计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

基于贡献度优化的异质智能体移动群智感知自适应 PPO 协同方法

蒋伟进[1),2)] 欧阳宇明[1),2)] 石顺究[1),2)] 许宇胜[3)] 刘溪溪[1),2)] 崔新雨[1),2)] 罗贞焱[1)] 蒋意容[2)]

[1)](湖南工商大学 计算机学院, 长沙 410205)

[2)](湘江实验室 长沙 410205)

[3)](北方导航控制技术股份有限公司 北京 100176)

摘 要 城市道路网络移动群智感知（Vehicular Mobile Crowd Sensing in Urban Road Networks, VMCS）旨在实现大规模、高时效性的城市数据采集，其核心效能取决于智能体的协同机制与多源传感器数据融合。然而，现有方法多基于智能体同质化假设，忽视了智能体在时空分布和能力上的异质贡献，导致策略震荡、收敛低效与扩展受限，严重制约了系统在动态开放环境中的实际部署。针对这一挑战，本文提出贡献感知自适应近端策略优化算法（Contribution-Aware Adaptive Proximal Policy Optimization, CA2-PPO），该算法通过贡献度加权熵实现差异化激励，利用动态 Kullback-Leibler（KL）散度裁剪自适应调节稳定性与收敛效率，并引入优势函数裁剪与归一化技术以提升策略更新的鲁棒性。算法协同路侧单元（Roadside Unit, RSU）实现任务调度与资源分配的全局优化。为全面评估算法性能，本文在网格仿真环境与 OpenStreetMap（OSM）真实路网数据集上开展了系统性实验。实验结果表明，在网格环境中，20×20 网格 4 智能体场景下，CA2-PPO 平均覆盖得分较 MADQN、MACPPO 分别提升 22.2%、28.3%；30×30 网格 16 智能体场景下分别提升 7.9%、4.8%。此外，在真实路网测试中，CA2-PPO 的平均覆盖得分稳定在 0.60~0.95 区间。综上，本文从理论与实证层面验证了贡献感知激励、动态裁剪与优势稳健处理的有效性，为异质智能体协同调度与信息物理融合的集群组网提供了可扩展的技术路径。

关键词 移动群智感知；多智能体强化学习；贡献度感知；近端策略优化；协同调度优化；异质智能体

中图法分类号 TP18

Adaptive PPO Collaboration Method for Heterogeneous Agent Mobile Crowd Sensing with Contribution Optimization

JIANG Wei-Jin[1),2)] OUYANG Yu-Ming[1),2)*] SHI Shun-Jiu[1),2)] XU Yu-Sheng[3)] LIU Xi-Xi[1),2)]

CUI Xin-Yu[1),2)] LUO Zhen-Yan[1)] JIANG Yi-Rong[2)]

[1)](School of Computer Science, Hunan University of Technology and Business, Changsha, Hunan 410205)

[2)](Xiangjiang Laboratory, Changsha, Hunan 410205)

[3)](Northern Navigation Control Technology Co., Ltd, Beijing 100176)

本课题得到国家自然科学基金(No.61772196)、湖南省自然科学基金(No.2020JJ4249)、湖南省教育厅科学研究重点项目(No.24A0446、No.24A0753)、湖南省研究生科研创新项目（LXBZZ2024347）、湖南工商大学研究生科研创新项目（No.CX2025YB028、No.CX2025YB029）等资助。**蒋伟进**，博士，二级教授，博士生导师，中国计算机学会（CCF）高级会员，主要研究领域为群智感知、多智能体强化学习、网络空间安全。**欧阳宇明**（**通信作者**），硕士研究生，CCF 学生会员，主要研究领域为群智感知、多智能体强化学习。**石顺究**，硕士研究生，CCF 学生会员，主要研究领域为资源调度、群智感知。**许宇胜**，博士，研究员级高工，主要研究领域为群智感知、多智能体强化学习、系统控制。**刘溪溪**，硕士研究生，CCF 学生会员，主要研究领域为群智感知、联邦学习。**崔新雨**，硕士研究生，CCF 学生会员，主要研究领域为联邦学习、群智感知。**罗贞焱**，博士，副教授，主要研究领域为群智感知、网络空间安全。**蒋意容**，博士，主要研究领域为群智感知、数字音乐。

Abstract Vehicular Mobile Crowd Sensing in Urban Road Networks (VMCS) is an emerging paradigm that aims to achieve large-scale, high-timeliness urban data collection by leveraging the mobility of vehicles as sensing carriers. As cities become increasingly instrumented and interconnected, the ability to continuously monitor road conditions, traffic dynamics, and environmental variables at scale has emerged as a foundational requirement for intelligent urban management. The core effectiveness of such systems depends on agent coordination mechanisms and multi-source sensor data fusion. However, existing methods predominantly assume agent homogeneity, overlooking the heterogeneous contributions of agents in spatiotemporal distribution and capability, leading to policy oscillation, inefficient convergence, and limited scalability, which severely constrain practical deployment in dynamic and open environments. In practice, the heterogeneity among agents is not merely a theoretical concern but a pervasive reality in urban road networks, where vehicles differ substantially in their sensing ranges, communication capabilities, energy constraints, and mobility patterns. Treating such inherently diverse agents as equivalent contributors inevitably introduces systematic biases into the learning process, causing well-performing agents to be under-rewarded while underperforming agents receive disproportionate influence over the shared policy. This mismatch between actual contribution and received incentive is a root cause of policy oscillation and convergence failure in existing multi-agent reinforcement learning frameworks applied to VMCS. To address this challenge, this paper proposes the Contribution-Aware Adaptive Proximal Policy Optimization (CA2-PPO) algorithm, a novel multi-agent reinforcement learning framework specifically designed for heterogeneous vehicular sensing scenarios. The proposed algorithm introduces three key technical innovations. First, it achieves differentiated incentivization through contribution-weighted entropy, which quantifies each agent's relative contribution to the overall sensing task and adjusts incentive signals accordingly. Second, CA2-PPO employs dynamic Kullback-Leibler (KL) divergence clipping to adaptively balance training stability and convergence efficiency, allowing the policy update process to respond flexibly to varying learning dynamics rather than relying on fixed clipping thresholds. Third, the algorithm introduces advantage function clipping and normalization techniques to enhance the robustness of policy updates, mitigating the risk of destabilizing gradient updates caused by outlier advantage estimates. The algorithm further cooperates with Roadside Units (RSUs) to achieve global optimization of task scheduling and resource allocation. To comprehensively evaluate the performance of the proposed algorithm, we conduct systematic experiments on both grid simulation environments and real road network datasets from OpenStreetMap (OSM). The grid simulation experiments provide controlled benchmarking across varying scales and agent configurations, while the OSM-based experiments validate the algorithm's generalization capability under realistic and topologically complex urban road structures. Experimental results demonstrate that in a 20×20 grid with 4 agents, CA2-PPO improves the average coverage score by 22.2% and 28.3% compared to MADQN and MACPPO, respectively; in a 30×30 grid with 16 agents, the improvements are 7.9% and 4.8%. Furthermore, in tests on real road networks, CA2-PPO's average coverage score remains stable in the range of 0.60 to 0.95. Overall, this paper verifies, from both theoretical and empirical perspectives, the effectiveness of contribution-aware incentives, dynamic clipping, and robust advantage handling as core components of an efficient multi-agent sensing framework. The proposed CA2-PPO algorithm provides a scalable and practical technical approach for coordinating heterogeneous agents and enabling cyber-physical integration in swarm-networked systems, laying a solid foundation for the next generation of intelligent urban sensing infrastructures that demand both adaptability and interpretability in complex real-world deployments.

Key words Mobile crowd sensing; Multi-agent reinforcement learning; Contribution awareness; Proximal policy optimization; Collaborative scheduling optimization; Heterogeneous agents

1 引言

随着城市化进程的持续推进与智慧城市建设的深入发展，城市感知系统对多智能体协同调度的需求日益迫切。移动群智感知（Mobile Crowd Sensing, MCS）以车辆、无人机等移动载体为感知单元，兼具灵活部署与广泛覆盖的优势，已在智慧交通[1]、环境监测[2]、公共安全[3]等领域得到广泛应用[4]。然而，移动载体固有的动态性与分布异质性，也给多智能体协同调度带来了新的挑战[5]。相较于传统固定传感器网络乃至其他主流感知方法，MCS 在覆盖广度与感知精度之间实现了更优的平衡[6]。在城市交通场景中，这一范式自然延伸为城市道路网络移动群智感知（Vehicular Mobile Crowd Sensing in Urban Road Networks, VMCS），并逐渐成为智慧城市感知领域的重要研究方向。

VMCS 由城市路网中的一组探测车辆构成，这些车辆既可采用集中调度与路线控制策略（如专用探测车）[7]，也可遵循其他行为范例（如出租车或公交车）[8][9]。作为MCS在交通场景中的典型应用，VMCS 以城市中大量异构移动参与者（如私家车、出租车、公交车及骑行设备等）为感知载体，能够实现灵活且经济的城市环境密集感知[10]。相比固定传感器网络，该类系统具备部署成本低、感知范围广及数据时效性强等优势，在智能交通管控与城市环境监测等场景中展现出广阔的应用前景。然而，VMCS 的实际部署面临两类结构性挑战。其一，城市路网结构复杂，车辆轨迹具有显著的随机性与时空不均匀性，系统难以实时掌握各移动载体的动态位置与覆盖状态，导致感知任务无法精准分配至合适的参与者，进而影响整体感知效率与覆盖质量。其二，不同类型参与者在感知能力、运动模式与资源禀赋上存在先天异质性，其对全局感知覆盖的实际贡献差异显著，且随空间位置、任务负载及历史决策动态演化，统一调度框架难以有效识别和利用这种差异。上述两类挑战相互交织，使得如何在轨迹不确定性与参与者贡献异质性的双重约束下，实现感知任务的精准分配与高效协同，同时兼顾全局感知覆盖率与资源利用效率，成为 VMCS 领域亟待突破的核心技术问题。为便于后续建模与调度优化，本文将上述异构移动参与者形式化为智能体。

为应对上述挑战，现有研究已从传统优化方法逐步演进至强化学习方法。早期研究主要依赖启发式算法与组合优化方法[11]，贪心算法常用于预算约束下的最大覆盖率优化[12]，整数规划方法则侧重参与者效用最大化与预算可行性保障[13]，但此类方法依赖确定性建模与全局状态预知，且将参与车辆视为同质感知单元，从根本上缺乏对异构参与者差异化贡献的表达能力[14]。Xiao 等人[15]将单智能体强化学习引入 VMCS 调度，缓解了轨迹不确定性问题，但忽略了车辆间的交互关系与个体差异。为此，Ma 等人[16]引入多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）框架，Lu 等人[17]将其扩展至空—地—人协同感知场景，表明 MARL 在 VMCS 动态调度中具有重要应用潜力。现有 MARL 研究主要沿信用分配与策略优化两条路线推进：信用分配方法通过价值分解[18-19]、单调约束或反事实基线[20-21]实现个体贡献归因，但通常依赖集中式训练或全局状态信息，且贡献估计基于静态网络结构，难以适应 VMCS 中参与者贡献动态演化的场景；策略优化方法利用概率比裁剪[22]、经验回放与目标网络[23]及确定性策略梯度[24]等机制稳定策略学习，但普遍采用固定裁剪阈值与统一更新系数，无法根据各智能体实时贡献差异自适应调整策略更新强度。两条路线的共同局限在于：在建模层面多依赖参数共享与同质化智能体假设[25]，未能显式表征参与者在感知能力、资源禀赋与任务适配性上的结构差异；在优化层面普遍对所有智能体施加相同激励信号[26]，激励机制与实际贡献之间存在系统性偏差，导致调度策略难以识别高贡献参与者并随其状态变化动态调整。因此，有必要在 MARL 框架内构建面向异质参与者贡献的显式建模与动态自适应调度机制。

在上述机制局限的基础上，VMCS 场景中还面临策略更新不稳定的问题：在高度动态的城市路网环境中，MARL 方法普遍面临训练震荡与收敛不稳的问题[27]，这与前述固定裁剪约束和优势函数敏感性问题一脉相承，尤其是近端策略优化（Proximal Policy Optimization, PPO）类方法在规模扩展时表现得尤为突出，使激励机制难以与智能体实际贡献形成有效对齐。针对上述问题，本文以 PPO 为基础框架，提出贡献感知自适应近端策略优化算法（Contribution-Aware Adaptive Proximal Policy Optimization, CA2-PPO）。算法包含三项核心机制：利用指数移动平均（Exponential Moving Average, EMA）动态评估各智能体贡献并加权熵项，实现差异化探索激励；构建基于 Kullback-Leibler（KL）散度的动态裁剪机制，使策略裁剪范围随策略分布

偏移自适应调整；对优势函数采用裁剪与归一化相结合的处理方式，抑制极端样本对梯度估计的干扰。上述设计在非独立同分布（Non-Independent and Identically Distributed, Non-IID）任务分布下实现贡献感知的自适应协同优化，有效提升复杂动态路网中多智能体训练的稳定性与鲁棒性，为 VMCS 的规模化部署提供可行路径。

本文主要贡献和创新点如下：

（1）提出贡献度加权熵机制：采用 EMA 稳定估计智能体动态贡献，并以贡献为权重自适应调节熵正则强度，使高贡献智能体保持探索、低贡献智能体更快收敛到协同策略，以缓解均等激励造成的协作低效。

（2）针对策略更新不稳定问题，设计动态 KL 散度裁剪策略，基于新旧策略间的相似度，自适应调节策略裁剪范围$\varepsilon(t)$，以在训练稳定性与收敛效率之间实现动态平衡。

（3）提出一种区别于仅归一化或仅裁剪的优势预处理方案：在策略更新前对优势函数先阈值裁剪、后标准化归一化，抑制极端回报导致的优势离群与梯度偏移，提升训练稳定性及策略更新的鲁棒性与数据利用效率。

（4）系统化实验验证与可扩展性评估：构建网格场景与 OpenStreetMap（OSM）两类基准测试框架，围绕平均覆盖得分与平均总奖励等指标与多智能体深度 Q 网络（Multi-Agent Deep Q-Network，MADQN）、多智能体裁剪近端策略优化(Multi-Agent Clipped Proximal Policy Optimization，MACPPO)等主流 MARL 方法完成对比评估，验证 CA2-PPO 在效率、训练稳定性及规模扩展方面的综合优势。

2 相关工作

本章首先全面梳理了多智能体强化学习的核心方法体系，重点围绕信用分配机制、策略梯度协同优化框架及 EMA 机制展开论述，以应对多智能体协同训练过程中的稳定性与样本效率挑战。接着，对移动群智感知领域的研究进展进行综合性回顾，聚焦覆盖优化、激励机制设计与异构智能体协同三个核心维度。上述两条研究主线并非相互独立：MARL 为分布式协同决策提供了坚实的算法理论基础，而移动群智感知场景中固有的参与者异质性与任务分布的 Non-IID 特性，则对算法在贡献建模精度与策略自适应能力方面提出了更为严苛的表达力诉求。正是这一算法能力边界与场景结构复杂性之间的深层矛盾，构成了后续 CA2-PPO 方法设计的根本动因与理论建构起点。

2.1 多智能体强化学习

在多智能体协同场景中，传统单智能体强化学习受限于单一智能体的感知与决策能力，在面对复杂目标、广泛空间分布和强任务异质性时表现不足[28]。多智能体系统中的智能体不仅需要权衡自身利益与环境风险，还需兼顾受其他智能体影响的任务目标达成[29]。相比之下，MARL 能够克服上述局限，在复杂任务中展现出更大潜力 **Error! Reference source not found.**。但多智能体间的信息共享与策略同步仍是关键挑战 **Error! Reference source not found.**31]。

从方法体系看，MARL 通常分为基于价值和基于策略两种范式。基于价值的方法以价值函数为核心，典型算法包括 Q 学习[32]及深度 Q 网络（Deep Q-Network，DQN）[33]，后者通过经验回放与目标网络提升 Q 值估计稳定性[34]，但设计初衷面向单智能体场景，对多智能体间交互耦合考虑不足。基于策略的方法直接优化策略函数，PPO 通过裁剪机制兼顾样本效率与训练稳定性，已在机器人集群[35]、电力优化[36]、传感网络[37]和智能交通调度等场[38]景中得到应用，但在移动多智能体协作任务中效率与稳定性之间仍存在权衡。在应用层面，MARL 已广泛用于路径规划与任务分配等协同场景[39]，在无人机追踪、边缘计算[40]、智能交通[41]、资源分配[42]等领域均取得了重要进展。在信用分配方面，值分解网络（Value Decomposition Network，VDN）[43]与单调值函数分解网络（Monotonic Value Function Factorisation，QMIX）[19]通过价值分解实现多智能体协作，双工对决多智能体 Q 学习（Duplex Dueling Multi-Agent Q-Learning，QPLEX）[21]进一步提升了联合价值分解的表达灵活性，基于 Shapley 值的方法[44]则从博弈论角度实现公平归因。然而，上述方法或依赖全局状态信息，或需额外网络结构，难以实时响应贡献变化。在 EMA 应用方面，现有研究通常将其作为参数平滑与训练稳定化工具。Mean Teacher 框架[45]通过对模型参数进行指数滑动平均构建教师模型，以提升训练过程的稳定性；相关研究也表明，EMA 能够在一定程度上产生隐式正则效应并改善模型泛化能力[46]。然而，上述方法主要作用于模型参数层面，尚未面向多智能体协同任务中的动态贡献差异进行建模。与之不同，CA2-PPO

