加法分解，将团队价值函数分解为各个智能体对应的价值函数。遵循个体-全局最大化（IGM）原则 [Son 等人， 2019], QMIX [Rashid 等人， 2018]通过学习非线性混合网络改进了价值函数的分解方式，该网络近似实现单调函数价值分解。QPLEX [Wang 等，2021b] 采用双路对决网络架构对联合价值函数进行分解，从而实现了 IGM 的完全表达能力。Wang 等人 [2021a] 最近通过应用多智能体拟合Q迭代算法, 对 IGM 进行了理论分析。有关多智能体强化学习的更多细节和最新进展，可参阅综述文献 [Zhang 等，2021，Canese 等，2021，Zhu 等， 2022]。

对抗训练在强化学习鲁棒性方面展现出良好的前景 [Moos 等人， 2022], 其中涉及在不同情况下发生的扰动，例如状态、奖励、策略等。随后，这些方法以对抗的方式训练强化学习策略，以便在最坏情况下获得鲁棒策略。鲁棒对抗强化学习（RARL） [Pinto 等人， 2017] 选取对手作用于其上的特定机器人关节，以找到极小极大目标的均衡点

采用替代学习对手。RARARL [Pan 等， 2019] 通过引入风险规避型鲁棒对抗强化学习，进一步训练风险规避主角和风险寻求对手；与未使用对手训练的智能体相比，这种方法在自动驾驶车辆控制器上显著减少了碰撞。上述方法仅学习单个对手，而且在对手的标准参数化设置下，这种方法无法始终实现对动力学变化的鲁棒性。随后，RAP [Vinitsky 等， 2020] 和 GC [Song 和 Schneider， 2022] 将基于群体的增强引入 RobustRL 框架。详细综述请参阅 [Ilahi 等，2021年，Moos 等， 2022] ；近期进展请参阅 [Smirnova 等，2019年，Zhang 等，2020a、b，Oikarinen 等，2021年，Xie 等，2022]。

鲁棒多智能体强化学习近年来受到了广泛关注 [Guo 等人， 2022]。M3DDPG [Li 等人， 2019] 学习了 MADDPG [Lowe 等人， 2017] 的极小极大扩展，并以对抗方式训练多智能体强化学习策略，这表明其在解决由对手策略变化引起的局部最优问题方面具有潜力。为了对模型知识不准确所导致的不确定性进行建模，R-MADDPG [Zhang 等，2020c] 引入了鲁棒纳什均衡的概念，并将不确定性视为一个自然智能体，在面对奖励不确定性时展现出显著优势。针对合作式多智能体强化学习中的观测扰动，Lin 等人 [2020] 学习了一种对抗观测策略来攻击系统，表明自我系统极易受到观测扰动的影响。RADAR [Phan 等人， 2021] 通过对抗价值分解学习韧性多智能体强化学习策略。Hu 和 Zhang [2022] 进一步设计了一种动作正则化器，以高效攻击合作式多智能体强化学习系统。Xue 等人 [2022最近考虑了多智能体对抗通信问题，在部分消息发送者受到污染时，学习鲁棒通信策略。据我们所知，先前尚无研究在 LPA-Dec-POMDP 框架下探索合作式多智能体强化学习，无论是在问题建模方面，还是在高效算法设计方面。

此外，另一些研究关注与不同队友协调时的鲁棒性，即临时团队协作 [Stone 等，2010，Gu 等，2022，Mirsky 等， 2022], 或零样本协调（ZSC） [Hu 等，2020，Lupu 等，2021，Xue 等，2022a]。前一类方法旨在创建一种自主智能体，使其能够在各团队成员都具备单独贡献能力的任务中，与此前未知的队友高效且稳健地协作。而在 ZSC 设定中，这种临时团队协作的特殊情形下，智能体朝着共同目标努力，并在每个步骤获得相同奖励。对抗攻击的引入使受害者成为常规智能体的未知队友，而情况更具挑战性，因为未知队友可能会执行破坏性行动。我们提出的方法进一步推动了鲁棒合作式多智能体强化学习朝这一方向发展。

3 问题表述

本文考虑在分散式部分可观测马尔可夫决策过程 [Oliehoek 和 Amato，2016], 框架下的合作式多智能体强化学习任务，其定义为一个元组 $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega\ O, R, \gamma \rangle$。其中， $\mathcal{N} = \{1, \ldots, n\}$ 是智能体集合， $\mathcal{S}$ 是全局状态集合， $\mathcal{A} = \mathcal{A}^1 \times ... \times \mathcal{A}^n$ 是联合动作集合，Ω 是观测集合， $\gamma \in [0, 1)$ 表示折扣因子。在每个时间步，智能体 i 接收观测 $o^i = O(s, i)$ 并输出动作 $a^i \in \mathcal{A}^i$。联合动作 $\boldsymbol{a} = (a^1, ..., a^n)$ 导致下一个状态 $s' \sim P(\cdot|s, \boldsymbol{a})$，并产生全局奖励 $R(s, \boldsymbol{a}, s')$。为缓解部分可观测性，我们将智能体 i 截至时间步 t 的历史 $(o_i^1, a_i^1, \ldots, o_i^{t-1}, a_i^{t-1}, o_i^t)$ 编码为 τ_i，然后结合 $\boldsymbol{\tau} = \langle \tau_1, \ldots, \tau_n \rangle$，形式化目标是寻找联合策略 $\boldsymbol{\pi}(\boldsymbol{\tau}, \boldsymbol{a})$，使全局价值函数 $Q_{tot}^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}) = \mathbb{E}_{s, \boldsymbol{a}} \left[\sum_{t=0}^{\infty} \gamma^t R(s, \boldsymbol{a}) \mid s_0 = s, \boldsymbol{a_0} = \boldsymbol{a}, \boldsymbol{\pi} \right]$ 最大化。

我们的目标是在团队中的部分协调者受到策略扰动时优化策略。合作式多智能体强化学习（CMARL）的脆弱性使其难以容忍无限数量的扰动。为了避免自我系统遭到彻底破坏，我们假设扰动的数量有限，并将这种设定形式化为一个 LPA-Dec-POMDP：

定义 *1*（有限策略对抗者去中心化部分可观测马尔可夫决策过程）。给定一个分散式部分可观测马尔可夫决策过程 $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega\ O, R, \gamma \rangle$，我们通过引入一个对抗攻击者 π_{adv}：$\mathcal{S} \times \mathcal{A} \times \mathbb{N} \to \mathcal{A}$，定义一个有限策略对抗者去中心化部分可观测马尔可夫决策过程（*LPA-Dec-POMDP*）$\hat{}\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O, R, \gamma \rangle$。该攻击者通过强制智能体执行联合动作 $\hat{}\boldsymbol{a} \sim \pi_{adv}(\cdot|s, \boldsymbol{a}, k)$ 来扰动自我智能体的策略，使得 $s' \sim P(\cdot|s, \hat{}\boldsymbol{a})$, $r = R(s, \hat{}\boldsymbol{a}, s')$。其中， $K \in \mathbb{N}$ 表示满足 $\sum_t \sum_{i \in \mathcal{N}} \mathbb{I}(\hat{a}_t^i \neq a_t^i) \leq K$ 的攻击次数， $k \leq K$ 表示当前剩余攻击次数。

为高效解决攻击问题，我们在定义2中通过将策略分解为两个部分——受害者选择和策略扰动——引入了一类解耦对抗攻击者策略。

定义2（解耦对抗攻击者策略）。 对于对抗攻击者策略 π_{adv}，如果存在受害者选择函数 v：$\mathcal{S} \times \mathbb{N} \to \Delta(\hat{}\mathcal{N})$ 和策略扰动函数 g：$\hat{}\mathcal{N} \times \mathcal{A} \times \mathbb{N} \to \Delta(\mathcal{A})$，使得以下两个等式成立：

$$\pi_{adv}(\hat{\boldsymbol{a}}|s, \boldsymbol{a}, k) = v(i|s, k) g(\hat{\boldsymbol{a}}|i, \boldsymbol{a}, k)$$
$$g(\hat{\boldsymbol{a}}|i, \boldsymbol{a}, 0) = g(\hat{\boldsymbol{a}}|null, \boldsymbol{a}, k) = \mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{a}),$$

其中 $\hat{}\mathcal{N} = \mathcal{N} \cup \{null\}$, $i \sim v(\cdot|s, k)$，则我们称 v 和 g 将 π_{adv} 解耦。

关于策略扰动函数，已有许多基于启发式的方法 [Pattanaik 等人，2018年；Tessler 等人，2019年；Sun 等人，2021] 被提出，用于为固定的强化学习策略寻找对抗性扰动。一种常见的攻击方式是在某些步骤中迫使受害者选择Q值最小的动作 [Pattanaik 等人，2018]。因此，对于策略扰动函数 g，一种有效且高效的形式可以是 $g(\hat{\boldsymbol{a}}|i,\boldsymbol{a},k) = g(\hat{a}^i,\boldsymbol{a}^{-i}|i,\boldsymbol{a},k) = \mathbb{I}(\hat{a}^i = \arg\min_{a^i} Q^i(\tau^i,a^i))$，如果 $k>0$ 且 $i \neq null$。为提高效率，在本文其余部分，我们只关注采用基于启发式的策略扰动的解耦攻击者策略 $\pi_{adv} = v \circ g$；并且不失一般性地，假设 g 通过 $\hat{\boldsymbol{a}} = g(i,\boldsymbol{a},k)$ 执行确定性扰动。

4 方法

本节将介绍 ROMANCE 的设计，这是一种在 LPA-Dec-POMDP 下学习鲁棒协调策略的新型框架。我们首先讨论每个攻击者的优化目标，然后展示如何通过专门设计的更新和选择机制，维护一组行为多样的高质量解。最后，我们提出 ROMANCE——一种交替训练范式，以提升策略扰动下 CMARL 智能体（自我系统）的鲁棒性。

4.1 攻击者优化目标

在本节中，我们将讨论如何在固定的联合自我系统策略 $\boldsymbol{\pi}$ 下训练对抗攻击者策略群体 $P_{adv} = \{\pi^j_{adv}\}_{j=1}^{n_p} = \{v^j \circ g\}_{j=1}^{n_p}$，其中 n_p 为种群规模。我们的目标是使攻击者种群具备较高的质量和多样性：质量目标要求其最小化自我系统回报，而多样性则鼓励攻击者采取不同的行为。为实现这一目标，我们首先说明如何在特定的马尔可夫决策过程设置下求解单个最优对抗攻击者，随后讨论具体的求解方法。

定理1. 给定一个 *LPA-Dec-POMDP* $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega, O, R, \gamma \rangle$、自我系统 $\bar{\quad}$ 的固定联合策略 $\boldsymbol{\pi}$ 以及一个基于启发式的策略扰动函数 g，存在一个马尔可夫决策过程 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$，使得 $\mathcal{M}$ 的最优对抗攻击者 $\hat{\bar{\pi}}^*_{adv}$ 可由 $\mathcal{M}$ 的最优策略 v^* 和 g 解耦，其中 $\mathcal{S} = \mathcal{S} \times \mathbb{N}$，$\bar{s} = (s,k)$ $\bar{s}' = (s',k')$ $k,k' \leq K$ 表示剩余攻击预算，$\mathcal{A} = \mathcal{N} \cup \{null\}$, $R(\bar{s},\bar{a},\bar{s}') = -R(s,\hat{\boldsymbol{a}},s')$,

$$\bar{P}(\bar{s}'|\bar{s},\bar{a}) = \begin{cases} 0 & k-k' \notin \{0,1\} \\ P(s'|s,\hat{\boldsymbol{a}})\mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{a}) & k-k' = 0 \\ P(s'|s,\hat{\boldsymbol{a}})\mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a}) & k-k' = 1 \end{cases},$$

其中，$\bar{d}(\bar{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$、$d$ 和 $\bar{d}$ 分别是 $\hat{\mathcal{M}}$ 和 $\bar{\mathcal{M}}$ 中初始状态的分布，而 $\boldsymbol{a} = \boldsymbol{\pi}(s)$、$\hat{\boldsymbol{a}} = g(\bar{a}, \boldsymbol{\pi}(s), k)$ 分别是自我系统的原始动作和强制动作。

