面向多智能体博弈策略鲁棒性的自适应融合评估方法

李骏唯[1],阮书岚[1],梁嘉旋[2],刘 瑜[3],何 友[1*]

(1. 清华大学深圳国际研究生院,广东深圳 518055;2. 哈尔滨工业大学(深圳)计算机科学与技术学院,广东深圳 518055;
3. 清华大学电子工程系,北京 100084)

摘 要: 随着多智能体强化学习算法的快速发展,智能体在博弈任务中的协作与竞争能力得到了显著提升。然而,面对实际场景中环境的动态变化,智能体策略在跨环境迁移中的性能波动问题日益凸显。尽管当前已涌现出对抗训练、域随机化等鲁棒性增强技术,但现有的鲁棒性评估体系仍存在明显局限。现有方法往往仅关注平均奖励等单一性能指标的变化,忽视了碰撞次数等反映安全性或稳定性的特征,难以全面衡量策略的稳定性。此外,由于缺乏统一的测试基准,不同研究常依赖特定的实验环境参数设定,导致算法难以在不同的场景条件下进行公平的横向比较。这些局限制约了博弈策略的实际落地与迭代优化。为此,本文提出了面向多智能体博弈策略的多维自适应融合鲁棒性评估方法,旨在通过数学形式化建模实现对策略稳定性的量化分析。首先,本文设计了基于条件变异系数(Conditional Coefficient of Variation, CondCV)的鲁棒性评分指标(Robustness Score, RS),用于精确捕捉并融合多种基础评测指标在环境扰动下的波动特征。通过消除指标间的量纲差异,该方法构建了一种标准化的通用度量,具备良好的自适应性与评估公平性,广泛适用于多智能体协作、对抗等各类环境下的策略评估。同时,针对多维指标权重的分配,本文提出基于α-Rank演化博弈的权重自适应融合框架。该框架将指标间的排序一致性建模为博弈过程,通过计算稳态分布获得客观权重,并与先验权重进行动态融合,有效平衡了指标的客观稳定性与专家先验知识。为验证方法的有效性,本文基于Isaac Sim平台自主构建了高度可配置的实验环境,涵盖对抗与协作两类典型的多智能体博弈场景,并集成多种主流算法开展了系统性的实验验证。实验结果表明,该评估方法可有效度量策略在不同环境设定下的稳定性,具备多维波动捕捉能力和跨任务通用性,为算法优化与评估提供了理论支持和参考。最后,本文探讨了评估方法在虚实迁移中的应用潜力,并提出了相应的可行方案,为未来研究提供了参考。

关键词: 多智能体强化学习;评测基准;鲁棒性评估;自适应融合;博弈策略

基金项目: 国家自然科学基金(No.62293544,No.62425117,No.62506205);中国博士后科学基金(No.2025T180426);国家资助博士后研究人员计划(No.GZB20250393)

中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2026)03-0912-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251264

Adaptive Fusion-Based Robustness Evaluation Method for Multi-Agent Game Strategies

LI Junwei[1], RUAN Shulan[1], LIANG Jiaxuan[2], LIU Yu[3], HE You[1*]

(1. *Shenzhen International Graduate School*, *Tsinghua University*, *Shenzhen*, *Guangdong* 518055, *China*;
2. *School of Computer Science and Technology*, *Harbin Institute of Technology* (*Shenzhen*), *Shenzhen*, *Guangdong* 518055, *China*;
3. *Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: With the rapid development of multi-agent reinforcement learning algorithms, agents' capabilities for cooperation and competition in game-based tasks have significantly improved. However, given the dynamic changes of real-world environments, performance fluctuations of strategies during cross-environment transfer have become increasingly prominent. Although robustness enhancement techniques such as adversarial training and domain randomization have emerged, existing robustness evaluation frameworks still exhibit evident limitations. Current methods often focus only on changes in a single performance metric such as average reward, while neglecting safety or stability metrics such as collision frequency, making it difficult to comprehensively evaluate strategy stability. In addition, the lack of unified evaluation benchmarks leads different studies to rely on specific experimental parameter settings, hindering fair comparisons across diverse scenarios. These limitations restrict the practical deployment and iterative optimization of game strategies. To address

收稿日期:2026-02-16;录用日期:2026-03-19;责任编委:吕金虎
*通信作者:何友

these issues, we propose a robustness evaluation method for multi-agent game strategies via multidimensional adaptive fusion, aiming to provide a quantitative analysis of strategy stability through mathematically formalized modeling. First, we design a robustness score (RS) based on the conditional coefficient of variation (CondCV) to accurately capture and fuse the fluctuation characteristics of base metrics under environmental perturbations. By eliminating dimensional differences among metrics, the method establishes a standardized and generalizable measurement with strong adaptability and evaluation fairness, making it broadly applicable to strategy evaluation in cooperative, competitive, and other multi-agent environments. To address the weight assignment for multidimensional metrics, we propose an adaptive weight fusion framework based on an adversarial α-Rank evolutionary game. This framework models ranking consistency among metrics as a game process, derives objective weights from the stationary distribution, and dynamically fuses them with expert prior weights, achieving a balance between objective metric stability and expert prior knowledge. To validate the effectiveness of our method, we develops highly configurable multi-agent environments based on Isaac Sim that cover typical adversarial and cooperative game scenarios, and conducts systematic experiments with various mainstream algorithms. Experimental results demonstrate that the evaluation method can effectively measure strategy stability under diverse environmental settings, exhibiting multidimensional fluctuation-capturing capability and cross-task generality, thereby providing theoretical support and reference for algorithm optimization and evaluation. Finally, we discuss the potential application of the evaluation method in sim-to-real transfer and propose corresponding feasible solutions, offering insights for future research.

Keywords: multi-agent reinforcement learning; evaluation benchmark; robustness evaluation; adaptive fusion; game strategy

Foundation Item(s): National Natural Science Foundation of China (No. 62293544, No. 62425117, No. 62506205); China Postdoctoral Science Foundation (No.2025T180426); Postdoctoral Fellowship Program of CPSF (No.GZB20250393)

0 引言

近年来,强化学习(Reinforcement Learning,RL)作为一种模拟智能体通过与环境交互学习最优策略的机器学习方法,已成为解决复杂决策问题的核心方法之一。早期的强化学习研究集中于单智能体的决策优化问题,经典算法如Q-Learning[1]和Sarsa[2]通过值函数迭代的方式学习智能体在特定状态下的最优行为。随着深度学习[3]的发展,深度强化学习进一步提升了强化学习在高维和连续空间中的适用性,并在游戏、控制系统、自动驾驶等领域展现出了强大的应用潜力[4-6]。然而,在许多实际场景中,决策问题不仅涉及单一智能体与环境的交互,还涉及多个智能体之间的协作或竞争,多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)[7]因此成为研究热点。MARL通过扩展单智能体强化学习的方法,使多个智能体能够在共享环境中优化其行为。在此基础上,博弈论作为建模多智能体策略的系统化数学工具,被广泛嵌入到多智能体的相关研究问题中[8-10],多智能体博弈已然成为一种流行的研究范式。

当前,国内外对多智能体博弈算法的研究正逐步从理论探索走向实际应用。在多个复杂任务中,这类算法展现出强大的问题求解能力。例如,DeepMind提出的AlphaStar在《星际争霸II》中成功超越人类玩家,展示了其在高维、动态、实时决策环境中的策略学习能力[11]。在城市交通领域,基于MARL的智能交通系统可动态优化红绿灯配时,从而有效缓解交通拥堵并降低碳排放[12]。在救援搜索场景中,基于MARL的无人机协同控制方法在复杂地形和动态目标任务中取得了优异表现[13],进一步验证了其在真实复杂环境中的应用潜力。

然而,随着强化学习算法应用于实际场景,其在稳定性方面的问题也日益凸显,成为制约其进一步推广的关键因素。当前主要存在两方面挑战:一是部分算法在面对场景变更、传感器噪声或外部干扰时性能出现大幅波动;二是一些算法在仿真环境中训练良好,但在真实环境中因环境复杂性等因素性能骤降[14]。为此,研究者们提出了诸多改进方法,例如,引入对手(adversary)模型,使智能体在对抗式训练中增强抗干扰能力[15];在奖励函数或动作空间加入噪声,使智能体学会应对各种外力干扰[16-17];通过域随机化(domain randomization)手段,在仿真训练阶段引入多样化环境扰动,从而提升其迁移到现实世界时的泛化性能[18-19]等。

虽然这些研究工作一定程度上提升了强化学习算法的鲁棒性,但现有研究仍缺乏统一、多维度的多智能体博弈策略评估基准。现有的方法存在多方面局限:首先,缺乏对鲁棒性的关注。如表1所示,现有主流测试环境多关注任务成功率或平均奖励等结果导向的指标,而对算法在动态扰动下的鲁棒性考虑不足。这导致许多算法虽然在固定设定的环境中得分很高,但一旦测试条件发生变化,其性能就大幅下跌,难以反映其在复杂多变环境下的真实可靠性。其次,评估指

