arXiv:2602.16062v1 [eess.SY] 17 Feb 2026

利用隐式合作：面向去中心化本地能源市场的多智能体强化学习方法

NelsonSalazar-Peña a、lejandra Tabares b 和 Andrés González-Mancera a,*

a 安第斯大学机械工程系，哥伦比亚波哥大 111711；b 安第斯大学工业工程系，哥伦比亚波哥大 111711

摘要

分布式能源资源的快速集成使集中式管理承压，因而需要在不损害隐私的前提下确保物理稳定性的去中心化控制范式。本文提出隐式合作，这一框架使去中心化智能体能够在无需显式点对点通信的情况下，于本地能源市场中逼近最优协调。我们将该问题表述为去中心化部分可观测马尔可夫决策问题，并通过多智能体强化学习任务加以求解；其中，智能体利用信息素信号（系统层面的关键绩效指标）来推断全局状态并作出响应。通过在IEEE 34节点拓扑上开展的 $_3 \times _3$ 析因设计，我们评估了三种训练范式（集中式训练与集中式执行、CTDE、DTDE）和三种算法（近端策略优化、异步近端策略优化、软演员-评论家）。结果表明，APPO-DTDE是最优配置，其协调得分达到理论集中式基准（集中式训练与集中式执行）的91.7%。然而，效率与稳定性之间出现了关键权衡：集中式基准以0.6的点对点交易比率实现了配置效率最大化，而完全去中心化方法（DTDE）则展现出更优的物理稳定性。具体而言，与混合架构相比，DTDE将电网平衡的方差降低了31%，从而形成高度可预测、偏进口的负荷曲线，简化了电网调节。此外，拓扑分析揭示了涌现的空间聚类，去中心化智能体自组织为稳定的交易社区，以最小化拥塞惩罚。尽管软演员-评论家在混合设置中表现优异，但由于熵驱动的不稳定性，它在分散式环境中未能成功。本研究证明，信息素信号机制为复杂的电网协调提供了充分的上下文，从而为昂贵的集中式通信基础设施提供了一种稳健且保护隐私的替代方案。

关键词隐式合作、多智能体强化学习、去中心化能源市场、分布式能源资源、本地能源市场

*通讯作者**电子邮箱：**na.salazar10@uniandes.edu.co（Nelson Salazar-Peña），a.tabaresp@uniandes.edu.co（Alejandra Tabares），angonzal@uniandes.edu.co（Andrés González-Mancera）

1 引言

1.1 研究动机

能源格局正在经历结构性转型，其驱动力来自脱碳、数字化和去中心化 [1]。这一转变的特征是分布式能源资源（DERs）的大量涌现，例如太阳能光伏、电池储能系统和电动汽车，它们正在将被动的消费者转变为主动的产消者 [1]。尽管这一转变有望提升电网韧性并减少碳排放，但它从根本上改变了传统的电网管理。

集中式控制范式适用于调度数量有限的大型发电机，但在尝试协调数千个地理位置分散、间歇运行的终端时，会面临难以处理的计算复杂性和单点故障风险 [2]。因此，本地能源市场（LEM）应运而生，成为管理这一复杂性的运行框架，能够实现能源和灵活性服务的去中心化交易 [3]。

然而，要成功实施本地能源市场，仍面临诸多挑战：需要实现计算可扩展性以管理多个智能体，保障数据隐私以维护智能体自主性，并平衡电网中的供需 [1]。传统解决方案无法同时满足这三个条件。例如，集中式优化在可扩展性和隐私方面表现不佳，而点对点（P2P）交易虽能提升隐私保护，却因通信开销随规模呈二次增长而面临可扩展性障碍 [1]。

1.2 问题陈述

尽管去中心化协调的必要性已得到充分认可，但当前最新研究进展中的方法在应用于真实世界能源系统时仍存在缺陷。主要问题在于，当前多智能体学习在能源市场中默认依赖集中式训练 [4]。多智能体深度确定性策略梯度（MADDPG）等算法虽然成功展示了仿真环境中的协调能力，却在训练阶段违背了隐私和去中心化约束。集中式训练需要一个能够访问全局状态的集中式评论家，其中包括所有智能体的私有成本函数、电池状态和偏好，以指导学习过程 [5]。

相反，去中心化训练遵循这些隐私约束，但会受到非平稳性的影响。由于所有智能体同时学习，环境对任一单个智能体而言都显得不可预测，从而导致学习不稳定性，并收敛到次优均衡 [6]。此外，现有隐式协调模型往往仅依赖价格信号，这可能诱发价格波动和负荷同步等系统性不稳定，从而威胁物理电网安全 [7]。

因此，目前仍缺乏一种框架，能够使完全去中心化的智能体学习用于能源平衡的稳定合作策略，而无需集中式训练数据，也无需依赖会造成不稳定的价格信号。

1.3 研究空白

本文所要解决的主要挑战是，在去中心化电网中通过隐式合作实现能源供需平衡，即不依赖集中式调度或显式通信。与传统的集中式控制不同——后者依赖单一实体对所有资源作出决策或进行直接协商——隐式合作要求自利型智能体在独立运行的同时学会协同工作，并通过智能体对共享环境信号的响应来实现系统整体协调性 [8]。该方法通过解耦决策并同时维持电网平衡，提供了一种潜在解决方案。随后，我们在本研究中建立隐式合作的理论基础，并将其提出为本地能源市场所必需的协调模型；在该模型中，智能体必须协同合作以维持电网平衡，同时不损害隐私或自主性。

本文基于我们此前的研究 [9],，该研究提出了一个仿真框架，用于研究本地能源市场中的多智能体交互，并将模块化市场机制与真实的物理网络约束（如能量流、拥塞）相结合，以利用学习型智能体验证隐式合作假设。这一隐式合作挑战带来了一系列约束，使其有别于传统的多智能体控制问题：

- **通信约束：** 智能体必须在没有显式双向通信的情况下协调其动作。协调只能通过观察共享环境信号（例如智能体声誉或电网拥塞指标）来实现，以保护隐私并确保可扩展性。

- **目标冲突：** 系统中充满了追求自身目标函数最大化的自利型智能体。挑战在于设计激励结构和信息反馈回路，使个体目标与电网平衡（供需平衡）这一系统级目标相一致，从而让智能体学会兼顾两者的策略。

• **非平稳性：** 由于其他智能体也在同时学习和调整，LEM环境具有非平稳性。一个智能体的最优策略取决于其他智能体的策略，由此产生移动目标问题：智能体必须适应不断变化的对手策略，同时维持协调。

• **部分可观测性：** 智能体在运行时只能获得有关系统整体状态的有限信息。它们可以观测自身的局部状态以及一些系统级信号，但无法直接观测其他智能体的状态或策略。这种部分可观测性限制了智能体理解自身动作对整个系统影响的能力。

• **连续空间：** 与离散动作选择问题不同，能源市场协调需要对价格和数量进行连续决策。智能体必须在连续动作空间中学习策略，因此需要采用复杂的策略梯度方法和探索策略。

• **不确定性与随机性：** 系统必须考虑各种不确定性来源，包括发电量和需求的预测误差、随机市场条件以及技术约束。智能体必须学习能够在不确定性下表现良好的鲁棒策略。