将 EMA 用于智能体贡献度的轻量级在线估计，无需额外网络结构或全局状态信息，直接将贡献估计映射为熵正则强度，从探索激励角度动态响应贡献异质性，将 EMA 从参数平滑工具扩展为自适应的贡献感知机制。

上述 MARL 方法虽在多智能体协同场景中取得了广泛应用，但仍面临关键局限。集中训练分散执行 (Centralized Training with Decentralized Execution，CTDE）框架下的多智能体 PPO 与 DQN 方法普遍隐含贡献均衡假设，并依赖固定更新约束，在 Non-IID 任务分布下难以刻画个体贡献差异，易导致协作失衡与训练不稳定[47]。本文提出的 CA2-PPO 算法从贡献异质性建模出发，通过在线贡献估计与自适应策略更新的协同设计，有效提升了复杂动态路网中多智能体协同调度的稳定性与效率。

2.2 移动群智感知

随着城市路网感知任务规模持续扩大与参与者异质性日益凸显，如何实现移动参与者的高效协同调度逐渐成为 VMCS 领域的研究热点。为此，越来越多的学者开始关注面向动态路网的自适应调度与贡献感知激励机制设计。早期 VMCS 研究多依赖单智能体强化学习，如 Xiao 等人基于 Q 学习将感知决策建模为马尔可夫决策过程，但未考虑智能体贡献差异且策略更新稳定性不足[15]。现有 VMCS 相关研究大致沿覆盖优化、激励机制与结构异构智能体协同三条主线推进。在覆盖优化方面，Starace 等人[48]在波尔图、罗马等复杂历史路网城市中系统评估了出租车作为感知探针的时空覆盖可行性，发现路网拓扑复杂度对覆盖分布具有显著影响，常见的网格化抽象难以反映真实路段级感知差异；针对数据时效性问题，He 等人[49]面向空—地协同移动群智感知场景，提出预测式多智能体深度强化学习方法，以增强无人机（Unmanned Aerial Vehicle，UAV）与地面参与者之间的协同任务执行能力。然而，该方法仍主要关注空—地协同效率与任务完成性能，缺乏将异质参与者动态贡献差异反馈至策略更新过程的建模方式。在激励机制方面，Luo 等人[50]将车辆社交网络中的社交收益引入激励框架，结合深度强化学习优化长期感知策略；Jaimes 等人[51]提出时间依赖的动态定价激励方案 SenseNow，通过引导车辆偏离预定路径访问感知兴趣点提升时间覆盖密度，但该方案对参与者行为模式仍依赖先验分布假设；针对预算约束下的异构任务分配问题，近期研究提出基于多头注意力机制的多智能体近端策略优化（Multi-Agent Proximal Policy Optimization，MAPPO）方法，将任务难度与车辆穿行意愿纳入联合优化。在异构智能体协同方面，Xu 等人[52]提出基于 MARL 的去中心化 MCS 任务分配框架，验证了多智能体强化学习处理异构任务的适用性；Deng 等人[53]针对 UAV 辅助的 MCS 网络提出基于异构智能体近端策略优化 (Heterogeneous-Agent Proximal Policy Optimization，HAPPO) 与柯尔莫哥洛夫 - 阿诺德网络（Kolmogorov-Arnold Networks，KAN）网络的异构多智能体协同框架，通过混合 Actor 网络提升了感知数据处理量，但该方法聚焦于空地协同，未能直接适配地面车辆路网场景下的贡献异质性建模；Monon 等人[54]进一步提出协作异构强化学习算法，利用图神经网络驱动的内在激励在去中心化训练条件下促进异构智能体协作，但其设计主要面向通用 MARL 场景，缺乏与 VMCS 时空感知任务的深度耦合。此外，Wang 等人[55]从覆盖感知角度提出高质量感知数据采集方法，针对节点贡献不均问题建立了精细化评估框架；Wu 等人[56]系统综述了 MCS 激励机制中贡献评估的研究现状，指出差异化激励与个体贡献对齐仍是核心待解问题。近期研究表明，量化个体贡献并动态调整是提升分布式协同系统效能的可行路径[57]。

上述 VMCS 研究虽在覆盖优化与激励机制方面取得了一定进展，但在贡献异质性建模与策略更新稳定性方面仍存在明显不足，普遍依赖均等激励假设与固定裁剪约束，导致激励分配与智能体实际贡献脱耦，难以在 Non-IID 任务分布下实现有效协同优化。对此，本文提出 CA2-PPO 算法，构建系统化的贡献异质性建模框架，识别移动智能体的差异化贡献，提升复杂城市路网场景下的覆盖率与训练稳定性。

3 问题建模

在城市道路网络感知场景中，VMCS 系统一般由 MCS 平台、任务请求方以及移动用户三类主体构成：平台由云端数据中心和感知服务器构成，负责收集任务需求并协调调度；移动用户依托车载设备、智能手机传感器及可穿戴设备采集数据，并通过蜂窝网络或短距通信上报。基于上述系统构成，本节对涉及的核心概念进行形式化定义，为后续模

型构建与算法设计奠定基础。

3.1 问题定义

为支撑感知系统的精细化建模与运行，本文首先给出贡献异质智能体集的形式化定义如下：

定义 1. 贡献异质智能体集. 设系统包含 N 个智能体，记智能体 i 在时刻 t 获得的即时奖励为$\hat{r}_i^t$。所有智能体采用统一的观测空间和动作空间，并采用同构的 Actor–Critic 策略网络。由于空间位置与任务负载的动态变化，各智能体在协同感知过程中所发挥的作用不尽相同。需要指出的是，本文所定义的异质性聚焦于行为层面的动态贡献差异，即在同构硬件配置的前提下，由空间分布、任务负载及历史决策轨迹共同决定的贡献异质性。考虑到感知半径、移动速度、能耗等硬件层面的异质性将显著增加建模复杂度，超出本文当前研究范围，相关扩展建模将作为后续工作加以推进。

基于上述动态贡献差异的建模需求，本文采用指数移动平均对智能体贡献度进行在线估计。智能体 i 在时刻 t 的贡献度$\tilde{C}_i^t$由由历史归一化即时奖励的 EMA 递推得到，经归一化后得贡献度权重C_i^t，其初始值均匀设定为$\tilde{C}_i^t = 1/N$，具体更新规则与归一化方式分别见公式（9）与公式（25），以保证$\sum_{i=1}^{N} C_i^t = 1$且$C_i^t \geqslant 0$。

若训练过程中智能体贡献度差异持续显著，则称该智能体集合为贡献异质智能体集。

为确保感知任务的有效执行和数据采集的精确性，本文进一步定义感知区域与任务，具体如下：

定义 2. 感知区域与任务. 依据感知任务的空间分布，将感兴趣区域划分为若干网格。感知任务的目标是在给定时间范围内获取各网格内目标感知现象的空间感知图。当智能体在时隙内对某一网格实现有效覆盖时，判定该网格的感知任务在该时隙完成；其对应感知数据随即更新，并作为后续决策与分析的输入。

定义 3. 移动多智能体任务分发与协作优化. 在城市道路网格场景中，采用多智能体强化学习框架对感知任务进行动态分发与自适应编排，利用任务时空异质性引导智能体的空间迁移与行为选择。在满足时间与任务分布约束的前提下，引入贡献度加权的多目标优化机制，综合权衡个体收益与群体协作效率，以最大化加权任务收益并最小化未覆盖价值。

定义 4. 单网格单时隙直接覆盖分数. 给定调度策略π，若网格g_k在时隙t内被至少一个智能体覆盖，则认为其在该时隙实现直接覆盖。其直接覆盖收益$Sc_t^0(\pi; g_k)$，其表达式见公式(1)。

$$Sc_t^0(\pi; g_k) = 1 - \prod_{m=1}^{M}\left(1 - \omega_t^{u_m}(\pi; g_k)\right)\prod_{n=1}^{N}\Big(1 - \omega_t^{v_n}(\pi; g_k)\Big) \quad (1)$$

其中，Sc_t^0为覆盖分数，$\omega_t^{u_m}$为第m个专用探测车在时隙t对网格g_k的覆盖权重；$\omega_t^{v_n}$为第 n 个普通车辆的覆盖权重，两者取值范围[0,1]，M为专用探测车总数，由传感器精度与距离共同决定。

在直接覆盖基础上，引入时间衰减与空间扩散因子，定义时空覆盖收益：

定义 5. 单网格单时隙时空覆盖收益. 给定调度策略π，网格g_k在时隙t的时空覆盖收益$C_t^{ST}(\pi; g_k)$由时间影响函数$T^T(t-t')$、空间影响函数$I^S(d_{kk'})$及历史直接覆盖状态$\{Sc_{t'}^0(\pi; g_{k'}) | t' \leq t\}$定义，其表达式见公式(2)。

$$C_t^{ST}(\pi; g_k) = 1 - \prod_{t'=1}^{t}\prod_{k'=1}^{K}\left(1 - T^T(t-t') \cdot I^S(d_{kk'}) \cdot Sc_{t'}^0(\pi; g_k)\right) \quad (2)$$

上述影响函数的形式化定义如下：时间影响函数采用指数衰减形式：$T^T(t-t') = e^{-\lambda_T(t-t')}$。其中，$\lambda_T$设为 0.01，对应约 100 个时间步的有效衰减窗口，当$t = t'$时T^T=1，表示当前时刻覆盖贡献最大；随时间差$t - t' \geqslant 0$ 增大，$T \in (0,1]$，贡献指数衰减趋近于 0，使超出窗口的历史覆盖影响可忽略不计。空间影响函数采用高斯衰减形式：$I^S(d_{kk'}) = e^{-d_{kk'}^2/(2\sigma_S^2)}$ $(\sigma_S = 1)$，其中，$d_{kk'}$为两网格的欧氏距离，$I^S(d_{kk'})$刻画网格$g_{k'}$的历史覆盖对目标网格g_k的空间影响强度，k遍历感知区域内所有网格，K为感知区域内的网格总数，与定义 2 中的网格划分一致。

上述时空覆盖收益仅建模了单一尺度下的时空关联，为单尺度基础形式；为进一步捕捉周期性需求波动与区域簇影响，将其推广至多尺度时空覆盖收益见定义 6：

定义 6. 多尺度时空覆盖收益. 在 VMCS 场景中，在定义 5 基础上，进一步引入周期影响项$I^P(t-t')$、区域簇影响$I^C(G, g_k)$，其表达式分别见公式(3)与公式(4)。

$$\mathrm{C_t^{ST,grid}}(\pi; \mathrm{g_k}) = 1 - \prod_{\mathrm{t'=1}}^{\mathrm{t}}\prod_{\mathrm{k'=1}}^{\mathrm{K}}\left(1 - \mathrm{T^T(t-t')} \cdot \mathrm{I^S(d_{kk'})} \cdot \mathrm{Sc_t^0}(\pi; \mathrm{g_{k'}}) \cdot \mathrm{I^P(t-t')}\right) \quad (3)$$

其中，周期影响函数采用余弦相似形式：$I^P(t-t')=\left(1+\cos(2\pi(t-t')/T_P)\right)/2$，其中，$T_P$为任务周期长度；$I^P\in[0,1]$，当时间差为$T_P$整数倍时$I^P$=1，表示历史时刻与当前时刻处于相同周期位置，参考价值最高；当时间差为T_P/2 时I^P=0 表示半周期错位，参考价值最低。上式刻画了单一网格g_k在时间与空间维度的覆盖收益。进一步引入网格簇G的区域影响，将覆盖建模扩展至簇级尺度，表示为：

$$C_t^{ST,\text{cluster}}(\pi;g_k)=\prod_{g_j\in G}\left(1-I^c(G,g_j)\cdot\max_{g_j\in G}\mathrm{Sc}_{t'}^{0}(\pi;g_j)\right)\#(4)$$

其中，$I^C(G,g_j)$为区域簇对网格g的影响强度；$\max_{g_j\in G}Sc_{t'}^{0}(\pi;g_j)$表示簇内最大直接覆盖收益。公式(3)与公式(4)分别对应网格级与簇级两个尺度，共同构成多尺度时空覆盖建模框架。

3.2 总体架构

本框架基于 CA2-PPO 算法，提出了三层协同的城市网格感知框架，包含底层强化学习模块、执行层网格区域和顶层需求层，形成闭环交互机制。底层逻辑层：智能体在状态感知与动作执行的持续循环中与环境交互，通过奖励信号驱动策略迭代优化，根据当前网格覆盖情况和任务完成度，通过奖励反馈（*R1*，*R2*）优化策略，实现自适应调度和高效执行。执行层：道路被划分为若干网格，并通过颜色区分高价值、中价值和已完成区域。车载智能体在网格中行驶，通过 WiFi、GPS、雷达等与路侧单元（Roadside Unit, RSU）通信，接收任务并根据优先级选择路径，优先覆盖高价值区域。数据中心和任务发布者负责任务发布、反馈接收及决策分析，与执行层和逻辑层协作，确保任务闭环。整体框架如图 1 所示。

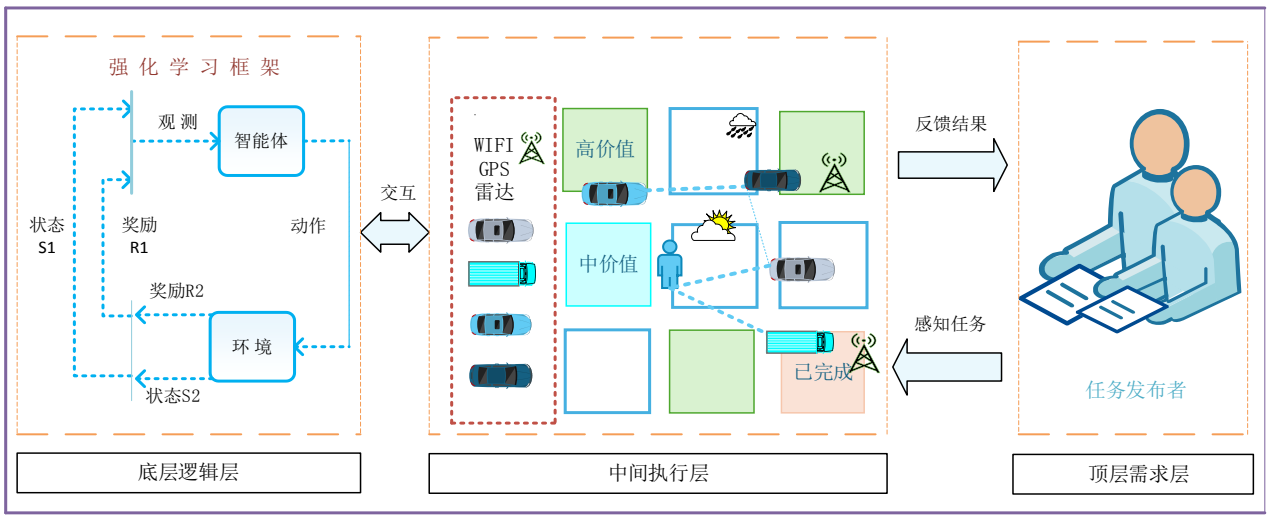

图 1　基于贡献度优化的异质智能体移动群智感知与自适应 PPO 协同框架

如图 1 所示，CA2-PPO 框架由底层逻辑层、中间执行层与顶层需求层三层架构构成，智能体在执行感知任务过程中涉及状态上报、奖励反馈与任务交互等多类数据流通环节。在群智感知场景下，上述数据流通过程面临通信链路窃听、服务器端轨迹关联推断以及恶意参与方奖励信号投毒攻击等数据隐私与安全威胁。针对上述威胁，本文从三个层面构建隐私保障体系：在 VMCS 系统的实际部署中，参与感知的车辆智能体在执行任务过程中会产生位置轨迹、传感器读数等敏感数据，其隐私保障主要体现在以下三个层面。数据传输层：智能体通过 RSU 上报感知数据时，采用传输层安全协议（Transport Layer Security, TLS）对通信链路加密，防止数据在上传过程中被窃听或篡改。身份隐匿层：系统在任务调度与贡献度计算中，仅需智能体上报归一化的即时奖励$\hat{r}_i^t$，该信号不含绝对坐标或时间戳，攻击者无法从中还原真实轨迹。模型参数层：CA2-PPO 集中训练架构中感知数据无需跨智能体共享，如需更强隐私保护，可结合联邦学习框架引入安全聚合协议，确保服务器仅能获得梯度聚合结果而无法逆向推断个体信息。在此基础上，CA2-PPO 的贡献度计算仅依赖奖励信号，从机制设计上规避了额外的隐私暴露；针对投毒威胁，可结合异常检测模块对历史奖励分布进行统计监控以保障系统鲁棒性；差分隐私标定与联邦聚合的深度集成将作为后续工作进一步推进。本文使用的符号总结于表 1。