标维度单一。现有鲁棒性评估标准往往仅关注单一性能指标(如平均奖励)的变化情况[20-21],忽略了如碰撞次数、目标逃逸次数或任务执行步数等反映安全性、稳定性和算法效率的基础指标,缺乏对算法多维度的整体评价。然而,若要实现对策略的多维度综合评价,不同指标之间的量纲差异处理以及多指标融合中的权重分配问题,将成为影响评估结果合理性的新的挑战。最后,评测基准缺乏统一性。现有方法对算法的评估依赖于具体实验设置,不同研究往往基于各自设定的测试环境与评价标准进行实验,使得算法之间难以直接对比。例如,某些算法在特定评估基准中展现出良好性能,但在换用另一平台时则表现不稳定[22]。这种评估标准的不一致性不仅阻碍了不同方法间的有效比较,也限制了整个领域的研究积累与成果推广。因此,构建一种多维度融合的标准化鲁棒性评估基准,已成为当前多智能体博弈研究中亟待解决的问题。

表1 主流多智能体评估环境

Table 1 Mainstream multi-agent evaluation environments

评估环境	任务描述	主要评估指标
SMAC[23]	RTS 博弈	胜率、平均回合数、击杀数
Hanabi Challenge[24]	合作博弈	平均得分、平均标准误差、完美对局百分比
MPE[9]	协作/竞争	任务完成率、平均奖励
GRF[25]	足球策略	进球数、最终奖励

在此背景下,本文构建了一种面向多智能体博弈策略的多维自适应融合鲁棒性评估方法,以数学形式化方式对算法在复杂环境中的性能进行量化分析。本文从基于条件变异系数的鲁棒性指标出发,使用α-Rank[26]算法的演化思想提出基于博弈论的多维自适应融合框架,结合主客观评估得到最合理的权重分配。值得注意的是,该方法的多维自适应融合特点不仅能为算法在不同任务间的横向比较提供标准化依据,还弥补了现有研究未能综合利用多维基础指标的不足。本文的方法能有效揭示博弈算法在跨环境迁移中存在的性能波动,为算法优化与实际应用提供理论支持,从而有助于促进多智能体博弈领域的算法发展与应用落地。

为支撑上述评估方法的实验验证,本文基于Isaac Sim平台[27],自主设计了涵盖对抗性围捕与协作性搜救两类典型博弈范式的实验环境。它们具备完整的任务逻辑与训练流程,支持对多个环境参数进行灵活配置,便于测试算法在不同设定下的表现,从而系统开展鲁棒性测试实验。环境接入了多种主流算法,本文对它们进行了全面测试,并基于提出的评估方法进行了多个定量评价和对比分析实验。最后,本文讨论了该评估方法在仿真-现实领域的潜在应用,提出一种面向虚实迁移泛化能力评估[28]的解决方案。

综上,本文的主要贡献如下。

(1)针对多智能体博弈领域鲁棒性评估基准的不足,提出了面向多智能体博弈策略的多维自适应融合鲁棒性评估方法。该方法融合了多维度基础评测指标,能够进行自适应计算,适用于各种场景下的算法评估。

(2)提出了基于条件变异系数的鲁棒性评分指标计算方法。该方法不依赖于特定任务设定或特定算法,可根据实验需求自由选择基础指标,为多智能体博弈策略的鲁棒性评估提供了统一的度量标准。

(3)提出了基于对抗演化博弈的多维自适应融合框架,创新性地引入指标间博弈的求解思路。该方法既利用博弈论的思想客观分配不同指标的权重,又引入专家先验或主观偏好信息进行灵活调整。

(4)构建了对抗性围捕和协作性搜救两类实验场景,支持对物理参数与难度等级的灵活配置,可形成多样的鲁棒性测试环境。在此基础上,选取多个主流算法,全面测试并计算了它们的鲁棒性评分,进行了深入的对比分析,验证了评估方法的有效性。此外,进一步探讨了其在虚实迁移场景中的潜在应用。

1 相关工作

1.1 多智能体强化学习

强化学习是一种机器学习范式,其核心在于智能体通过与环境的持续交互,逐步学习并优化策略,以实现长期累积奖励的最大化。如图1所示,强化学习算法可以分为三类。基于价值的方法通过估计状态价值函数或状态-动作对价值函数来指导决策,例如,DQN(Deep Q-Network)[29]在Atari游戏中的成功展示了深度学习结合强化学习的强大潜力。基于策略的方法直接参数化策略,通过策略梯度方法更新策略。如PPO(Proximal Policy Optimization)[30]算法限制新旧策略的概率比防止过大的更新,结合优势函数估计提高样本效率。基于演员-评论家的方法则结合评估状态价值的价值网络和决策动作的动作网络,例如DDPG(Deep Deterministic Policy Gradient)[31]结合了确定性策略梯度和深度Q网络,成功解决了连续动作空间中的控制问题。SAC(Soft Actor-Critic)算法[32]则进一步引入最大熵框架,从而提升了算法的探索能力和鲁棒性。

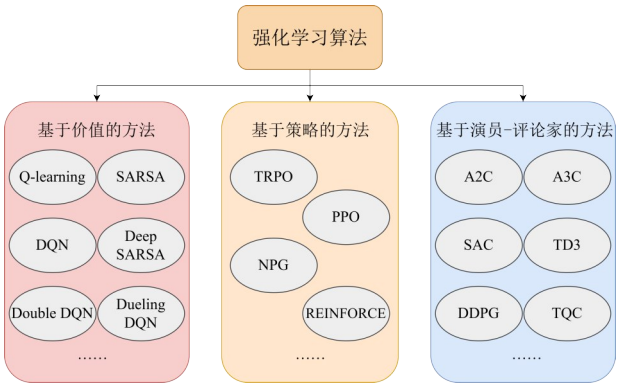


图1　经典强化学习算法分类

Figure 1　Classification of classic reinforcement learning algorithms

多智能体强化学习扩展了单智能体强化学习的方法,其目标是使多个智能体在共享环境中优化行为以最大化累积奖励。与单智能体强化学习相比,多智能体强化学习面临更复杂的挑战,如智能体的可扩展性、环境的部分可观测性以及通信限制等[33]。早期的算法结合博弈论和时序差分的思想更新动作-状态价值函数,进而得到最优策略[8,34]。随着深度神经网络的发展,研究者们联合状态空间和动作空间,将深度强化学习算法拓展到多智能体场景,提出集中训练集中执行(centralized training with centralized execution)框架。为应对部分可观测和维数灾难问题,研究者们进一步提出集中训练分布执行(centralized training with decentralized execution)框架,典型的例子包括基于值函数分解的QMIX算法[35]、基于策略梯度的MAPPO[36](Multi-Agent Proximal Policy Optimization)和MADDPG[9](Multi-Agent Deep Deterministic Policy Gradient)以及进一步优化协作效率的信赖域方法HAPPO[37],这些算法在不同任务中均展示了卓越性能。为解决多智能体算法在不同环境表现不稳定的问题,研究者们还陆续提出了鲁棒强化学习方法,如将鲁棒协作问题建模为贝叶斯博弈的EIR-MAPPO算法[38]、基于随机对抗体的对抗算法的ATSA[39]以及防御协同攻击的Wolfpack算法[40],这些方法显著提升了智能体策略的稳定性和适应性。

1.2　评估体系构建

随着多智能体博弈学习的不断发展,智能体在交互过程中需要处理更加复杂的行为模式,并在动态环境中优化策略,这对算法的适应性和稳定性提出了更高的要求。因此,如何科学评估这些智能体的学习能力、协作与竞争效率成为关键问题。

在传统机器学习中,模型评估主要依赖于测试集上的准确率、召回率等指标。随着深度学习技术的兴起,尤其在大语言模型和生成模型[41]领域,模型测评逐渐发展为更加多样化和复杂化的体系,出现了多种综合基准测试框架[42]。如GLUE[43]用于衡量模型的自然语言理解能力并被广泛用于测试BERT系列模型;BIG-bench[44]用于评估模型在多类型任务中的表现。这些基准测试框架通过设计多样化的任务和评价指标,能够更全面地评估模型的性能。

在多智能体博弈领域,经典的测试基准为算法性能评估提供了标准化环境。例如,SMAC(Starcraft Multi-Agent Challenge)通过模拟《星际争霸II》中的复杂对战场景,测试智能体在高维状态空间和动态环境中的协作与竞争能力[23]。Hanabi Challenge聚焦不完全信息下的协作,重点评估有限通信下的推理与策略协调能力[24]。MPE(Multi-agent Particle Environment)则提供了轻量化的多智能体任务场景,如协作导航和竞争追逐,为算法的快速验证和迭代提供了便利[9]。近年来,研究者们提出了更复杂的评估基准以应对更高的研究需求。例如,BenchMARL[45]构建了涵盖多种经典任务的标准化评估框架,提升了实验的可复现性与公平性。PyMARLzoo+[46]通过扩展PyMARL填补了高维图像观测与完全协作任务的评估空白。MOSMAC[20]则针对实际应用中的长视距与多目标权衡问题,在SMAC基础上引入了时序性多目标任务。此外,Li等人[47]提出了大规模鲁棒性评估基准,通过在多种合作场景中引入多类型的不确定性扰动,系统地衡量了智能体的鲁棒性与恢复力。这些基准推动了多智能体博弈算法的发展,促进了领域内的可比性和可复现性。