1.4 创新与贡献

为解决这一挑战，我们采用多智能体强化学习（MARL）作为主要研究方法，并分析了有助于深入理解协调机制的互补研究：

1. **训练范式：** 不同的多智能体强化学习训练范式，具体包括集中式训练与集中式执行（CTCE）、集中式训练与去中心化执行（CTDE），以及分散式训练与分散式执行（DTDE），如何影响隐式合作的形成与稳定性。本研究探讨协调质量与部署可行性之间的权衡，分析有效协调是否必须依赖集中式训练，或完全分散式学习能否取得可比结果。 2. **算法有效性：** 哪些多智能体强化学习算法最适合在去中心化能源市场中学习合作策略。本研究探讨促进或阻碍协调的算法特定特征，为实际部署中的算法选择提供指导。 3. **机制设计：** 将系统层面的关键绩效指标（KPI）纳入观测空间和奖励函数，如何实现隐式协调。本研究深入分析隐式合作的机制，理解信息结构与激励设计如何为协调的形成创造条件。 4. **可扩展性与鲁棒性：** 随着智能体数量增加，隐式合作如何扩展。本研究关注实际部署中的相关问题，确保协调机制能够超越实验设置并有效运行。

本研究提出，可通过设计观测空间并纳入印迹式环境信号，解决完全去中心化训练（DTDE）中的非平稳性问题。具体而言，我们将系统级关键绩效指标（例如社会福利、电网拥塞水平和声誉分数）直接嵌入智能体的局部观测中，构建一个反馈回路，作为集中式控制器的替代机制。这样一来，智能体便能将自身对电网的影响纳入考量，并学习隐式平衡电网的行为。本研究的贡献如下：

• 我们证明，采用去中心化实现方式（DTDE），并结合设计的观测空间和奖励塑形，其性能可接近集中式训练（CTCE/CTDE），从而验证了一条保护隐私的实际部署路径。

• 我们提出并验证了一种机制，将声誉分数和系统级关键绩效指标用作连续状态变量。我们展示了这些非价格信号如何缓解标准交易型能源模型中的同步风险，使智能体能够学习稳定的负载均衡策略，同时遵守电网物理约束。

• 我们在一个高保真度仿真环境中，对不同训练范式和多智能体强化学习算法配置进行了系统评估。该仿真涉及具有连续动作空间的中等规模智能体。本分析确定了在竞争性能源市场中促进涌现合作的最佳算法特征。

1.5 论文结构

本文从去中心化协调的理论基础出发，逐步推进至所提出的隐式合作模型的计算验证。本文其余部分结构如下：第 2 节回顾能源系统中协调演进的最新研究进展，分析本地能源市场中显式控制的局限性

以及多智能体强化学习的潜力，并明确隐式合作与隐私方面的具体研究空白。第 3 节详细阐述方法论，包括对稀缺信息共享约束建模的数学基础、隐式合作的概念架构——将系统级关键绩效指标纳入观测空间和奖励函数，使其发挥信息素信号的作用——三种训练范式（集中式训练与集中式执行、CTDE 和 DTDE）的理论分析，以及多智能体强化学习算法的数学公式，为比较这些算法奠定基础。第 4 节概述实验设置，介绍中等规模智能体的框架配置、电网拓扑和市场规则，以及用于评估解空间的具体指标。第 5 节分析实验结果，比较各训练范式促进协调的效果，评估特定算法在连续控制任务中的鲁棒性，并分析智能体的涌现行为。本节提供证据，揭示协调质量与部署可行性之间的权衡，并验证完全去中心化的智能体能否通过隐式机制逼近集中式基线的性能。最后，第 6 节综合研究发现并总结全文，概述隐式合作可行性方面的贡献，为设计保护隐私的能源市场提供实践指南，并指出未来将这些机制扩展至规模更大、异质性更强的电网环境的研究方向。

2 文献综述

2.1 隐式合作与涌现式协调

电力系统控制的发展，源于电网从集中式层级结构向由自主分布式能源资源构成的分布式网络转型，这从根本上改变了电网的控制理念 [1, 10]。传统的集中式控制虽然理论上能够实现全局优化，但在应用于数千个自主终端时，会面临难以克服的计算复杂性（可扩展性有限），并造成单点故障 [2, 11]。因此，研究格局已转向去中心化范式，将控制智能分布到电网边缘 [12]。然而，必须对这些去中心化智能体的协调方式作出关键区分：它们究竟是通过显式通信信道（涉及直接协商和消息传递）进行协调，还是通过隐式通信信道，依靠环境观测涌现出协调。

2.1.1 协调分类法：显式控制的局限

为了解本地能源市场格局，我们采用 [1], 提出的结构分类法，该分类法根据决策方式和信息流架构对策略进行分类。其根本区别在于直接控制与间接控制：前者由中央实体下达设定值，后者则由产消者保留自主性 [13]。

在直接控制和集中式瓶颈范式下，中央控制器会汇总所有参与智能体的数据，并计算全局最优设定值 [12]。输电系统运营商（TSO）管理的协调模式便是一个例子：中央 TSO 全权负责验证和调度分布式能源资源服务，而配电系统运营商（DSO）仅被视为数据提供方 [14]。尽管这种模式具有较高的可控性和理论上的最优性，却受到维数灾难的阻碍 [1]。随着分布式能源资源数量增加，计算负担呈指数级增长，而传输大量实时数据的要求会造成严重的隐私漏洞和通信瓶颈 [2]。相比之下，配电系统运营商管理模式将验证和调度工作交由 DSO，借助其本地网络知识，但在资产相关问题上可能引发利益冲突 [14]。混合模型则试图弥合这一差距，将预认证交由 DSO，将调度交由 TSO。

为解决这些局限性，研究转向了间接控制，并分化为中介式和双边式架构。中介式协调依赖中心辐射式拓扑，由中央聚合商在电网与产消者之间进行接口连接 [15]。尽管这通过全局优化（例如混合整数线性规划（MILP）或交替方向乘子法（ADMM））促进了价值共创，但仍存在单点故障和严重的隐私漏洞，因为智能体必须向协调方披露详细的负荷曲线 [13]。此外，求解复杂随机优化问题所需的计算负担，为实时规划造成了可扩展性瓶颈 [16]。双边协调，即点对点（P2P）交易，则试图通过利用区块链等技术实现直接的成对协商来解决这一问题 [3, 17]。这些市场通常采用双向拍卖或基于博弈论的协商等复杂机制 [18]。然而，这会带来可扩展性障碍，因为在网状网络中，通信开销会随着参与者数量的增加呈二次增长 [10]。依赖显式偏好匹配的方法通常会产生难以承受的协商延迟，使其不适用于数千项资产所需的精细化日内规划 [1]。

2.1.2 用于隐式协调的纯价格机制

控制系统的架构选择决定了本地能源市场的运行能力和故障模式。文献1）概述了从完全集中式到完全去中心化的范围（见表，每种架构都在系统层面的最优性与智能体层面的自主性之间呈现出不同的权衡。

在控制范式的一端是集中式控制架构，通常由中央控制器 [12] 进行协调。在这一范式中，中央控制器汇总所有参与智能体（如发电机、负荷和储能单元）的数据，并计算全局最优设定值，以最大限度地降低运行成本或维持稳定性 [12]。尽管理论上这种方法能够确保全局最优性和较高的可控性，但如今越来越多的人认为，对于分布式能源渗透率高的大规模系统而言，这种方法在计算上难以实现 [2]。此外，集中式架构会引入单点故障。中央控制器一旦遭到破坏或发生故障，就可能引发整个系统的连锁崩溃；而传输大量实时数据所需的通信基础设施也会造成严重瓶颈，并带来隐私漏洞 [2]。