定理 1 背后的直觉是，对抗攻击者旨在最大限度地降低自我系统获得的奖励。在固定的基于启发式的策略扰动函数 g 下，对抗攻击者希望找到最优受害者策略 v，以帮助确定攻击哪个智能体以及何时实施攻击。证明见附录。$\bar{\quad}$ 的构造使得可以应用现成的深度强化学习算法来求解 $\mathcal{M}$ 的最优智能体 v，从而推导出 LPA-Dec-POMDP 的最优对抗攻击者。

请注意，攻击次数是有限的，只能执行 K 次，这使得动作在智能体和时间维度上都具有稀疏性。这种稀疏动作通常在获得高奖励方面发挥着关键作用，因此会被积极利用。然而，采取稀疏动作的机会是有限的。如果攻击者在一开始就耗尽这些机会，将导致次优结果。为了引导智能体更加谨慎地采取稀疏动作，稀疏性先验正则化Q学习（SPRQ） [Pang 等人，2021] 基于参考分布为稀疏动作分配较低概率，从而构建正则化马尔可夫决策过程，并通过提出以下正则化目标来解决该任务：

$$\max_v \mathbb{E}[\sum_{t=0}^{T} \gamma^t (\bar{R}_t - \lambda D_{KL}(v(\cdot|\bar{s}_t), p_{ref}(\cdot)))], \tag{1}$$

其中，$\lambda \in \mathbb{R}$ p_{ref} 是参考分布，它为稀疏动作“攻击”分配较小的概率 δ，为“不攻击”分配概率 $1-\delta$（即在此情况下，$p_{ref}(\cdot) = (\frac{\delta}{|\mathcal{N}|}, \cdots \frac{\delta}{|\mathcal{N}|}, 1-\delta)$），而 D_{KL} 是 Kullback-Leibler（KL）散度。正如 [Pang 等人 2021],在命题 3.1 和 3.2 中所述，正则化最优策略可通过以下方式获得：

$$v(\bar{a}|\bar{s}) = \frac{p_{ref}(\bar{a}) \exp(\frac{\bar{Q}_v(\bar{s},\bar{a})}{\lambda})}{Z(\bar{s})}, \tag{2}$$

其中， $\bar{Q}_v(\bar{s},\bar{a})$是 v下的正则化 Q 函数， $Z(\bar{s}) = \sum_{\bar{a}\in\bar{\mathcal{A}}} p_{ref}(\bar{a})\exp\bar{Q}_v(\bar{s},\bar{a})\lambda$是配分函数。按照所提出的正则化贝尔曼最优算子，我们用 ϕ 对 Q 函数进行参数化，因此 SPRQ 损失函数定义如下:

$$L_{opt}(\phi) = \mathbb{E}[(\bar{Q}_\phi(\bar{s}_t,\bar{a}_t) - y)^2], \tag{3}$$

其中 $y = \bar{r}_t + \gamma\lambda$对数（$\mathbb{E}_{\bar{a}'_{t+1}\sim p_{ref}(\cdot)}$[指数（$\bar{Q}_{\phi^-}(\bar{s}_{t+1},\bar{a}'_{t+1})\lambda$）]），且 ϕ^-是目标网络的参数 .

应用上述技术可以使攻击者具有较高的攻击效率，但单个攻击者很容易对特定类型的自我系统过拟合，仍然会导致其在未见过的情境下泛化能力较差。受基于种群的训练（PBT）已被广泛证明的能力启发 [Jaderberg 等人， 2019], 我们引入多样性正则化目标，以确保整个种群具有有效的行为多样性。在相同状态下动作分布的散度，是衡量策略之间多样性的可靠指标。因此，我们使用詹森-香农散度（JSD） [Fuglede 和 Topsøe， 2004] 来反映种群的多样性$P_{adv}(\boldsymbol{\phi}) = \{\pi^{\phi_j}_{adv}\}^{n_p}_{j=1} = \{v^{\phi_j}\circ g\}^{n_p}_{j=1}$，其计算方式如下:

$$\text{JSD}(\{v^{\phi_j}(\cdot|\bar{s})\}^{n_p}_{j=1}) = \frac{1}{n_p}\sum_{j=1}^{n_p} D_{KL}(v^{\phi_j}(\cdot|\bar{s}),\bar{v}(\cdot|\bar{s})), \tag{4}$$

其中 $\bar{v}(\bar{a}|\bar{s}) = \frac{1}{n_p}v^{\phi_j}(\bar{a}|\bar{s})$ 是平均策略。因此，对于该种群，正则化目标定义如下:

$$L_{div}(\boldsymbol{\phi}) = \mathbb{E}_{\bar{s}\sim S_a}[\text{JSD}(\{v^{\phi_j}(\cdot|\bar{s})\}^{n_p}_{j=1})], \tag{5}$$

其中 $S_a = \bigcup_{j=1}^{n_p} S_a^j = \bigcup_{j=1}^{n_p}\{\bar{s}|k>0,\bar{a}\neq null,\bar{a}\sim v^{\phi_j}(\cdot|\bar{s})\}$ 是对抗攻击者选择进行攻击的状态（攻击点）并集。

考虑上述稀疏攻击和行为多样性，可推导出我们的完整损失函数 :

$$L_{adv}(\boldsymbol{\phi}) = \frac{1}{n_p}\sum_{j=1}^{n_p} L_{opt}(\phi_j) - \alpha L_{div}(\boldsymbol{\phi}), \tag{6}$$

其中， L_{opt} 和 L_{div} 分别在公式（3）和公式（5）中定义， α 是一个可调超参数，用于平衡质量与行为多样性。

4.2 攻击者的进化生成

尽管以式（6）为目标的 PBT 行之有效，自我系统仍可能对当前种群中的某些特定类型的攻击者过拟合，从而遗忘训练早期阶段出现的攻击模式。为避免这一灾难性结果，我们尝试进一步提升对手策略空间的覆盖范围。

在各种方法中，质量-多样性（QD）算法 [Cully 和 Demiris，2017，Chatzilygeroudis 等人， 2021]能够高效地获得一组高质量且行为多样的解。近年来，这类算法已被用于在强化学习中发现多样化策略 [Wang 等人， 2022], 生成环境 [Bhatt 等人， 2022] 以及伙伴 [Xue 等，2022a] 。作为一种特定类型的进化算法 [B¨ack, 1996], QD 算法通常会维护一个档案库（即迄今生成的一组性能优异且行为多样的解），并通过迭代更新和选择来模拟自然进化过程。

受质量多样性（QD）算法的启发，我们设计了专门的更新与选择机制，以生成所需的辅助对抗攻击者。我们维护一个档案库 Arc_{adv}，其最大容量为 n_a，并为每个个体（即攻击者）分配质量分数和行为。具体而言，给定一个自我智能体联合策略 $\boldsymbol{\pi}$，对抗攻击者 $\pi^{\phi_i}_{adv}$ 的质量分数定义为该攻击者的折扣累积回报:

$$\text{Quality}(\pi^{\phi_i}_{adv}) = \mathbb{E}_{\bar{\boldsymbol{\tau}}}[\sum_t \gamma^t \bar{R}(\bar{s},\bar{a})|\boldsymbol{\pi}], \tag{7}$$

其中， $\bar{\boldsymbol{\tau}}$是攻击者 $\pi^{\phi_i}_{adv}$的轨迹，而 $\bar{R}$ 在定理 1 中定义。为了描述 $\pi^{\phi_i}_{adv}$的行为，我们计算它与另一位攻击者 $\pi^{\phi_j}_{adv}$之间的距离:

$$\text{Dist}(\pi^{\phi_i}_{adv},\pi^{\phi_j}_{adv}) = \mathbb{E}_{\bar{s}\sim S_a^{i,j}}[\text{JSD}(v^{\phi_i}(\cdot|\bar{s}),v^{\phi_j}(\cdot|\bar{s}))], \tag{8}$$

其中，$S^{i,j}_a = S^i_a \cup S^j_a$ 是攻击点集合。为简便起见，我们分别用 Quality(i) 和 Dist(i, j) 表示 $\pi^{\phi_i}_{adv}$ 的质量分数以及 $\pi^{\phi_i}_{adv}$ 与 $\pi^{\phi_j}_{adv}$之间的距离，而 i, j 是它们在档案库中的索引。

在每次迭代中，我们根据质量分数采用基于适应度的选择 [Blickle 和 Thiele, 1996] 从档案库中选取 n_p对抗攻击者，并使其与自我系统进行交互。接下来，我们将式（6）中描述的优化步骤作为隐式变异算子，以生成新的攻击者。随后，通过逐一添加新生成的攻击者来更新档案库。为了保持档案库的多样性，我们避免添加具有相似行为的攻击者 [Cully 和 Mouret，2013年，Wang 等人，2022]。具体而言，每当我们想要向档案库中添加新的攻击者时，首先选择档案库中与其最相似的攻击者，并计算二者之间的行为距离。如果该距离超过某个阈值，则添加新的攻击者；否则，随机保留其中一个。需要注意的是，如果添加新攻击者后当前档案库大小超过容量 n_a，则删除最早加入的攻击者。上述过程的完整流程如附录中的算法1所示，我们将这样的一次迭代称为一代。

4.3 鲁棒性训练范式

在获得一组攻击能力强且行为多样性高的攻击者后，我们旨在针对现有攻击者学习一个鲁棒的自我系统。我们首先研究具有固定对抗攻击者 π_{adv} 的 LPA-Dec-POMDP，以及如何优化自我智能体联合策略。

定理 2. 给定一个 *LPA-Dec-POMDP* $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O, R, \gamma \rangle$，以及一个固定的确定性对抗攻击者策略 π_{adv}，存在一个 *Dec-POMDP* $\tilde{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega\ O, R, \gamma \rangle$，使得 $\tilde{\mathcal{M}}$ 的最优策略是 $\tilde{\mathcal{M}}$，即给定 π_{adv} 时 $\hat{\mathcal{M}}$ 的最优策略，其中 $\mathcal{S} = \mathcal{S} \times \mathbb{N}$, $\tilde{s} = (s, k)$, $\tilde{s}' = (s', k')$, $d(\tilde{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$, $\tilde{O}(\tilde{s}, i) = O(s, i)$, $\tilde{R}(\tilde{s}, \boldsymbol{a}, \tilde{s}') = R(s, \hat{\boldsymbol{a}}, s')$,

$$\tilde{P}(\tilde{s}'|\tilde{s}, \boldsymbol{a}) = \begin{cases} 0 & k - k' \notin \{0, 1\} \\ P(s'|s, \hat{\boldsymbol{a}}) & otherwise \end{cases},$$

其中，$\tilde{d}$和 d 分别是 $\tilde{M}$和$\hat{\mathcal{M}}$中初始状态上的分布，而$\hat{\boldsymbol{a}} = \pi_{adv}(s, \boldsymbol{a}, k)$ 表示自我系统执行的联合动作。

定理 3。 给定$\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O, R, \gamma \rangle$以及一个随机对抗攻击者策略 π_{adv}，存在一个分散式部分可观测马尔可夫决策过程 $\tilde{\mathcal{M}} = \langle \mathcal{N}, \tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \Omega\ \tilde{O}, \tilde{R}, \gamma \rangle$，使得 $\forall \boldsymbol{\pi}$，我们有 $\tilde{V}_{\boldsymbol{\pi}}(\tilde{s}) \leq \hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$，其中 $\tilde{s} = (s, k)$、$\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$ 表示原始 *LPA-Dec-POMDP* $\hat{\mathcal{M}}$ 中的状态价值函数，对任意 $\forall s \in \mathcal{S}, \forall k \in \{0, 1, ..., K\}$ 均成立。

定理 5 背后的直觉是，在给定固定的 π_{adv} 的情况下，所构造的 Dec-POMDP M 在功能上与 LPA-Dec-POMDP 完全相同。该定理揭示了，LPA-Dec-POMDP 可以被视为 Dec-POMDP 的一种特殊版本，其中的策略应当在“遭受攻击”时的转移函数和奖励函数下保持鲁棒性。该定理也很容易扩展到群体版本，其中 $\pi_{adv} \sim p(P_{adv})$，而 p 是对抗攻击者种群上的某个分布 P_{adv}。定理 3 表明，在 π_{adv} 是随机策略的情况下，新 Dec-POMDP 中的价值函数是原始 LPA-Dec-POMDP 中相同联合策略的价值函数的下界。相关证明见附录。该定理揭示，通过在所构造的 $\tilde{\mathcal{M}}$ 下优化自我系统，我们可以在有限对手去中心化部分可观测马尔可夫决策过程 $\hat{\mathcal{M}}$下获得一个具有鲁棒性的自我系统。