然而,现有评估方法在多智能体博弈场景中仍存在局限性:缺乏对鲁棒性的关注,如BenchMARL和MOSMAC主要关注奖励和胜率,未考虑算法在环境扰动下的鲁棒性;评估指标过于单一,难以全面反映智能体的鲁棒性,如IMAP[21]主要通过对抗攻击导致的奖励下降来衡量鲁棒性;不同算法基于不同环境和评价指标进行测试,缺乏统一的评估标准,如对抗攻击的攻击力度、训练步数等设置不同,结果就会不同。因此,构建一种统一的多维度自适应评估体系,从多个角度系统性地衡量算法性能,已成为当前研究的迫切需求。本研究将在现有研究的基础上,进一步探索多维度自适应评估体系的构建,为多智能体博弈算法的优化与应用提供理论支持。

2　面向多智能体博弈策略的多维自适应融合鲁棒性评估方法

现有多智能体博弈评测研究较少关注算法在动态环境中的稳定性,且缺乏多维统一的鲁棒性评价指标。针对上述问题,本文提出一种多维自适应融合的评估方法。如图2所示,该方法主要包括指标计算与

权重计算两个部分：指标计算部分基于原始实验数据，通过条件变异系数刻画算法在不同环境条件下的波动特征；权重计算部分利用博弈演化方法求解客观权重，并进一步与专家先验权重进行融合。最终，通过加权求和得到综合鲁棒性评分，形成由实验数据到评估结果的完整评价流程。

本章首先基于条件变异系数构建鲁棒性评分指标，具体内容见2.1节；随后提出权重融合框架，将博弈演化生成的客观权重与专家先验权重相结合，以确定最终融合权重，具体内容见2.2节。通过数学形式化建模，本章构建了一套标准的面向多智能体博弈算法的鲁棒性评估框架。

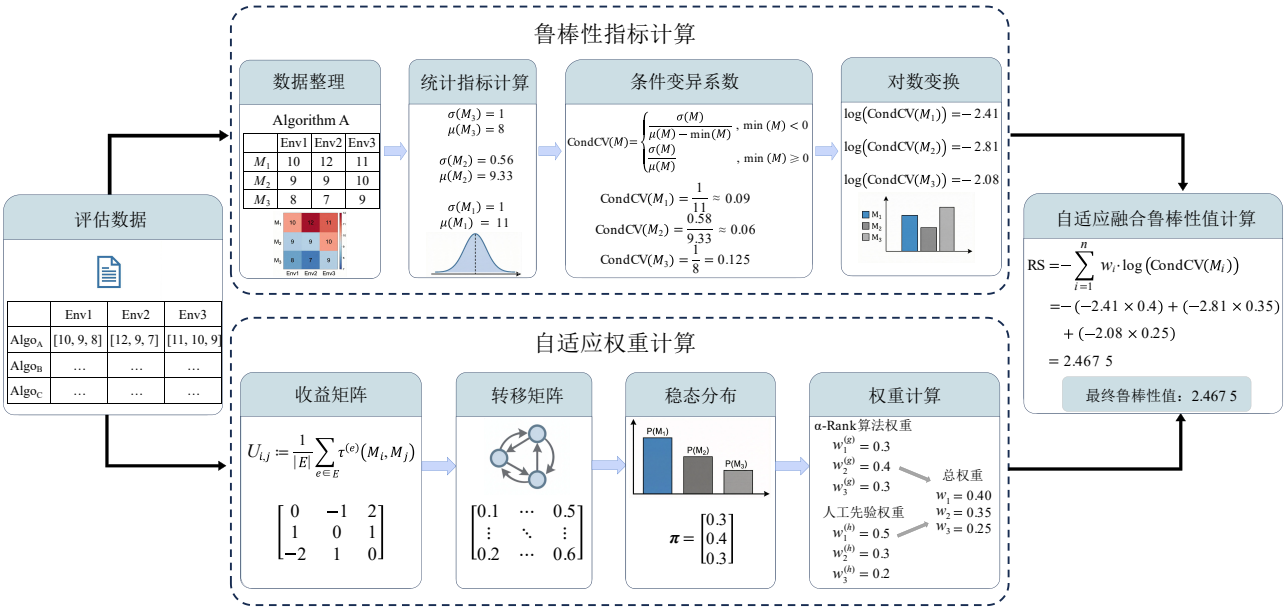


图2 鲁棒性评分计算流程示意图

Figure 2 Workflow of robustness score computation

2.1 基于条件变异系数的鲁棒性评分指标

鲁棒性是指智能体策略在不同环境或环境扰动下的稳定性程度。因此，鲁棒性对于确保智能体在复杂动态场景中的可靠部署至关重要。为此，本文提出了一种基于变异系数的鲁棒性评估方法。

鲁棒性评分指标的核心思想是衡量策略在不同环境配置下各基础评测指标的波动程度，并通过加权归约的方式进行整体评价。设有n个基础评测指标，每个指标在不同环境下的测量值构成一个随机变量，鲁棒性评分（Robustness Score，RS）定义如下：

$$\mathrm{RS}=-\sum_{i=1}^{n}w_i\cdot\log\left(\mathrm{CondCV}(M_i)\right) \tag{1}$$

其中：M_i代表第i个基础评测指标；w_i为第i个基础评测指标对应的权重，满足$\sum_{i=1}^{n}w_i$的约束条件。$\mathrm{CondCV}(M_i)$为第i个基础评测指标的条件变异系数（Conditional Coefficient of Variation，CondCV），表示为

$$\mathrm{CondCV}(M)=\begin{cases}\dfrac{\sigma(M)}{\mu(M)-\min(M)}, & \text{if } \min(M)<0\\ \dfrac{\sigma(M)}{\mu(M)}, & \text{if } \min(M)\geqslant 0\end{cases} \tag{2}$$

其中：$\sigma(M)$表示标准差；$\mu(M)$表示均值；$\min(M)$表示数据集M的最小值。

条件变异系数是鲁棒性评分的核心计算单元，用于评估策略在不同环境下的波动性。相比于直接计算标准差，变异系数通过归一化处理，使得不同量纲的基础评测指标在计算时具有可比性，避免了量纲影响。例如，成功率通常在[0, 1]之间，而平均步数可能高达几十甚至数百，若直接相加，后者的影响力将远远大于前者，导致后者主导最终结果，损害了评分的公平性。此外，针对变异系数无法处理负值的局限，条件变异系数通过整体偏移实现了数据集的正值化。

同时，为了更精确地反映策略的稳定性，本文采用了对数变换，使得较小的变异系数对应较高的鲁棒性评分，增强对小幅波动的区分能力。最后，加权求和确保了各指标在综合评估中的影响力可控，能够根据不同任务的重要性灵活设定权重，以适应不同的评测需求。权重的取值策略将在2.2节详细介绍。

在上述评估框架下，分值越高，表示策略在环境变化时表现越稳定，即鲁棒性越强；反之则表示策略鲁棒性较差。这一设计不仅客观反映了策略抵抗环境干扰的能力，也通过统一的数值标度，为后续多维

指标的自适应融合提供了标准化的输入基础。

更重要的是，本文的方法具有广泛的适应性，不受特定任务类型、环境配置或算法类型的限制，是一种标准化的评估方法。这主要得益于变异系数的归一化特性，使得不同任务的基础指标都能统一处理，无需针对特定任务进行专门调整。因此，无论智能体系统的规模、任务目标或交互机制如何，研究者都可以将其作为通用的度量标准，并根据实际需求灵活选取最合适的基础指标。示例如下。

（1）多智能体围捕任务：可选择平均奖励、成功率作为核心指标，同时引入平均碰撞次数（智能体之间及与障碍物碰撞次数）、平均逃逸次数（逃逸者逃出包围圈次数）等安全性与可靠性指标。

（2）多智能体对抗任务：可选择平均碰撞次数、拦截率、击打成功率等，以评估智能体的对抗能力。

（3）多智能体协作任务：可选择成功率、平均协作时间、团队奖励等，以衡量智能体间的协作效率。

无论选择哪种任务场景，研究者只需根据实验需求选定合适的指标即可完成自适应计算。这种通用性和灵活性使本文的方法适用于各种多智能体场景，为鲁棒性评估提供了统一的度量标准。

此外，该方法具备显著的模型无关性（model-agnostic）与评估公平性。由于条件变异系数的计算仅依赖于智能体与环境交互的最终测试结果，而无需访问策略网络的内部参数或梯度信息，这使得该指标能够跨越算法架构的差异，在统一的标准下公平地衡量不同类型算法（如基于价值的方法与基于策略的方法）的鲁棒性。这种“黑盒”式的评估特性，规避了因算法内部机制不同（如是否使用经验回放、是否为异构智能体等）而可能产生的评估偏差，从而确保了评价结果能够真实、客观地反映策略在执行层面的最终性能。

总体而言，本节提出的评估方法在度量多智能体博弈策略的鲁棒性方面具有较强的通用性。其核心思想是基于条件变异系数计算不同基础评测指标的波动性，并通过对数变换和加权求和的方式得到最终的鲁棒性评分。

2.2　基于对抗演化博弈的自适应融合框架

在鲁棒性指标的最终评分中，如何合理设置各个指标的权重是决定评分系统有效性的重要因素。若直接采用等权平均，容易导致关键指标与次要指标对最终结果贡献相同，削弱指标系统的区分力与评价效果。为此，本文提出一种结合对抗演化博弈与专家先验的融合方法，自适应计算指标排序能力的同时，也引入专家经验对指标重要性的指导，从而构建出兼具客观性与主观性的权重分配策略。

自适应融合评估的核心思路是：一个有效的基础评测指标，应当在多环境下稳定合理地对策略进行排序，且这一排序结果应被所有其他指标所“认可”；同时，不同任务背景下，某些指标可能天然更为关键或在主观上更重要。因此，本文将指标排序一致性建模为对抗博弈过程以获得客观权重，并进一步结合主观先验实现自适应融合。