相反，分散式控制范式将决策权分配给个体智能体或本地控制器 [12]。这种架构与分布式能源资源的物理分布特性相契合，使智能体能够依据局部测量（如电压、频率）或局部目标作出决策 [19]。去中心化的主要优势在于增强了韧性，因为系统不存在单点故障；同时还具备即插即用可扩展性，无需重新设计中央控制逻辑即可集成新资产 [2, 19]。然而，如果缺乏协调，完全去中心化的决策可能导致次优全局结果或局部不平衡 [20]。点对点控制进一步拓展了这种自主性：智能体作为对等方行动，不存在层级关系，并由此构成直接能源交易的技术基础 [12, 21]。

系统综述表明，该领域正呈现出转向去中心化协调方案的趋势 [20]。由于数据隐私问题和计算负荷，集中式优化越来越被视为在实践中难以实现。去中心化方法通常采用博弈论或分布式优化，因能够显著降低复杂度和运行成本而更受青睐，即使与理想的集中式基准相比，这些方法得出的资源分配在理论上并非最优 [20]。这一现实中的结构性趋势进一步凸显了本地能源市场需要可扩展、低通信开销的协调方法。

为了管理本地能源市场中产消者的去中心化动作，协调策略已从直接的物理指令转向利用经济信号的基于市场的机制。例如，需求响应是这一转变的基础层，通过价格信号或可靠性需求激励消费者改变消费模式 [11]。尽管基于激励的需求响应有助于削峰，但要为消费建立准确的反事实基线仍面临技术挑战，而参与者可能会利用这一点谋取策略性利益 [22, 23]。

交易型能源通过利用利润作为关键运行参数，在整个基础设施中建立供需动态平衡，从而拓展了这些原则 [24]。交易型能源使产消者能够主动买卖能源，提高效率并整合间歇性可再生能源 [25, 26]。

点对点能源交易是交易型能源最去中心化的表现形式，促成智能体之间的直接交换 [3]。这些市场通常采用双向拍卖、博弈论协商或基于优化的出清等复杂机制 [18]。然而，这些方法主要依赖显式协调：它们要求使用结构化且具有约束力的通信协议（例如提交正式的买价和卖价），并采用复杂的市场出清流程 [27]。这种对显式协商的依赖会造成大量通信开销、出清过程中的计算延迟，以及因披露产消者偏好而产生的潜在隐私问题 [3, 28]。

表 1：分布式能源资源（DERs）控制范式的比较分析。

范式	控制器	协调信号	主要优势函数	主要劣势	可扩展性	韧性
集中式控制	单一集中式控制器	直接命令与控制信号	全局优化，高可控性	单点故障，高计算负担	Low	Low
去中心化控制	个体智能体	本地测量值（电压、频率）	无单点故障，即插即用能力	全局结果次优，系统不平衡	High	High
交易型能源	个体智能体	经济信号（价格、买价、卖价）	促进产消者自主性与经济效率	通信基础设施，潜在的市场力量滥用	High	High

2.1.3 理论基础：从显式协商到柱头式协作

仅价格信号带来的挑战要求我们转向更鲁棒的隐式协调形式，而这建立在柱头式协作的概念之上。柱头式协作描述了一种间接协调机制：智能体采取的行动会在环境中留下痕迹，进而促使同一或其他智能体执行后续行动 [29, 30, 31]。

在本地能源市场的情境中，市场平台本身充当痕量协同环境 [31]。智能体无需直接通信（分布式拓扑），也无需向中央控制器报告（集中式拓扑），而是与共享环境进行交互。智能体的行为（例如注入或提取电力）会改变共享环境变量。

这些变化充当环境标记 [30, 32]。其他智能体会感知这些变化，并据此调整其策略 [33]。这一过程使系统级秩序得以从局部规则中涌现，从而有效地将系统复杂性与个体代理的复杂性解耦 [34, 35]。

这种交互的目标是实现涌现式协调：形成一种全局属性（例如供需平衡），它并未被明确编程到任何单个智能体中，而是由智能体之间的集体交互产生 [35]。然而，涌现具有非确定性，且并非天然有益；若缺乏谨慎设计，自利型局部规则的汇总可能导致电网不稳定 [36]。因此，研究挑战转向通过基于定量标记的协同觅食来设计涌现 [30, 37]。将系统级关键绩效指标嵌入每个智能体的局部观测空间后，环境便能有效引导学习过程。通过多智能体强化学习训练的智能体可以学会将这些环境标记解读为未来奖励的预测指标，即使没有显式合作协议，也能涌现出合作策略 [38]。

2.1.4 涌现式协调与非预期行为的挑战

隐式合作的目标是实现涌现式协调：形成一种连贯的全局属性，例如供需平衡。它并未被明确编程到任何单个智能体中，而是自然地源自智能体之间的集体交互。这一原则表明，对于大规模系统，明确协调每种情境下的每一项决策既不可行，也无必要。相反，通过设计恰当的局部规则和信息结构，可以诱导出理想的系统级行为 [35]。然而，要实现有益的涌现，必须应对复杂多变的协调架构；显式控制的局限性凸显了采用隐式机制的必要性。

在中介式协调架构中，由中央实体管理系统，计算负担会变得难以承受。 [39] 通过一种连接批发市场和本地市场的随机优化框架对此进行了说明；尽管该框架在数学上具有鲁棒性，但通过蒙特卡洛模拟管理不确定性的需求会导致场景数量激增，使集中式优化问题对于实时大规模应用而言在计算上难以处理。同样，[40] 强调了拉格朗日分解法固有的延迟问题。尽管这些迭代定价机制理论上能够实现收敛，但智能体计算并上传响应所需的通信延迟，加之迭代交换最终会暴露私有效用函数这一隐私挑战，使其不适用于现代电网所要求的精细化响应。

通过双边协调（点对点）实现控制去中心化，虽然解决了单点故障问题，却引入了严重的通信开销。 [3] 提出了一种受网络约束的点对点方案，利用灵敏度系数（电压灵敏度系数和功率转移分布因子）依据电网物理特性对交易进行验证。尽管该方案成功整合了物理约束，但验证每笔双边交易的要求造成了新的计算瓶颈。此外，协商过程本身也会引发显著的延迟。 [41] 在一个博弈论模型中展示了这一点，其中买方和卖方参与嵌套式演化博弈与斯塔克尔伯格博弈。这些迭代算法收敛到稳定策略所需的时间，使其不适用于高频日内规划。即使机制发展到纳入异质偏好，例如 [42], 中智能体依据声誉和位置等属性进行交易，对显式匹配算法的依赖仍然带来了沉重的通信负担，其协商复杂性与之相当。

因此，研究挑战从显式统筹转向设计隐式协调。在这种模式下，智能体不会进行协商，只会观察并作出反应。然而，文献提醒我们，未加定义的涌现具有非确定性，而且并非天然有益。如果缺乏精心设计，自利型局部规则的汇总可能导致电网不稳定。如果数千个智能体在没有协调逻辑的情况下，独立地对简单的广播信号（如低价）作出反应，它们可能会同步动作，形成新的涌现峰值，进而破坏网络稳定性 [35, 36]。