因此，可以应用许多合作式多智能体强化学习算法。具体而言，我们采用 QMIX [Rashid 等人, 2018] 作为问题求解器，其中每个智能体 i 都有一个 Q 网络 $Q_i(\tau^i, a^i)$，以及一个混合网络，该网络将每个 Q 值连同全局状态作为输入，并生成 Q_{tot}的值。在定理 3 提出的分散式部分可观测马尔可夫决策过程 $\mathcal{M}$ 下，我们使用 θ 对 QMIX 进行参数化，并通过最小化以下目标进行训练:

$$L_{ego}(\theta) = \mathbb{E}[(Q_{tot}(\tilde{\boldsymbol{\tau}}, \boldsymbol{a}, \tilde{s}; \theta) - y_{tot})^2], \tag{9}$$

其中 $y_{tot} = \tilde{r} + \gamma \max_{\boldsymbol{a}'} Q_{tot}(\tilde{\boldsymbol{\tau}}', \boldsymbol{a}', \tilde{s}'; \theta^-)$，并且 θ^- 是周期性更新的目标网络的参数。

在我们的 ROMANCE 框架中，我们从档案库中选择一组对抗攻击者，通过固定其中一方交替优化对抗攻击者或自我系统，并相应地更新档案库。我们 ROMANCE 的完整算法见附录中的算法 2。

5 实验

在本节中，我们通过大量实验来回答以下问题：1）与其他基线方法相比，ROMANCE[2] 能否在不同场景中实现较高的鲁棒性？2）ROMANCE 能否获得一组攻击质量高且

[2]Code 可在 https://github.com/zzq-bot/ROMANC 获取E

具有多样性的攻击者？3）ROMANCE 能否集成到多种合作式多智能体强化学习方法中？各个超参数会如何影响 ROMANCE 的性能？

我们在 SMAC [Samvelyan 等人， 2019], 这一广泛使用的《星际争霸II》单位微操任务战斗场景上开展实验。在该场景中，我们训练友方单位击败由采用未知策略的内置人工智能控制的敌方单位。在每个时间步，智能体可以移动或攻击任意敌方单位，并获得等于对敌方单位造成的总伤害的全局奖励。本文考虑了多张地图，包括 2s3z、3m、3s 对 3z、8m、MMM 和 1c3s5z。详细描述见附录。

为确保评估公平，我们使用五个随机种子进行所有实验，并以 95% 置信区间呈现结果。ROMANCE 的详细网络架构和超参数设置见附录。



图 1：未见动作攻击下动作选择的可视化结果。第一行和第二行分别展示了未使用和使用 ROMANCE 学习得到的策略。

5.1 竞争结果与分析

我们基于 QMIX [Rashid 等人， 2018] 实现了 ROMANCE，因为 QMIX 已被广泛证明具有协调能力。随后，我们将 ROMANCE 与四种基线方法进行比较：原始 QMIX（无需任何对抗训练即可完成任务）、随机（在训练期间加入随机攻击），以及两个强基线方法 RARL [Pinto 等人， 2017] 和 RAP [Vinitsky 等人， 2020]。

RARL [Pinto 等人， 2017] 在存在对系统施加扰动的对抗攻击者时训练鲁棒智能体。通过迭代训练，攻击者学习成为最优破坏稳定策略，随后智能体学习完成原始任务，同时抵御对抗攻击者。

RAP [Vinitsky 等人， 2020] 通过引入基于种群的训练扩展了 RARL。在每次 rollout 中，RAP 从种群中均匀采样一个攻击者，并像 RARL 那样迭代训练智能体和攻击者。引入种群能够迫使智能体抵御各种各样的攻击者，从而提升泛化能力，避免被某些特定攻击者利用。

鲁棒性可视化

首先，我们在 MMM 地图上开展实验，研究训练框架如何影响协调策略在遭受不可预测攻击时的行为。如图 1 所示，当一个训练良好的协调策略中的某个协调器遭受动作攻击时，未经对抗训练的策略仍会采取原有的协调模式，却忽视突发情况，从而导致策略未能达到最优。在战斗开始时，原始 QMIX 训练出的幸存者仍试图发动攻击，却忽视了遭受攻击后被恶意攻击引开的掠夺者，导致团队遭受严重伤害。相比之下，我们的 ROMANCE 能够学得一种策略，让幸存者撤退并等待掠夺者归队，以实现高效协调。在战斗中期，当一些陆战队员濒临死亡，却因医疗运输机的动作遭受攻击而无法得到治疗时，ROMANCE 训练出的满血幸存者会冲上前替换濒死的陆战队员并掩护他，使他得以接受治疗；而原始 QMIX 学得的策略仍会在地图上漫无目的地游走，却无视队友。

训练阶段评估

图2展示了在面对固定未知攻击者时，3s 对 3z 和 2s3z 地图上不同方法的学习曲线。在两个地图中，ROMANCE 在每一代的表现都优于所有基线方法，无论是在收敛速度还是渐近性能方面。RANDOM 几乎在每一代都取得了最差的表现，这表明增加

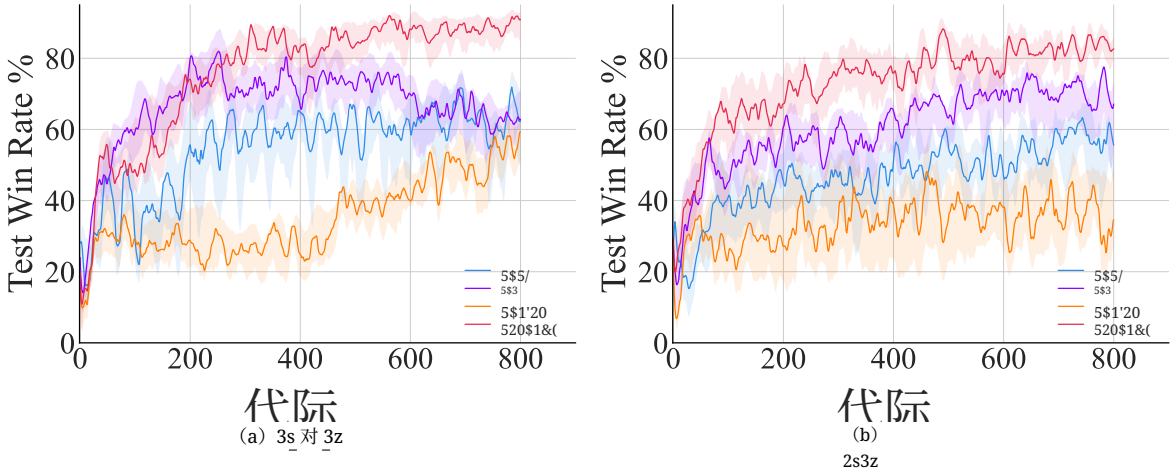
方法 \ 地图名称		2s3z $K=8$	3m $K=4$	3s_对 3z $K=8$	8m $K=5$	MMM $K=8$	1c3s5z $K=6$	+/ − / ≈
自然	原始 QMIX	92.8 ± 1.62	$\mathbf{97.9 \pm 1.02}$	98.3 ± 0.78	98.2 ± 0.45	95.8 ± 1.59	88.8 ± 2.13	1/1/4
	RARL	96.4 ± 1.19	86.0 ± 5.38	80.6 ± 27.5	95.3 ± 3.31	89.3 ± 7.01	76.9 ± 9.85	0/4/2
	RAP	$\mathbf{98.1 \pm 0.76}$	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
	随机	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	$\mathbf{97.1 \pm 1.49}$	$\mathbf{93.9 \pm 1.24}$	
随机攻击	原始 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
	随机	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	原始 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
	随机	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	$\mathbf{81.6 \pm 0.84}$	$\mathbf{89.7 \pm 1.52}$	$\mathbf{90.5 \pm 1.97}$	$\mathbf{86.2 \pm 5.11}$	$\mathbf{84.0 \pm 11.5}$	$\mathbf{66.5 \pm 3.24}$	

表1：不同攻击设置下各方法的平均测试胜率，其中 K 表示训练期间的攻击次数，“自然”表示测试期间没有攻击，“随机攻击”表示自我系统中的每个智能体都可能被随机攻击，“EGA”表示我们提出的基于进化生成的攻击者。每列的最佳结果以粗体标出。符号‘+、‘–’和‘≈’分别表示根据 Wilcoxon秩和检验 [Mann和Whitney,1947] 在置信水平0.05下，该结果显著优于、显著劣于或几乎等同于 ROMANCE。

方法	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	$K=11$	$K=12$	$K=14$
原始 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
随机	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	$\mathbf{89.9 \pm 1.19}$	$\mathbf{86.4 \pm 1.87}$	$\mathbf{81.6 \pm 0.84}$	$\mathbf{75.1 \pm 0.58}$	$\mathbf{66.7 \pm 1.56}$	$\mathbf{57.4 \pm 1.61}$	$\mathbf{48.6 \pm 2.60}$	$\mathbf{41.5 \pm 2.17}$

表 2：测试攻击次数 K 变化时，各方法在地图 2s3z 上的平均测试胜率。每列的最佳结果以粗体突出显示，训练次数（即 $K=8$）所在的列以灰色突出显示。

随机策略扰动在无攻击环境下或许能提升探索能力，但对存在攻击者的任务没有影响。ROMANCE 优于 RARL 和 RAP，分别证明了对抗种群和多样化种群训练的必要性。在图 8 中，我们展示了在地图 2s3z 上基于 QPLEX 和 VDN 实现的不同方法的学习曲线。曲线表明，集成 ROMANCE 后，基于价值的多智能体强化学习算法的鲁棒性能够得到显著提升。