表 1 主要符号及其含义

符号	含义
B	训练批次大小
$\pi_{\theta_{old}}$	策略更新前的旧策略
π_θ	待更新的新策略
s_b	批次 b 中的状态样本
$\mathcal{L}_{ent}$	总熵损失（用于平衡策略探索与利用）
D_{KL}	新旧策略间的 KL 散度（衡量策略更新幅度）
$\bar{D}_{\mathrm{KL}}$	各智能体 KL 散度按C_i^t加权的全局 KL 散度估计
$\varepsilon(t)$	时刻t的策略裁剪范围
$A(s,a)$	状态 s 下执行动作a的优势函数
$H(\pi)$	策略π的熵
γ	折扣因子
λ	GAE 优势函数计算的权衡系数
ε_{adv}	优势函数裁剪阈值
$\rho^{(t)}(s,a)$	时刻 t 状态 s 下执行动作 a 的新旧策略概率比
C_{cnt}	熵损失系数，控制整体探索强度
$\mathcal{L}_{value}$	价值函数损失，衡量 Critic 网络估计误差
$\mathcal{L}_{total}$	总损失函数，由策略损失、价值损失与熵正则化项联合构成
$\mathcal{L}_{clip}$	PPO 裁剪策略损失函数
α_{lr}	Adam 优化器学习率
θ	策略网络参数
C_v	价值函数损失系数，控制 Critic 网络更新强度
$a_t^{(i)}$	智能体i在时刻 t执行的动作
o_t^i	智能体i在时刻 t的局部观测
α	贡献度平滑系数
τ	KL 散度阈值
$V(s)$	状态s的价值函数估计
N	智能体总数
T	训练时间步数
$\tilde{A}_t$	处理后的优势估计
$\hat{r}_i^t$	智能体 i 在时刻 t 的归一化即时奖励
$\tilde{C}_i^t$	智能体 i在时刻 t经 EMA 平滑后的原始贡献得分，无归一化约束
C_i^t	由$\tilde{C}_i^t$经公式(25)归一化后得到的贡献度权重
$\boldsymbol{C^t}$	所有智能体归一化贡献度构成的概率向量，满足 $\sum_i C_i^t = 1$
λ_1	时空关联奖励的平衡系数
λ_2	协作覆盖奖励的平衡系数
$\hat{R}_b$	批次 b 中样本的折扣回报目标
$\mathcal{D}$	经验回放集合
b	单次迭代取出的单个小批量样本子集
θ_{old}	旧策略网络的参数
K_{KL}	每轮更新全局加权 KL 散度记录集合
L_1	KL 裁剪计算时概率比与优势相乘的中间临时分项
L_2	KL 裁剪计算时概率比与优势相乘的中间临时分项
B	全部小批量样本集合，批量总大小
$\widehat{E}_b$	对单个小批量样本子集 b 内全部样本求取经验均值
$L(\theta)$	Adam 更新使用
r_i^t	原始即时奖励
s_t	全局环境状态

4 算法设计与分析

为求解城市网格环境下多智能体感知任务的协作优化问题，本节介绍 CA2-PPO 算法：构建网格化城市道路环境并定义状态与动作空间；提出融合贡献度加权熵、动态 KL 裁剪与优势函数归一化裁剪三项核心机制，分别实现差异化激励、稳定策略更新与极端样本抑制；结合算法流程与理论推导说明实现细节，为后续实验奠定基础。

4.1 环境构建

为了模拟真实城市道路网络的特性，本文构建 20×20 与 30×30 网格状道路网络，节点代表交叉口，链路旅行时间设为 30 秒（s）至 300s，以模拟不同路段的通行效率差异，为多智能体任务分配与优化提供环境基础。

4.2 状态与动作空间

记时 t 的全局状态为$s_t = [X_t, V_t]$，其中$X_t = \left\{x_i^{(t)}\right\}_{i=1}^{N}$为 N 个智能体在道路网络中的位置；V_t为 RSU 广播的网格价值场。

动作空间定义为 A={东，南，西，北}。当智能体选择动作$a_t^{(i)} \in A$时，系统根据道路网络拓扑从当前节点到目标节点规划最短路径并执行迁移，以保证动作的可行性和执行效率。

4.3 奖励函数设计

奖励函数决定智能体如何评估行动好坏。本文为引导智能体优先覆盖高价值区域并实现全局协作优化，基于第 3.1 节中感知区域与任务定义即定义 2、协作优化目标即定义 3 及贡献异质情境下的智能体集合即定义 1，将感知任务完成度与未覆盖价值显式建模为回报信号；进一步基于定义 4、5、

6 构建层次化奖励函数，将感知数据的时空关联价值融入即时反馈，指导智能体训练。

4.3.1 即时覆盖奖励

当智能体i在时隙t到达网格g_k时，基于定义 4 的直接覆盖分数，其即时覆盖奖励定义为：

$$r_t^{cover}(i) = Sc_t^0(\pi; g_k) \cdot V(g_k) \quad (5)$$

其中，$V(g_k)$为网格g_k的感知价值权重，由 RSU 依据区域感知需求动态维护。实验设置中任务分布模式包括聚集分布（Gathered Distribution，GD）、均匀分布（Uniform Distribution，UD）、随机分布（Random Distribution，RD）和动态分布（Dynamic Distribution，DD）等任务分布模式。$Sc_t^0(\pi; g_k)$由定义 4 的公式（1）给出，网格g_k在时隙 t 内首次被覆盖时收益为 1，重复覆盖因历史状态累积使收益衰减，引导智能体优先覆盖高价值网格。

4.3.2 时空关联奖励

考虑感知数据的时空相关性（定义 5），智能体可通过邻近网格的历史覆盖获得关联收益，时空关联奖励定义为移动带来的全局时空覆盖收益增量：

$$r_t^{ST}(i) = \sum_{g_k \in \mathcal{N}(p_{pos})} [C_t^{ST}(\pi_t; g_k) - C_{t-1}^{ST}(\pi_t; g_k)] \#(6)$$

其中，$\mathcal{N}(p_{pos})$为智能体 i 当前位置的邻域网格集合，$C_t^{ST}(\pi_t; g_k)$由定义 6 的公式（3）计算。该奖励引导智能体最大化时空覆盖增量，充分利用感知数据的时空关系，提升覆盖效率。

4.3.3 协作覆盖奖励

为优化多智能体协作效率，避免资源浪费的重复覆盖，引入协作覆盖奖励：

$$r_t^{\text{coop}}(i) = \begin{cases} \beta \cdot V(g_k), & \text{若网格} g_k \text{在时隙} t \text{首次被覆盖} \\ -\eta \cdot V(g_k), & \text{若网格} g_k \text{已被其他智能体覆盖} \\ 0, & \text{其他情况} \end{cases} \quad (7)$$

其中，β=1.0 为首次覆盖奖励系数，η=0.2 为重复覆盖惩罚系数。两者比值（5:1）的设计需在两类错误间取得平衡：比值过大（如 10:1）时，惩罚系数η偏小，对重复覆盖的抑制力度不足，智能体倾向于反复覆盖少数高价值网格，协作分散效果消失；比值过小（如 2:1）时，η偏大，强惩罚使智能体过度回避已覆盖区域，破坏多智能体协同覆盖大型价值簇时的必要协作行为。参考多智能体覆盖类任务中奖惩比的常见设置范围 3:1～8:1，本文将β固定为 1.0 作为基准单位，η设为 0.2，比值 5:1 处于该区间中位，在协作分散效果与必要重叠覆盖之间取得合理平衡。该项奖励促进智能体分散覆盖不同区域，减少协作过程中的资源重叠，与定义 6 中多尺度覆盖收益的设计目标一致。

4.3.4 总奖励函数

综合上述三项，智能体 i 在时隙 t 的总即时奖励为：

$$r_t(i) = r_t^{\text{cover}}(i) + \lambda_1 \cdot r_t^{\text{ST}}(i) + \lambda_2 \cdot r_t^{\text{coop}}(i) \#(8)$$

其中，λ_1=0.3、λ_2=0.5 为平衡系数，分别控制时空关联与协作覆盖的相对重要性。两者的设定遵循即时覆盖为主、辅助项渐进增强的原则，并需满足不等式约束 $1>\lambda_2>\lambda_1$，以保证即时覆盖奖励始终占主导地位。为确定最优取值，本文在 20×20 网格、4 智能体的 GD 环境下，对$\lambda_1 \in \{0.1,0.3,0.5\}$与$\lambda_2 \in \{0.3,0.5,0.7\}$的组合进行网格搜索，以 5 个随机种子下的平均覆盖得分为评估指标。实验结果表明，λ_1=0.3、λ_2=0.5 时综合表现最优：λ_1过大（0.5）会导致智能体过度依赖时空扩散收益而忽视即时高价值网格；λ_2过大（0.7）则因协作惩罚过强抑制必要的重叠覆盖行为。该设置与多目标奖励设计的常见经验范围辅助项系数 0.2～0.5 一致，并经预实验验证在四种任务分布下均保持稳定，最终作为默认配置。

该设计通过即时覆盖奖励公式（5）引导智能体优先覆盖高价值区域，时空关联奖励公式（6）利用感知数据的时空相关性促进全局最优决策，协作覆盖奖励公式（7）实现多智能体分工协作、避免资源重复消耗，三项奖励协同驱动智能体高效提升覆盖性能。

4.4 CA2-PPO 算法

CA2-PPO 在标准 PPO 框架基础上针对贡献异质智能体协同调度问题进行改进，以缓解协作低效与策略更新不稳定问题。算法 1 给出整体训练流程，环境状态与动作空间基于定义 2、3 构建，即时奖励由定义 4 至 6 的多尺度时空覆盖收益综合计算得到。算法核心遵循经验收集、多目标优化、动态适配的闭环过程：初始化均匀贡献度向量与动态裁剪范围，迭代收集交互轨迹并结合广义优势估计（Generalized Advantage Estimation，GAE）计算优势函数，构造联合损失函数以 Adam 优化器更新策略网络，同时引入基于 KL 散度反馈的动态裁剪机制平衡探索与利用，最终输出最优策略、智能体贡献度及训练历史，实现感知任务完成率最大化与未

覆盖价值最小化。融合贡献度机制的 CA2-PPO 算法描述如下所示。

算法 1. CA2-PPO 算法.

输入：网格 $H\times W$、智能体数 N、任务分布 GD/UD/RD/DD、超参（γ，λ，α，τ）、总步数 T、训练批次大小 B、Critic 参数ϕ、经验缓冲区$\mathcal{D}$

输出：策略θ*、贡献度C_i^t、训练历史$\mathcal{H}$

$\tilde{C}_i^t \leftarrow 1/N, \forall i \in \{1,\ldots,N\}$

$\varepsilon(t) \leftarrow 0.2; \theta_{old} \leftarrow \theta; \mathcal{D} \leftarrow \phi; t \leftarrow 0; \mathcal{H} \leftarrow \phi$

$t \leftarrow 0$,WHILE $t<T$ DO

FOR *step*=1 TO 2048：

RSU 更新并广播价值场 $V_t; s_t \leftarrow (X_t, V_t)$

$a_t^{(i)} \sim \pi_{\theta_{old}}(\cdot|o_t^i), \forall i \in \{1,\ldots,N\}$;执行 a_t 与环境单步交互,得 s_{t+1} 与 r_t^i

$\mathcal{D} \leftarrow \mathcal{D} \cup \{(o_t^i, a_t^{(i)}, r_t^i, o_t^{i+1})\}$

按公式(9)更新$\tilde{C}_i^t$，经公式(25)归一化得C_i^t

由 GAE(γ,λ)基于$\mathcal{D}$计算优势函数 $A(s,a)$

END FOR

FOR *epoch*=1 TO 10 DO

FOR each $b \subseteq \mathcal{D}$ DO

按公式(29)(30)对 $A(s,a)$裁剪并归一化,得$\tilde{A}_t$

$\rho^{(t)}(s,a) \leftarrow \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}$

$\mathcal{L}_{clip}(\theta) \leftarrow -\frac{1}{B}\sum_{b=1}^{B} \min(\rho^{(t)}(s_b,a_b)\cdot \tilde{A}_b,\ clip(\rho^{(t)}(s_b,a_b), 1-\varepsilon(t), 1+\varepsilon(t))\cdot \tilde{A}_b)$

$\mathcal{L}_{value} \leftarrow \hat{E}_b\left[\left(V(s_b)-\hat{R}_b\right)^2\right]$

按公式(11)计算贡献度加权熵正则项$\mathcal{L}_{ent}$

$\mathcal{L}_{total}(\theta) \leftarrow \mathcal{L}_{clip}(\theta) + C_v \cdot \mathcal{L}_{value}(\theta) + C_{cnt} \cdot \mathcal{L}_{ent}$

按公式(12)计算新旧策略 KL 散度,经公式(26)加权得 D_{KL}

IF $D_{\mathrm{KL}} > \tau$ THEN $\varepsilon(t) \leftarrow \max(\varepsilon(t)-0.05, 0.05)$

ELSE IF $D_{\mathrm{KL}} < \tau/2$ THEN $\varepsilon(t) \leftarrow \min(\varepsilon(t)+0.05, 0.3)$

END FOR

END FOR

$\theta_{old} \leftarrow \theta$

IF 当前测试得分>历史最佳 THEN $\theta^* \leftarrow \theta$

$\mathcal{H} \leftarrow \mathcal{H} \cup${本轮得分与训练统计}

END WHILE

RETURN θ*、C_i^t、$\mathcal{H}$

本节以下涉及的算法将是对 CA2-PPO 算法的更进一步细化和讨论。算法 1 是主框架，算法 2、3、4 是对算法 1 中三个核心创新机制的详细展开。

4.4.1 智能体贡献度加权熵算法

算法 2 为算法 1 第 1、8、17 行的详细展开，核心是动态调整探索权重以匹配个体贡献与群体协作。初始分配均等贡献度，每步通过 EMA 融合即时奖励与历史贡献度并归一化；批次优化时以贡献度为权重计算加权熵损失，联合策略与价值损失反向传播，使高贡献智能体增强高价值探索激励、低贡献智能体减少无效探索，提升整体覆盖效率。具体流程如算法 2 所示。

算法 2. 基于贡献度的自适应多智能体加权熵正则化算法.

输入：智能体数量 N、贡献度平滑系数 $\alpha \in$（0，1）、熵损失系数C_{cnt}，控制探索强度、每个回合的时间步数 T、训练批次大小 B、Critic 参数ϕ

输出：更新后的参数θ、归一化贡献权重C_i^t、训练历史$\mathcal{H}$

$\tilde{C}_i^t \leftarrow 1/N, \forall i \in \{1,\ldots,N\}$

FOR t=1 TO T DO

与环境交互,采集$(o_i^t, a_i^t, r_i^t, o_i^{t+1}), \forall i \in \{1,\ldots,N\}$

$\hat{r}_i^t \leftarrow \frac{clip(r_i^t, -5, 5) - r_{\min}^t}{r_{\max}^t - r_{\min}^t}$

$\tilde{C}_i^t \leftarrow \alpha\, \hat{r}_i^t + (1-\alpha)\, \tilde{C}_i^{t-1}$

$C_i^t = \mathrm{Normalize}(\tilde{C}_1^t, \ldots, \tilde{C}_N^t)$

END FOR

FOR each b of size B DO

$H_i \leftarrow H(\pi_\theta(\cdot|o_i^t)), \forall i \in \{1,\ldots,N\}$

$\mathcal{L}_{ent} \leftarrow -\hat{E}_b(\sum_{i=1}^{N} C_i^t \cdot H_i)$

$\mathcal{L}_{total}(\theta) \leftarrow \mathcal{L}_{\mathrm{clip}}(\theta) + C_{\mathrm{v}} \cdot \mathcal{L}_{\mathrm{value}} + C_{\mathrm{cnt}} \cdot \mathcal{L}_{ent}$

$\theta \leftarrow \mathrm{Adam}(\theta, \nabla_\theta L(\theta))$

END FOR

为实现对高贡献智能体的差异化激励，本文基于即时奖励构建贡献度并引入 EMA 进行时序平滑，贡献度的更新形式如下：

$$\tilde{C}_i^t = \alpha \hat{r}_i^t + (1-\alpha)\tilde{C}_i^{t-1} \#(9)$$

其中，α=0.1 为贡献度平滑系数，用于平衡历史贡献与当前贡献，避免贡献度因短期奖励波动而剧烈变化。选择合适的 α 值需在历史稳定性与表现响应速度之间平衡：当 $\alpha \geq 0.5$ 时，即时奖励主导贡献度，单步奖励波动将直接引发熵正则权重的剧烈变化，破坏差异化激励的平稳性；当 $\alpha \leq 0.01$ 时，更新过慢，失去差异化效果。本文在 20×20 网格 4 智能体情境下对 $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$进行预实验扫描，$\alpha$=0.1 时贡献度能够在约 20 个时间步内响应智能体间的性能分化，同时保持平滑的更新轨迹，综合表现最优，最终选择为本文默认值。此外，α=0.1 与

每 2048 步更新策略的训练节奏匹配，确保贡献度估计的一致性与可靠性。$\hat{r}_i^t$为智能体 i 在时刻 t 的归一化即时奖励，随后按公式(25)归一化得到C_i^t。

$$\hat{r}_i^t = \frac{\mathrm{clip}(r_i^t, -5, 5) - r_{\min}^t}{r_{\max}^t - r_{\min}^t} \#(10)$$

其中，$r_{\min}^t$与$r_{\max}^t$分别表示时刻 t 所有智能体裁剪后奖励的最小值与最大值，通过 min–max 归一化将奖励缩放至[0，1]区间，该归一化约束防止单个智能体因短期高贡献导致资源过度倾斜，同时保证低贡献智能体仍能获得必要的探索资源。

基于 EMA 的贡献度机制通过历史奖励累积保留智能体长期优势，并借助归一化约束维持分配公平性，在差异化激励与全局合理性间形成动态平衡，进而以贡献度为权重构建总熵损失：

$$\mathcal{L}_{ent} = -\sum_{i=1}^{N} C_i^t \, H\left(\pi_\theta(\cdot \mid s_i^t)\right) \#(11)$$

其中，$H\left(\pi_\theta(\cdot \mid s_i^t)\right)$表示智能体 i 在状态s_i^t下策略π_θ的熵，用于衡量策略的探索程度。基于归一化后的贡献度向量，本文以均匀贡献基准 1/N 作为判定阈值，对智能体进行高/低贡献区分：

$$\text{贡献类型}_i^t = \begin{cases} \text{高贡献智能体}, & \mathrm{C_i^t} > 1/N \\ \text{低贡献智能体}, & \mathrm{C_i^t} < 1/N \\ \text{均等贡献智能体}, & \mathrm{C_i^t} = 1/\mathrm{N} \end{cases}$$，该判定依据具有直观的物理意义：若智能体 i 的贡献度高于均等分配水平，则认为其在当前时刻对全局感知覆盖的实际贡献超出平均水平，算法应给予更强的探索激励；反之，则引导其向已验证的高价值策略收敛，以减少无效探索。该阈值随智能体总数 N 自动调整，在不同规模场景下保持判定标准的一致性，无需人工设定额外超参数。上述分类结果直接作用于公式(11)的贡献度加权熵项，高贡献智能体获得更大的熵正则化权重，低贡献智能体权重相对收缩，从而在损失函数层面实现差异化探索调控。

4.4.2　动态 KL 散度裁剪范围调整

为解决传统 PPO 固定裁剪范围导致的策略更新不稳定问题，CA2-PPO 提出动态 KL 散度裁剪机制，对应算法 1 第 2、19–21 行。该机制根据新旧策略间的 KL 散度动态调整裁剪范围：KL 散度过大会导致剧烈更新影响收敛，过小则更新缓慢降低效率。通过动态调节使 KL 散度保持在合理范围内，从而提升训练稳定性与效率。动态 KL 散度裁剪机制具体流程如下。

算法 3.　KL 散度驱动的动态 PPO 裁剪范围算法.