为缓解这一问题，必须设计环境的信息结构，使局部最优动作与全局稳定性相一致。 [13] 利用多智能体强化学习在这一领域取得了突破。其框架表明，通过使用边际奖励（用于量化个体对全局状态具体贡献的信号）训练智能体，智能体能够学会使自身行为去同步化。通过Q学习，智能体学会将环境标记解读为集体系统状态的预测指标，而不只是即时激励，从而学会在电网拥堵时主动等待或减少消费 [38]。

该领域的前沿方向，是将观测空间扩展到简单的价格信号之外，纳入显式模型所识别的丰富非价格属性。[3] 识别出的物理灵敏度，以及 [42] 采用的声誉与偏好指标，必须从验证工具转化为被动的环境信号。将这些系统级目标嵌入局部观测空间后，环境便能有效引导学习过程，使复杂的协作策略得以涌现，同时避免直接通信带来的可扩展性约束。

2.1.5 基于KPI的协调与声誉在能源市场中的作用

在本地能源市场中实现定量共识主动性的一种主要方式，是使用关键绩效指标和声誉指标。从显式协商转向隐式协调，需要一种特定机制来充当环境信号。虽然价格信号已被证明不稳定，但声誉是一种鲁棒的参数，可将历史表现编码为环境标记 [30]。通过将可信度转化为具体、以数据为基础的衡量指标，声誉系统使智能体能够隐式协调；它们并非通过直接指令，而是通过追求更高的声誉以及由此带来的经济优势，与系统目标保持一致。

在多智能体系统中，有必要区分声誉与信任。信任通常被定义为通过直接成对交互形成的一种主观、私人的信念 [43]。早期模型依赖主观推荐（例如，智能体彼此评价），但这种方法容易受到偏差和串谋的影响 [44]。相反，声誉是对智能体地位的一种集体性、公开衡量，源自可观察到的历史表现 [45]。对于本地能源市场，鲁棒的范式依赖于基于定量标记的协调，其中声誉是根据客观关键绩效指标计算得出的、可验证的事实 [30]。这种方法与纳入社交网络拓扑的研究相契合，在这类研究中，智能体的相对电气距离等因素会影响其声誉，而不依赖于直接反馈 [45]。

要使声誉成为有效的协调信号，就必须通过多维指标对其进行量化，以捕捉智能体行为的不同方面，涵盖从物理可靠性到网络安全的各个维度。例如，承诺分数通过追踪已成交市场计划与实际功率交付之间的偏差来衡量智能体的可靠性 [46]。该分数根据这些偏差的移动平均值计算得出，可提供物理履约情况的直接历史记录，用以判断产消者是否会实际提供其承诺的灵活性。此外，可信性得分通过对信息物理数据（如通信数据包元数据和功率数据）进行异常检测，评估智能体的完整性，从而防范节点遭到入侵 [46]。

声誉也可以直接与电网约束相关联。例如，电压影响得分会对导致局部电压升高的产消者进行惩罚，同时奖励那些缓解电压升高的产消者，从而有效地将电网物理特性内化到社会地位之中[47]。同样，系统关键绩效指标（如电网拥塞水平）可作为可观测信号 [47]。文献提出了诸如韧性得分 [46], 等整体性指标，该得分汇总了承诺得分与可信性得分；或R360得分，其综合了服务质量、防御能力与资源可用性 [48]。

一旦量化，这些得分便从被动指标转变为主动控制变量，从而在隐式协调中形成闭环。在去中心化点对点市场中，它们充当伙伴选择的过滤器，智能体更倾向于与高声誉对等方进行交易以降低风险 [47]。这有效地边缘化了不可靠的智能体，并防止价格上的逐底竞争。先进模型甚至提出声誉调整定价，使高声誉智能体获得优惠费率，从而为电网友好行为创造直接的经济激励 [49]。

对于多智能体强化学习应用而言，更重要的是，这些指标作为状态变量，能够消解环境的非平稳性。在完全去中心化设置中，其他智能体是环境的一部分，因此环境难以预测。通过将声誉分数和系统关键绩效指标纳入观测空间，智能体能够学会将这些高层级信号与成功交易或电网约束违例的概率相关联，从而使环境更具可预测性。

然而，依赖声誉也会带来新的脆弱性。系统抵御策略性操纵的鲁棒性仍是一项主要研究挑战。文献指出了剧本攻击等风险：智能体通过大量低价值交易积累高声誉，随后利用这种信任，在一笔高价值交易中大举背叛 [50, 51]。承诺分数等简单的移动平均指标容易受到此类攻击。因此，鲁棒的指标设计必须根据动作的影响和关键程度加权，而不能只考虑其频率。

此外，智能体可能串谋以人为抬高分数；系统也可能出现自发的集中化现象，即少数智能体因高声誉带来更多互动机会而主导市场 [50]。具有韧性的设计必须在效率与多样性之间取得平衡，例如通过激励与新智能体或中层级智能体互动，来增强系统级韧性。

当前分析证实，尽管声誉为隐式协调提供了必要的信任层，但简单的实现并不足够。最先进水平的不足在于，缺乏通过信息素机制专门解决实时供需平衡问题的隐式协调模型。现有方法大多将声誉局限于显式交易或故障检测中的筛选机制，而未将其用作驱动协同负载均衡行为的主要环境信号 [24]。此外，目前缺乏系统性比较，尚不清楚不同训练范式（集中式与去中心化）会如何影响智能体在共享关键绩效指标信号驱动下涌现的这种隐式合作。

2.2 能源系统中的多智能体强化学习

现代电力系统去中心化所推动的结构性转型，使传统的集中式优化范式难以为继，主要原因是它们无法有效应对分布式能源资源的随机性和规模[52, 53]。随着可再生能源渗透率不断提高，电网正从被动的配电网转变为由产消者、微电网和聚合商构成的主动生态系统，因此需要鲁棒的本地能源市场，以促进点对点交易和社区交易 [53]。

在此背景下，多智能体强化学习已成为一种关键的计算框架。与凸优化或动态规划等集中式优化方法不同，这些方法难以应对非线性动力学，也难以承担协调数千个异构智能体所需的计算负担；多智能体强化学习则使去中心化智能体能够通过与环境交互来学习最优策略 [53]。然而，将多智能体强化学习应用于本地能源市场会带来复杂的挑战，尤其是在不确定性下的协调方面。环境的非平稳性——即一个智能体的最优策略取决于其他智能体不断变化的策略——要求采用超越简单独立学习的复杂学习范式 [53]。此外，电网的物理约束以及市场参与者的隐私要求对信息共享施加了严格限制，因此需要采用隐式合作模型，在不直接通信的情况下实现协调 [54]。

2.2.1 从独立学习到集中式训练

强化学习在能源系统中的应用随着电网本身的复杂化而不断演进。早期研究主要采用独立学习方法，例如独立Q学习 [1]。在这一范式中，每个智能体（例如家庭能源管理系统或电池控制器）都将其他智能体仅视为环境的一部分。尽管独立学习在计算上效率较高，且能够随智能体数量线性扩展，但它从根本上违背了收敛所需的马尔可夫性质：当智能体同时更新策略以最大化局部奖励时，从任何单个智能体的角度来看，环境都会变得非平稳 [53]。