（a）3s_对 3z　　（b）2s3z

图 2：两张地图上的平均测试胜率。

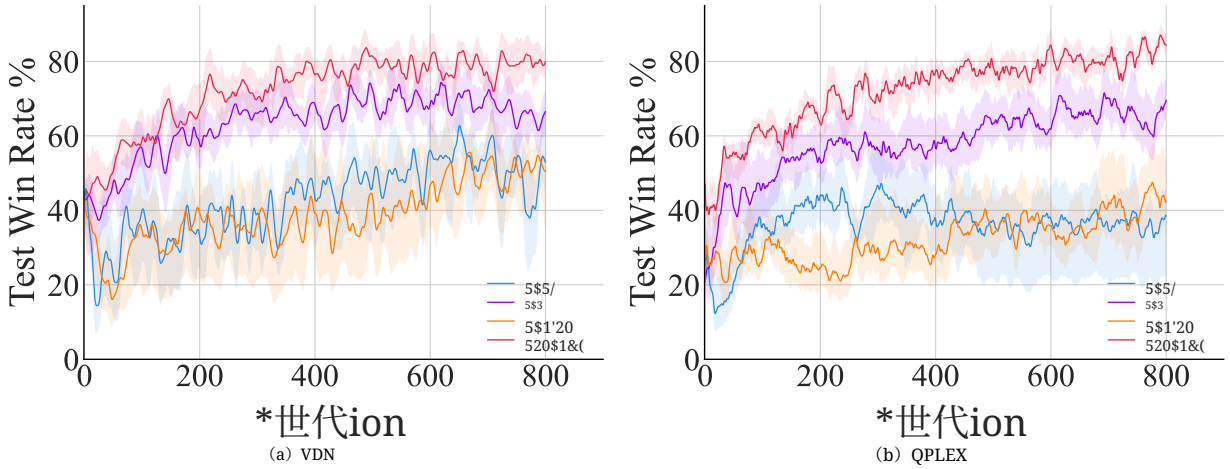


图 3：训练阶段 VDN 和 QPLEX 在地图 2s3z 上的平均测试胜率。

鲁棒性比较

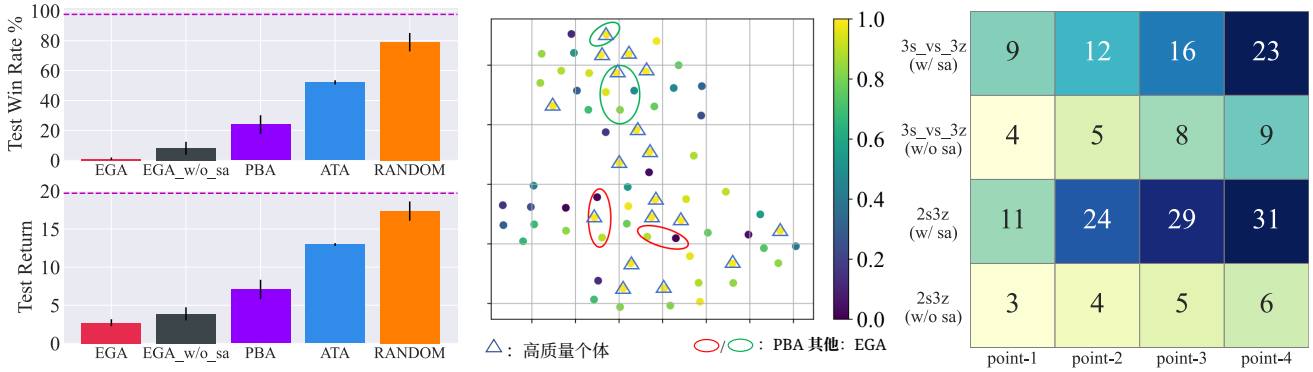
在这里，我们展示 ROMANCE 与多个基线方法相比，是否能够在不同的不可预测攻击下提升协调能力。如表 1 所示，我们设置了三种场景，其中，“Natural”表示测试期间不存在攻击者，或攻击数量为 $K = 0$；“Random Attack”表示自我系统中的每个智能体在每个时间步都可能受到随机攻击；“EGA（基于进化生成的攻击者）”是指通过我们的方法在多次额外运行中生成的、具有较高性能和多样性的未见过的攻击者，可将其视为针对不同方法的分布外攻击者。在“Natural”场景下，ROMANCE 的性能与其他基线方法相当或更优。RARL 的表现逊于其他基线方法，因为其目标是学习最坏情况下的性能，从而导致其协调能力呈现出悲观的结果。在某些地图中，例如 2s3z，RAP 和 RANDOM 的表现优于原始 QMIX。我们认为，这是因为随机攻击或训练期间较弱的对抗种群能够在自然环境下促进多智能体强化学习的探索。此外，在测试期间遭受随机攻击时（即“Random Attack”场景），原始 QMIX 在大多数地图中表现出最显著的性能下降，这表明在没有任何对抗训练的情况下，多智能体协调策略具有较高的脆弱性。基于对抗训练的方法，如 RANDOM、RARL 和 RAP，其表现均优于原始 QMIX，这表明对抗训练能够提升多智能体强化学习策略的鲁棒性。我们进一步发现，当遭遇强大的攻击者时（即“EGA”场景），所有基线方法的性能都出现了严重下降。在不同攻击模式下，所提出的 ROMANCE 在大多数地图上的表现均显著优于其他基线方法，这表明它确实能够在不同的策略扰动条件下学习到鲁棒的协调策略。

超越有限预算的评估

由于本研究考虑的是训练阶段攻击次数固定的设置，因此我们评估了测试时改变攻击预算后的泛化能力。我们在 map2s3z 上进行实验，训练时的攻击次数为 $K = 8$。如表 2 所示，当预算与训练阶段不同时，原始 QMIX 和随机方法学习到的策略即使在预算略有变化时，性能也会严重下降，这表明这两种方法可能过拟合于训练情境，缺乏泛化能力。RARL 和 RAP 优于原始 QMIX 和随机方法，说明对抗训练能够缓解过拟合问题；但随着预算增加，它们的表现仍不及 ROMANCE，这体现了 ROMANCE 的训练范式所带来的高泛化能力。

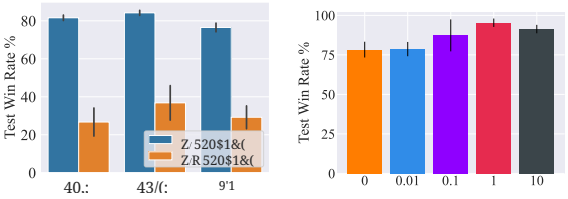
5.2 攻击者种群验证

由于我们的方法需要维护一组对抗攻击者，因此我们设计了实验来考察该方法生成的攻击者。如图 4(a) 所示，多种攻击方法均能相应地降低己方系统的性能，而我们的 EGA（基于进化生成的攻击者）在回报和测试胜率方面都显著优于其他方法，证明了所提出训练范式的有效性。EGA、EGA（不含 sa）和 PBA（基于种群的攻击者）均优于 ATA（交替训练攻击者）和随机，表明种群训练确实能够提升攻击能力。然而，PBA 的效率较低，这表明仅...
...



(a) 3m地图上的攻击质量　　(b) 种群的行为多样性　　(c) ROMANCE 学到的稀疏攻击

图 4：攻击者质量验证，其中 EGA 和 EGA（不含 sa）分别表示使用和不使用稀疏动作正则化器的攻击者的进化生成；PBA 和 ATA 分别指基于种群的攻击者和交替训练攻击者；RANDOM 表示我们随机选择一个协调器进行攻击；(a) 攻击质量。(b) 攻击者在 3m 地图上的 t-SNE 投影。(c) 使用和不使用稀疏动作正则化器时在两张地图上生成的攻击点。



(a) 整合能力　　(b) α的敏感性

图 5：关于 ROMANCE 的更多实验。

种群中随机初始化的个体不足以进行高效的种群训练。EGA 优于其消融版本 EGA（不含 sa），这证明了稀疏动作正则化器的有效性。

此外，我们使用 t-SNE 方法 [Van der Maaten 和 Hinton， 2008]，在二维平面上分析各攻击方法所学习到的行为表示。如图 4(b) 所示，我们发现，传统的基于种群的训练范式只能生成多样性和质量都非常有限的攻击者，大多数点聚集在少数几个特定区域；而我们的 EGA 则能找到分布广泛且兼具高质量与多样性的点。图 4(c) 显示，我们的稀疏动作正则化器能够有效促使稀疏攻击点在一个回合内尽可能分散，这也能通过防止攻击者一开始就耗尽攻击机会，间接提高种群中攻击者的多样性。

5.3 整合能力与参数敏感性研究

ROMANCE 不依赖于特定的价值分解多智能体强化学习方法。我们可以将其视为一种插件模型，并将其与现有的多智能体强化学习价值分解方法（如 QPLEX [Wang 等人，2021b], QMIX [Rashid 等人，2018], 和 VDN [Sunehag 等人， 2018]）相结合。如图 5(a) 所示，与 ROMANCE 结合后，基线方法在地图 2s3z 上的表现大幅提升，这表明所提出的训练范式能够显著增强这些基于价值的多智能体强化学习方法的鲁棒性。

我们框架的关键要素之一是超参数 α，它控制着攻击质量和多样性。我们在此进行实验，研究公式 (6) 中 α 对整个框架的敏感性。如图 5(b) 所示，在地图 3s 对 3z 上，当我们设置 $\alpha = 0$ 或 0。01 时，性能受到的影响更大；这意味着系统仅优化攻击质量，而几乎忽略了多样性目标，表明行为多样性对于种群训练的重要性。尽管如此，其他取值对性能的影响可以忽略不计。我们认为，这是因为进化机制中定制的选择和更新算子能够帮助平衡这两个目标，即使 α稍大一些也是如此。更多实验结果（例如各参数对 ROMANCE 的影响）见附录。

6 结论

本文研究了鲁棒合作式多智能体强化学习问题，其中部分协调者会受到不可预测的策略扰动。我们首先将该问题形式化为 LPA-Dec-POMDP，在该模型中，团队中的部分协调者可能会意外且不可预测地遭受动作扰动。随后，我们提出了 ROMANCE，这是一种通过进化生成辅助对抗攻击者来学习鲁棒多智能体协调的高效方法。针对鲁棒性和泛化测试的实验结果验证了 ROMANCE 的有效性，更多的分析结果也从多个方面证实了这一点。由于我们的方法旨在学习一种解耦对抗攻击者策略，这需要一种基于启发式的策略扰动函数，因此，未来研究更合理、更高效的方法（例如观测扰动，以及针对不同任务自动搜索最佳预算）将具有重要价值。此外，如何为开放环境设定设计高效且有效的鲁棒多智能体强化学习算法，对于多智能体强化学习领域而言同样具有重要价值 [Zhou, 2022]。

致谢

本研究得到国家重点研发计划（2020AAA0107200）、国家自然科学基金（61921006、62022039）以及南京大学优秀博士生B计划的资助。我们感谢Jingcheng Pang、Chengxing Jia以及匿名评审人提供的宝贵讨论和支持。

7 LPA-Dec-POMDP 的证明

引理1（固定 $\boldsymbol{\pi}$ 和 π_{adv} 时的贝尔曼方程）。给定 $\mathcal{M}=\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O,\ R,\ \gamma\rangle$，$\boldsymbol{\pi}$ 和 π_{adv}，我们有

$$\begin{aligned}\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,k) =& \mathbb{E}_{\boldsymbol{a},\hat{\boldsymbol{a}}\sim\boldsymbol{\pi}\circ\pi_{adv}(s,k)}[\mathbb{E}_{s'\sim P(\cdot|s',\boldsymbol{a})}[R(s,\hat{\boldsymbol{a}},s') \\ &+\gamma\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s',k-\mathbb{I}(\hat{\boldsymbol{a}}\neq\boldsymbol{a}))]],\end{aligned} \tag{10}$$

其中 $\boldsymbol{\pi}(\boldsymbol{a}|s)=\prod_{i\in\mathcal{N}}\pi^i(a^i|O(s,i))$，$\mathbb{E}_{\boldsymbol{a},\ \hat{\boldsymbol{a}}\sim\boldsymbol{\pi}\circ\pi_{adv}(s,\ k)}[f]=\sum_{\boldsymbol{a}\in\mathcal{A}}\boldsymbol{\pi}(\boldsymbol{a}|s)\sum_{\hat{\boldsymbol{a}}\in\mathcal{A}}\pi_{adv}(\hat{\boldsymbol{a}}|s,\ \boldsymbol{a},\ k)f$。

证明。根据 $\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,k)$ 的定义：

$$\begin{aligned}&\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,k)\\ &=\mathbb{E}_{\boldsymbol{\pi}\circ\pi_{adv}}[\sum_{n=0}^{\infty}\gamma^n r_{t+n+1}|s_t=s,k_t=k]\\ &=\mathbb{E}_{\boldsymbol{\pi}\circ\pi_{adv}}[r_{t+1}+\sum_{n=0}^{\infty}\gamma^n r_{t+n+2}|s_t=s,k_t=k]\\ &=\sum_{\boldsymbol{a}\in\mathcal{A}}\boldsymbol{\pi}(\boldsymbol{a}|s)\sum_{\hat{\boldsymbol{a}}\in\mathcal{A}}\pi_{adv}(\hat{\boldsymbol{a}}|s,\boldsymbol{a},k)\sum_{s'\in\mathcal{S}}P(s'|s,\hat{\boldsymbol{a}})[r_{t+1}+\\ &\gamma\mathbb{E}_{\boldsymbol{\pi}\circ\pi_{adv}}[\sum_{n=0}^{\infty}\gamma^n r_{t+n+2}|s_{t+1}=s',k_{t+1}=k-\mathbb{I}(\hat{\boldsymbol{a}}\neq\boldsymbol{a})]]\\ &=\mathbb{E}_{\boldsymbol{a},\hat{\boldsymbol{a}}\sim\boldsymbol{\pi}\circ\pi_{adv}(s,k)}[\mathbb{E}_{s'\sim P(\cdot|s',\hat{\boldsymbol{a}})}[R(s,\hat{\boldsymbol{a}},s')\\ &+\gamma\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s',k-\mathbb{I}(\hat{\boldsymbol{a}}\neq\boldsymbol{a}))]].\end{aligned} \tag{11}$$

状态价值函数 $\hat{Q}_{\boldsymbol{\pi}\circ\boldsymbol{\pi}_{\boldsymbol{adv}}}(s,\boldsymbol{a},k)$ 可以用类似的方法推导。

定理*4.* 给定一个 *LPA-Dec-POMDP* $\hat{\mathcal{M}}=\langle\mathcal{N},\mathcal{S},\mathcal{A},P,K,\Omega\ O,R,\gamma\rangle$、自我系统 $\bar{\quad}\ \bar{\quad}\ \bar{\quad}\ \bar{\quad}\ \bar{\quad}$ 的固定联合策略 $\boldsymbol{\pi}$ 以及基于启发式的策略扰动函数 g，存在一个马尔可夫决策过程 $\mathcal{M}=(\mathcal{S},\mathcal{A},P,R,\gamma)$，使得 *LPA-Dec-POMDP* 的最优对抗攻击者 π^*_{adv} 可由 $\mathcal{M}$ 的最优策略 v^* 与 g解耦，其中 $\mathcal{S}=\mathcal{S}\times\mathbb{N}$,$_{_\ \ \ \