输入：网络参数Ω、智能体数量 N、网格（H×W）、超参（γ，λ，α，τ）、RSU 广播的网格价值场V_t、Critic 参数ϕ、经验缓冲区$\mathcal{D}$

输出：训练后策略参数θ*、最终裁剪范围ε_{final}、裁剪范围演化轨迹E_ε、KL 散度记录 K_{KL}

1. $\varepsilon(t) \leftarrow \varepsilon_0; E_\varepsilon \leftarrow \phi; K_{\mathrm{KL}} \leftarrow \phi$
2. FOR each $b \subseteq \mathcal{D}$ DO：
3. $\rho^{(t)}(s,a) \leftarrow \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}$
4. $L_1 \leftarrow \rho^{(t)}(s,a)\ \tilde{A}_t$
5. $L_2 \leftarrow \mathrm{clip}(\rho^{(t)}(s,a), 1-\varepsilon(t), 1+\varepsilon(t))\ \tilde{A}_t$
6. $\mathcal{L}_{clip} \leftarrow \hat{E}_b[\min(L_1, L_2)]$
7. $\theta \leftarrow \mathrm{Adam}(\theta, \nabla_\theta L(\theta))$
8. 按公式(12)估计各智能体 KL 散度$D_{\mathrm{KL}}^{(i)}$,i=1,…,N
9. $D_{\mathrm{KL}} \leftarrow \sum_{i=1}^{N} C_i^t \ D_{\mathrm{KL}}^{(i)}$
10. IF $D_{\mathrm{KL}} > \tau$ THEN
11. $\varepsilon(t) \leftarrow \max(\varepsilon(t) - \Delta\varepsilon, \varepsilon_{\min})$
12. ELSE IF $D_{KL} < \tau/2$ THEN
13. $\varepsilon(t) \leftarrow \min(\varepsilon(t) + \Delta\varepsilon, \varepsilon_{\max})$
14. END IF
15. $E_\varepsilon \leftarrow E_\varepsilon \cup \{\varepsilon(t)\}; K_{\mathrm{KL}} \leftarrow K_{\mathrm{KL}} \cup \{D_{KL}\}$
16. END FOR
17. RETURN　θ、E_ε、K_{KL}

CA2-PPO 通过动态 KL 散度裁剪机制，将裁剪范围与策略更新幅度及智能体贡献度动态关联，KL 过大时收紧以抑制剧烈更新，过小时放宽以提升效率，由此推导定理 1，即多智能体 PPO 稳定性与可控收敛，在保证稳定收敛的同时兼顾协作效率。

策略稳定性条件：采用与策略更新幅度动态关联的 KL 散度裁剪范围$\varepsilon(t)$，且该范围满足随当前新策略$\pi_{\theta_{new}}$与旧策略$\pi_{\theta_{old}}$，则两者之间的 KL 散度$D_{KL}\left(\pi_{\theta_{old}} \| \pi_\theta\right)$动态调整的分段函数关系；新老策略的平均 KL 散度D_{KL}由下式无偏估计：

$$D_{\mathrm{KL}} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{a\sim\pi_{\theta_{old}}}\left[log\frac{\pi_{\theta_{old}}(a|s_b)}{\pi_\theta(a|s_b)}\right] \#(12)$$

动态裁剪范围的分段调节规则如下：基础裁剪范围$\varepsilon(t) \in [\varepsilon_{min}, \varepsilon_{max}]$，满足以下分段函数关系。设定 KL 阈值$\tau$=0.02，处于 PPO 标准经验范围 0.01～0.05 的中位。该取值兼顾两端权衡：过小会导致裁

剪范围频繁收缩、训练过于保守，过大则裁剪调整滞后、震荡抑制效果减弱。每次根据 KL 反馈对$\varepsilon(t)$进行最多±0.05 的离散调整。具体规则为：

$$\varepsilon(t+1) = \begin{cases} \max(\varepsilon(t)-0.05,\ 0.05), & \text{if } D_{\mathrm{KL}} > \tau \\ \varepsilon(t), & \text{if } \tau/2 \le D_{\mathrm{KL}} \le \tau \\ \min(\varepsilon(t)+0.05,\ 0.3), & \text{if } D_{\mathrm{KL}} < \tau/2 \end{cases} \#(13)$$

其中，$\varepsilon(t)\in[\varepsilon_{\min},\ \varepsilon_{\max}]$，本文取$\varepsilon_{\min}$=0.05，$\varepsilon_{\max}$=0.3，每次调整步长$\Delta\varepsilon$=0.05。调整步长$\Delta\varepsilon$=0.05 与 KL 阈值的量级匹配性。由公式(15)，在 Lipschitz 策略假设下 KL 散度满足$D_{\mathrm{KL}}\le C\varepsilon_t^2$，因此裁剪范围的变化$\Delta\varepsilon$所引发的 KL 约束变化量为：$\Delta D_{\mathrm{KL}}^{\max}=C\left[(\varepsilon_t+\Delta\varepsilon)^2-\varepsilon_t^2\right]=C\left(2\varepsilon_t\Delta\varepsilon\pm(\Delta\varepsilon)^2\right)$ 在 $\varepsilon_t=\varepsilon_0$=0. 2 处代入$\Delta\varepsilon$=0.05，得$\Delta D_{\mathrm{KL}}^{\max}\approx$0.02C，与 KL 裁剪阈值$\tau$=0.02 恰在同一量级（取 C=1 时$\Delta D_{\mathrm{KL}}^{\max}\approx\tau$），确保每次调整对 KL 约束产生恰当响应。步长过小则调节效果仅为 0.004C，响应迟缓；步长过大则单步变化达 0.04C，超过阈值两倍，易引发震荡。需要说明的是，C 为与策略参数化方式相关的 Lipschitz 常数，此处取 C=1 仅为说明量级关系的保守估计，关键结论在于$\Delta D_{\mathrm{KL}}^{\max}$与阈值$\tau$同阶这一量级关系，该关系对 C 的具体取值不敏感。因此，$\Delta\varepsilon$=0.05是与 KL 阈值τ量级相匹配的较优选择。$\varepsilon_{\min}$=0.05 下界的有效性保障。由公式（18）与定理 1 的约束条件，策略更新的有效性要求$D_{\mathrm{KL}}\le\varepsilon(t)$始终成立且$c$>0。当$\varepsilon_t$降至$\varepsilon_{\min}$=0.05 时，概率比率仍被约束在[0.95,1.05]内，策略仍有空间执行局部有效更新，不会完全停滞。同时，由公式（20）中的 Pinsker 不等式，此时两策略总变差距离上界为$\sqrt{2\times0.05}\approx$0.316，仍属于可控范围，不会引发策略发散。相比之下，若$\varepsilon_{\min}$设为 0，则算法可能陷入零更新的退化状态；若设为 0.1，则在高 KL 情形下无法充分收紧约束。因此$\varepsilon_{\min}$=0.05 在保证最低更新幅度与防止过度放松之间取得了合理平衡。参数取值在有效区间内具有鲁棒性。定理 1 已证明，只要$\varepsilon(t)\in$[0.05,0.3]，算法在期望意义下收敛至一阶驻点邻域且收敛速度上界可达 O(1/k)，说明算法性能对区间内$\varepsilon(t)$的具体取值不敏感。$\Delta\varepsilon = 0.05$的步长正是$\varepsilon(t)$始终维持在此有效区间内的关键保障。

策略更新约束性与 KL 有界性：PPO 的概率比率$\rho^{(t)}(s,a)$被限制在$[1-\varepsilon(t),1+\varepsilon(t)]$区间内：

$$\rho^{(t)}(s,a)=\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}\in[1-\varepsilon(t),1+\varepsilon(t)]\#(14)$$

由裁剪约束$\rho^{(t)}(s,a)\in[1-\varepsilon(t),1+\varepsilon(t)]$，可得：

$$\left|\log\rho^{(t)}(s,a)\right|\le\log\frac{1+\varepsilon(t)}{1-\varepsilon(t)}\le\frac{2\varepsilon(t)}{1-\varepsilon(t)}\#(15)$$

结合 KL 散度定义与假设 3 的 Lipschitz 条件，当$\varepsilon(t)\le$0.3 时，存在仅依赖于$\varepsilon_{\max}$与 Lipschitz 常数 L 的正常数K_L，使得：

$$D_{\mathrm{KL}}\le K_L\varepsilon(t)^2\#(16)$$

其中，K_L为与策略分布 Lipschitz 常数相关的正常数，当$\varepsilon_{\max}$=0.3 时，KL 随 L 的增大而增大，收敛界随之放宽。由公式（12）与（16）可得：

$$0\le D_{\mathrm{KL}}\le\tau_{\max}=K_L\varepsilon_{\max}^2\#(17)$$

上述的证明保证了策略更新过程中的全程有界性，以防止策略发散。

策略更新过程中，新策略与旧策略的 KL 散度被限制在[0，$\varepsilon(t)$]范围内，通过引入裁剪目标函数梯度，CA2-PPO 能够根据当前训练进度和智能体之间的协作情况动态调整策略裁剪范围$\varepsilon(t)$，避免过大的更新幅度导致的策略震荡，同时也避免过小的调整策略裁剪范围$\varepsilon(t)$导致的训练停滞。

基于上述推导，本文给出定理 1，形式化证明 CA2-PPO 在迭代更新意义下的稳定性。

定理 1. 基于动态 KL 裁剪与贡献加权熵的多智能体 PPO 稳定性与可控收敛.

假设 1. 奖励有界性. 存在常数$R_{\max}$>0，使得对所有状态 s 和动作 a，满足$|r(s,a)|\le R_{\max}$。

假设 2. 状态动作空间紧致性. 状态空间 S 和动作空间 A 是紧致的。

假设 3. 策略平滑性. 对任意状态s和动作a，$log\pi_\theta(a\mid s)$关于参数θ是 L-Lipschitz 连续的。即存在常数L>0 使得对任意θ_1,θ_2:$\left|log\pi_{\theta_1}(a\mid s)-log\pi_{\theta_2}(a\mid s)\right|\le L\parallel\theta_1-\theta_2\parallel$。该假设在常见的 softmax 或 Gaussian 参数化策略网络中通常成立。

假设 4. 优势函数有界性. 经过裁剪和归一化处理后的优势函数满足$\tilde{A}_t\in$[-5,5]，即$A_{\max}=5$。该上界由假设 1 与公式(29)的阈值裁剪操作共同保证。

假设 5. 局部平稳性. 在每个 PPO 更新周期内，各智能体的状态访问分布D_K相对于策略变化近似平稳，即满足$\parallel D_{k+1}-D_k\parallel_1\le\delta_D$，其中，$\delta_D$为与裁剪范围$\varepsilon(t)$同阶的小量。

记第k次策略更新中，固定旧策略$\pi_{\theta_{old}}=\pi^{(k)}$，通过 PPO 优化得到新策略$\pi_\theta=\pi^{(k+1)}$，其 KL 散度的经验估计为：

$$D_{\mathrm{KL}}^{(k)}=\mathbb{E}_{s\sim D_k}\left[KL\left(\pi_{\theta_{old}}(\cdot\mid s)\parallel\pi_\theta(\cdot\mid s)\right)\right]\#(18)$$

其中，$D_{\mathrm{KL}}^{(k)}$为第 k 次更新周期对状态分布D_k取期望的 KL 估计，用于定理证明；D_{KL}为每批次经验估计，用于实际动态裁剪调节。$\pi_{\theta_{old}}$在当前 PPO 更新周期内参数固定，仅用于计算策略比例与 KL 散度；D_k为第 k 次策略更新时的状态访问分布。

若对所有更新步 k 均满足动态裁剪约束：

$$D_{\mathrm{KL}}^{(k)} \le \varepsilon(t), \quad \varepsilon(t) \in [0.05,0.3] \#(19)$$

则由 CA2-PPO 生成的策略序列$\{\pi_k\}$在期望意义下收敛至目标函数的一阶驻点邻域内，其收敛速度可以通过裁剪区间$\varepsilon(t)$与阈值参数的设定实现调控。

证明. 动态 KL 裁剪保证策略更新有界。具体而言，在实现中，当前裁剪范围$\varepsilon(t)$由经验 KL 散度$D_{\mathrm{KL}}^{(k)}$自适应更新：当$D_{\mathrm{KL}}^{(k)} > \tau$时，减小，从而减小下一轮更新幅度；当$D_{\mathrm{KL}}^{(k)} < \tau/2$时，增大，以避免更新过于保守；同时在实现中对$\varepsilon(t)$施加上下界约束[0.05，0.3]，当$\varepsilon(t)$取值在该区间内时，算法性能对其变化不敏感，能够在稳定性与收敛速度之间取得良好平衡。进一步地，公式(13)的裁剪调整基于上一轮 KL 反馈，存在一步滞后，但由于调整步长固定为 0.05 且$\varepsilon(t)$∈[0.05,0.3]有界，单步偏差不超过 0.05，不影响收敛结论的有效性。因此，对所有迭代 k 步如下：

$$0 \le D_{\mathrm{KL}}^{(k)} \le \varepsilon(t) \le 0.3 \#(20)$$

由 Pinsker 不等式可得，对任意状态 s，两策略在动作分布上的总变差距离满足：

$$\| \pi^{(k+1)}(\cdot \mid s) - \pi^{(k)}(\cdot \mid s) \|_1 \le \sqrt{2 D_{\mathrm{KL}}^{(k)}}$$
$$\le \sqrt{2\varepsilon(t)} \le \sqrt{0.6} \#(21)$$

KL 有界⇒性能差异有界与近似单调改进。记折扣回报目标为：

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_t\right] \#(22)$$

代理目标为 PPO 常用形式。在奖励有界、优势函数裁剪与归一化的前提下，优势$\tilde{A}_t$有统一上界$\left|\tilde{A}_t\right| \le A_{\max}$=5，利用性能差异引理及状态分布对策略分布的 Lipschitz 依赖性，可推出真实目标与代理目标之间的差异满足：

$$\left|\mathrm{J}(\pi^{(k+1)}) - \mathrm{J}(\pi^{(k)}) - \left(\tilde{\mathrm{J}}_{\mathrm{k}}\pi^{(k+1)} - \tilde{\mathrm{J}}_{\mathrm{k}}(\pi^{(k)})\right)\right|$$
$$< K_R\sqrt{\varepsilon(\mathrm{t})} \#(23)$$

其中，常数K_R仅与折扣因子、奖励与优势的上界有关。上述误差界在假设 5 的局部平稳条件下成立。由公式(21)，单步策略变化满足$\| \pi^{(k+1)}(\cdot \mid s) - \pi^{(k)}(\cdot \mid s) \|_1 \le \sqrt{0.6}$，结合状态分布对策略的 Lipschitz 依赖性，可得$\| D_{k+1} - D_k \|_1 \le L_D\sqrt{0.6}$，即$\delta_D$与$\varepsilon(t)$同阶有界，假设 5 在动态 KL 裁剪的约束下实践中成立。