这种非平稳性会导致不理想的行为：智能体追逐不断变化的目标，进而引发策略振荡。例如，如果多个产消者都学会在价格飙升时放电，集体行动可能会消除套利机会，使它们刚学到的策略失效 [53]。此外，部分可观测环境中的独立学习者往往会收敛到次优的纳什均衡，无法实现全局帕累托最优，甚至可能引发电网不稳定问题，例如电压振荡 [1]。此外，在合作场景中，如果没有将个体贡献分离的机制，就很难确定哪个智能体的行动促成了全局奖励（例如电网平衡） [53]。

为缓解这些不稳定性，同时避免退回到难以求解的集中式控制，研究界逐渐形成了 CTDE 范式 [55]。CTDE 承认该问题具有双重特性：训练时可以获取全局状态信息，而执行时则必须仅依赖局部观测。这些信息用于训练集中式评论家，随后由其指导去中心化演员更新；这些演员执行时仅使用局部观测 [6]。

诸如多智能体深度确定性策略梯度等著名算法使用集中式评论家，基于联合状态与动作空间估计Q值，从而稳定分散式行动者的学习过程 [1, 56]。多智能体深度确定性策略梯度采用集中式评论家，以联合动作与状态空间作为输入来估计Q值，而执行者保持去中心化。从根本上说，集中式评论家是一种仅用于训练的神经网络，能够访问完整的全局状态和联合动作，而Q值量化了在给定状态下采取特定动作的预期长期回报。研究表明，多智能体深度确定性策略梯度在连续动作空间（对能源竞价至关重要）中具有有效性，表明它比独立学习者更能有效地平缓社区需求曲线并降低成本 [5]。

虽然多智能体深度确定性策略梯度属于异策略（即它使用由智能体旧版本或完全不同的探索策略收集的存储数据来更新其策略），但多智能体近端策略优化将同策略PPO算法应用于多智能体场景，要求智能体根据其在遵循同一策略时所采取的动作来更新当前策略 [57]。这一差异会影响收敛。工业制造和电网控制领域的近期比较研究表明，与多智能体深度确定性策略梯度相比，多智能体近端策略优化通常表现出更优的稳定性和超参数鲁棒性，尤其是在异策略学习可能遭受分布偏移的高维环境中 [58]。

同样，诸如QMIX之类的价值分解方法已成为离散控制任务的标准方法；QMIX通过将全局团队奖励分解为各个智能体贡献来解决多智能体信用分配问题。它们通过将全局团队奖励分解为各个智能体效用来发挥作用，从而确保局部价值的最大化能够带来联合价值的最大化 [5]。

然而，标准的 CTDE 框架在本地能源市场中面临数据隐私和稀缺信息共享方面的挑战。集中式评论家的要求意味着，在训练期间，所有智能体都必须将其私有状态（例如电池电量、消费模式、成本函数）共享给中央学习器 [24]。这与许多

放松管制的本地能源市场以及智能体可能相互竞争或归不同实体所有的实际运行情况相矛盾，因此有必要探索隐式合作模型，在这些模型中，即使在训练阶段也能最大限度地减少全局信息交换 [1]。

2.2.2 多智能体强化学习的训练范式

尽管文献综述中首选 CTDE，但批判性分析表明，将其应用于隐私敏感本地能源市场时存在显著局限性。训练范式的选择决定了智能体之间的信息流，并限制了可实现的协调类型（显式或隐式）。对三种主要范式进行系统比较后，可以发现本研究旨在解决的具体空白：

- **CTCE：** 该范式将多智能体问题视为具有联合动作空间的单智能体问题。中央控制器（即单个神经网络）接收包含所有智能体全部私有信息的完整状态，并优化联合策略（即每个智能体的动作） [55]。尽管这可作为性能的理论上界，但其应用于本地能源市场时面临的挑战在于可扩展性瓶颈（智能体数量增加时，规模呈指数级增长），以及隐私侵犯问题，因为产消者越来越不愿意向中央机构分享详细的用电数据和成本函数 [55]。
- **CTDE：** 在解决执行时通信问题的同时，CTDE 假设训练期间存在丰富的信息共享。智能体拥有各自的行动者神经网络，可根据部分观测在本地执行。不过，训练期间会使用一个集中式评论家神经网络 [6]。集中式评论家的要求意味着，在学习阶段，所有智能体都必须与中央学习器共享其私有观测（例如电池状态、消费偏好、内部成本函数） [6]。不过，CTDE 允许评论家训练行动者，使其了解自身的局部动作如何影响总体奖励，从而促进隐式合作。行动者仅使用局部数据来学习近似全局价值函数。然而，对于智能体之间可能存在竞争的本地能源市场，共享私有观测以供集中式评论家使用往往难以实现 [24]。

- **DTDE：** 这种范式最符合在不假设存在全局视图的情况下实现有效训练的要求 [55]。在 DTDE 中，智能体仅根据各自的轨迹（观测、动作和奖励）更新策略。虽然这遵循去中心化部分可观测马尔可夫决策过程（Dec-POMDP）的约束（即限制智能体只能获取局部部分观测，无法访问真实的全局状态），但也重新带来了非平稳性以及在缺乏通信时进行协调的挑战 [58]。研究的前沿正在于此：开发相应机制，使 DTDE 智能体在不依赖集中式信息的情况下，实现与 CTDE 相当的协调 [1]。近期研究，例如 [5], 尝试通过仅共享策略信息而非原始数据，在 CTDE 与 DTDE 之间搭建桥梁，在提升隐私的同时优化决策。这代表着向联邦式或隐私保护型多智能体强化学习迈进 [59], ，但针对连续本地能源市场的严格比较仍然有限。

去中心化部分可观测马尔可夫决策过程框架尤其适用于本地能源市场，因为它明确考虑了隐私约束和分布式地理所导致的部分可观测性 [53]。智能体知道自身的负荷和电池状态，但无法观察其他智能体的内部状态 [60]。全局状态（网络电压、总需求）由联合动作共同决定，但任何单一实体都无法完全观测到该状态 [61]。求解去中心化部分可观测马尔可夫决策过程需要通过多智能体强化学习采用近似解，而非精确的动态规划 [61]。近期研究已将点对点交易明确建模为去中心化部分可观测马尔可夫决策过程，并利用强化学习在传统启发式方法失效的连续空间中近似求得最优策略 [62]。

2.2.3 多智能体强化学习中的隐式合作机制

在信息稀缺且显式通信受限的场景中，多智能体强化学习智能体必须通过隐式合作进行协调。这依赖于智能体学会将环境反馈解读为协调信号。电网的物理特性和市场的经济机制构成了天然的迹象媒介。电网频率或电压的偏差可作为隐式广播信号。频率下降会向所有智能体发出信号，提示它们增加发电量或削减负荷。多智能体强化学习智能体可以学会解读这些环境反馈，从而协调纠正措施（例如电压恢复），而无需显式数据交换 [63]。类似地，本地能源市场中的出清价格汇总了所有智能体的偏好和资源稀缺程度。高出清价格隐式传达了需求高或供应低的状态。研究表明，多智能体强化学习智能体可以学会利用价格波动作为协调信号，例如推断价格急剧上涨意味着邻近智能体正在放电，从而促使该智能体保留自身容量 [64]。通过对综合价格信号作出反应，智能体便能实现负荷平衡 [1]。