其中 $\bar{d}(\bar{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$，$d$ 和 $\bar{d}$ 分别是 $\hat{\mathcal{M}}$ 和 $\bar{\mathcal{M}}$上初始状态的分布，$\boldsymbol{a} = \boldsymbol{\pi}(s)$、$\hat{\boldsymbol{a}} = g(\bar{a}, \boldsymbol{\pi}(s)\ k)$分别是自我系统的原始动作和强制动作。

证明。为了证明该定理，我们首先证明：给定固定的联合自我智能体策略 $\boldsymbol{\pi}$ 和基于启发式的策略 $\hat{}$扰动函数 $g, \forall v$，有 $V_{\boldsymbol{\pi}\circ\pi_{adv}}(s,k) = -V_v(\bar{s})$，其中 $\pi_{adv} = v \circ g$, $\bar{s} = (s,k)$。

根据 V 的贝尔曼方程，我们可以推导出：

$$
\begin{aligned}
&\bar{V}(\bar{s}) \\
&= \sum_{\bar{a}\sim\bar{\mathcal{A}}} v(\bar{a}|\bar{s}) \sum_{\bar{s}'\sim\bar{\mathcal{S}}} P(\bar{s}'|\bar{s},\bar{a})[\bar{R}(\bar{s},\bar{a},\bar{s}') + \gamma\bar{V}(\bar{s}')] \\
&= \sum_{i\in\hat{\mathcal{N}}} v(i|s,k)[\sum_{s'\sim\mathcal{S}} P(s'|s,\hat{\boldsymbol{a}})\mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{v}(s))[-R(s,\hat{\boldsymbol{a}},s') + \\
&\quad \gamma\bar{V}_v((s',k))] + \sum_{s'\sim\mathcal{S}} P(s'|s,\hat{\boldsymbol{a}})\mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{v}(s))[-R(s,\hat{\boldsymbol{a}},s') \\
&\quad - \gamma\bar{V}_v((s',k-1))]] \\
&= \sum_{i\in\hat{\mathcal{N}}} v(i|s,k)[\sum_{s'\sim\mathcal{S}} P(s'|s,\hat{\boldsymbol{a}})[-R(s,\hat{\boldsymbol{a}},s') + \\
&\quad \gamma\bar{V}_v((s',k-\mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a})))] \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) \sum_{\hat{\boldsymbol{a}}\in\mathcal{A}} \pi_{adv}(\hat{\boldsymbol{a}}|s,\boldsymbol{a},k) \sum_{s'\in\mathcal{S}} P(s'|s,\hat{\boldsymbol{a}})[-R(s, \\
&\quad \hat{\boldsymbol{a}},s') + \gamma\bar{V}((s',k-\mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a}))).
\end{aligned}
\tag{12}
$$

然后，我们讨论 $\forall s \in \mathcal{S}$，$k = 0$ 的情况，其中 $\pi_{adv}(\hat{\boldsymbol{a}}|s,\ \boldsymbol{a},\ 0) = \mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{a})$:

$$
\begin{aligned}
&\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,0) \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a})[R(s,\hat{\boldsymbol{a}},s') + \gamma\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s',0)] \\
&= V_{\boldsymbol{\pi}}(s) \\
&\quad \bar{V}_v((s,0)) \\
&= \sum_{\bar{a}\sim\bar{\mathcal{A}}} v(\bar{a}|\bar{s}) g(\bar{a},\boldsymbol{\pi}(s),0) \sum_{s'\sim\mathcal{S}} P(s'|s,\hat{\boldsymbol{a}})[-R(s,\hat{\boldsymbol{a}},s') + \\
&\quad \gamma\bar{V}_v((s,0))] \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a})[-R(s,\hat{\boldsymbol{a}},s') + \gamma\bar{V}_v((s,0))] \\
&= -V_{\boldsymbol{\pi}}(s),
\end{aligned}
\tag{13}
$$

其中，$V_{\boldsymbol{\pi}}(s)$ 是原始分散式部分可观测马尔可夫决策过程下的价值函数。

因此，我们有 $k = 0$，$\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,\ s,\ s,\ s = ,\ 0) = -\bar{V}_v((,\ 0))\ \forall s \in \mathcal{S}$。然后，根据公式(11)和公式(12)，我们可以推导出 $\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(,\ k) - V_v((s,\ k))$，$\forall s \in \mathcal{S}$，$\forall k \in \{0,1,...,K\}$。

此外，对于 M，我们有 $\forall(\bar{s},\ \bar{a},\ \bar{s}')$,

$$
-\max_{s,\hat{\boldsymbol{a}},s'} R(s,\hat{\boldsymbol{a}},s') \leq \bar{R}(\bar{s},\bar{a},\bar{s}') \leq -\min_{s,\hat{\boldsymbol{a}},s'} R(s,\hat{\boldsymbol{a}},s').
$$

根据马尔可夫决策过程的基本性质，存在一个针对 M 的最优策略 v^*，使得 $V_{v^*}(\bar{s}) \geq V_v(\bar{s})$，$\forall \bar{s} \in \mathcal{S}$，$\forall v$。

现在，我们得到，LPA-Dec-POMDP 的最优对抗攻击者 π^*_{adv} 可由 M和 g的最优策略 v^* $\hat{}$ 解耦，即 $\pi^*_{adv} = v^* \circ g$。通过优化 $\mathcal{M}$中的 v，我们能够优化 $\mathcal{M}$中的对抗攻击者。

定理 5. 给定 $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O, R, \gamma \rangle$以及一个固定的确定性对抗攻击者策略 π_{adv}，存在一个 Dec-POMDP $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega\ O, R, \gamma \rangle$，使得 $\mathcal{M}$ 的最优策略是在给定 π_{adv}时 LPA-Dec-POMDP 的最优策略，其中 $\mathcal{S} = \mathcal{S} \times \mathbb{N}$, $d(\tilde{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$, $O(\tilde{s}, i) = O(s, i)$, $R(\tilde{s}, \boldsymbol{a}, \tilde{s}') = R(s, \hat{\boldsymbol{a}}, s')$,

$$\tilde{P}(\tilde{s}'|\tilde{s}, \boldsymbol{a}) = \begin{cases} 0 & k - k' \notin \{0, 1\} \\ P(s'|s, \hat{\boldsymbol{a}}) & otherwise \end{cases},$$

其中，$\tilde{d}$ 和 d 分别是 $\tilde{\mathcal{M}}$ 和 $\hat{\mathcal{M}}$上的初始状态分布，$\tilde{s} = (s, k)$，$\tilde{s}' = (s', k')$，以及$\hat{\boldsymbol{a}} = \pi_{adv}(s, \boldsymbol{a}, k)$。

证明。沿用定理4证明中的相同思路，我们旨在证明：给定确定性 π_{adv} $\forall \boldsymbol{\pi}$，有$\tilde{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k) = V_{\boldsymbol{\pi}}(\tilde{s})$, $\forall s \in \mathcal{S}, k \in \{0, 1, ..., K\}$，其中 $\tilde{s} = (s, k)$。

对于 $\mathcal{M}$，我们有贝尔曼方程：

$$\begin{aligned} &\tilde{V}_{\boldsymbol{\pi}}(\tilde{s}) \\ &= \sum_{\boldsymbol{a} \sim \mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s}) \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{P}(\tilde{s}'|\tilde{s}, \boldsymbol{a})[\tilde{R}(\tilde{s}, \boldsymbol{a}, \tilde{s}') + \gamma \tilde{V}_{\boldsymbol{\pi}}(\tilde{s}')] \\ &= \sum_{\boldsymbol{a} \sim \mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s}) \sum_{s' \in \mathcal{S}} P(s'|s, \hat{\boldsymbol{a}})[R(s, \hat{\boldsymbol{a}}, s') + \gamma \tilde{V}_{\boldsymbol{\pi}}((s', k \\ &- \mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a})))], \end{aligned} \tag{14}$$

其中$\hat{\boldsymbol{a}} = \pi_{adv}(s, \boldsymbol{a}, k)$。

将式（11）中的 $\pi_{adv}(\hat{\boldsymbol{a}}|s, \boldsymbol{a}, k)$替换为确定性策略，可得：

$$\begin{aligned} &\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k) \\ &= \sum_{\boldsymbol{a} \sim \mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) \sum_{s' \in \mathcal{S}} P(s'|s, \hat{\boldsymbol{a}})[R(s, \hat{\boldsymbol{a}}) + \gamma \hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}((s', k \\ &- \mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a})))]. \end{aligned} \tag{15}$$

我们不区分 $\boldsymbol{\pi}(\boldsymbol{a}|\tilde{s})$ 和 $\boldsymbol{\pi}(\boldsymbol{a}|s)$，因为 $\boldsymbol{\pi}(\boldsymbol{a}|\tilde{s}) = \prod_{i \in \mathcal{N}} \pi^i(a^i|\tilde{O}(\tilde{s}, i)) = \prod_{i \in \mathcal{N}} \pi^i(a^i|O(s, i)) = \boldsymbol{\pi}(\boldsymbol{a}|s)$。

与定理（4）的证明类似，我们可以推导出 $\forall \boldsymbol{\pi}, k = 0$，有 $\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, 0) = \tilde{V}_{\boldsymbol{\pi}}(\tilde{s})$，$\forall s \in \mathcal{S}$，其中$\tilde{s} = (s, 0)$。结合公式（14）和公式（15），对于 $\forall \boldsymbol{\pi}$，我们有 $\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k) = \tilde{V}_{\boldsymbol{\pi}}(\tilde{s})$，$\forall s \in \mathcal{S}, k \in \{0, 1, ..., K\}$，其中 $\tilde{s} = (s, k)$。$\tilde{M}$ 中的最优策略 $\boldsymbol{\pi}$ 也是原始 LPA-Dec-POMDP $\hat{\mathcal{M}}$中的最优策略。

定理 6. 给定$\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega\ O, R, \gamma \rangle$以及随机对抗攻击者策略 π_{adv}，存在一个分散式部分可观测马尔可夫决策过程 $\tilde{\mathcal{M}} = \langle \mathcal{N}, \tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \Omega\ \tilde{O}, \tilde{R}, \gamma \rangle$，使得 $\forall \boldsymbol{\pi}$，我们有 $\tilde{V}_{\boldsymbol{\pi}}(\tilde{s}) \leq \hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$，其中 $\tilde{s} = (s, k)$、$\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$ 表示原始 *LPA-Dec-POMDP* 中的状态价值函数，$\forall s \in \mathcal{S}, \forall k \in \{0, 1, ..., K\}$。

证明。与定理（5）类似，我们将奖励函数和转移函数定义如下：

$$\tilde{R}(\tilde{s}, \boldsymbol{a}, \tilde{s}') = \begin{cases} R(s, \boldsymbol{a}, s') & k - k' = 0 \\ \hat{R}(s, \boldsymbol{a}, s', k) & otherwise \end{cases}$$

$$\tilde{P}(\tilde{s}'|\tilde{s}, \boldsymbol{a}) = \begin{cases} 0 & k - k' \notin \{0, 1\} \\ P(s'|s, \boldsymbol{a})\pi_{adv}(\boldsymbol{a}|s, \boldsymbol{a}, k) & k - k' = 0 \\ \hat{P}(s'|s, \boldsymbol{a}, k) & k - k' = 1 \end{cases},$$

其中 $\hat{P}(s'|s, \boldsymbol{a}, k) = \sum_{\hat{\boldsymbol{a}} \sim \mathcal{A}, \hat{\boldsymbol{a}} \neq \boldsymbol{a}} P(s'|s, \hat{\boldsymbol{a}}), \hat{R}(s, \boldsymbol{a}, s') = \sum_{\hat{\boldsymbol{a}} \sim \mathcal{A}, \hat{\boldsymbol{a}} \neq \boldsymbol{a}} R(s, \hat{\boldsymbol{a}}, s')$。