由于 CA2-PPO 沿代理目标上升方向优化，当代理目标改进量满足$\tilde{\mathrm{J}}_{\mathrm{k}}(\pi^{(\mathrm{k}+1)}) - \tilde{\mathrm{J}}_{\mathrm{k}}(\pi^{(\mathrm{k})}) \ge 0$时，结合公式(23)可得：

$$J(\pi^{(k+1)}) - J(\pi^{(k)}) \ge -K_R\sqrt{\varepsilon(t)} \#(24)$$

其中右端为性能退化的显式下界。由$\varepsilon(t) \le 0.3$，退化上界为有限常数$K_R\sqrt{0.3}$，排除策略发散。随训练推进$\varepsilon(t)$收紧，退化上界单调递减。综上，动态 KL 裁剪将每步性能退化严格约束在$K_R\sqrt{\varepsilon(t)}$以内，显著降低训练中大幅回退与发散的风险。

进一步地，贡献度平滑与归一化在提升 KL 估计稳健性方面亦起到关键作用。基于公式(9)的 EMA 更新所得原始贡献得分$\tilde{C}_i^t$无归一化约束，经公式(25)归一化后得到用于策略加权的贡献度C_i^t，后者构成概率单纯形，如公式(25)所示。

$$C_i^t = \frac{\tilde{C}_i^t + \delta}{\sum_{i=1}^{N}(\tilde{C}_i^t + \delta)} \#(25)$$

其中，δ为防止分母为零的平滑项，防止所有智能体贡献得分同时为零时分母为零导致的数值溢出。该取值与深度学习中 Adam 优化器、批归一化等标准操作中数值稳定项的量级一致，远小于正常贡献得分的数量级，不影响归一化结果的实际分布。由此可得$C_i^t \geqslant 0$ 且$\sum_{i=1}^{N} C_i^t = 1$。在此基础上，将多智能体 KL 散度估计视作各智能体 KL 的加权组合则有公式（26）：

$$\bar{D}_{\mathrm{KL}} = \sum_{i=1}^{N} C_i^t \, D_{\mathrm{KL}}^{(i)} \#(26)$$

其中，C_i^t与公式(11)中所用贡献度一致，$\bar{D}_{\mathrm{KL}}$为贡献度加权的全局 KL 散度估计，其非均等分布经由(26)直接作用于全局 KL 估计，使策略更新幅度与智能体实际贡献水平动态适配，从而实现贡献感知驱动的差异化协同调度。

由方差的线性性质可得公式（27）：

$$Var[\bar{D}_{\mathrm{KL}}] \le \sum_{i=1}^{N} (C_i^t)^2 \, Var\left[D_{\mathrm{KL}}^{(i)}\right] \#(27)$$

由于 $C_i^t \in [0,1]$，故 $(C_i^t)^2 \le C_i^t$，从而 $\sum_i (C_i^t)^2 \le \sum_i C_i^t = 1$，上述不等式表明，$\bar{D}_{\mathrm{KL}}$的估计

更平滑、方差更小，进一步增强了动态裁剪机制的稳定性。基于此，CA2-PPO 的收敛性可从以下四个维度得到保证与调节：

（1）通过公式（13）动态调整策略裁剪范围$\varepsilon(t)$。当 KL 过大时，自动减小$\varepsilon(t)$以抑制更新幅度；当 KL 过小时，适度放大$\varepsilon(t)$以加快收敛速度，并保持$\varepsilon(t)$在 0.05 至 0.3 的范围内；

（2）通过公式（9）和公式（25）调整熵正则化权重，增强高贡献智能体的探索激励，抑制低贡献智能体的探索，使得熵正则化机制在多智能体系统中实现差异化调节；

（3）结合优势裁剪与归一化以及参数梯度裁剪，整体损失函数的梯度始终有界；

（4）收敛速度分析：由假设 3 与假设 4，采用衰减学习率$\alpha_{\mathrm{lr},k}=\alpha_{\mathrm{lr}}/\sqrt{k}$时，由非凸随机梯度下降标准收敛结论：

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla J\left(\pi^{(k)}\right)\right\|^2\right]\le\frac{C_{\mathrm{sgd}}}{\sqrt{K}}=\mathrm{O}\left(\frac{1}{\sqrt{K}}\right)\#(28)$$

其中，C_{sgd}为与初始目标函数值、梯度噪声方差及学习率相关的综合常数。即策略序列以$\mathrm{O}\left(\frac{1}{\sqrt{K}}\right)$速率收敛至一阶驻点。在常数学习率与有界梯度噪声条件下，结合动态 KL 裁剪可进一步得$\mathrm{O}\left(\frac{1}{K}\right)$邻域收敛界，邻域半径由$\varepsilon(t)$上界与噪声方差共同决定，可通过调节动态裁剪超参数实现收敛精度与速度的联合控制。

证毕.

综上，CA2-PPO 通过动态 KL 裁剪、贡献加权熵与优势裁剪的协同约束，在非平稳环境下保证了策略序列的稳定收敛性，从理论层面验证了算法的鲁棒性与可扩展性。

4.4.3 优势函数归一化裁剪

在本文中，提出了 1 种带裁剪的优势函数归一化算法旨在预处理原始优势函数提升策略梯度稳定性。该算法是根据算法 1 第 12 行详细展开，该过程如算法 4 所示。算法 4 说明优势函数归一化裁剪的过程，先将 A 按默认阈值 5.0 裁剪至[-5，5]以过滤异常值，再经均值归一化消除整体偏移，最后通过标准差归一化统一数据尺度，输出$\widetilde{\mathrm{A}}_{\mathrm{t}}$。该过程抑制噪声与极端值干扰，使分布更稳定，为后续策略优化提供可靠梯度信号，适配网格环境下多智能体在各类任务分布中的协作训练。

其中，裁剪阈值的选择需要在保留有效梯度信息与抑制极端样本噪声之间取得平衡。参考标准 PPO 实现中的经验设置，并结合本文任务特性，本文将裁剪阈值设定为[-5，5]，主要基于以下考虑：（1）阈值过小会过度压缩优势幅度，导致有效梯度被抑制，降低策略优化效率；（2）阈值过大则难以过滤离群样本，极端优势可能引发策略更新震荡；（3）在本文场景中，网格环境与真实路网的奖励尺度差异较大，而[-5，5]能够较好覆盖两类场景的优势分布范围，从而在不同任务设置下保持稳定且一致的梯度更新行为。为验证裁剪阈值±5.0 的合理性，本文在网格环境与 OSM 真实路网场景下，对$\varepsilon_{adv}\in\{\pm1,\pm3,\pm5,\pm8,$无裁剪$\}$五组设置进行系统消融实验，其余超参数保持不变，结果如表 2 所示。

表 2 不同裁剪阈值下的梯度质量对比实验结果

ε_{adv}	网格均值/OSM 均值	网格标准差/OSM 标准差
±1	0.3098/0.0361	0.0241/0.0058
±3	0.3251/0.0407	0.0198/0.0044
±5	0.3364/0.0441	0.0203/0.0039
±8	0.3289/0.0418	0.0267/0.0053
无裁剪	0.3153/0.0374	0.0358/0.0082

由表 2 可知，裁剪阈值对梯度质量的影响呈现先改善后劣化的规律。当ε_{adv}=±1 时，两类场景均值低于无裁剪基线，说明过度裁剪压制了有效优势信号；随阈值放宽至±3，均值提升至 0.3251/0.0407，稳定性有所改善；当ε_{adv} ±5 时均值达到峰值，均值分别为 0.3364 与 0.0441，标准差同时降至最低，梯度更新实现最优平衡。阈值继续扩大至±8 乃至取消裁剪后，均值下降而标准差持续攀升，训练稳定性明显劣化。综上，ε_{adv}=±5 是兼顾梯度有效性与训练稳定性的最优选择。带裁剪的优势函数归一化算法过程具体如算法 4。

算法 4. 带裁剪的优势函数归一化算法.

输入：原始优势函数 *A*、裁剪阈值ε_{adv}，取值±5

输出：归一化后的优势函数$\tilde{A}_t$

1. $A_{clip}(s,a)\leftarrow\max(\min(A(s,a),\varepsilon_{adv}),-\varepsilon_{adv})$
2. $\mu_A\leftarrow\mathrm{mean}(A_{clip})$
3. $\sigma_A\leftarrow\mathrm{std}(A_{clip})$
4. $\tilde{A}_t\leftarrow\frac{A_{clip}(s,a)-\mu_A}{\sigma_A+10^{-8}}$
5. RETURN $\tilde{A}_t$

传统 PPO 算法中，优势函数易受异常值与噪声干扰，影响策略梯度精度。优势函数衡量状态下执

行某动作相对随机动作的价值优势，极端样本会导致梯度计算偏差。为此，CA2-PPO 采用先裁剪后归一化的两步处理：抑制异常值的同时统一数值尺度，使优势函数分布更加稳定，提升梯度估计精度并增强训练收敛性。

先将优势函数A(s,a)裁剪至[-5，5]区间：

$$A_{clip}(s,a) = \max(\min(A(s,a),5),-5)\#(29)$$

再进行标准化：

$$\widetilde{\mathrm{A}}_{\mathrm{t}} = \frac{A_{clip}(s,a) - \mu_A}{\sigma_A + 10^{-8}}\#(30)$$

其中，μ_A与σ_A为批次优势函数的均值与标准差，分母添加10^{-8}避免标准差为 0 的计算误差。这与 CA2-PPO 的机制设计密切相关。具体而言，CA2-PPO 的贡献度机制仅用于调节熵正则化强度，而非直接参与奖励重构，这一设计选择具有明确的动机：将贡献度引入奖励分配将向每个智能体的训练信号引入对其他智能体行为的间接估计误差，在 Non-IID 分布下可能放大协作噪声，破坏收敛稳定性。然而，上述设计并不意味着系统对搭便车（Free-riding, FR）行为缺乏约束。防止 FR 的机制已内生于奖励函数结构中：协作覆盖奖励对重复覆盖显式施加$-\eta \cdot V(g_k)$惩罚；即时覆盖奖励中的直接覆盖分数因历史状态累积而对重复覆盖衰减；时空关联奖励基于覆盖增量计算，路径重叠导致增量为零时不产生有效收益。上述三项机制协同保证：低贡献智能体即使跟随高贡献者的路径，也无法从路径重合中获得额外奖励，反而因协作惩罚而受到激励约束，从而在奖励结构层面天然排除了奖励过度分配的可能。关于高密度场景下贡献度向量陷入局部最优的风险，CA2-PPO 通过以下机制加以规避：其一，EMA 的快速响应性确保贡献度能够在约 20 步内捕捉智能体间的性能分化，避免长期固化；其二，归一化约束使贡献度构成概率单纯形，从结构上限制了单一智能体贡献度的无限扩张；其三，即便低贡献智能体贡献度趋近于零，其熵正则化权重仍保持正值，保证最低探索强度，防止策略退化为确定性行为；其四，RSU 持续广播的全局价值场提供了独立于贡献度机制的外部分散信号，在高密度区域引导智能体向低覆盖价值区域迁移，与贡献度动态形成互补的调节闭环。上述机制的综合作用使贡献度向量在训练过程中保持动态更新，在正常训练条件下不会收敛至退化分布。

5 实验仿真与结果分析

在实验中，本文构建了 20×20 与 30×30 的网格道路网络，并选取多种主流算法作为对比基线，以验证 CA2-PPO 在不同环境复杂度与任务分布下的性能；同时，基于 OSM 的北京与湖南多区域路网开展仿真测试，评估多智能体协作调度的有效性。

5.1 实验设置

刻画城市道路拓扑结构与通行特性，在 4.1 节所构建的两种规模网格路网基础上，进一步模拟主干道与支路的通行效率差异。每个回合开始时车辆随机初始化，网格感知价值从[0,1]均匀分布采样归一化，刻画路网的交通异质性。每个回合持续 3600s，决策步长 180s，共 20 个时间步，采用多智能体 PPO 联合优化策略与价值函数。

在此基础上，进一步构建局部通行效率演化与拥堵传播机制，模拟瓶颈扩散与连锁拥堵过程，形成高现实一致性的多智能体仿真环境。实验设计 GD、UD、RD、DD 四种典型任务分布模式以模拟不同城市交通环境，以平均覆盖得分为核心评价指标与多种主流算法进行对比，具体的实验环境见表 3。

表 3 实验环境

软硬件配置	版本
CPU	Intel Core i9-13900K
GPU	NVIDIA RTX 4090
内存	64GB
操作系统	Ubuntu 22.04 LTS
PyTorch	1.13
Python	3.9

本文中智能体执行感知任务的规则设定如下：

将感兴趣区域离散为网格，各网格初始价值按任务需求赋值并可随环境动态更新；智能体进入网格且满足有效覆盖条件即判定任务完成，系统实时更新进度并反馈。任务周期设为 300/600s，决策频率为每 30/180s；长周期情境下每 20 次决策自适应调整任务配置以刻画需求演化并评估稳定性。为保证可比性，每回合结束后重置环境；各回合同步随机初始化智能体位置与任务状态。训练后通过任务回放验证数据与环境一致性，并在迭代中提升复杂交通情境下的协同调度能力。

本文选取了五种主流算法作为对比，并在保持基本参数一致的前提下开展实验评估。具体的实验

参数设置见表 4，对比算法说明详见表 5。

表 4 实验参数取值

参数名称	数值
批次大小	64 个
每次更新步数	2048 步
每次更新轮数	10 轮
回合时长	3600s
决策时间步长	180s
每回合决策次数	20 次

表 5 算法说明

算法名称	算法说明
Random	每个时隙，各智能体随机选择东、南、西、北四个方向移动，作为参照验证算法有效性的基线
MADQN	主流基于价值的多智能体强化学习方法
QMIX	价值分解型多智能体协作算法，引入了通信模块实现了有效的全局协调
MAPPO	基于 CTDE 的多智能体 PPO，集中式 critic、分散式执行
MACPPO	未加入 CA2-PPO 模块的基础版本，用于对比验证创新机制效果

在本文实验设置下，平均覆盖得分、平均任务完成率与平均奖励具有一致的单调性，其中平均奖励可视为覆盖效果在强化学习框架下的等价回报表达。

5.2 核心机制消融与扩展实验

本文在多智能体调度任务中开展了系统性的消融实验，实验环境设定为 20×20 网格地图，共 400 个位置点。在每个实验回合中，4 辆自主调度智能体在 3600s 的时长内以 180s 为决策间隔执行调度，即每辆车每回合进行 20 次决策，总计 80 次决策。

训练使用标准 PPO 参数配置：每收集 2048 个时间步数据更新一次、每次更新 10 轮、GAE 参数 λ=0.95，为 PPO 的标准经验值，接近 1 时偏差小但方差大，偏小时方差小但偏差增大，0.95 在两者间取得平衡、熵系数C_{cnt}=0.1、折扣因子γ=0.99、学习率$\alpha_{\mathrm{lr}}=3\times10^{-4}$。在此基础上额外设置：初始裁剪范围 ε_0 =0.2、价值损失系数C_v=0.5。

本节通过设计消融实验，对 CA2-PPO 中在线贡献感知机制及其关键模块在多智能体协同优化中的作用进行系统评估，并构造如下三种消融变体算法进行对比。

（1）CA2-PPO/WE：在完整方法 CA2-PPO 中，仅保留贡献度加权熵机制；

（2）CA2-PPO/DC：在完整方法 CA2-PPO 中，仅使用动态 KL 散度裁剪；

（3）CA2-PPO/AC：在完整方法 CA2-PPO 中，仅使用优势函数归一化裁剪。

实验的评估指标为任务完成度得分即 1 减去剩余未收集权重的比例，取值范围[0，1]，数值越高代表覆盖率越高。为确保结果的统计可靠性，每个实验配置在 5 个随机种子下独立重复运行，最终结果以均值±标准差的形式报告。

表 6 CA2-PPO 算法及其 3 个变体算法的检测性能与统计显著性分析

类别	方法名称	均值±标准差	统计显著性 p
基线算法	Random	0.3107±0.0186	—
完整算法	CA2-PPO	0.3364±0.0203	0.0035＜0.05
CA2-PPO 变体算法	CA2-PPO/WE	0.3273±0.0157	0.0060＜0.05
	CA2-PPO/DC	0.3246±0.0130	0.0110＜0.05
	CA2-PPO/AC	0.3259±0.0123	0.0171＜0.05

实验结果如表 6 所示。所提出的各创新机制在性能上均优于基线。具体来说，CA2-PPO 相比基线提升 8.3%，验证了算法的有效性。三个创新机制的消融版本均优于基线：CA2-PPO/WE 提升 5.3%、CA2-PPO/DC 提升 4.5%、CA2-PPO/AC 提升 4.9%，表明各机制独立有效。

从稳定性角度来看，CA2-PPO 的标准差为 0.0203，虽然略高于其他方法，但其性能的显著提升使得该波动性增加处于可接受范围内。而贡献度加权熵和动态 KL 散度裁剪等方法则展现了较小的标准差，表明它们在不同实验条件下具有更高的稳定性。