近期的多智能体强化学习研究探讨了在本地能源市场等混合动机博弈中，自利型智能体如何学会互惠利他主义。智能体可能会认识到，毫无节制的贪婪会导致市场崩溃或电网不稳定，并因此招致长期惩罚。通过重复交互，智能体可以发现可持续策略，例如在峰值时自愿减少消费，以

降低所有人的出清价格 [5]。为促进这些行为，通常会采用奖励塑形，调整智能体的奖励函数，使其不仅考虑个体利润，还纳入对系统的边际贡献或全局稳定性项 [65]。这种技术使个体贪婪与全球福利相协调，从而让削峰等协同行为能够通过自利学习涌现 [66, 67]。

在缺乏直接通信的情况下，智能体必须构建关于其他智能体的内部模型。 [68] 提出的深度隐式协调图等技术，使智能体能够根据对等方可观察到的动作，推断其隐藏状态或未来意图 [69]。例如，智能体可能会推断，竞争对手的激进竞价意味着其荷电状态较高，进而调整自身策略以避免价格竞争。这种推断使智能体能够基于局部观测，就系统状态达成共识 [54]。

2.2.4 本地能源市场中现有多智能体强化学习应用综述

多智能体强化学习在能源系统中的应用已拓展至应对特定市场结构和物理约束。点对点市场的研究采用共识多智能体强化学习，使智能体能够制定交易策略，在没有集中调度的情况下实现动态平衡和电压稳定性 [70]。其他研究采用由局部策略驱动的多智能体深度确定性策略梯度，使单栋建筑的需求与社区目标相协调，通过共享策略信息有效降低能源成本并平缓需求曲线 [5]。在微电网中，多智能体双延迟深度确定性策略梯度（MATD3）已用于优化储能单元的竞价行为，在市场建模方面比标准多智能体深度确定性策略梯度表现出更好的稳定性 [71]。研究者还提出了分层方法，由上层智能体（聚合商）学习协调下层智能体（家庭用户），但这类方法往往又趋向于集中式控制 [72]。

文献区分了不同规模的智能体，而这会影响学习形式。尽管许多研究聚焦于居民家庭，但中等规模智能体代表了本地能源市场中一个独特且至关重要的参与者类别。中等规模智能体指规模大于单个居民家庭、但小于公用事业级发电厂的实体，例如商业建筑和工业微电网 [73, 74, 75, 76]。与作为价格接受者的小型家庭用户不同，中等规模智能体通常拥有足够的容量，能够成为价格制定者 [73]。这会引发市场力问题：智能体的动作会直接影响市场价格的状态转移，从而加剧环境的非平稳性 [77]。例如，掌控大规模电动汽车车队的聚合商，可以通过扣留或大量涌入市场的方式显著改变出清价格。多智能体强化学习在此尤其有价值，可用于学习能够将价格影响纳入考量的策略性竞价行为，并优化交易量与价格影响力之间的权衡 [78]。

对于中等规模智能体而言，控制本质上涉及连续变量（功率数量和价格），智能体必须决定交易多少能源以及以何种价格交易。早期方法采用离散化动作的深度Q网络（深度Q网络），但会导致维数灾难和精度损失，进而造成市场出清效果不佳 [53]。在工业领域，多智能体强化学习已应用于基于部分可观测马尔可夫决策过程框架优化高能耗场景中的能源消费 [79]。对于商业建筑，多智能体强化学习智能体控制暖通空调和储能系统，以降低动态定价下的成本。值得注意的是，在这些场景中对比多智能体近端策略优化和多智能体深度确定性策略梯度的研究表明，尽管后者在连续动作操控方面表现出色，前者通常能在复杂的多模态环境中实现更鲁棒的收敛 [80]。此外，中等规模智能体运行着受到严格物理限制的关键基础设施。安全（或受约束）多智能体强化学习是一个不断发展的子领域，智能体在其中学习能够满足安全约束并最大化奖励的策略。Safe-MADDPG 等技术会将障碍函数或拉格朗日乘子纳入损失函数，以便在学习过程中强制执行约束（例如电压限制）[81]。这与无显式通信的协调有关，可确保隐式合作不会违反电网安全限制 [82]。

近期研究倾向于采用多智能体深度确定性策略梯度和软演员-评论家（SAC）等策略梯度方法，这些方法能够直接输出连续动作 [1]。近期文献对连续能源问题中的算法进行了比较分析。软演员-评论家凭借其最大熵框架日益受到青睐，该框架有助于鼓励探索；对于非平稳本地能源市场而言，这一特性至关重要，因为智能体必须不断适应竞争者不断变化的策略 [83]。研究表明，在微电网能源管理方面，软演员-评论家在鲁棒性和样本效率上优于 DDPG [84]。

尽管现有研究已付出诸多努力，但目前仍缺乏针对训练范式（CTCE、CTDE、DTDE）的系统比较，尤其缺乏对其在信息稀缺约束下促进协调的能力进行评估的研究 [58]。大多数研究默认采用 CTDE（例如 MADDPG），却未探讨训练期间共享观测所带来的隐私影响 [53]。CTDE 的最优性与 DTDE 的隐私性和可扩展性之间的权衡仍未得到充分研究 [58]。此外，大多数文献要么关注微观规模的家庭用户，要么关注宏观尺度的输电系统运营商运行。通过隐式合作实现中等规模智能体之间的协调，相关研究仍然不足 [1]。这些智能体拥有市场力，也就是说，它们的动作会直接影响价格形成，进而加剧环境的非平稳性，因此需要专门设计的多智能体强化学习建模方法（例如 Safe-MADDPG 或约束SAC），而非通用的 P2P算法 [77, 85]。此外，尽管去中心化部分可观测马尔可夫决策过程是处理部分可观测性的理论标准 [53], , 应用研究往往会将问题简化为离散

动作空间，或假设具有完全可观测性。亟需鲁棒的多智能体强化学习算法，在严谨的连续型去中心化部分可观测马尔可夫决策过程建模框架下应用于日内市场，具体而言，要通过隐式机制而非显式消息传递，在不进行直接通信的情况下实现协调 [60]。

2.3 训练范式：理论与实践

从集中式调度模型转向去中心化电网运行与控制，会引入随机性和复杂性，而传统优化技术因计算负担和隐私约束，难以应对这些问题 [1]。因此，多智能体强化学习（MARL）应运而生，成为一种使自主智能体能够在这些复杂环境中学习最优策略的框架。然而，MARL 在本地能源市场（LEM）中的部署受到信息共享和决策权限架构的限制。这种架构选择决定了系统在协调分布式能源资源（DERs）以进行能源管理时，如何权衡计算可扩展性、性能最优性和数据隐私等复杂因素 [86, 87]。文献概述了多种训练范式（集中式训练与集中式执行、CTDE 和 DTDE），各自具有不同的理论特征，并在应对这些挑战时存在不同的实际权衡（见表 2）。

2.3.1 集中式训练与集中式执行：最优性的理论基准

集中式训练与集中式执行代表了经典的自上而下方法。在这一范式中，单一且全知的控制器拥有完整的全局视图，收集每个智能体的所有私有观测、奖励和状态信息，以求解全局优化问题 [86]。在本地能源市场中，集中式训练与集中式执行作为理论性能上界基准 [86]。它通常通过混合整数线性规划（MILP）或单智能体强化学习来实现，用于量化与拥有完全信息的中央规划者相比，去中心化系统所固有的效率损失 [86, 88]。尽管集中式训练与集中式执行在理论上适用于计算社会福利上界，但文献一致认为，由于存在三项致命缺陷，它不适用于实际部署[89]:

- **计算不可解性：**联合状态-动作空间会随智能体数量呈指数级增长，这被称为维数灾难。除规模较小的问题外，这使得优化在计算上难以处理 [1]。
- **隐私侵犯：**集中式训练与集中式执行要求产消者向一家垄断机构交出精细、高分辨率的数据，例如内部消费模式、电池荷电状态以及经济效用函数。在放松管制且注重隐私的市场中，这一要求并不可取 [90]。
- **延迟与基础设施限制：**整个系统的可靠性取决于单一中央控制器能否持续、无故障地运行 [53]。这种架构十分脆弱，容易受到通信故障影响。研究表明，微电网的故障保护通常要求在 3–5 毫秒内作出响应；依赖标准遥测的集中式控制器往往无法满足这些严格的限制，从而给关键能源基础设施带来风险 [91]。

2.3.2 CTDE：稳定性的混合标准

为解决集中式执行的局限性，研究界已形成共识，将 CTDE 视为当前的学术标准 [92]。以多智能体深度确定性策略梯度和多智能体双延迟深度确定性策略梯度等算法为代表，这一范式区分了集中式训练阶段与去中心化执行阶段 [86]。

CTDE 算法通过在训练期间采用集中式评论家，解决多智能体学习中固有的非平稳性 [5]。该评论家接收所有智能体的联合观测和动作，从其视角有效地使环境保持平稳。通过观测完整的全局状态，评论家可以明确解决信用分配问题，准确判定在其他智能体行为的干扰下，某个智能体的动作如何影响全局奖励 [86]。这使分散式演员能够学习协调策略，例如在 [1], 中，智能体会隐式地错开各自的负荷曲线，以避免在储能市场中发生价格竞争。

然而，集中式评论家的神经网络架构通常具有固定的输入层维度，而该维度取决于智能体的数量。如果有新的产消者加入本地能源市场，或现有智能体退出，输入向量的大小就会改变，使训练后的模型失效。系统无法适应。因此，必须将系统下线，重新设计架构，并从头开始重新训练整个系统 [86]。此外，训练假设存在一个高保真模拟器，可在其中获取全局状态（包括产消者隐私数据）。在隐私敏感的住宅场景中，构建这样的集中式训练环境往往和集中式执行一样不可行 [86, 93]。因此，尽管 CTDE 解决了执行时间和单点故障问题，却无法满足动态能源市场所需的即插即用可扩展性和隐私保护要求。

2.3.3 DTDE：去中心化前沿

集中式范式不再适用，这迫使我们得出结论：DTDE（也称为独立学习）是实现可扩展、保护隐私且鲁棒的本地能源市场的可行架构 [94]。在 DTDE 中，智能体仅依据局部数据进行学习和行动，从而确保线性可扩展性和数据主权 [86]。

尽管其在架构上具有适用性，朴素的DTDE在实践中仍会因非平稳性问题而失败。从任一单个智能体的视角来看，环境是不一致的，因为所有其他智能体都在同时更新其策略 [86]。这种对马尔可夫性质的违反会导致信用分配失败，即智能体无法区分某项奖励（例如一笔盈利交易）是源于自身的智能竞价，还是仅仅源于其他智能体同时动作所产生的随机噪声 [86],，同时也会导致雪崩效应，即针对相似电价进行优化的独立智能体会不可避免地同步其行为。例如，在住宅能源管理中，独立学习者常常识别出相同的低价时段，并同步进行电池充电。这种集体行动会产生新的最大负荷，影响套利机会并违反网络限制 [1, 95]。

DTDE的失败源于部分可观测环境，在该环境中，智能体无法察觉其动作所带来的系统性后果。为保障DTDE的可扩展性，智能体必须进行协调。然而，显式通信并不被采纳，因为它会重新引入高带宽成本与隐私风险 [96]。

所提出的解决方案是隐式协调，即多智能体强化学习智能体可以通过扩充观测来实现协调 [86]。这种柱头式协作提供了共同的参照框架，从而有效解决环境的非平稳性。智能体不再对混乱且难以解释的价格信号做出反应，而是感知相关情境（例如，高价格与高拥塞并存）。这使合作得以仅通过本地学习涌现（例如，在拥塞程度较高时减少负荷），从而弥合 DTDE 的隐私保护与 CTDE 的协调能力之间的差距 [86]。

表2：本地能源市场中多智能体强化学习训练范式的技术分类体系。

特征	CTCE	CTDE	DTDE	Ref
架构	1对N	N对N（集中式评论家）	N对N（独立）	[96, 97, 98]
信息（训练）	全局状态	全局状态（集中式评论家）	局部观测	[5, 86]
信息（执行）	全局状态	局部观测	局部观测	[86]
协调机制	集中式优化器	集中式评论家	独立学习者	[5, 86]
优势	最优性	性能	可扩展性	[5, 86]
劣势	维数灾难	维数灾难	非平稳性	[89, 97, 98]
最优性	High	中	Low	[94, 99]
协调	High	中	Low	[1]
通信开销	High	中	Low	[94, 99]
隐私保护	Low	中	High	[53, 92, 93, 100]
可扩展性	Low	Low	High	[1, 53, 89, 93]
非平稳性处理	High	中	Low	[5, 89, 101]
权衡	计算复杂度与最优性	隐私与性能	非平稳性与可扩展性	-
强化学习算法	单智能体 DQN、近端策略优化	多智能体深度确定性策略梯度、多智能体近端策略优化、QMIX、多智能体双延迟深度确定性策略梯度	独立 Q 学习、独立深度确定性策略梯度、独立近端策略优化、迟滞 Q	[1, 6, 5, 86]
优化模型	MILP	ADMM	启发式方法/局部控制	[102, 103]
博弈论模型	斯塔克尔伯格领导者	混合深度强化学习/演化博弈论	独立智能体	[104, 105]

2.4 用于连续空间控制的多智能体强化学习算法

随着本地能源市场从集中式的自上而下调度模式演变为点对点和社区交易结构，决策范式已从静态优化转向不确定性下的动态序贯决策过程 [57]。在这一环境中，多智能体强化学习已成为一种计算框架，使自主智能体能够通过与环境及彼此交互，学习最优的竞价和调度策略。与将决策简化为有限选项（例如负荷、卸载或保持不变）的离散动作空间不同，连续控制能够精确调节有功功率、定价和能源调度。这种精细化控制对于满足日前和日内市场严格的平衡要求至关重要，因为功率注入或提取的偏差可能导致电网不稳定或电压越限。

2.4.1 优化、博弈论与强化学习的比较分析

为了论证在现代本地能源市场中选择多智能体强化学习的合理性，我们首先分析传统控制范式的局限性，具体包括基于优化的模型和博弈论模型，这些模型是电力系统分析中首选的方法。表3中给出了对这些范式的比较概述。

基于优化的模型，例如混合整数线性规划或凸优化，一直是解决调度问题的标准方法。这些模型通常假设存在一个具备完美预见能力且能够完整访问系统信息的中央实体 [3]。其主要目标通常是最大化社会福利或最小化系统总成本，并隐含地假设所有市场参与者都会采取合作行动，以实现全局最优 [106]。然而，这种系统最优方法无法捕捉放松管制的市场中的行为特征，因为市场参与者是寻求最大化个体收益而非系统效率的自利型智能体 [107, 108]。