将 π_{adv} 替换为随机版本并代入式（14），可得

$$\begin{aligned}
&\tilde{V}_{\boldsymbol{\pi}}(\tilde{s}) \\
&= \sum_{\boldsymbol{a}\sim\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s})\pi_{adv}(\boldsymbol{a}|s,\boldsymbol{a},k) \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a})[R(s,\boldsymbol{a},s') \\
&+ \gamma\tilde{V}_{\boldsymbol{\pi}}((s',k))] + \sum_{\boldsymbol{a}\sim\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s}) \sum_{s'\in\mathcal{S}} \sum_{\hat{\boldsymbol{a}}\neq\boldsymbol{a}} \pi_{adv}(\hat{\boldsymbol{a}}|s,\boldsymbol{a},k) \\
&P(s'|s,\hat{\boldsymbol{a}})[\sum_{\hat{\boldsymbol{a}}'\neq\boldsymbol{a}} \pi_{adv}(\hat{\boldsymbol{a}}'|s,\boldsymbol{a},k)R(s,\hat{\boldsymbol{a}}',s')+ \\
&\quad \gamma\tilde{V}_{\boldsymbol{\pi}}((s',k-1))].
\end{aligned} \tag{16}$$

基于 $\mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a})$对式(11)进行拆分，可得:

$$\begin{aligned}
&\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,k) \\
&= \sum_{\boldsymbol{a}\sim\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s})\pi_{adv}(\boldsymbol{a}|s,\boldsymbol{a},k) \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a})[R(s,\boldsymbol{a},s') \\
&+ \gamma\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s',k)] + \sum_{\boldsymbol{a}\sim\mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|\tilde{s}) \sum_{s'\in\mathcal{S}} \sum_{\hat{\boldsymbol{a}}\neq\boldsymbol{a}} \pi_{adv}(\hat{\boldsymbol{a}}|s,\boldsymbol{a},k) \\
&P(s'|s,\hat{\boldsymbol{a}})[R(s,\hat{\boldsymbol{a}},s') + \gamma\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s',k-1)].
\end{aligned} \tag{17}$$

与定理(4)的证明类似，可得 $\forall\boldsymbol{\pi}, k=0$，并且有$\hat{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,\ 0) = \tilde{V}_{\boldsymbol{\pi}}(\tilde{s}),\ \forall s\in\mathcal{S}$，其中$\tilde{s}=(s,\ 0)$。

根据期望的基本性质，即 $\mathbb{E}_X[f(X)g(X)] \geq \mathbb{E}_X[f(X)]\mathbb{E}_X[g(X)]$,，可得 $\forall\boldsymbol{\pi}$，并且$\hat{}$

$\tilde{}$ $\bar{V}_{\boldsymbol{\pi}\circ\pi_{adv}}(s,\ k) \geq V_{\boldsymbol{\pi}}(\bar{s}),\ \forall s\in\mathcal{S},\ k\in\{0,1,...,K\}$，其中 $\bar{s}=(s,\ k)$。通过在分散式部分可观测马尔可夫决策过程 $\mathcal{M}$中优化 $\boldsymbol{\pi}$，我们实际上是在优化原始LPA-Dec-POMDP中 $\boldsymbol{\pi}$ 的状态价值函数下界。

8 算法的补充说明

8.1 相关算法的描述

如主文稿所述，我们的工作主要包括两部分：攻击者种群生成和 ROMANCE 的整个过程，详细伪代码分别如算法 1 和算法 2 所示。

我们在算法1中描述了攻击者种群的进化生成过程。在每次迭代中，我们首先根据质量分数从档案库中选择 n_p 个对抗攻击者，作为当前种群。自我系统将按照函数 CollectTraj 中所述的方式，与种群中的攻击者交替进行交互。收集到的轨迹用于优化种群，从而获得 n_p 个新的对抗攻击者。为了保持档案库中攻击者的质量和多样性，我们采用了函数 Update 中的专用更新机制。

在算法2中，我们提出了交替训练范式 ROMANCE，其中针对对抗攻击者种群的最佳响应基于算法1中的攻击者进化生成进行优化。

对于两个比较基线方法 RARL 和 RAP，我们分别在算法 3 和算法 4 中给出了详细伪代码。

9 SMAC 的详细描述

SMAC [Samvelyan 等人， 2019] 是一个《星际争霸II》单位微操任务的战斗场景。我们考虑部分观测设置：智能体只能看到周围一个圆形区域，其半径等于视野范围，设为 9。我们使用强化学习算法训练友方单位，以击败由内置人工智能控制的敌方单位。每个回合开始时，友方单位和敌方单位会在地图上的特定区域生成。每个智能体在每个时间步从离散动作空间中选择动作，包括以下动作：空操作、向 [方向],移动、攻击 [敌方单位编号], 以及停止。在这些动作的控制下，智能体可以在连续地图上移动和攻击。多智能体强化学习智能体在每个时间步都会获得全局奖励，其数值等于对敌方单位造成的总伤害。每击杀一个敌方单位以及赢得战斗（击杀所有敌人）都会分别获得 10 和 200 的额外奖励。我们在表 3 中简要介绍了本文使用的 SMAC 地图，各地图的截图见图 6。

算法1： 攻击者的进化生成**输入：** 联合自我智能体策略 $\boldsymbol{\pi}$，档案库 Arc_{adv}，种群规模 n_p，档案库容量 n_a。

1 $P_{adv} \leftarrow$**选择(**Arc_{adv}, n_p**)；2 对** $\pi_{adv}^{\phi_j} \in P_{adv}$ **执行 3** $\mathcal{D}_{\boldsymbol{\pi}}$，$\mathcal{D}_{adv}^{j} \leftarrow$ **轨迹收集(**$\boldsymbol{\pi}, \pi_{adv}^{\phi_j}$**)；4 结束 5 基于公式(6)** $\{\phi_j\}_{j=1}^{n_p}$ **优化；6** $Arc_{adv} \leftarrow$ **更新(**Arc_{adv}, P_{adv}**)；7 函数 更新（档案库**$_{adv}$，$\mathbf{P}_{adv}$**）：8 对** $\pi_{adv}^{\phi_j} \in P_{adv}$ **执行 9 对** $\pi_{adv}^{\phi_i} \in Arc_{adv}$ **执行 10 基于公式(8)计算距离(**i, j**)；11 结束 12 如果 min**$_i$ **距离(**i, j**)** $\geq threshold$ **则 13 添加(**$\pi_{adv}^{\phi_j}$，Arc_{adv}**)；14 否则 15 保留(**$\pi^{\arg\min_i}$ **距离(**$_{i,j}$**)** $_{adv}$，π_{adv}^{j}, Arc_{adv}**)；16 结束 17 结束 18 返回** Arc_{adv} **19 函数 轨迹收集（**$\boldsymbol{\pi}$,$\pi_{adv}^{\phi_j}$**）：20** $k \leftarrow K$**；21** $\mathcal{D}_{\boldsymbol{\pi}}, \mathcal{D}_{adv}^{j} \leftarrow \{\}, \{\}$**；22 对** $t = 0$**到** T **执行 23** $\{o_t^i\}_{i=1}^{N} = \{O(s_t, i)\}_{i=1}^{N}$**；24** $\boldsymbol{a}_t = \boldsymbol{\pi}(\boldsymbol{\tau}_t)$**；25** $\hat{i} \sim v^j(s_t, k)$**；26 如果** $i \neq null$ **且** $k > 0$**则 27** $\hat{a}_t^i =$**使(**$_{a^i} Q^i(\tau_t^i, a^i)$**) 最小；28** $\hat{\boldsymbol{a}}_t = (\hat{a}_t^{\hat{i}}, \boldsymbol{a}_t^{-\hat{i}})$**；29** $k \leftarrow k - 1$**；30 否则 31** $\boldsymbol{a}_t = \boldsymbol{a}_t$**32 结束 33** $s_{t+1}, \ r_t, \ done \leftarrow env.step(\hat{\boldsymbol{a}}_t)$**；34** $\mathcal{D}_{\pi} \leftarrow \mathcal{D}_{\pi} \cup \{s_t, \boldsymbol{o}_t, \boldsymbol{a}_t, r_t, done\}$**；35** $\mathcal{D}_{adv}^{j} \leftarrow \mathcal{D}_{adv}^{j} \cup \{s_t, \hat{i}, \ -r_t, \ done\}$**；36 如果完成为真则 37 中断；38 结束 39 结束 40 返回** $\mathcal{D}_{\boldsymbol{\pi}}, \ \mathcal{D}_{adv}^{j}$

10 ROMANCE 的架构、基础设施与超参数选择

由于主实验中的 ROMANCE 构建在 QMIX 之上（主文稿“整合能力”部分中的 VDN 和 QPLEX 亦然），因此我们在此介绍具体设置，包括网络架构和超参数选择。局部智能体网络与 QMIX 采用相同的架构，包含一个维度为 64 的 GRU单元，用于编码历史信息，以及两个全连接层，用于计算局部 Q 值。混合网络按照现有的多智能体强化学习方法进行设置。对抗攻击者采用一个包含 64 个隐藏单元的多层感知机（MLP）作为受害者选择函数，以选出受害者，并根据基于启发式的策略扰动函数迫使其执行局部最差动作。我们采用 RMSProp 作为优化器，其参数为 $\alpha = 0.99, \epsilon = 1 \times 10^{-5}$

算法 2： ROMANCE

输入： 环境 $\mathcal{E}$，种群规模 n_p，档案库容量 n_a，迭代次数 N_{gen}。

初始化容量为 n_a的档案库 Arc_{adv}；
对于 $gen = 1$ **到** N_{gen} **执行**
  <escape>$P_{adv} \leftarrow$ select(Arc_{adv}, n_p);</escape>
  当 $n = 1$ **到** N_{adv} **时**
    对于 $\pi_{adv}^{\phi_j} \in P_{adv}$
      $\mathcal{D}_{\boldsymbol{\pi}}$, $\mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集($\boldsymbol{\pi}^{\theta}$, $\pi_{adv}^{\phi_j}$);
    end
    根据式（6）使用 $\mathcal{D}_{adv}$ 优化 $\boldsymbol{\phi}$；
  end
  对于 $n = 1$ **到** N_{ego} **do**
    对 $\pi_{adv}^{\phi_j} \in P_{adv}$ **执行**
      $\mathcal{D}_{\boldsymbol{\pi}}$, $\mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集($\boldsymbol{\pi}^{\theta}$, $\pi_{adv}^{\phi_j}$);
    end
    基于式（9）使用 $\mathcal{D}_{\boldsymbol{\pi}}$ 优化 θ；
  end
  $Arc_{adv} \leftarrow$ 更新(Arc_{adv}, P_{adv});
结束

算法3： RARL

输入： 环境 $\mathcal{E}$，迭代次数 N_{gen}。

初始化攻击者策略 π_{adv}^{ϕ}；
对于 $gen = 1$**到** N_{gen} **执行**
  对于 $n = 1$**到** N_{adv} **执行**
    $\mathcal{D}_{\boldsymbol{\pi}}$, $\mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集($\boldsymbol{\pi}^{\theta}, \pi_{adv}^{\phi_j}$);
    根据式（3），使用 $\mathcal{D}_{adv}$ 优化 $\boldsymbol{\phi}$；
  end
  对于 $n = 1$ **到** N_{ego} **do**
    $\mathcal{D}_{\boldsymbol{\pi}}$, $\mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集($\boldsymbol{\pi}^{\theta}, \pi_{adv}^{\phi_j}$);
    根据式（9），使用 $\mathcal{D}_{\boldsymbol{\pi}}$ 优化 θ；
  end
结束

分别用于自我系统和攻击者。具体而言，自我系统的学习率在 2s3z、3m 和 3s 对 3z 中设为 4×10^{-4}，在其他场景中设为 2×10^{-4}。$\delta = 5 \times 10^{-2}$ 和 $\lambda = 4 \times 10^{-2}$ 分别是 SPRQ 的参考分布参数和正则化因子。此外，我们在动作分布上设置平滑常数 $b = 2 \times 10^{-2}$，以防 KL 散度趋于无穷大。整个框架在 NVIDIA GeForce RTX 2080 Ti GPU 上，利用收集到的回合数据进行端到端训练，共进行 800 次迭代（世代）。

11 额外实验结果

除主文稿中已展示的内容外，我们还在图 7 中展示了基于 QMIX 的不同方法在另外四张地图上的学习曲线。我们可以观察到，在强对抗攻击下，ROMANCE 在其他地图上都取得了最佳性能。我们还可以观察到，在地图 MMM 和 1c3s5z 上，RARL 的表现甚至不如 RANDOM。我们认为，当存在许多异构智能体时，自我系统可能很容易对某一特定类型的攻击者过拟合。在图 8 中，我们展示了不同方法在地图 2s3z 上基于 QPLEX 和 VDN 实现的学习曲线。这些曲线表明，将 ROMANCE 集成其中后，可以显著增强基于价值的多智能体强化学习算法的鲁棒性。