在基于博弈的权重部分，计算过程被建模为一个零和博弈（zero-sum game）。玩家1（指标提议者）与玩家2（指标验证者）的策略空间$\mathcal{S}_1$、$\mathcal{S}_2$定义为全体指标集合，博弈空间$\mathcal{S}:=\{\mathcal{S}_1\times\mathcal{S}_2\}$。玩家1倾向最大化收益，希望选择的指标$M_i\in\mathcal{M}$的排序结果能被其他指标认可；玩家2倾向于切换到收益更低的指标对指标体系进行“攻击”，也即试图找到指标$M_i\in\mathcal{M}$使得当前指标M_i的排序显得不合理。这种对抗性设计模拟了指标间的竞争与制衡，充分挖掘了指标的排序稳定性和合理性。

博弈的收益函数（payoff function）$\boldsymbol{U}\in\mathbb{R}^{n\times n}$定义为指标间的排序一致性，即衡量两个指标对策略优劣的判断是否相似。对于任意两个指标M_i与M_j，计算在多环境测试下，指标M_i对算法的排序与指标M_j对算法排序之间的一致性程度。本文采用Kendall's τ相关系数[48]作为一致性衡量，这是一种通过比较所有数据对的协同性衡量两个变量排序一致性的非参数统计量。综上，收益矩阵定义如下：

$$U_{i,j}:=\frac{1}{|E|}\sum_{e\in E}\tau^{(e)}(M_i,M_j)\tag{3}$$

其中：$i,j=1,2,\cdots,n$，n为指标数；$U_{i,j}$是矩阵$\boldsymbol{U}$的第i行第j列元素；E是全体环境集合；τ为Kendall's τ函数。

在获得矩阵$\boldsymbol{U}$后，本文采用α-Rank算法计算每个指标在博弈中的稳态概率分布，表示该指标在排序能力博弈中的重要性程度。α-Rank算法是由Omidshafiei等人[26]提出的博弈求解方法，用于多智能体策略评估和排序。该算法通过模拟种群演化过程中的选择压力，计算策略的稳态分布从而得到各策略的强度或稳定性排序。本文受α-Rank算法演化排序思想启发，对其进行了适应性调整，将原算法中的收益值替换为各指标之间的相对排序信息，通过α-Rank的马尔可夫链收敛过程提取每个指标在长期演化中表现出的“生存优势”，这种优势可以直接反映指标排序能力的稳定性。该稳态概率经归一化处理后作为每个指标权重的博弈部分$w_i^{(g)}$。博弈权重计算的具体步骤如算法1所示。

另一方面，考虑到实验者在具体场景中对指标重要性的主观判断，本文引入了一组先验权重$w_i^{(h)}$，该主观分配可来源于任务背景经验、行业标准或应用导向。例如，在多智能体击打任务中，设计者可能更关

算法1 基于α-Rank的博弈权重计算

输入：基础评测指标集合 $M=\{M_1,\cdots,M_i,\cdots,M_n\}$，环境集合 E，演化强度参数 α

输出：权重向量 $\boldsymbol{w}^{(g)}=(w_1,w_2,\cdots,w_n)$

初始化收益矩阵 $\boldsymbol{U}\in\mathbb{R}^{n\times n}$

for $i=1$ to n

  for $j=1$ to n

    $U_{i,j}\leftarrow\dfrac{1}{|E|}\sum_{e\in E}\tau^{(e)}(M_i,M_j)$

定义状态空间 $S=\{(i,j)\mid i,j=1,2,\cdots,n\}$

初始化转移矩阵 $\boldsymbol{T}\in\mathbb{R}^{|S|\times|S|}$

for each (i,j) in S:

  for $i'=1$ to n

    if $i'\neq i$:

      $T_{(i,j)\to(i',j)}\leftarrow\exp\left(\alpha(U_{i',j}-U_{i,j})\right)$

  for $j'=1$ to n

    if $j'\neq j$:

      $T_{(i,j)\to(i,j')}\leftarrow\exp\left(\alpha(U_{i,j}-U_{i,j'})\right)$

归一化 $\boldsymbol{T}$ 的每行

求解 $\boldsymbol{\pi T}=\boldsymbol{\pi}$，得到稳态分布 $\boldsymbol{\pi}$

for $i=1$ to n

  $w_i^{(g)}\leftarrow\sum_j\dfrac{\pi_{i,j}}{\sum_{(i,j)\in S}\pi_{i,j}}$

返回 $\boldsymbol{w}^{(g)}$

注击打成功率而非平均奖励。为使最终权重既能体现数据的一致性，又保留人为调控能力，本文设计如下融合机制：

$$w_i=\lambda\cdot w_i^{(g)}+(1-\lambda)\cdot w_i^{(h)} \tag{4}$$

其中，$\lambda\in[0,1]$，表示融合超参数，用于控制主客观权重的平衡程度。

综上所述，本节提出了一种融合客观博弈与主观偏好的多维指标融合框架，创新性地引入博弈论视角，在指标间建立起基于排序认可度的双边对抗关系，通过α-Rank算法计算出反映指标稳定性与群体共识的客观博弈权重。同时，考虑到实际任务中某些指标的重要性往往源于业务逻辑或专家经验，本节进一步引入专家先验权重。这一融合机制与鲁棒性评分指标共同构建起了多维自适应融合鲁棒性评估方法，确保了基础指标权重的合理性、稳定性与可解释性，为多智能体策略的鲁棒性评估提供了统一的度量标准。

3 实验验证与结果分析

3.1 实验设置

为了验证鲁棒性评估方法的有效性，本文自主开发了两个典型的多智能体任务场景：围捕任务与搜救任务。在围捕场景中，三名追捕者需协作追捕一名逃逸者；在搜救场景中，两名搜救智能体需在复杂环境中快速寻找被困者位置。两个任务涵盖了对抗与协作这两大基本范式，能够有效反映评估框架在真实复杂场景下的通用性。

实验平台基于Isaac Sim构建，整合了其底层API和自定义扩展的Gym[49]环境框架，是一套具备高拓展性、支持自动评估和难度调节的多智能体仿真系统。平台支持速度定义、障碍物布置、采样数据记录，可用于多种类型的智能体算法训练与测试。

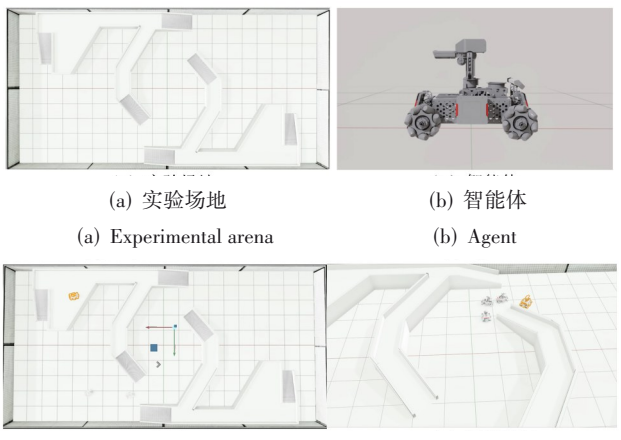
实验场景复刻真实测试场地，整体尺寸为20 m × 10 m，四周设置物理边界以限定任务区域。场地内布置斜坡、高台等大型结构，并在此基础上设计可调节的环境难易度模式。如表2所示，在简单模式下，高台与斜坡区域被封闭，场地转化为带固定障碍物的平地环境；在困难模式下，相关地形保持开放，智能体需要在复杂起伏环境中完成任务。难易度模式的灵活切换，有效提升了实验场景的多样性与灵活性。

实验采用四轮全向驱动小车作为智能体模型，其具备灵活的前进、横移和原地旋转能力。通过仿真接口，每台小车均配置了传感器，以模拟真实感知。图3展示了实验场地、智能体和任务场景。

表2 场景难度模式定义

Table 2 Definition of scenario difficulty modes

难度模式	定义
简单模式	高台与斜坡封闭，形成带障碍物的平地环境
困难模式	高台与斜坡开放，形成复杂起伏的地形环境



(a) 实验场地 (b) 智能体

(a) Experimental arena (b) Agent

(c) 救援场景 (d) 围捕场景

(c) Rescue scenario (d) Capture scenario

图3 仿真测试场地与智能体

Figure 3 Simulation test scene and agents

实验选取了TD3[50]、SAC、PPO、DDPG、CrossQ[51]、TRPO[52]和TQC[53]七种主流算法，在表3所示的多种环境下进行测试。为全方位衡量博弈策略的表现，实验选取了成功率、平均奖励、平均碰撞次数等多个评估指标，涵盖了任务完成度、协作效率及安全性等多个方面。

表3　测试环境参数配置

Table 3　Configuration of test environments

环境名称	逃逸者速度	初始位置	障碍物数量	摩擦系数
default	默认(0.8)	场地随机生成	0	0.2(静),1.0(动)
speed0	增大两倍(1.6)	场地随机生成	0	0.2(静),1.0(动)
speed1	减小一半(0.4)	场地随机生成	0	0.2(静),1.0(动)
obstacle0	默认(0.8)	场地随机生成	3	0.2(静),1.0(动)
obstacle1	默认(0.8)	场地随机生成	1	0.2(静),1.0(动)
poses0	默认(0.8)	初始位置较远、对立状	0	0.2(静),1.0(动)
poses1	默认(0.8)	初始位置较近、对立状	0	0.2(静),1.0(动)
friction0	默认(0.8)	场地随机生成	0	0.5(静),2.0(动)
friction1	默认(0.8)	场地随机生成	0	0.1(静),0.5(动)