综合分析表明，实验结果充分验证了各项创新机制在多智能体协同任务中的有效性，有力印证了算法的有效性与稳定性，也凸显了其在实际协同调度场景中的应用潜力。上述结果的统计显著性通过双侧独立样本 t 检验计算，显著性水平设为 0.05。

图 2 展示了不同智能体贡献度的组间差异化特征。图 3 展示了不同熵正则化策略的奖励指标对比结果及其对差异化奖励机制的影响。

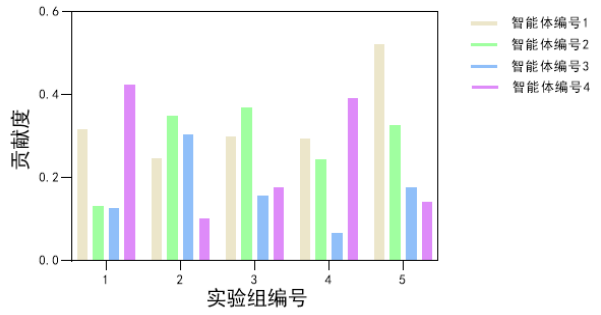

图 2　智能体贡献度的组间差异化特征

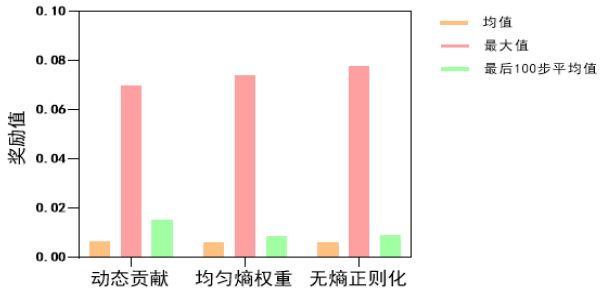

图 3　不同熵正则化策略的奖励指标对比结果

为验证贡献度加权熵机制的有效性，本文在 20×20 网格、4 智能体环境下对比了 CA2-PPO、PPO 和无熵正则化三种策略。结果显示，CA2-PPO 的平均奖励为 0.0779，明显优于均匀熵权重的 0.0739 和无熵正则化的 0.0697，并在训练后期稳定性上表现最佳。该机制使高贡献智能体优先覆盖高价值区域，同时促使低贡献智能体持续探索，在提升平均奖励的同时减少了训练后期的波动，验证了贡献感知探索机制的有效性。上述现象的背后，贡献度与实际贡献的匹配性首先可从机制设计上得到保证：贡献度基于归一化即时奖励的 EMA 估计，而即时奖励由公式（5）直接计算自智能体的实际覆盖行为，覆盖高价值网格则奖励高，重复覆盖则奖励衰减。因此，贡献度本质上是对智能体有效覆盖行为的历史累积度量，高贡献度意味着该智能体在过去时间步内持续产生了高质量的感知覆盖，而非统计噪声或随机涌现。从实验数据来看，这一机制特性得到了充分体现：不同智能体的贡献度差异显著，第 5 组中智能体 1 贡献度为 0.522，而智能体 4 仅为 0.142；第 1 组中智能体 4 的贡献度则达到 0.425。贡献度差异真实反映了智能体间的任务分工差异，高贡献智能体覆盖高价值网格的频率普遍高于低贡献智能体，表明贡献度不仅反映覆盖数量，还与覆盖质量保持一致，有效避免了传统算法的平均化问题。从宏观层面来看，如图 5 所示，GD 分布下贡献度高度集中于少数智能体，UD 分布下各智能体贡献度趋于均等，进一步验证了贡献度分配与真实任务分布的动态一致性。

5.2.1　不同任务分布下智能体贡献度

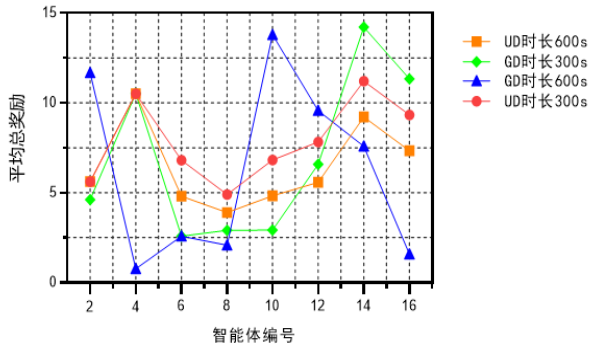

图 4　不同时长智能体性能对比结果

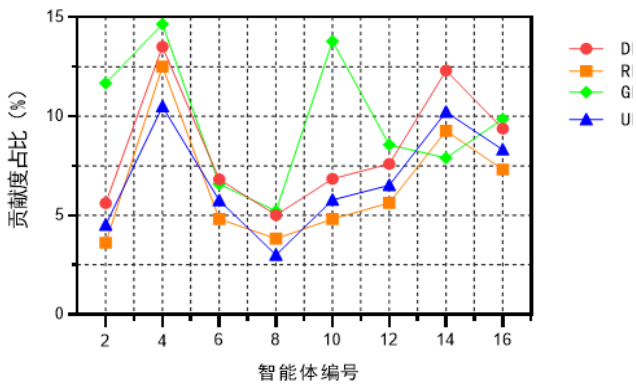

图 5　不同环境下的智能体贡献度对比结果

图 4 对比了 GD 与 UD 两种任务分布在 300s 与 600s 时长下各智能体的平均总奖励分布。结果显示，在 600s 的 GD 环境中，智能体 2、10、12 的平均总奖励明显高于其他智能体，贡献度呈现显著分化，表明 CA2-PPO 能够依据任务空间分布精准识别并持续激励高价值智能体，同时保证低价值区域任务不完全停滞；而在 UD 分布下，各智能体奖励差异相对均衡，贡献度分布趋于一致，符合均匀任务分布下资源均等分配的预期特征。

图 5 展示了 CA2-PPO 在 DD、RD、GD、UD 四种任务分布下随智能体数量变化的贡献度占比曲线。在 GD 分布中，贡献度向少数高价值智能体

高度集中，峰值智能体贡献度占比超过 10%，反映出算法对任务聚集区域的强感知能力；在 UD 分布中，各智能体贡献度趋于均等，整体波动较小；DD 与 RD 分布下的贡献度曲线介于两者之间，波动幅度随任务随机性增大而略有上升。上述结果直观验证了贡献度加权熵机制能够依据实际任务分布自适应调整激励强度，在不同环境下均与智能体真实贡献水平保持动态一致。

5.2.2 不同奖励分配策略对 CA2-PPO 性能的影响

为了验证不同奖励模式对算法性能的影响，本文对比了四种典型的奖励分配策略：贪婪策略（greedy）、总和奖励（sum）、贪婪均值（greedy-mean）和团队精神（team-spirit）。实验采用最佳 100 回合平均覆盖率作为核心评估指标，全面地评估了 CA2-PPO 算法在不同激励机制下的性能表现。如图 6 至图 9 所示。

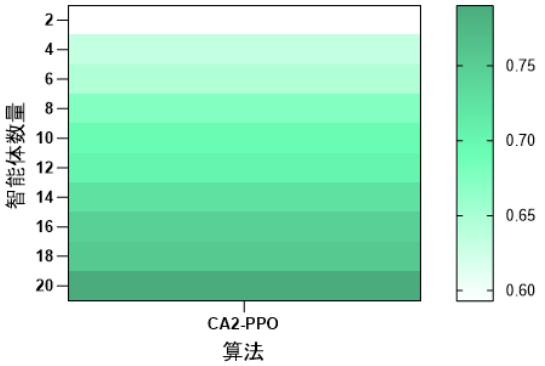

图 6 在 greedy 下的性能对比结果

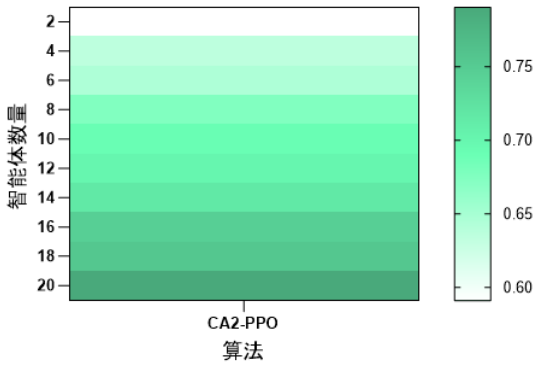

图 7 在 sum 下的性能对比结果

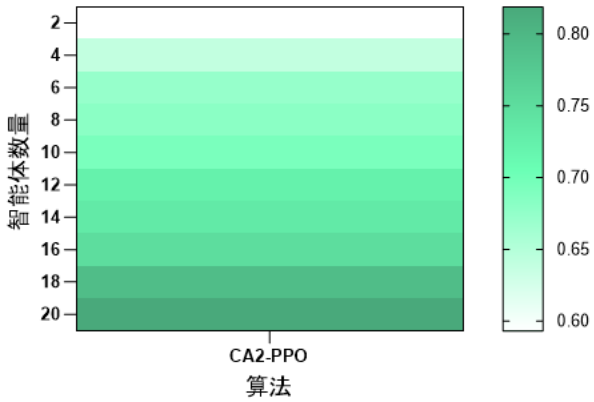

图 8 在 greedy-mean 下的性能对比结果

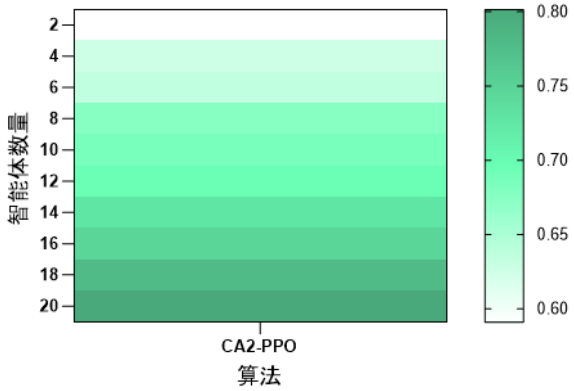

图 9 在 team-spirit 下的性能对比结果

本文评估了 CA2-PPO 算法在 4 种奖励机制下的性能。结果显示，随智能体规模从 2 扩展至 20，四种策略的平均覆盖得分均呈显著增长态势，但增幅各有差异。其中，greedy-mean 模式下性能最优，平均覆盖得分达 0.8185，较 greedy、sum、team-spirit 三种机制分别提升约 33.2%～35.6%；sum 模式与 team-spirit 模式表现居中，覆盖得分集中在 0.70～0.75 区间；greedy 模式在小规模场景下提升幅度相对有限，覆盖得分多集中于 0.60～0.70 区间。上述结果表明，greedy-mean 激励机制能够在个体收益与群体协作之间取得更优平衡，验证了多智能体奖励机制设计对 CA2-PPO 整体性能的重要影响。

5.2.3 不同算法贡献度分配

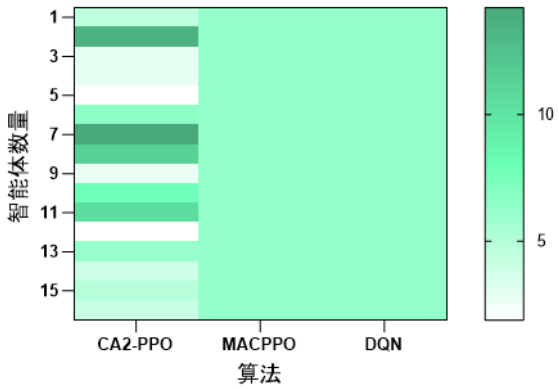

图 10　不同算法的贡献度分配对比结果

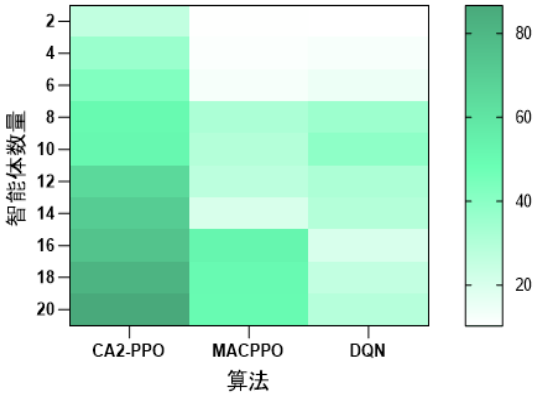

图 11　GD 环境下的平均任务完成率

为进一步可视化算法在空间维度的贡献效果，图 10 以热力图形式对比了 CA2-PPO、MACPPO 和 DQN 三种算法的智能体贡献度分布。从色深分布可以观察到，CA2-PPO 中智能体 2、7、8 的贡献度分别为 13.52%、14.21%、11.33%，对应色块明显深于其余智能体，且与 MACPPO、DQN 相比，CA2-PPO 的高贡献区域更为集中、对比度更强，说明其能够更精准地识别并持续强化高价值智能体的激励，而 MACPPO 和 DQN 的贡献度分布则相对分散均匀，缺乏有效的差异化调节能力，进而导致协作效率偏低。

通过引入贡献度加权机制，CA2-PPO 突破传统均等激励范式，在线量化个体贡献差异并自适应实施差异化激励，有效缓解协作低效，在理论创新性与实际性能上均展现出良好的可扩展性。

为进一步评估贡献度加权机制对任务完成率的影响，图 11 展示了 GD 环境下智能体数量从 2 增至 20 时各算法平均任务完成率的变化趋势。结果显示，随智能体规模扩大，CA2-PPO 的任务完成率始终高于 MACPPO 与 DQN，在智能体数量为 2 时平均任务完成率为 35.98%，增至 20 时已达 86.54%，且在智能体数量较多时优势更为显著；这表明贡献度加权机制在大规模协作场景下能够有效抑制智能体间的资源竞争与重复覆盖，保持稳定的任务调度效率，验证了该机制在不同规模下的可扩展性与有效性。