在图 9 中，我们展示了不同方法在全部六张地图上生成的攻击者质量。对于 EGA、EGA（不含 sa）和 PBA 这三种在单次运行中生成多个攻击者的方法，我们取五个最佳攻击者的平均值，作为每次运行的质量。EGA 在全部六张地图上相较于其他方法的优势，证明了其有效性。

算法 4： RAP

输入： 环境 $\mathcal{E}$，种群规模 n_p，迭代次数 N_{gen}。

以容量 n_p 初始化种群 P_{adv}；
对于 $gen = 1$ **到** N_{gen}，**执行 do**
  对于 $n = 1$ **到** N_{adv} **执行**
    for $\pi_{adv}^{\phi_j} \in P_{adv}$ **do**
      $\mathcal{D}_{\boldsymbol{\pi}}, \mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集$(\boldsymbol{\pi}^{\theta}, \pi_{adv}^{\phi_j})$;
    end
    优化 ϕ with $\mathcal{D}_{adv}$ based on Eq. (3);
  end
  从 $n = 1$ **到** N_{ego} **do**
    对于 $\pi_{adv}^{\phi_j} \in P_{adv}$ **做,**
      $\mathcal{D}_{\boldsymbol{\pi}}, \mathcal{D}_{adv}^{j} \leftarrow$ 轨迹收集$(\boldsymbol{\pi}^{\theta}, \pi_{adv}^{\phi_j})$;
    end
    优化 θ，使用基于公式 (9) 的 $\mathcal{D}_{\boldsymbol{\pi}}$；
  end
结束

Map	友方单位	敌方单位	Type
2s3z	2个追猎者 3个狂热者	2个追猎者， 3个狂热者	对称， 异构
3m	3名陆战队员	3名陆战队员	对称， 齐次
3s_对 3z	3名追猎者	3个狂热者	微操技巧， 风筝战术
8m	8 名陆战队员	8名陆战队员	对称， 齐次
MMM	1 架医疗运输机， 2 名掠夺者， 7名陆战队员	1 架医疗运输机， 2 名掠夺者， 7 名陆战队员	对称， 异构
1c3s5z	1名巨像， 3名追猎者， 5 个狂热者	1 个巨像， 3 个追猎者， 5 个狂热者	对称， 异构

T表 3：所进行的 6 个 SMAC 场景的属性 ios.

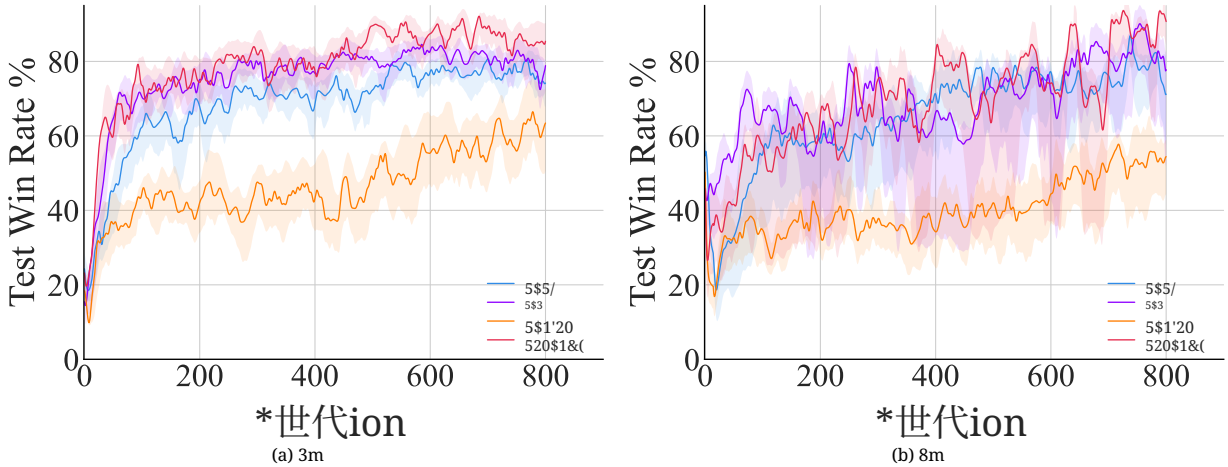


(a) 2s3z　(b) 3米　(c) 3s_vs_3z

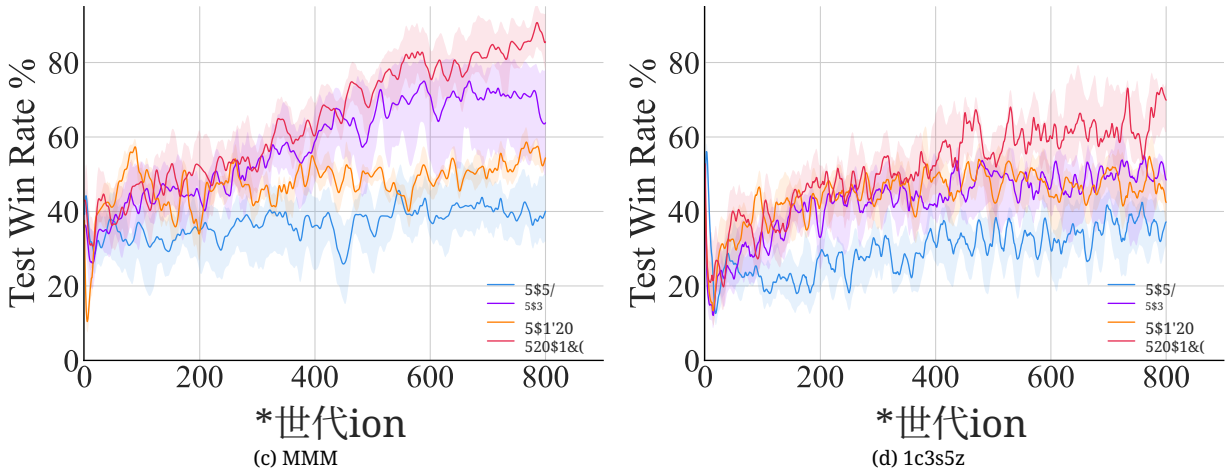
(d) 8米　(e) MMM　(f) 1c3s5z

图 6：我们选定的星际争霸II场景快照。

在表 4-6 中，我们研究不同类型的超参数如何影响 ROMANCE 的性能。我们分别调整档案库的大小、种群规模以及训练期间的攻击次数，然后在三种设置下测试自我系统。

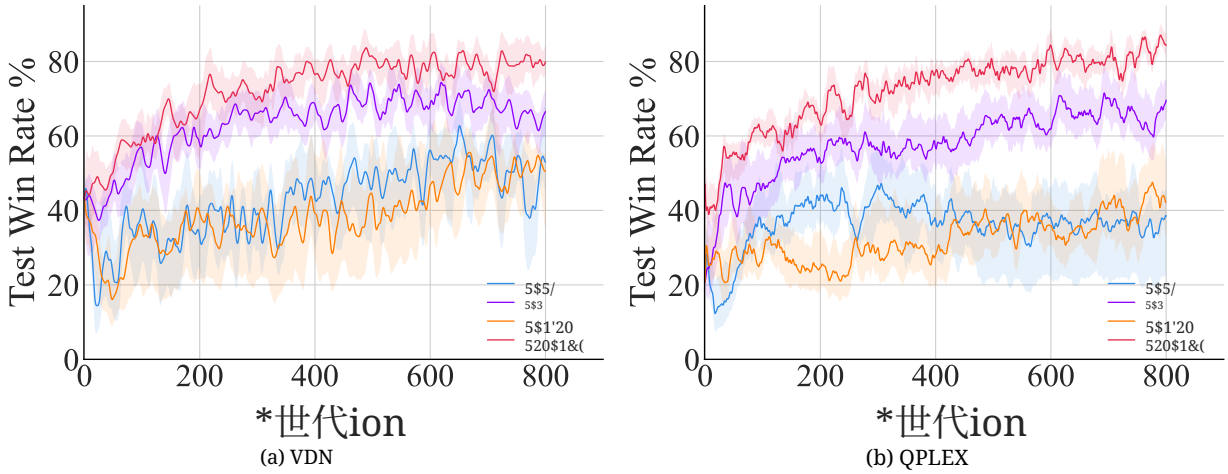


(a) 3m (b) 8m



(c) MMM (d) 1c3s5z

图 7：另外四张地图上的平均测试胜率。



(a) VDN (b) QPLEX

图 8：VDN 和 QPLEX 在地图 2s3z 上的平均测试胜率，QMIX 的结果见正文。

档案大小。 档案库用于记录迄今生成的具有高性能和多样化行为的攻击者。对于 ROMANCE 而言，这至关重要，因为新的攻击者是基于从档案库中选出的攻击者生成的。档案大小过小会导致多样性下降，而过大则会使选择高性能攻击者的过程效率低下。



(a) 2s3z (b) 3m

(c) 3s 对 3z (d) 8m

(e) MMM (f) 1c3s5z

图9：不同方法在另外六张地图上生成的攻击者质量。

如表 4 所示，我们发现，适当增大档案大小有助于提升鲁棒性，因为更大的容量能够保存更多表现优异且多样化的攻击者。如上文所述，如果档案过大，性能也可能下降，同时效率也会降低。

档案大小	自然	随机攻击	EGA
11	94.0 ± 9.06	83.7 ± 10.3	71.2 ± 12.1
13	98.5 ± 0.68	88.9 ± 1.59	76.5 ± 4.16
15	97.9 ± 1.34	89.1 ± 1.97	81.6 ± 0.84
17	98.1 ± 0.62	88.2 ± 0.94	83.6 ± 3.02
19	96.8 ± 1.14	86.6 ± 1.06	78.8 ± 7.40

表 4：档案大小变化时，ROMANCE 在地图 2s3z 上的平均测试胜率。

种群规模。 表 5 描述了种群规模变化时 ROMANCE 的鲁棒性。种群规模指自我系统在一代中会遇到的攻击者数量。较大的种群规模有助于提升自我系统在最终阶段的鲁棒性，但也可能生成行为相似的攻击者，从而损害效率。尽管档案大小相同，针对规模较小、仅有 2 或 3 个攻击者的种群训练最佳响应，往往会导致过拟合，从而削弱其对多样化攻击者的泛化能力。表 5 中的结果表明，在地图 2s3z 中，默认种群规模 4 较为合适。

种群规模	自然	随机攻击	EGA
2	94.7 ± 4.56	83.4 ± 7.40	66.2 ± 9.44
3	92.6 ± 5.47	82.0 ± 9.78	68.8 ± 12.9
4	97.9 ± 1.34	89.1 ± 1.97	81.6 ± 0.84
5	98.4 ± 0.83	88.7 ± 2.74	79.2 ± 5.25
6	98.8 ± 0.45	89.6 ± 1.07	80.9 ± 1.34

表 5：种群规模变化时 ROMANCE 在地图 2s3z 上的平均测试胜率。

训练攻击预算。 我们还在表 6 中展示了 ROMANCE 在不同训练攻击数量预算下的平均测试胜率。当训练预算低于测试预算时，我们的方法凭借泛化能力保证了不错的表现。然而，其表现仍不及训练攻击数量预算与测试时一致的自我系统。训练攻击数量过多也可能导致性能下降，因为过强的攻击者会使自我系统变得保守。

攻击数量	自然	随机攻击	EGA
6	97.3 ± 1.42	87.2 ± 2.42	71.6 ± 5.59
7	98.0 ± 1.05	87.2 ± 3.28	76.2 ± 2.80
8	97.9 ± 1.34	89.1 ± 1.97	81.6 ± 0.84
9	97.8 ± 1.38	88.2 ± 3.29	84.4 ± 2.68
10	98.3 ± 0.86	87.4 ± 2.91	80.1 ± 1.99

表6：攻击者数量变化时，ROMANCE 在地图 2s3z 上的平均测试胜率 .