3.2　鲁棒性评分指标评估实验与分析

为建立鲁棒性的直观基准,实验首先以围捕环境为例,利用雷达图(如图4所示)对各算法在跨环境下的表现进行可视化分析。图中展示了各算法在不同环境设定下各指标的表现,其中可分为以下几类。

(1)高稳定性组(如DDPG、CrossQ):DDPG与CrossQ算法的雷达图轮廓最为紧凑且规则。这意味着无论环境参数如何变化,其各项指标均未发生剧烈波动,表现出极强的适应能力。

(2)低稳定性组(如PPO、TRPO):相比之下,PPO和TRPO的雷达图呈现出明显的"尖峰"与"凹陷"特征,且覆盖面积在不同环境间差异巨大,这表明其性能对环境参数高度敏感。

(3)高收益局部波动组(如TQC):值得注意的是,TQC算法虽然在成功率等维度的覆盖面积较大、波动

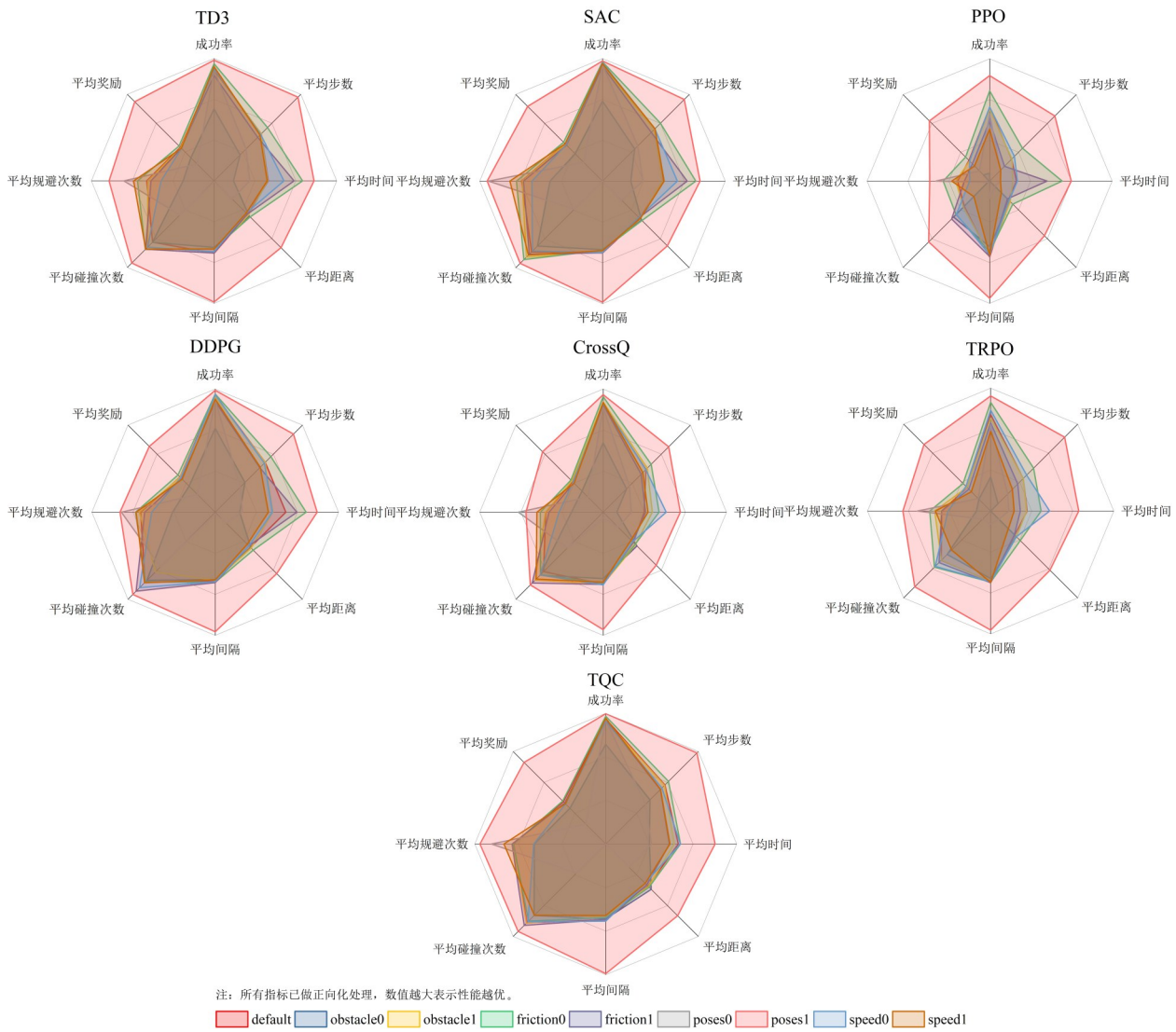


图4　不同环境下各算法多维基础指标雷达图

Figure 4　Radar chart of multidimensional basic metrics for algorithms in different test environments

较小，但在另一些维度（如碰撞次数）的边缘表现出参差不齐的锯齿状，呈现一定的方差。

基于算法1对实验数据进行权重计算，所得结果如表4所示。结果显示，平均碰撞次数在两个任务中均获得了较低的博弈权重（分别为0.124和0.107），而成功率、平均奖励与平均逃逸次数则获得了较高的权重分配（均在0.20以上）。这说明从数据的内在规律来看，平均碰撞次数与其他基础指标的排序一致性较弱，即单纯的低碰撞并不一定代表策略的高性能。基于博弈的融合机制捕捉了测试数据间的一致性差异，自适应地降低了离群指标在总评估中的比重，确保了最终结果能够反映策略在多维视角下的真实性能。

表4 各基础评测指标的权重分配结果

Table 4 Weight allocation for basic evaluation metrics

环境类型	指标名称	基于博弈的权重	专家先验权重	总权重
围捕	成功率	0.205	0.350	0.278
	平均步数	0.217	0.100	0.159
	平均碰撞次数	0.124	0.150	0.137
	平均逃逸次数	0.228	0.100	0.164
	平均奖励	0.225	0.300	0.263
搜救	成功率	0.262	0.400	0.331
	平均步数	0.317	0.100	0.208
	平均碰撞次数	0.107	0.150	0.128
	平均奖励	0.314	0.350	0.332

进一步依据式（1）对各算法进行鲁棒性评分计算，结果如表5所示。结果显示，DDPG与CrossQ获得了最高的鲁棒性评分，而形态波动最大的PPO与TRPO评分最低。这一定量排名与雷达图的形态表现高度契合，表明本文的评估方法能准确地将策略在不同环境下的稳定性转化为量化指标。

表5 各指标条件变异系数值及鲁棒性值结果

Table 5 Conditional coefficient of variation values and robustness scores for each metric

环境类型	算法	成功率	平均步数	平均奖励	平均碰撞次数	平均逃逸次数	鲁棒性值
围捕	TD3	0.071	0.223	0.890	0.280	0.312	1.369
	SAC	0.061	0.208	0.739	0.389	0.379	1.395
	PPO	0.173	0.192	0.952	0.200	0.170	1.274
	DDPG	0.054	0.209	0.693	0.216	0.245	1.594
	CrossQ	0.071	0.154	0.737	0.226	0.198	1.579
	TRPO	0.129	0.213	1.101	0.283	0.236	1.198
	TQC	0.040	0.208	0.667	0.383	0.339	1.559
搜救	TD3	0.068	0.128	0.540	0.315	—	1.669
	SAC	0.016	0.052	0.246	0.554	—	2.526
	PPO	0.028	0.026	1.066	0.249	—	2.106
	DDPG	0.068	0.154	0.542	0.673	—	1.532

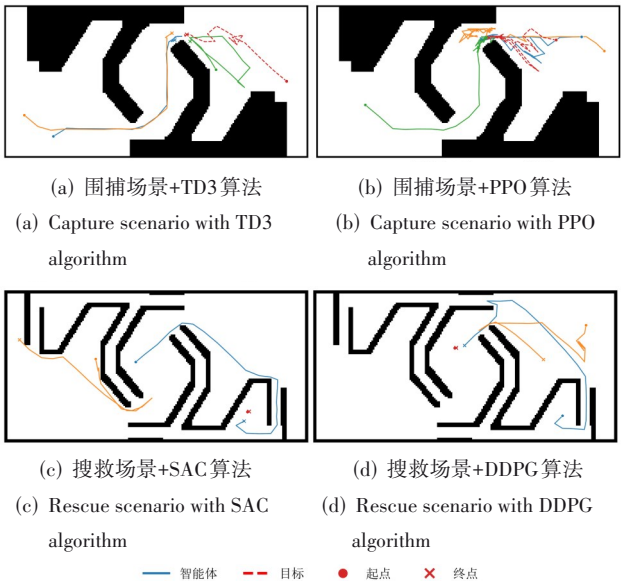
注：“—”表示该场景下此指标不适用。

另一方面，本文提出的多维融合方式能有效解决单一维度评估的局限性。以围捕任务中的TQC算法为例，尽管其在成功率和平均奖励等传统核心指标上表现出极高的稳定性（CondCV值按升序排名均为第一），但在平均碰撞次数等维度上的条件变异系数却显著高于均衡型算法DDPG。在传统评价体系下，TQC本应凭借核心指标的优势优于DDPG，但多维自适应融合机制修正了对其鲁棒性的过度高估，特别是对数变换显著增强了指标对波动性的敏感度，使得最终TQC的鲁棒性值低于DDPG。上述结果表明，本文的方法具备敏锐的多维波动捕捉能力，能够有效识别并修正单一指标的评估偏见，确保了评估结果能够真实、全面地反映策略在复杂环境下的整体稳定性。