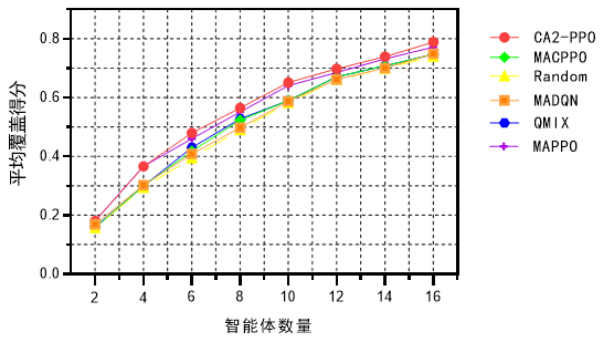

图 12　20×20 网格路网中不同智能体数量下各算法性能比较

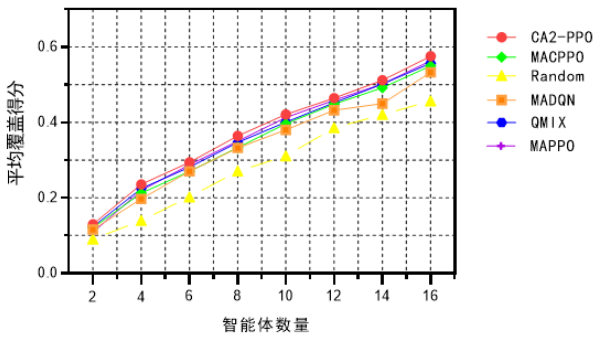

图 13　30×30 网格路网中不同智能体数量下各算法性能比较

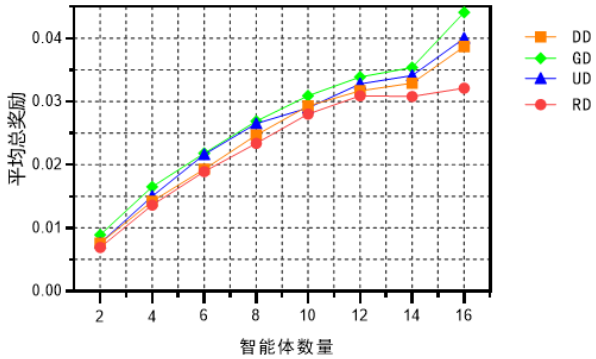

图 14 CA2-PPO 算法不同智能体数量各环境性能比较

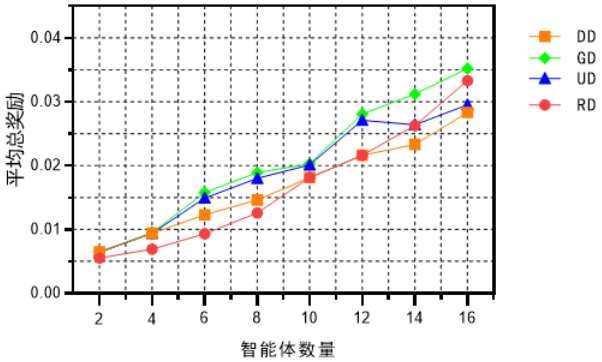

图 15 Random 算法不同智能体数量各环境性能比较

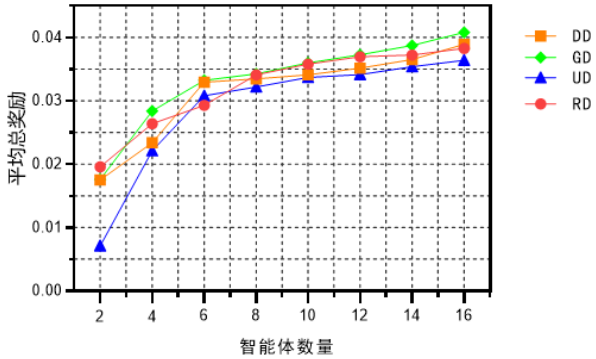

图 16 MADQN 算法不同智能体数量各环境性能比较

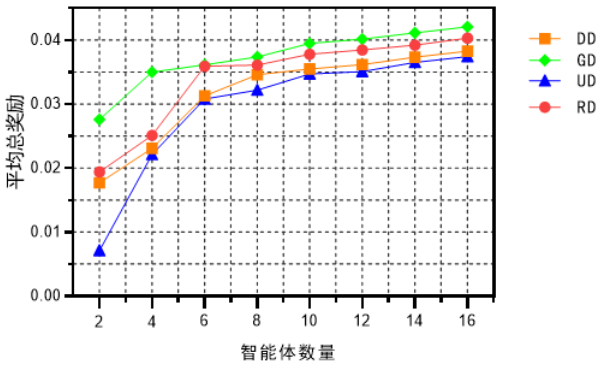

图 17 MACPPO 算法不同智能体数量各环境性能比较

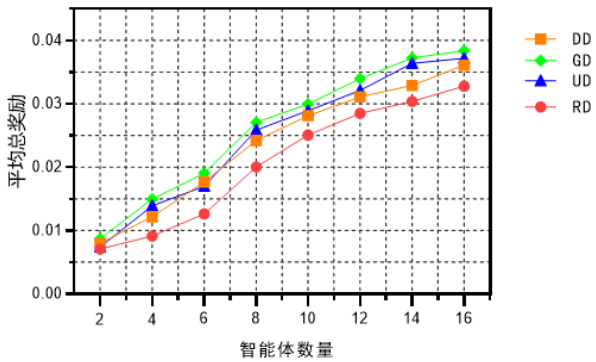

图 18 QMIX 算法不同智能体数量各环境性能比较

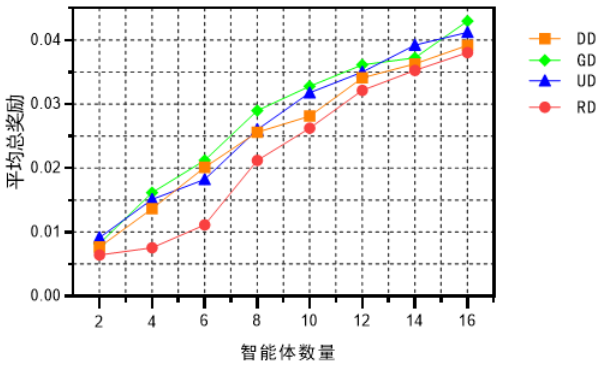

图 19 MAPPO 算法不同智能体数量各环境性能比较

实验从多个评估维度对 CA2-PPO 算法的有效性进行了全面验证与系统化分析。与其他五种常见强化学习算法展开比较，如图 12 所示，在 20×20 网格环境中，CA2-PPO 在本文 Non-IID 测试场景中均表现优异。在中等规模场景下，算法相较于 MADQN 和 MACPPO 分别提升了 22.2%和 28.3%；在大规模场景下，覆盖得分达到 0.7845，取得了最优表现。如图 13 所示，在 30×30 大规模环境中，尽管状态空间扩展了 2.25 倍，CA2-PPO 依然保持优良的性能优势。在 16 个智能体的场景下，相比 MADQN 和 MACPPO，算法分别提升了 7.9%和 4.8%，并且超越了 MAPPO 和 QMIX 的表现，尽管在末端，MADQN 略高，但这一局部波动并不影响 CA2-PPO 的整体性能优势。从机制设计角度分析，即使智能体数量进一步增加，CA2-PPO 的贡献感知机制仍能动态调整各智能体的探索力度与任务分配，有效抑制协作冗余；而 MAPPO 采用均匀激励方式，随智能体数量持续增加，重复覆盖与协作失衡问题会愈发突出，性能提升空间受限。因此，随着智能体规模的持续扩大，MAPPO 的平均奖励预计将持续低于 CA2-PPO。这一趋势判断与图 12、图 13 所呈现的实验结果高度吻合，表明 CA2-PPO

的性能优势并非仅在特定规模下成立，而是具备一定的规模鲁棒性。为进一步验证这一判断，我们在 20×20/30×30 网格环境下智能体数量为 18、20 的扩展实验。结果显示，在 18 个智能体场景下，CA2-PPO 平均覆盖得分为 0.8012/0.6086，MAPPO 为 0.7823/0.5812，CA2-PPO 分别领先 2.4%和 4.7%；在 20 个智能体场景下，CA2-PPO 覆盖得分为 0.8132/0.6525，MAPPO 为 0.7904/0.6221，CA2-PPO 仍分别保持 2.9%和 4.9%的优势，验证了该优势的规模鲁棒性。

如图 14 至 19 所示，CA2-PPO 在 GD、UD、RD、DD 四种任务分布下均呈现稳定的性能提升，环境间性能差异控制在 15%以内，其中 GD 场景下平均总奖励达 0.044，较其他分布提升约 10%～15%。上述结果表明，CA2-PPO 通过显式建模个体贡献差异并将贡献感知机制融入熵正则与策略裁剪约束，有效解决了传统多智能体 PPO 在 Non-IID 分布下贡献均衡假设导致的协作失衡问题，在不同任务分布、运行周期及任务规模条件下均具备良好的鲁棒性与泛化能力。

5.3 复杂城市路网下 CA2-PPO 算法有效性验证

为对所提方法开展实证性能验证，本研究基于 OSM 数据集构建了真实路网仿真实验平台，对 CA2-PPO 的性能进行评估。实验环境数据由 OSM 全球地理信息中定向裁剪得到的特定区域局部数据构成，相关区域的具体信息如表 7 和表 8 所示。在此基础上，实验设置智能体数量为 48 辆车，该规模综合考虑了所选 OSM 区域的路网节点数量及典型高峰时段的活跃车辆密度，能够在覆盖效率与通信负载之间取得较为合理的平衡。

表 7 北京区域

数据集	北京地区	地区说明
OSM	国贸 CBD	重要商业核心地带，具有典型的高密度城市特点
	石景山区	超大区域，拥有复杂的道路和较高的路网拓扑复杂性
	中关村	科技创新中心，具有较高的复杂度和密集的交通流量
	天安门	历史文化中心，交通流量较为集中
	通州副中心	重要核心区域，具有较高的路网复杂度和交通流量
	东城区	核心城区之一，交通密集且具有一定的路网复杂性

表 8 湖南区域

数据集	湖南地区	地区说明
OSM	长沙市	核心商业聚集区，是消费活力与人口密度的城市核心板块
	株洲市	交通枢纽与老工业基地，核心城区兼具产业底蕴与通达优势
	湘潭市	长株潭城市群核心，兼具紧凑布局与高效调度特性
	衡阳市	湘南地区核心枢纽，路网覆盖广且节点密度适中
	怀化市	湘西南生态与交通枢纽，山区路网特征显著
	永州市	市域面积 22441 平方公里，路网稀疏

为检验算法在真实城市路网环境下的性能表现，本研究依托 OSM 数据资源开展了仿真验证。相较规则网格，真实路网具有更多交叉路口、复杂交通流向及不规则布局，更能反映算法的实际表现，如表 7、8 所示。实验选取北京和湖南两地共 12 个典型区域，覆盖高密度商业核心如国贸 CBD、长沙市，交通枢纽如株洲市，科创中心如中关村，以及大规模稀疏区域如怀化、永州，以适配不同规模与复杂度的调度测试需求。

本节选取北京和湖南两地代表性城市场景进行对比实验，重点分析智能体数量与调度覆盖效果的关联规律，实验结果如图 20 至图 23 所示。

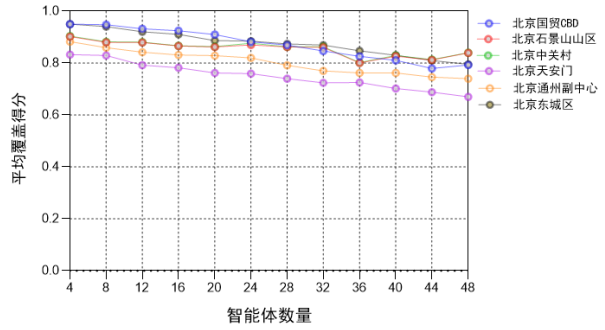

图 20 北京不同区域多智能体车辆平均覆盖得分对比分析

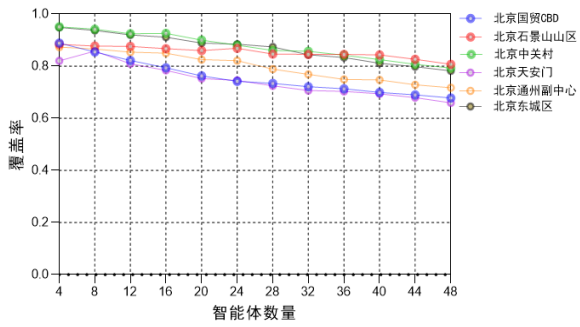

图 21 北京不同区域多智能体覆盖率对比分析

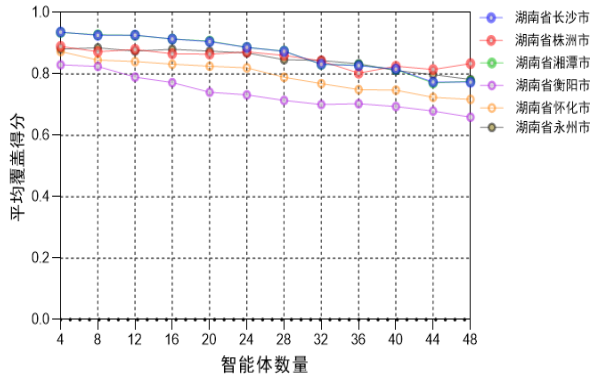

图 22 湖南不同区域多智能体平均覆盖得分对比分析

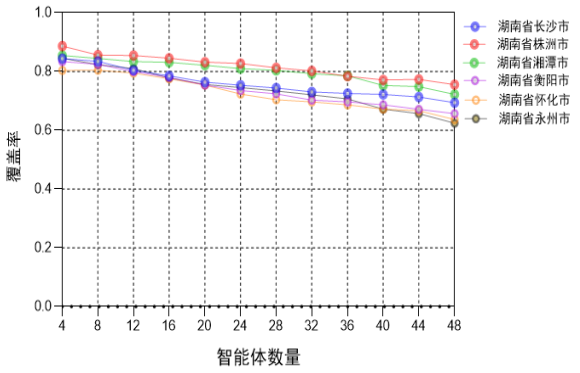

图 23 湖南不同区域多智能体覆盖率对比分析

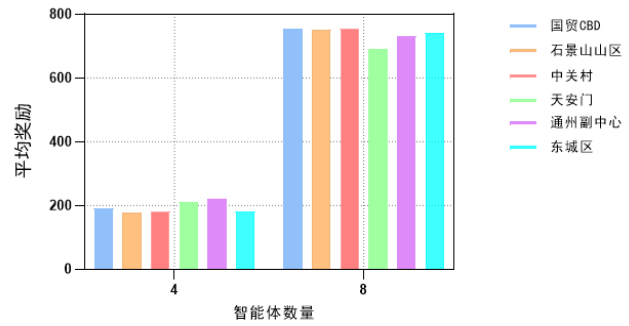

图 24 北京不同区域多智能体的平均奖励

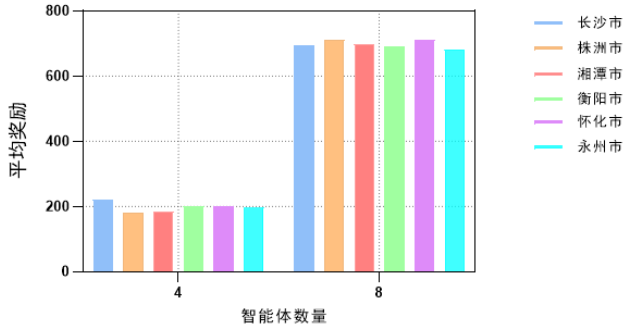

图 25 湖南不同区域多智能体的平均奖励

图 20 至 23 展示了北京与湖南不同区域的平均覆盖得分与覆盖率随智能体数量变化的趋势。在北京区域，各区域曲线随智能体数量增加整体呈下降态势，边际收益递减规律明显。国贸 CBD 初始覆盖得分最高 0.9502，增至 48 辆时缓降至 0.7924；中关村、天安门得分平稳下降但维持相对高位；石景山、通州曲线平滑、下降幅度较小；东城区得分整体偏低。覆盖率方面，各区域在智能体数量为 4 时普遍高于 0.85，随规模扩大逐步收敛至 0.61～0.85 区间。在湖南区域，路网异质性特征更为突出：长沙初始得分最高，下降趋势符合边际递减规律；株洲、湘潭表现居中；怀化山地路网鲁棒性较优；永州稀疏路网对智能体密度最为敏感，得分下降最为明显。覆盖率方面，各城市在 48 辆规模下整体收敛至 0.60～0.75 区间，区域间曲线间距较北京更为集中。综上，CA2-PPO 在两地不同路网结构下均实现稳定收敛，跨区域泛化能力与边际收益递减特征得到一致验证。

如图 24、25 所示，在平均奖励方面，北京与湖南呈现显著的区域差异。北京以中关村最高，平均奖励达 753.65，石景山、通州相对较低。湖南以长沙表现最优，平均奖励为 220.26，株洲次之，为 180.24。整体呈现核心区域发展成熟、其他区域相对滞后的特征。综上，CA2-PPO 能够在不同路网结构与需求分布条件下实现稳定收敛并保持一致的调度收益提升，进一步验证了该方法在跨区域场景

下的适应性，并为实际部署提供了一定的参考依据。

表 9 北京和湖南区域多智能体调度性能对比

区域	降幅	特征分类
国贸 CBD	-21.2%	边际递减区
石景山山区	-22.0%	边际递减区
中关村	-18.1%	高性能区域
天安门	-19.7%	资源重点区
通州副中心	-17.8%	高性能区域
东城区	-17.6%	高性能区域
长沙市	-19.5%	边际递减区
株洲市	-12.5%	高性能区域
湘潭市	-15.5%	高性能区域
衡阳市	-21.2%	异构适应性
怀化市	-16.4%	异构适应性
永州市	-20.5%	异构适应性

表 9 归纳了 12 个代表性区域的性能下降幅度特征及其分类归属。从区域类型来看，边际递减区如国贸 CBD、石景山山区、长沙市的降幅普遍较大，多在 19%～22%之间；高性能区域如中关村、通州副中心、东城区、株洲市、湘潭市的降幅相对均衡，集中在 12%～18%区间；异构适应性区域如衡阳市、怀化市、永州市降幅在 16%～21%之间。值得注意的是，天安门作为唯一的资源重点区，降幅达 19.7%。上述结果表明，CA2-PPO 在所选 12 个仿真区域中均能实现一定程度的调度性能提升，但不同路网结构下的性能差异较为明显，算法在高密度规则路网中的表现优于稀疏异构路网，其跨区域泛化能力有待在更多类型路网中进一步验证。本文方法在各类区域均实现有效调度性能提升，验证了其普适性与稳健性。

5.4 算法多维度扩展性与时间复杂度分析

在强化学习领域，系统的可扩展能力与时间复杂度特性构成了评估算法效率的核心指标。CA2-PPO 在计算、内存与通信三方面具备良好扩展性：计算开销随智能体数量近似线性增长，即使在大规模多智能体场景下，单步训练时间增幅仍处于可控范围，这得益于算法高效的并行化设计和轻量级卷积神经网络（Convolutional Neural Network, CNN）与全连接层（Fully Connected Layer, FC）结构，以及参数共享机制带来的并行效率；内存占用合理，显存压力主要集中于 RolloutBuffer；通信开销较低，参数广播与数据回传的时间占比均较小，能够满足城市级场景的扩展需求。CA2-PPO 算法的时间复杂度主要取决于智能体数量 N、训练时间步 T、训练轮次 E 及网络结构复杂度（C+F）。如表 10、表 11 所示，算法由数据采集、训练与测试三部分组成，其中训练模块为主要耗时环节。CA2-PPO 的总体时间复杂度关于智能体数量 N 与时间步 T 近似呈线性增长。其中训练模块为主导项，其复杂度主要由策略网络的前向与反向计算开销决定。

表 10 CA2-PPO 算法扩展性与计算性能

维度	描述	详细数值
计算开销	与基线算法的计算开销对比	仅增加不到 0.5%，近似线性趋势
内存占用	RolloutBuffer 的显存占用情况	346MB 支撑 50 车×20×20 网格规模
显存限制	单卡显存理论上限	单卡 24GB 显存可支持约 2000 车
通信量	参数广播量与数据回传量	参数广播量：2.1MB；单次 rollout 数据回传：328MB
通信时间	PCIe 接口通信时间占比	小于 5%的墙钟时间

表 11 CA2-PPO 算法的时间复杂度分析

维度	时间复杂度公式
数据采集	$O(T \times N \times (C + F))$
训练模块	$O(E \times T \times N \times (C + F))$
测试模块	$O(K \times T_{test} \times N \times (C + F))$
其他开销	—
算法总时间	$O(E \times T \times N \times (C + F) + K \times T_{test} \times N \times (C + F))$