参考文献

T. B¨ack。《理论与实践中的进化算法：进化策略、进化编程、遗传算法》。牛津大学出版社，1996年。Varun Bhatt、Bryon Tjanaka、Matthew C. Fontaine 和 Stefanos Nikolaidis。深度代理辅助环境生成。发表于*NeurIPS*，2022。Tobias Blickle 和 Lothar Thiele。进化算法中所用选择方案的比较。《进化计算》，4(4):361–394，1996年。Lorenzo Canese、Gian Carlo Cardarilli、Luca Di Nunzio、Rocco Fazzolari、Daniele Giardino、Marco Re 和 Sergio Span`o。多智能体强化学习：挑战与应用综述。《应用科学》，11(11):4948，2021年。Konstantinos Chatzilygeroudis、Antoine Cully、Vassilis Vassiliades 和 Jean-Baptiste Mouret。质量-多样性优化：随机优化的一个新分支。载于《黑箱优化、机器学习与无免费午餐定理》，第109–135页。施普林格，2021年。

Filippos Christianos、Georgios Papoudakis、Muhammad A Rahman 和 Stefano V Albrecht。通过选择性参数共享扩展多智能体强化学习。载于*ICML*，第 1989–1998 页，2021年。

A. Cully 和 J.-B. Mouret。《机器人学中的行为库学习》。载于*GECCO*，第175–182页，2013年。Antoine Cully 和 Yiannis Demiris。《质量与多样性优化：统一的模块化框架》。*IEEE*进化计算汇刊，22(2):245–259，2017年。Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。《反事实多智能体策略梯度》。载于*AAAI*，第2974–2982页，2018年。Bent Fuglede 和 Flemming Topsøe。《詹森-香农散度与希尔伯特空间嵌入》。载于*ISIT*，2004年。Nir Greshler、Ofir Gordon、Oren Salzman 和 Nahum Shimkin。《合作式多智能体寻路：超越路径规划与避碰》。载于*MRS*，第20–28页，2021年。Sven Gronauer 和 Klaus Diepold。《多智能体深度强化学习：综述》。《人工智能评论》，55(2):895–943，2022。Pengjie Gu、Mengchen Zhao、Jianye Hao 和 Bo An。《部分可观测性下的在线临时团队协作》。载于*ICLR*，2022。Jun Guo、Yonghong Chen、Yiha ng Hao、Zixin Yin、Yin Yu 和 Simin Li。《面向合作式多智能体强化学习鲁棒性的全面测试》。*arXiv:2204.07932*，2022。Pablo Hernandez-Leal、Bilal Kartal 和 Matthew E Taylor。《多智能体深度强化学习的综述与批评》。《自主智能体与多智能体系统》，33(6):750–797，2019年。Hengyuan Hu、Adam Lerer、Alex Peysakhovich 和 Jakob N. Foerster。《用于零样本协调的"他者博弈"》。载于*I CML*，第4399–4410页，2020年。Yizheng Hu 和 Zhihua Zhang。《多智能体强化学习中的稀疏对抗攻击》。*arXiv:2205.09362*，2022。Inaam Ilahi、Muhammad Usama、Junaid Qadir、Muhammad Umar Janjua、Ala Al-Fuqaha、Dinh Thai Hoang 和 Dusit Niyato。《深度强化学习对抗攻击的挑战与应对措施》。*IEEE*人工智能汇刊，3(2):90–109，2021年。Max Jaderberg、Wojciech M Czarnecki、Iain Dunning、Luke Marris、Guy Lever、Antonio Garcia Castaneda、Charles Beattie、Neil C Rabinowitz、Ari S Morcos、Avraham Ruderman 等。《基于种群的强化学习在3D多人游戏中的人类水平表现》。《科学》，364(6443):859–865，2019年。Shihui Li、Yi Wu、Xinyue Cui、Honghua Dong、Fei Fang 和 Stuart Russell。《通过极小极大深度确定性策略梯度实现鲁棒多智能体强化学习》。载于*AAAI*，第4213–4220页，2019年。Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。《利用深度强化学习进行连续控制》。载于*ICLR*，2016。Jieyu Lin、Kristina Dzeparoska、Sai Qian Zhang、Alberto Leon-Garcia 和 Nicolas Papernot。《关于合作式多智能体强化学习的鲁棒性》。载于*SPW*，第62–68页，2020年。Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb Pieter Abbeel 和 Igor Mordatch。《混合合作竞争环境下的多智能体行动者-评论家》。载于*NeurIPS*，第6379–6390页，2017年。Andrei Lupu、Brandon Cui、Hengyuan Hu 和 Jakob N. Foerster。《零样本协调的轨迹多样性》。载于*ICML*，第7204–7213页，2021年。Henry B Mann 和 Donald R Whitney。《关于检验两个随机变量中一个是否随机大于另一个》。《数学统计年刊》，第50–60页，1947年。Reuth Mirsky、Ignacio Carlucho、Arrasy Rahman、Elliot Fosong、William Macke、Mohan Sridharan、Peter Stone 和 Stefano V Albrecht。《临时团队协作综述：定义、方法与开放问题》。*arXiv:2202.10450*，2022。Janosch Moos、Kay Hansel、Hany Abdulsamad、Svenja Stark、Debora Clever 和 Jan Peters。《鲁棒强化学习：基础与最新进展综述》。《机器学习与知识提取》，4(1):276–315，2022。Tuomas Oikarinen、Wang Zhang、Alexandre Megretski、Luca Daniel 和 Tsui-Wei Weng。《通过对抗损失实现鲁棒深度强化学习》。载于*NeurIPS*，第26156–26167页，2021年。Frans A Oliehoek 和 Christopher Amato。《去中心化部分可观测马尔可夫决策过程简介》。施普林格，2016。Xinlei Pan、Daniel Seita、Yang Gao 和 John Canny。《风险规避的鲁棒对抗强化学习》。载于*ICRA*，第8522–8528页，2019年。Jing-Cheng Pang、Tian Xu、Sheng-Yi Jiang、Yu-Ren Liu 和 Yang Yu。《用于稀疏动作任务的稀疏先验正则化Q学习》。*arXiv:2105.08666*，2021年。

Georgios Papoudakis、Filippos Christianos、Arrasy Rahman 和 Stefano V Albrecht。处理多智能体深度强化学习中的非平稳性。*arXiv:1906.04737*，2019年。Anay Pattanaik、Zhenyi Tang、Shuijing Liu、Gautham Bommannan 和 Girish Chowdhary。通过对抗攻击实现鲁棒深度强化学习。载于*AAMAS*，第 2040–2042 页，2018年。Zhenghao Peng、Quanyi Li、Ka Ming Hui、Chunxiao Liu 和 Bolei Zhou。通过协调策略优化学习模拟自驱动粒子系统。载于*NeurIPS*，第 10784–10797 页，2021年。Thomy Phan、Lenz Belzner、Thomas Gabor、Andreas Sedlmeier、Fabian Ritz 和 Claudia Linnhoff-Popien。通过对抗价值分解实现具有韧性的多智能体强化学习。载于*AAAI*，第 11308–11316 页，2021年。Lerrel Pinto、James Davidson、Rahul Sukthankar 和 Abhinav Gupta。鲁棒对抗强化学习。载于*ICML*，第 2817–2826 页，2017年。Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于 *ICML*，第 4295– 4304 页，2018年。Mikayel Samvelyan、Tabish Rashid、Christian Schroeder de Witt、Gregory Farquhar、Nantas Nardelli、Tim GJ Rudner、Chia-Man Hung、Philip HS Torr、Jakob Foerster 和 Shimon Whiteson。星际争霸多智能体挑战赛。载于*AAMAS*，第 2186–2188 页，2019年。Elena Smirnova、Elvis Dohmatob 和 J´er´emie Mary。分布鲁棒强化学习。*arXiv:1902.08708*，2019年。Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。Qtran：通过变换学习分解合作式多智能体强化学习。载于*ICML*，第 5887–5896 页，2019年。Yeeho Song 和 Jeff Schneider。通过遗传课程实现鲁棒强化学习。载于 *ICRA*，第 5560–5566 页，2022。Peter Stone、Gal A. Kaminka、Sarit Kraus 和 Jeffrey S. Rosenschein。临时自主智能体团队：无需预先协调的协作。载于*AAAI*，2010年。Yanchao Sun、Ruijie Zheng、Yongyuan Liang 和 Furong Huang。谁是最强的敌人？迈向深度强化学习中最优且高效的规避攻击。载于*ICLR*，2021年。Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vin´ıcius Flores Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z Leibo、Karl Tuyls 和 Thore Graepel。基于团队奖励的合作式多智能体学习价值分解网络。载于*AAMAS*，第 2085–2087 页，2018年。Chen Tessler、Yonathan Efroni 和 Shie Mannor。连续控制中的动作鲁棒强化学习及其应用。载于*ICML*，第 6215–6224 页，2019年。Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。*Journal of Machine Learning Research*，9 (11):2579–2605，2008年。Eugene Vinitsky、Yuqing Du、Kanaad Parvate、Kathy Jang、Pieter Abbeel 和 Alexandre Bayen。使用对抗种群实现鲁棒强化学习。*arXiv:2008.01825*，2020年。Jianhao Wang、Zhizhou Ren、Beining Han、Jianing Ye 和 Chongjie Zhang。理解具有价值分解的合作式多智能体 Q 学习。载于*NeurIPS*，第 29142–29155 页，2021a。Jianhao Wang、Zhizhou Ren、Terry Liu、Yang Yu 和 Chongjie Zhang。QPLEX：双重决斗多智能体 Q 学习。载于*ICLR*，2021b。Yihan Wang、Beining Han、Tonghan Wang、Heng Dong 和 Chongjie Zhang。DOP：离策略多智能体分解策略梯度。载于*ICLR*，2021c。Yutong Wang、Ke Xue 和 Chao Qian。用于强化学习的基于聚类选择的进化多样性优化。载于*ICLR*，2022。Annie Xie、Shagun Sodhani、Chelsea Finn、Joelle Pineau 和 Amy Zhang。在多个不确定性集合上进行鲁棒策略学习。*arXiv:2202.07013*，2022。Mengdi Xu、Zuxin Liu、Peide Huang、Wenhao Ding、Zhepeng Cen、Bo Li 和 Ding Zhao。应对内在脆弱性的可信强化学习：鲁棒性、安全性与泛化性。*arXiv:2209.08025*，2022。Ke Xue、Yutong Wang、Lei Yuan、Cong Guan、Chao Qian 和 Yang Yu。通过协同进化实现异构多智能体零样本协调。*arXiv:2208.04957*，2022a。Ke Xue、Jiacheng Xu、Lei Yuan、Miqing Li、Chao Qian、Zongzhang Zhang 和 Yang Yu。多智能体动态算法配置。载于*NeurIPS*，2022b。

Wanqi Xue、Wei Qiu、Bo An、Zinovi Rabinovich、Svetlana Obraztsova 和 Chai Kiat Yeo。《说错还是误导：实现多智能体通信强化学习的鲁棒性》。发表于*AAMAS*，第 1418–1426 页，2022c。Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。《PPO 在合作多智能体博弈中的惊人有效性》。发表于*NeurIPS*，2022。Jing Yu、Clement Gehring、Florian Sch¨afer 和 Animashree Anandkumar。《鲁棒强化学习：一种约束博弈论方法》。发表于*L4DC*，第 1242–1254 页，2021。Won Joon Yun、Soohyun Park、Joongheon Kim、Myungjae Shin、Soyi Jung、Aziz Mohaisen 和 Jae-Hyun Kim。《通过自主多无人机控制实现可靠监控的合作式多智能体深度强化学习》。*IEEE*工业信息学汇刊，2022。Huan Zhang、Hongge Chen、Duane S Boning 和 Cho-Jui Hsieh。《利用学习得到的最优对手对状态观测进行鲁棒强化学习》。发表于*ICLR*，2020a。Huan Zhang、Hongge Chen、Chaowei Xiao、Bo Li、Mingyan Liu、Duane S. Boning 和 Cho-Jui Hsieh。《针对状态观测中对抗性扰动的鲁棒深度强化学习》。发表于*NeurIPS*，2020b。Kaiqing Zhang、Tao Sun、Yunzhe Tao、Sahika Genc、Sunil Mallya 和 Tamer Basar。《具有模型不确定性的鲁棒多智能体强化学习》。发表于*NeurIPS*，第 10571–10583 页，2020c。Kaiqing Zhang、Zhuoran Yang 和 Tamer Bas¸ar。《多智能体强化学习：理论与算法的选择性综述》。载于《强化学习与控制手册》，第 321–384 页，2021。Yuan Zhang、Jianhong Wang 和 Joschka Boedecker。《通过不确定集正则化实现连续控制任务中的鲁棒强化学习》。*arXiv:2207.02016*，2022。Zhi-Hua Zhou。《开放环境机器学习》。《国家科学评论》，9(8)，2022。Changxi Zhu、Mehdi Dastani 和 Shihan Wang。《多智能体通信强化学习综述》。*arXiv:2203.08975*，2022。