此外，如表5所示，通过对比围捕和搜救这两个难度不同的任务，可以发现PPO算法在复杂的围捕任务中排名靠后，但在相对简单的搜救任务中却以2.106的高分跃居第二。这客观反映出PPO更适合处理环境简单的协作任务，而在复杂的对抗环境中则显得力不从心。这说明，鲁棒性评分不仅能适配不同类型的任务，还能针对场景难度给出差异化评价，从而有效衡量算法在不同环境下的适应性，为实际应用中的算法选择提供参考。

最后，图5展示了围捕（简单模式）与搜救（困难模式）任务中各算法的智能体运动轨迹。对比围捕场景下的图5（a）与图5（b），可以观察到，鲁棒性评分较高的TD3算法呈现出平滑且果断的追踪轨迹，智能体之间协作紧密，能够以较优路径逼近目标；相比之下，评分较低的PPO算法虽然也能完成任务，但其轨迹充斥着大量曲折与无意义的抖动，直观解释了其在平均步数等指标上的高变异系数。同理，如图5（c）

与图5(d)所示，搜救场景中，鲁棒性评分略优的SAC算法在路径规划上也表现得更为干练，相比于DDPG算法轨迹中出现的冗余迂回，SAC的搜索路径更加平直高效。上述的算法行为表现差异与定量评分结果互为支撑，进一步验证了评估方法在衡量策略稳定性时的直观有效性，确保了鲁棒性量化结果不仅具有数学上的严谨性，更具备实际物理场景下的说服力。



(a) 围捕场景+TD3算法　(b) 围捕场景+PPO算法

(a) Capture scenario with TD3 algorithm　(b) Capture scenario with PPO algorithm

(c) 搜救场景+SAC算法　(d) 搜救场景+DDPG算法

(c) Rescue scenario with SAC algorithm　(d) Rescue scenario with DDPG algorithm

图5　不同任务场景和算法下智能体运动轨迹

Figure 5　Trajectories of agents under different task scenarios and algorithms

综上所述，通过雷达图分析、定量评分及轨迹可视化的共同验证，所提出的鲁棒性评估方法能够在不同策略表现接近的情况下体现稳定性差异，并对高波动策略形成有效区分，从而为策略鲁棒性比较提供了可靠的分析视角。

3.3　消融实验

为分析权重优化与多维融合模块对评估结果的影响，在围捕场景下设计如下三组对比实验，以验证多维自适应融合评估方法的合理性。

Ours：采用本文提出的框架，包含基于CondCV的鲁棒性评分、基于博弈框架的自适应权重计算及多维融合机制；

Equal-W：去除了权重计算环节，将所有指标权重强制设为均等($w_i = 1/n$)，模拟缺乏融合框架下的朴素评估；

SR-Only：仅计算成功率这一单一核心维度的稳定性，即$-\log(\mathrm{CondCV(SR)})$。

图6展示了不同设置下的鲁棒性评分对比。首先，Ours与Equal-W的对比说明了合理分配权重的必要性。在Equal-W的等权重机制下，由于缺乏对指标优先级的区分，核心指标的贡献被次要指标拉平，从而丧失了对算法优劣的有效辨识。以TD3与PPO的对比为例，TD3在成功率、平均奖励等核心维度上极具稳定性，但在次要指标上波动较大；PPO的表现则相反。在Equal-W方法下，二者的优劣势相互中和，导致两者的综合评分非常接近。而在Ours方法中，博弈框架依据指标间的一致性赋予了指标不同的权重，凸显了TD3在高权重指标上的优异表现，使其最终评分显著高于PPO。这证实了自适应权重机制能够有效克服等权重分配导致的区分度不足问题，确保评估结果与任务的核心目标保持一致。

随后，Ours与SR-Only的对比则揭露了单维度评估的不足。SR-Only完全依赖成功率，容易得出片面的高分评价。以TRPO为例，仅观察成功率指标时，其评分高于PPO，似乎表现更佳。然而，一旦引入多维视角的考量(无论在Ours还是Equal-W中)，TRPO的排名会跌落至PPO之后。这说明TRPO在成功率上的表面优势，并未能掩盖其在其他维度上的剧烈波动。本文的多维融合机制有效弥补了单一视角的盲区，避免了仅依赖单一指标可能产生的评估偏差，有力证明了构建多维融合体系在全面衡量策略鲁棒性方面的必要性。

综上所述，消融实验从两个维度验证了本文方法的严谨性：Equal-W组的对比证实了权重优化机制提升了评估的区分度与合理性，SR-Only组的对比则表明多维融合机制能有效规避单一指标的局限性。两者的有机结合，使得本文提出的方法能够在核心性能与综合稳定性之间取得平衡，为多智能体博弈策略提供了一个既具区分度又具全面性的标准化度量。

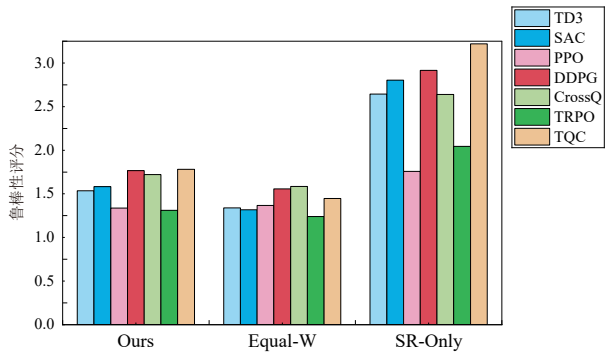


图6　消融实验的鲁棒性评分对比

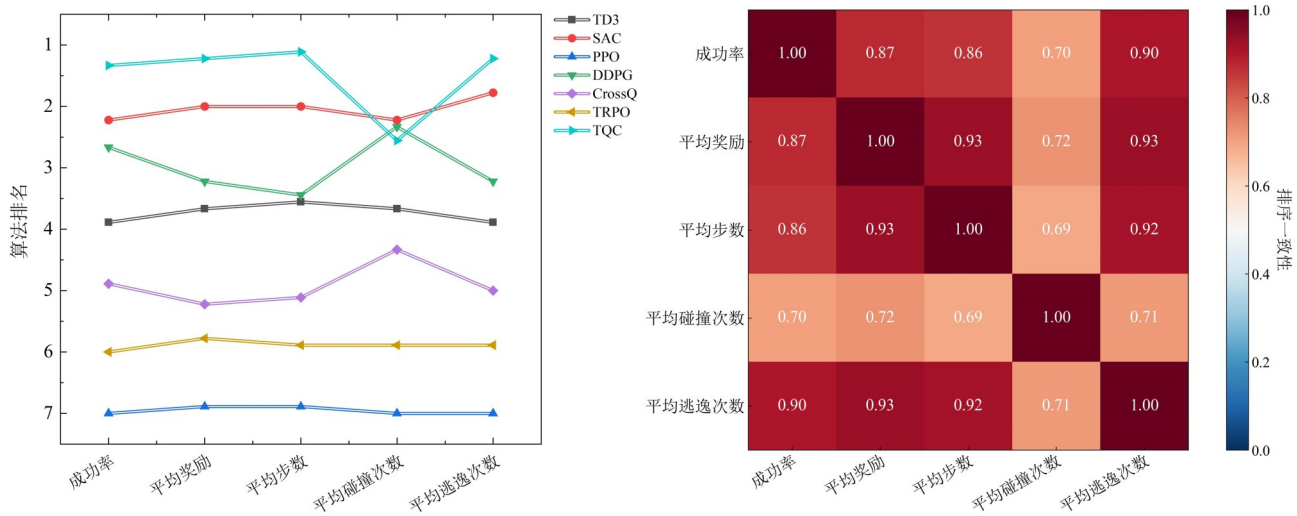
Figure 6　Ablation study results of robustness scores

3.4　自适应博弈权重的合理性分析

为了验证本文提出的自适应融合框架的合理性，本节将分析各基础评测指标对算法排序的一致性与

相关性。图7以围捕任务为例，展示了各算法在不同指标下的排名演变图及指标间的Kendall's τ相关性热力图。

根据博弈模型的定义，若某指标对策略的排序与其他指标相似，则该指标在博弈权重的计算中应被赋予较高的权重；反之则应被分配较低的权重。如图7(a)所示，在成功率、平均奖励及平均逃逸次数等指标之间，各算法的排名连线呈现出稳定的平行趋势。例如，TQC算法在这些指标上始终保持前列(排名1~2)，而PPO算法则稳定处于末位。这说明这些指标对策略性能的刻画具有较高一致性。与此相对应，在对抗式零和博弈的演化过程中，这些指标逐步形成了较高权重(成功率0.205，平均奖励0.225，平均逃逸次数0.228)。