表 10 从计算开销、内存占用、显存限制、通信量与通信时间五个维度对 CA2-PPO 的扩展性进行量化分析。在本文实验配置下，计算开销较基线算法仅增加不到 0.5%，随智能体数量增长呈近似线性趋势；RolloutBuffer 显存占用为 346MB（50 车×20×20 网格规模），基于线性外推估算单卡 24GB 显存理论上限约可支持 2000 辆车，实际部署受路网规模与硬件配置影响；参数广播量为 2.1MB，单次 rollout 数据回传为 328MB，PCIe 通信时间占比小于 5%的墙钟时间，通信开销处于可控范围。表 11 对各模块时间复杂度进行了理论分析，算法总时间复杂度为 $O(E \times T \times N \times (C + F) + K \times T_{test} \times N \times (C + F))$，主要开销集中于训练模块，其他开销因与主项相比可忽略而未单独列出。上述分析表明，CA2-PPO 在本文实验规模内具备合理的计算效率，但更大规模场景下的实际性能有待进一步验证。

5.5 通用性和局限性分析

CA2-PPO 通过引入贡献度加权熵、动态 KL 散度裁剪和优势归一化裁剪，在多场景下展现出明显优势：不仅能在动态与静态任务中精准识别高价值区域，在北京国贸 CBD、中关村等真实路网中也保持良好适应性。但该方法仍存在局限：在极端非结构化路网中，道路拓扑规则性低、节点密度剧烈波动，性能下降较为明显；算法对人工设计的状态特征依赖较强，在异构智能体场景中对感知能力差异的刻画仍不够精细。需要指出的是，本文当前的贡献度建模以归一化即时奖励为输入，未显式纳入感知半径、移动速度、能耗等硬件属性，在硬件能力强但位置分布不佳的智能体场景下可能存在贡献低估问题。未来可将硬件能力参数引入贡献度初始化权重或奖励归一化策略，以实现对硬件异质性的精细建模。同时，相比 MACPPO，训练时间有所增加，不过考虑到性能收益，该计算开销仍在可接受范围内。此外，在实际部署中还需关注贡献度向量的分布稳定性问题。具体而言，在高密度智能体场景下，若初始阶段少数智能体因位置优势持续获得高奖励，贡献度向量可能过早极化，导致低贡献智能体探索空间被持续压缩，系统陷入次优协作模式。对此，CA2-PPO 从以下三个层面加以抑制：（1）EMA 历史平滑：平滑系数 α=0.1 使贡献度估计对短期奖励波动保持钝感，单次高奖励事件对贡献度的影响被约束在 10%的增量以内，从而避免因局部偶然收益引发贡献度的快速极化。（2）归一化的竞争约束：贡献度归一化$C_i^t = \frac{\hat{C}_i^t}{\sum_j \hat{C}_j^t}$使各智能体贡献度始终处于相对竞争关系，单一智能体贡献度的提升必然对应其他智能体贡献度的相对下降，从结构上杜绝了赢者通吃式的极化现象。值得说明的是，CA2-PPO 并未对奖励信号本身进行差异化分配，低贡献智能体在路径重合时仍可获得与高贡献者相同的即时奖励，短期内存在一定的 FR 风险。然而，该风险在训练过程中可被自然抑制：低贡献智能体的熵正则权重持续偏低，探索激励受到长期压制；与此同时，竞争性归一化机制进一步压缩其贡献度估计，二者共同形成隐性的长期惩罚效应，从机制层面逐步瓦解 FR 行为的持续性。（3）低贡献智能体的基础探索保留：即使 C_i 偏低，其熵权重 $C_i \cdot \mathcal{H}(\pi_\theta)$仍为正值而非零，确保低贡献智能体始终保留基础探索能力，具备从次优路径中主动脱离的驱动力，从而在机制层面提供了对局部最优的自我纠偏能力。

尽管如此，在极端高密度场景下，上述机制的抑制效果可能有所减弱，贡献度向量仍存在陷入局部最优分布的风险。针对这一问题，可从两个维度加以强化：一是引入贡献度周期性重置机制，以均匀分布 1/N 为基准定期对贡献度向量进行校正，从而打破已固化的分布偏置，恢复探索多样性；二是引入多样性正则化约束，对贡献度方差设置显式上界，在优化目标层面主动抑制贡献度的过度集中趋势。上述两类扩展机制相辅相成，分别从分布校正与目标约束层面对现有方法形成有效补充，将在后续工作中进一步验证与完善。

6 总结和展望

针对 Non-IID 场景下多智能体协作中贡献失衡与训练不稳定问题，本文提出 CA2-PPO 算法，通过贡献感知的差异化激励与自适应策略更新提升协作效率与训练稳定性。实验在网格仿真环境与北京、湖南等多个 OSM 真实路网区域中验证了算法的有效性与跨区域适应性，但当前实验规模与路网拓扑类型仍有局限，结论的普适性有待进一步拓展。

本研究在异质智能体任务调度与群智感知协同方面取得了一定进展，但仍有若干方向值得进一步研究与完善。今后可从以下几方面进一步拓展工作：一是探索 CA2-PPO 在无人机集群、物联网边缘计算等更多异构多智能体群智感知场景中的适用性，针对智能体在感知能力、移动速度及传感器配置等方面存在差异的场景，设计适应性更强的贡献度建模与差异化策略更新机制，以进一步提升算法在真实复杂环境中的泛化能力；二是扩展实验，涵盖更多复杂且异质的真实路网数据，以进一步验证算法的适用性；三是在贡献度计算与任务分配机制中引入能力权重参数，对具有不同感知半径、移动速度、能耗等硬件配置的智能体进行差异化激励与调度，以进一步提升算法在异构智能体场景下的适应性与泛化能力。四是结合差分隐私或联邦学习机制，探索在保障参与方数据隐私的前提下实现贡献度估计与策略优化的协同，为城市级 VMCS 系统的安全合规部署提供技术支撑。

参 考 文 献

[1] Zhang Y, Li C, Li K, et al. High-resolution large-scale urban traffic speed estimation with multi-source crowd sensing data. IEEE Transactions on Vehicular Technology, 2024, 73(9): 12345-12357

[2] El Hafyani H, Abboud M, Zuo J W, et al. Learning the micro-environment from rich trajectories in the context of mobile crowd sensing: application to air quality monitoring. GeoInformatica, 2024, 28(2): 177-220

[3] Kim M, Kim Y. Multi-blockchain structure for a crowdsensing-based smart parking system. Future Internet, 2020, 12(5): 90

[4] Jiang Wei-Jin, Wang Hai-Juan, Zhou Wei, et al. A trajectory privacy protection scheme for crowdsensing based on adaptive continuous-time. Acta Electronica Sinica, 2023, 51(10): 2894-2901 (in Chinese)
(蒋伟进,王海娟,周为,等. 基于自适应连续时间的群智感知轨迹隐私保护方案. 电子学报, 2023, 51(10): 2894-2901)

[5] Budiman A, Wu W, Naula-Duchi E A, et al. Dynamic sensor nodes distribution with coordinated autonomous vehicles for environment pollution monitoring and modeling//Proceedings of the 7th International Conference of Recent Trends in Environmental Science and Engineering. Ottawa, Canada, 2023: 1-8

[6] Jiang Wei-Jin, Nie Cai-yan, Liu Qian, et al. Hybrid task allocation for mobile crowdsensing based on mobility prediction. Acta Electronica Sinica, 2025, 53(6): 1979-1995 (in Chinese)
(蒋伟进,聂彩燕,刘茜,等. 基于移动性预测的群智感知混合式任务分配. 电子学报, 2025, 53(6): 1979-1995)

[7] Petrov V, Samuylov A, Begishev V, et al. Vehicle-based relay assistance for opportunistic crowdsensing over narrowband IoT (NB-IoT). IEEE Internet of Things Journal, 2018, 5(5): 3710-3723

[8] Chen X, Xu S, Han J, et al. PAS: Prediction-based actuation system for city-scale ridesharing vehicular mobile crowdsensing. IEEE Internet of Things Journal, 2020, 7(5): 3719-3734

[9] Xu S, Chen X, Pi X, et al. ILocuS: Incentivizing vehicle mobility to optimize sensing distribution in crowd sensing. IEEE Transactions on Mobile Computing, 2020, 19(8): 1831-1847

[10] Tang B, Li Z, Han K. Multi-agent reinforcement learning for mobile crowdsensing systems with dedicated vehicles on road networks//Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference. Indianapolis, USA, 2021: 3584-3589

[11] Fan G, Jin H, Liu Q, et al. Joint scheduling and incentive mechanism for spatio-temporal vehicular crowd sensing. IEEE Transactions on Mobile Computing, 2021, 20(4): 1449-1464

[12] Yin Z, Wang Y, Liu L, et al. Maximizing coverage quality with budget constrained in mobile crowd-sensing network for environmental monitoring applications. Sensors, 2019, 19(10): 2399

[13] Azhar S, Chang S, Liu Y, et al. Utility-aware participant selection with budget constraints for mobile crowd sensing//Proceedings of the 15th EAI International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, Shenzhen, China, 2020: 27-39

[14] Liu Y, Yu Z, Zhang F, et al. Wisdom of crowds: a human-machine-things cooperative scheduling method for heterogeneous mobile crowdsensing. Proceedings of the ACM on Human-Computer Interaction, 2024, 8(CSCW2): 419:1-419:25

[15] Xiao L, Chen T, Xie C, et al. Mobile crowdsensing games in vehicular networks. IEEE Transactions on Vehicular Technology, 2018, 67(2): 1535-1545

[16] Ma J, Yu Y, Liu G, et al. Freshness-aware vehicular crowdsensing with multi-agent reinforcement learning. Computer Networks, 2025, 257: 110978

[17] Lu W, Zhu Z, Zhao Y, et al. A multi-agent reinforcement learning approach for cooperative air-ground-human crowdsensing in emergency rescue. arXiv preprint arXiv:2505.06997, 2025

[18] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden, 2018: 2085-2087

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

[20] Foerster J N, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2974-2982

[21] Wang J, Ren Z, Liu T, et al. QPLEX: duplex dueling multi-agent Q-learning//Proceedings of the 9th International Conference on Learning Representations. Virtual, 2021: 1-18

[22] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017

[23] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6379-6390

[24] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 24611-24624

[25] Kuba J G, Chen R, Wen M, et al. Trust region policy optimisation in multi-agent reinforcement learning//Proceedings of the 10th International Conference on Learning Representations. Virtual, 2022: 1-23

[26] Li C, Wang T, Wu C, et al. Celebrating diversity in shared multi-agent reinforcement learning//Proceedings of the 35th Conference on Neural Information Processing Systems. Virtual, 2021: 3991-4002

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

[28] O'Keeffe K P, Anjomshoa A, Strogatz S H, et al. Quantifying the sensing power of vehicle fleets. Proceedings of the National Academy of Sciences, 2019, 116(26): 12752-12757

[29] Ding Shi-Fei, Du Wei, Zhang Jian, et al. Research progress on multi-agent deep reinforcement learning. Chinese Journal of Computers, 2024, 47(7): 1547-1567 (in Chinese)
(丁世飞,杜威,张健,等. 多智能体深度强化学习研究进展. 计算机学报, 2024, 47(7): 1547-1567)

[30] Shi P, Yan B. A survey on intelligent control for multiagent systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(1): 161-175

[31] Abdulghani A M, Abdulghani M M, Walters W L, et al. Multi-agent reinforcement learning system using multi-agent proximal policy optimizer algorithm in SMAC environment//Proceedings of the 2023 Congress in Computer Science, Computer Engineering, & Applied Computing. Las Vegas, USA, 2023: 357-360

[32] Dong C Z, Che R, Tang C M. Design of cloud task scheduling simulation software based on multi-agent deep reinforcement learning//Proceedings of the 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference. Chengdu, China, 2024: 650-655

[33] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013

[34] Fan S, Chen X. Application of attention-based multi-agent deep reinforcement learning in team sports collaboration//Proceedings of the 2025 International Conference on Artificial Intelligence and Digital Ethics. Guangzhou, China, 2025: 252-255

[35] Gu S, Kuba J G, Chen Y P, et al. Safe multi-agent reinforcement learning for multi-robot control. Artificial Intelligence, 2023, 319: 103905

[36] Zhang Y, Yue M, Wang J, et al. Multi-agent graph-attention deep reinforcement learning for post-contingency grid emergency voltage control. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3340-3350

[37] Wang H, Liu C, Yang H, et al. Ensuring threshold AoI for UAV-assisted mobile crowdsensing by multi-agent deep reinforcement learning with transformer. IEEE/ACM Transactions on Networking, 2024, 32(1): 566-581

[38] Chu T, Wang J, Codecà L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(3): 1086-1095

[39] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839

[40] Lee N H, Uhm T Y, Park J H, et al. Reinforcement learning based sequential multi-robot task allocation considering weight of objects and payload of robots//Proceedings of the 23rd International Conference on Control, Automation and Systems. Yeosu, Republic of Korea, 2023: 1858-1861

[41] Khanzhahi N, Masoumi B, Karasfi B. Deep reinforcement learning issues and approaches for the multi-agent centric problems//Proceedings of the 2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium. Kish Island, Iran, 2018: 87-95

[42] Li Y, Zhang X, Zeng T, et al. Task placement and resource allocation for edge machine learning: a GNN-based multi-agent reinforcement learning paradigm. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(12): 3073-3089

[43] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv: 1706.05296, 2017

[44] Wang J, Zhang Y, Kim T K, et al. Shapley Q-value: a local reward approach to solve global reward games//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 7285-7292

[45] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised dee

p learning results//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 1195-1204

[46] Morales-Brotons D, Vogels T, Hendrikx H. Exponential moving average of weights in deep learning: dynamics and benefits. arXiv preprint arXiv:2411.18704, 2024

[47] Krishna V, Dommari S, Tiwari S, et al. Multi-agent deep reinforcement learning for adaptive traffic signal control in smart cities//Proceedings of the 2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity. Bhubaneswar, India, 2025: 1-6

[48] Starace G, Piro G, Grieco L A. Vehicular crowdsensing with high-mileage vehicles: investigating spatiotemporal coverage dynamics in historical cities with complex urban road networks. Journal of Advanced Transportation, 2023, 2023: 8668473

[49] He H, Peng J, Cai L, et al. Air-ground collaborative mobile crowdsensing by predictive multi-agent deep reinforcement learning//Proceedings of the 2025 IEEE 102nd Vehicular Technology Conference. Chengdu, China, 2025: 1-6

[50] Luo G, Zhan S, Liang C, et al. An incentive mechanism for joint sensing and communication vehicular crowdsensing by deep reinforcement learning. Computer Networks, 2025, 260: 111099

[51] Jaimes L G, Chintakunta H, Abedin P. SenseNow: a time-dependent incentive approach for vehicular crowdsensing. IEEE Open Journal of Intelligent Transportation Systems, 2024, 5: 307-321

[52] Xu C, Song W. Decentralized task assignment for mobile crowdsensing with multi-agent deep reinforcement learning. IEEE Internet of Things Journal, 2023, 10(18): 16564-16578

[53] Deng X, Liu W, Fu Y, et al. Heterogeneous multi-agent collaboration in UAV-assisted mobile crowdsensing networks. arXiv preprint arXiv:2509.25261, 2025

[54] Monon J S, Dutta Barua D, Khan M M. Enhancing heterogeneous multi-agent cooperation in decentralized MARL via GNN-driven intrinsic rewards. arXiv preprint arXiv:2408.06503, 2024

[55] Wang Y, Gao H, Ngai E, et al. A coverage-aware high-quality sensing data collection method in mobile crowd sensing. IEEE Transactions on Mobile Computing, 2025, 24(4): 3025-3040

[56] Wu E, Peng Z. Research progress on incentive mechanisms in mobile crowdsensing. IEEE Internet of Things Journal, 2024, 11(14): 24621-24633

[57] Jiang Wei-Jin, Yang Xuan, Li Bi-Xia. Federated learning poisoning attack defense method based on interpretable contribution anomaly detection and dynamic pruning. Chinese Journal of Computers, 2025, 48(12): 2855-2874 (in Chinese)
(蒋伟进, 杨璇, 李碧霞. 基于可解释贡献异常检测与动态裁剪的联邦学习投毒攻击防御方法. 计算机学报, 2025, 48(12): 2855-2874)



JIANG Wei-Jin,Ph. D., Professor. His main research interests include crowd sensing, multi-agent reinforcement learning and cyberspace security.

OUYANG Yu-Ming, M.S. candidate. His main research interests are crowd sensing and multi-agent reinforcement learning.

SHI Shun-Jiu, M.S. candidate. Her main research interests are resource scheduling and crowd sensing.

XU Yu-Sheng, Ph.D. His main research interests are crowd sensing, multi-agent reinforcement learning, and system control.

LIU Xi-Xi, M.S. candidate. Her main research interests are crowd sensing and federated learning.

CUI Xin-Yu, M.S. candidate. Her main research interests are federated learning and crowd sensing.

LUO Zhen-Yan, Ph.D. His main research interests are crowd sensing and cyberspace security.

JIANG Yi-Rong, Ph.D. His main research interests are crowd sensing and digital music.