(a) 算法在不同指标下的排名稳定性

(a) Ranking stability of algorithms across different metrics

(b) 指标间排序一致性热力图

(b) Heatmap of ranking consistency among metrics

图7 不同环境下指标排序稳定性分析

Figure 7 Stability analysis of metric rankings under different environments

相反，在平均碰撞次数指标处，排名连线呈现交叉与翻转的趋势。例如，TQC算法虽然在任务完成度上表现最优，但在碰撞指标上却跌至中游；而部分在其他指标表现平平的算法(如DDPG)在此处排名却异常上升。这表明该指标与其他评价维度的一致性较弱，博弈算法因此给予了其较低的权重，以避免对综合评估产生过大影响。

图7(b)的热力图进一步量化了指标间的相关性结果。结果显示，核心指标之间(如成功率与平均奖励)的Kendall's τ相关系数高达0.87以上，呈现深红色(高相关区)；而次要指标(平均碰撞次数)与其他指标的相关系数仅为0.70左右，颜色显著变浅(低相关区)。该相关性分布与自适应权重结果保持一致，从数据统计层面支持了权重分配的合理性。

综上，自适应融合框架根据指标间的一致性差异进行调整，在突出核心指标作用的同时，保留其他指标在综合评价中的参考价值。通过对抗式零和博弈建模，不同指标在相互制衡与竞争中形成稳定权重，减少了人为设定带来的主观偏差，也增强了权重分配过程的自适应性与解释性。相比固定权重或单一指标评价方式，该自适应评价体系能够更加客观、全面地反映多智能体策略在复杂环境下的整体鲁棒性表现。

3.5 训练过程中的鲁棒性演化分析

为了深入分析多智能体博弈策略在不断强化迭代过程中的鲁棒性变化规律，本文选取了两种具有代表性的算法：TD3(代表高稳定性、off-policy架构)和PPO(代表低稳定性、on-policy架构)在多智能体围捕任务下进行实验，记录了从早期训练到最终收敛的多个训练节点上的表现。

实验结果如图8所示。从整体趋势来看，随着训练步数的增加，TD3与PPO两种算法的鲁棒性评分均呈现出总体上升的态势。这表明随着智能体与环境交互次数的累积，算法在学习如何协作完成任务以最大化奖励的同时，应对环境扰动与动态变化的能力也在逐步增强。在算法对比方面，TD3算法的鲁棒性评分曲线总体位于PPO算法之上，这一结果与表5中基于最终策略的评估结果一致，说明不同算法间的鲁棒性差异并非训练后期的偶然结果，验证了本文提出的评估方法在反映策略鲁棒性动态规律方面，同样具有一致性与可靠性。

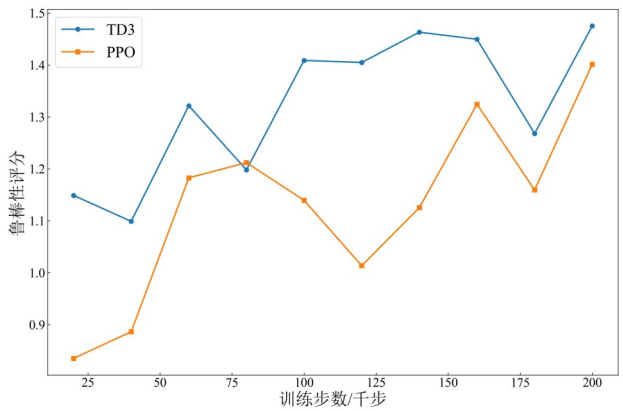


图8　训练过程中鲁棒性评分演化曲线

Figure 8　Evolution curves of robustness scores during training

值得注意的是,两条鲁棒性演化曲线并非单调上升,而是伴随着不同程度的波动。例如,PPO算法在120 000步附近出现了显著的评分回落,TD3算法在180 000步时也出现了短暂的性能震荡。这种波动反映了算法训练过程中的不稳定性:在策略更新的特定阶段,智能体可能为了在训练环境中追求更高的回报而过度拟合特定状态,导致其在跨环境测试中的泛化能力暂时下降。在这种情况下,传统的平均奖励等指标通常仍呈现大体上升的趋势;而本文方法通过捕捉多维度的波动特征,能够敏锐识别出此类鲁棒性失准。因此,该方法不仅能为训练过程中的过拟合监测提供预警,也为选取更具稳定性的策略节点提供了科学的量化依据。

4　讨论

4.1　虚实迁移的潜在应用探讨

本文提出的方法虽主要用于评估多智能体博弈策略在仿真环境中的鲁棒性,但其设计思路具备扩展至衡量多智能体博弈策略的仿真到现实(sim-to-real)泛化能力的潜在价值。在多智能体系统中,策略往往先在仿真环境中进行训练与评估,随后部署至现实环境中。但由于现实环境中的物理不确定性、感知误差与动态扰动,策略的实际表现常常与仿真结果存在一定偏差。近年来,关于sim-to-real的研究逐渐成为强化学习领域的重要分支,然而目前仍缺乏统一且多维度的标准化指标用于系统衡量该差异。因此,设计一套能合理刻画仿真到现实迁移差异的指标体系,是推动策略泛化评估的重要方向。

为衡量策略在模拟环境与现实环境之间的表现偏差,可在延续多维自适应融合评估思想的基础上,对虚实迁移差异进行统一建模,并通过比较策略在不同环境下的相对性能变化进行量化。具体而言,采用归一化的相对性能偏差$D(M)$作为差异度量指标:

$$D(M)=\frac{|M_{\text{sim}}-M_{\text{real}}|}{|M_{\text{sim}}|+|M_{\text{real}}|+\varepsilon} \tag{5}$$

其中,M_{sim}和M_{real}分别表示策略在虚拟环境与真实环境中的基础评测指标。进一步定义泛化性评分(Generalization Score,GS):

$$\text{GS}=-\log\sum_{i=1}^{n}q_i D(M_i) \tag{6}$$

其中,n为基础评测指标数,q_i为第i个基础评测指标的权重。

该评分机制继承了自适应融合的思想,因此同样具有全面性、通用性以及标准化衡量的特点。在未来的研究中,该方法可进一步结合现实部署测试,构建真实反馈数据集,并通过上述指标对策略迁移的可靠性进行建模与优化。这将为多智能体策略从虚拟世界向真实世界的部署提供一种可扩展的评估思路,并为sim-to-real泛化能力的量化分析提供参考框架。

4.2　研究局限性与未来展望

尽管本研究提出并构建了一套针对多智能体博弈策略的评估方法,但该方法仍存在以下局限性与待改进之处。

(1)环境与算法的局限性:尽管实验的场景(如围捕和搜救场景)、智能体策略(如TD3、PPO等)已具有一定多样性,但对于模拟真实场景下环境和策略的丰富变化上仍存在一定局限,无法完整反映算法在所有真实环境中的鲁棒性。未来可以考虑引入更多的测试环境和智能体策略,以进一步验证方法的有效性。

(2)缺乏针对性的改进指导:目前的鲁棒性评估仅提供定量分析结果,尚未形成直接指导算法优化的具体策略。未来可基于计算所得的鲁棒性评分值,进一步构建反馈机制,为算法参数调整或模型改进提供更具针对性的优化建议,以提升模型的实际应用效果。

未来,该方法可进一步与现有主流评测平台结合,支持在更复杂环境设定下开展多维度鲁棒性评估。随着仿真环境复杂度与算法能力的持续提升,所提出的评估框架有望在跨场景性能分析与算法比较中发挥更大作用,并为多智能体强化学习在具身智能、无人系统和智能交通等实际应用中的可靠部署提供评价依据。

5　结论

本文围绕多智能体博弈策略的鲁棒性评估问题,提出了一种多维自适应融合评估方法。首先,基于条件变异系数构建了鲁棒性评分机制,形成了较为全面的评估指标体系;其次,设计了基于α-Rank对抗演化博弈的多维自适应融合框架,使指标权重既能够融入专家先验,又可通过自适应博弈过程实现相对客观、

稳定的调整。最后,基于Isaac Sim平台构建了对抗与协作两类多智能体博弈场景,并对多种主流算法进行了评估。实验结果表明,所提出的评估方法能够在多种环境设定下对算法鲁棒性进行稳定、综合的量化分析,为不同算法之间的比较提供统一且可解释的评价依据。最后,本文探讨了该评估方法在虚实迁移中的潜在应用。未来工作将进一步扩展算法与测试环境规模,并探索将鲁棒性指标用于模型训练与性能优化的可能性。

参考文献

[1] Watkins C J C H, Dayan P. Q-learning[J]. Machine Learning, 1992, 8(3/4): 279-292.

[2] Rummery G A, Niranjan M. On-line Q-learning using connectionist systems[R]. Cambridge: University of Cambridge, 1994.

[3] Liu H J, Ruan S L, Liu Q, et al. Global structure-aware and feature-augmented graph neural network for heterophilic graphs[J]. ACM Transactions on Information Systems, 2026, 44(2): 1-28.

[4] Wang X, Wang S, Liang X X, et al. Deep reinforcement learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5064-5078.

[5] 顾健华, 冯建华, 许辉阳, 等. 基于有向图与卷积网络强化学习的端侧协同算力资源分配方法[J]. 电子学报, 2025, 53(6): 1771-1783.
Gu Jianhua, Feng Jianhua, Xu Huiyang, et al. Directed graph and convolutional network reinforcement learning for terminal-side collaborative computing resource allocation scheme[J]. Acta Electronica Sinica, 2025, 53(6): 1771-1783. (in Chinese)

[6] 王为念, 苏健, 陈勇, 等. 基于多智能体深度强化学习的车联网频谱共享[J]. 电子学报, 2024, 52(5): 1690-1699.
Wang Weinian, Su Jian, Chen Yong, et al. Multi-agent reinforcement learning enabled spectrum sharing for vehicular networks[J]. Acta Electronica Sinica, 2024, 52(5): 1690-1699. (in Chinese)

[7] 文鹏, 叶苗, 王勇, 等. SDWN中基于多智能体图强化学习的多对多通信路由方法[J]. 电子学报, 2025, 53(6): 1885-1905.
Wen Peng, Ye Miao, Wang Yong, et al. A multi-agent graph reinforcement learning method for many-to-many communication routing in SDWN[J]. Acta Electronica Sinica, 2025, 53(6): 1885-1905. (in Chinese)

[8] Littman M L. Value-function reinforcement learning in Markov games[J]. Cognitive Systems Research, 2001, 2(1): 55-66.

[9] Lowe R, Wu Yi, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing System. New York: Curran Associates, Inc., 2017: 6382-6393.

[10] Sun H R, Wu Y S, Cheng Y K, et al. Game theory meets large language models: A systematic survey[C]//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2025: 10669-10677.

[11] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[12] Huang H, Hu Z Q, Li M Y, et al. Cooperative optimization of traffic signals and vehicle speed using a novel multi-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2024, 73(6): 7785-7798.

[13] Zhu Y J, Chen M Z, Wang S H, et al. Collaborative reinforcement learning based unmanned aerial vehicle (UAV) trajectory design for 3D UAV tracking[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 10787-10802.

[14] Dimitropoulos K, Hatzilygeroudis I, Chatzilygeroudis K. A brief survey of Sim2Real methods for robot learning[M]//Advances in Service and Industrial Robotics. ChamSpringer International Publishing, 2022: 133-140.

[15] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017: 2817-2826.

[16] Tessler C, Efroni Y, Mannor S. Action robust reinforcement learning and applications in continuous control[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 6215-6224.

[17] Lee X Y, Ghadai S, Tan K L, et al. Spatiotemporally constrained action space attacks on deep reinforcement learning agents[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 4577-4584.

[18] Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2017: 23-30.

[19] Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization[C]//2018 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2018: 3803-3810.

[20] Geng M H, Pateria S, Subagdja B, et al. MOSMAC: A multi-agent reinforcement learning benchmark on sequential multi-objective tasks[J]. Proceedings of the Third International Joint Conference on Autonomous Agents and

Multiagent Systems - Volume 1, 2025: 867-876.

[21] Zheng X, Ma X J, Wang S J, et al. Toward evaluating robustness of reinforcement learning with adversarial policy[C]//2024 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Piscataway: IEEE, 2024: 288-301.

[22] 林谦, 余超, 伍夏威, 等. 面向机器人系统的虚实迁移强化学习综述[J]. 软件学报, 2024, 35(2): 711-738.
Lin Qian, Yu Chao, Wu Xiawei, et al. Survey on sim-to-real transfer reinforcement learning in robot systems[J]. Journal of Software, 2024, 35(2): 711-738. (in Chinese)

[23] Samvelyan M, Rashid T, Schroeder de Witt C, et al. The StarCraft multi-agent challenge[J]. Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1, 2019: 2186-2188.

[24] Bard N, Foerster J N, Chandar S, et al. The Hanabi challenge: A new frontier for AI research[J]. Artificial Intelligence, 2020, 280: 103216.

[25] Kurach K, Raichuk A, Stańczyk P, et al. Google research football: A novel reinforcement learning environment[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 4501-4510.

[26] Omidshafiei S, Papadimitriou C, Piliouras G, et al. α-Rank: Multi-agent evaluation by evolution[J]. Scientific Reports, 2019, 9: 9937.

[27] NVIDIA. NVIDIA Isaac sim[EB/OL]. [2026-02-14]. https://developer.nvidia.com/isaac-sim.

[28] Wang J D, Lan C L, Liu C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 8052-8072.

[29] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[30] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[PP/OL]. V2. arXiv (2017-08-28)[2026-02-14]. https://doi.org/10.48550/arXiv.1707.06347.

[31] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[PP/OL]. V6. arXiv (2019-07-05)[2026-02-14]. https://doi.org/10.48550/arXiv.1509.02971.

[32] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 1861-1870.

[33] Ning Z P, Xie L H. A survey on multi-agent reinforcement learning and its application[J]. Journal of Automation and Intelligence, 2024, 3(2): 73-91.

[34] Hu Junling, Wellman M P. Nash q-learning for general-sum stochastic games[J]. The Journal of Machine Learning Research, 2003, 4: 1039-1069.

[35] Rashid T, Samvelyan M, Schroeder C, et al. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement Learning[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 4295-4304.

[36] Bayen A, Gao J X, Velu A, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]//Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 24611-24624.

[37] Kuba J G, Chen Ruiqing, Wen Muning, et al. Trust region policy optimisation in multi-agent reinforcement learning[C/OL]//Proceedings of the 10th International Conference on Learning Representations, 2022: 1-27[2026-02-15]. https://openreview.net/forum?id=EcGGFkNTxdJ.

[38] Li Simin, Guo Jun, Xiu Jingqiao, et al. Byzantine robust cooperative multi-agent reinforcement learning as a Bayesian game[C/OL]//Proceedings of the 12th International Conference on Learning Representations, 2024: 1-27[2026-02-15]. https://openreview.net/forum?id=z6KS9D1dxt.

[39] Zhou Z Y, Liu G J, Zhou M C, et al. Robust multi-agent reinforcement learning with stochastic adversary[C]//Proceedings of the 42nd International Conference on Machine Learning. New York: ACM, 2025: 79004-79027.

[40] Lee S, Hwang J, Jo Y, et al. Wolfpack adversarial attack for robust multi-agent reinforcement learning[C]//Proceedings of the 42nd International Conference on Machine Learning, 2025: 33025-33056.

[41] Ruan S L, Zhang Y, Zhang K, et al. DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 13940-13949.

[42] Ruan S L, Liu H J, Chen Z, et al. CPWS: Confident programmatic weak supervision for high-quality data labeling[J]. ACM Transactions on Information Systems, 2025, 43(4): 1-26.

[43] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg: ACL, 2018: 353-355.

[44] Srivastava A, Rastogi A, Rao A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[J]. Transactions on Machine Learning Research, 2023, 2023(5): 1-95.

[45] Bettini M, Prorok A, Moens V. BenchMARL: Benchmarking multi-agent reinforcement learning[C]//New York: ACM, 2024: 10557-10566.
[46] Papadopoulos G, Kontogiannis A, Papadopoulou F, et al. An extended benchmarking of multi-agent reinforcement learning algorithms in complex fully cooperative tasks[J]. Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1, 2025: 1613-1622.
[47] Li Simin, Mao Zihao, Li Hanxiao, et al. Empirical study on robustness and resilience in cooperative multi-agent reinforcement learning[C]//Advances in Neural Information Processing Systems 38. New York: Curran Associates, Inc., 2025.
[48] Kendall M G. A new measure of rank correlation[J]. Biometrika, 1938, 30(1/2): 81-93.
[49] Brockman G, Cheung V, Pettersson L, et al. OpenAI gym[PP/OL]. V1. arXiv (2016-06-05) [2026-02-14]. https://doi.org/10.48550/arXiv.1606.01540.
[50] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 1587-1596.
[51] Bhatt A, Palenicek D, Belousov B, et al. CrossQ: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity[C/OL]//The 12th International Conference on Learning Representations, 2024: 1-19[2026-02-15]. https://openreview.net/forum?id=PczQtTsTIX.
[52] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015: 1889-1897.
[53] Kuznetsov A, Shvechikov P, Grishin A, et al. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 5556-5566.

作者简介

李骏唯 男，2002年9月出生于广东省广州市。现为清华大学深圳国际研究生院硕士研究生。主要研究方向为大模型持续学习、多智能体系统等。
E-mail: li-jw25@mails.tsinghua.edu.cn

阮书岚 男，1996年9月出生于安徽省芜湖市。现为清华大学深圳国际研究生院助理研究员、博士后。主要研究方向为多模态理解、多智能体系统智能决策等。中国电子学会会员编号：E190188544M。
E-mail: slruan@sz.tsinghua.edu.cn

梁嘉旋 男，1998年12月出生于广东省深圳市。现为哈尔滨工业大学（深圳）博士研究生。主要研究方向为算法评估、深度强化学习、多智能体系统等。
E-mail: 24B951014@stu.hit.edu.cn

刘 瑜 男，1986年12月出生于湖南省邵阳市。现为清华大学电子工程系研究员，博士生导师。长期从事多模态信息智能融合、无人系统智能博弈等方向的研究。发表学术论文80余篇，获省部级科技一等奖4项、二等奖3项。中国电子学会会员编号：E190198505M。
E-mail: liuyu_thu@mail.tsinghua.edu.cn

何 友 男，1956年10月出生于吉林省磐石市。现为中国工程院院士。主要研究方向为信号检测、信息融合、智能技术与应用研究。以第一完成人获国家科技进步奖二等奖4项、国家教学成果一、二等奖各1项，获省部级一等奖11项。
E-mail: heyou@mail.tsinghua.edu.cn