从建模角度看，优化方法面临可扩展性与保真度方面的挑战。许多研究采用铜板假设，即将网络视为具有无限容量，从而忽略关键的网络约束，并在拥塞网络中导致可能不可行的调度决策 [3]。此外，关于需求与可再生能源发电的完美预测假设正变得越来越难以成立。虽然随机或鲁棒优化可通过纳入不确定性集合或情景来缓解这一问题，但这些扩展往往会产生计算复杂模型，并给出过于保守的解 [106, 109]。从根本上说，它们仍是静态求解器，因为它们不允许智能体根据不断变化的市场政策或不断演变的竞争对手策略实时调整自身行为 [3]。

博弈论模型（例如斯塔克尔伯格博弈或纳什均衡分析）通过显式建模竞争型智能体之间的交互，弥补了纯优化方法的战略缺口。这些方法为理解市场动态以及识别没有任何智能体有动机偏离的稳定运行点提供了理论基础 [41]。然而，它们往往受限于完全理性与完全信息等强假设，例如假定智能体掌握竞争对手成本函数与约束的准确模型 [41]。在隐私保护型本地能源市场中，此类信息很少可得。此外，在具有连续动作空间的复杂动态市场中寻找均衡会带来显著的维数挑战。这些模型的计算复杂度往往使其难以用于大规模仿真或实时控制 [41],，从而将其效用限制在离线结构分析而非运行决策。

多智能体强化学习通过实现去中心化协调，弥合了这些范式之间的差距，而无需显式建立不确定性或智能体动态的模型。与需要预先定义不确定性集合的鲁棒优化不同，多智能体强化学习智能体通过与环境重复交互，利用不确定性表征，将可再生能源需求和发电量的随机性直接融入神经网络的隐藏层 [86]。这样一来，即使面临预测误差，也能保持鲁棒性能，同时避免基于情景的树所带来的计算量激增 [86]。此外，与优化的静态特性不同，多智能体强化学习的设计本就能够应对非平稳性。在多智能体环境中，随着其他智能体更新策略，环境也会发生变化；这违反了马尔可夫性质，进而导致标准的单智能体学习失效 [110]。多智能体强化学习算法，尤其是采用 CTDE 的算法，通过允许智能体在训练期间利用全局信息学习协调策略，同时仅依据局部观测执行策略，来解决这一问题 [111]。

表3：本地能源市场中市场建模与控制方法的比较分析。

特征	优化模型	博弈论模型	MARL
目标	系统最优（例如，全球福利）	纳什均衡	自适应利润最大化
不确定性处理	鲁棒或随机	确定性	信念状态（基于经验）
市场动态	完全竞争下的静态情形	静态均衡	非平稳性与涌现性
可扩展性	高（线性/凸）	低（组合型）	高（观测与奖励塑造）
局限性	需要完美预见	计算不可解性	非平稳性与训练稳定性

2.4.2 近端策略优化

近端策略优化已成为应用于电力系统的强化学习中的基础算法。它作为信赖域策略优化在计算上更高效的替代方案被提出，在简化优化过程的同时保留了信赖域方法的优势，主要在于防止可能使学习过程失稳的破坏性策略更新。在本地能源市场中，智能体必须应对由波动的可再生能源发电和动态定价驱动的非平稳环境，近端策略优化所提供的稳定性至关重要 [57]。

近端策略优化（PPO）是一种同策略梯度方法，直接根据当前策略收集的数据优化策略。其主要创新在于裁剪代理目标函数，它限制了策略更新的幅度。这种机-制在数学上表示为一个裁剪比率，用于限制单次更新步骤中新策略相对于旧策略的偏离程度 [57]。在能源应用中，这种裁剪机制起到了调节器的作用。在电力系统中，奖励景观往往高度非凸，且遍布这样的区域：策略的微小变化（例如功率注入略有增加）就可能因约束违反而触发严重惩罚，如线路过载或电压尖峰。标准的策略梯度方-

法在这些区域有策略崩溃的风险，而 PPO 的裁剪目标函数可避免此类剧烈变化，从而确保策略单调改进 [112]。

PPO 在 LEM 中的应用已展现出相较于 MILP 等传统优化方法的优势，尤其是在优先考虑计算速度和适应性时。研究表明，基于PPO的控制器能够在经济调度策略中实现近乎最优的性能，同时大幅降低每个时间步求解复杂优化问题所需的计算负荷。例如，在整合电动汽车和光伏单元的家庭能源管理系统中，与传统方案相比，PPO 已被证明可将每日能源成本降低约 54%；它能够有效学习随机输入与最优控制动作之间复杂的非线性关系，而无需显式建立环境动力学模型 [113]。此外，在同时涉及现货能源交易和调频的联合市场中，即使存在电池储能系统，PPO 也依然有效。在这些场景下，智能体必须持续决定将多少容量分配给套利或调频，这需要持续且精确的监测。研究已证实，PPO 智能体能够学习联合竞价策略，其表现显著优于单一市场策略，从而在不确定的市场条件下最大化总利润 [114]。

近端策略优化（PPO）的一项关键比较优势，是其在训练阶段具有稳定性。与 DDPG 或 TD3 等异策略算法相比，PPO 的奖励收敛方差较低，并且能够鲁棒地缓解功率不平衡 [62]。然而，这种保守特性可能导致算法对超参数调优较为敏感。关于电力系统调度的比较研究指出，与异策略算法相比，PPO 可能需要显著提高惩罚系数，才能严格执行功率平衡约束 [62]。此外，作为一种同策略算法，PPO 的采样效率通常较低，因为它会在策略更新后丢弃经验数据。在高保真仿真计算成本高昂的本地能源市场环境中，这种低效率可能成为瓶颈 [112]。

2.4.3 异步近端策略优化

随着能源系统扩展至包含数千名产消者，训练多智能体强化学习智能体的计算需求呈指数级增长。异步近端策略优化（APPO）通过将数据收集过程与策略优化过程解耦，解决了标准近端策略优化的可扩展性限制。对于采用类似分散式部分可观测马尔可夫决策过程（Dec-POMDP）建模的本地能源市场，这种架构尤为适用；在此类市场中，智能体依据局部观测运行，并需要高效机制来探索大型状态-动作空间 [115]。

APPO 引入了一种高性能架构，可促进分布式学习。与 PPO 不同，APPO 允许多个智能体异步地与各自的环境实例交互，并将收集到的轨迹发送给中央学习器。这种异步采样会造成一定程度的策略滞后，因为行为策略可能落后于目标策略。为纠正这一问题并确保数学上的严谨性，APPO 采用重要性采样技术，具体而言是 V-trace：它估计价值函数，并根据策略间的似然比对更新加权，从而修正策略梯度 [115]。

在电力系统中，这种架构能够并行模拟多种运行条件。不同的工作进程可以同时模拟不同日期、天气模式、网络拓扑或智能体配置，显著缩短训练复杂多智能体策略所需的时间 [116]。这项能力已成功应用于网络化微电网中的自主点对点交易。通过将交易过程建模为部分可观测马尔可夫博弈，APPO 智能体能够在竞争性环境中有效运行，在优化经济目标的同时遵守输电损耗和网络约束 [117]。在这种情况下，APPO 的异步特性具有优势，因为它避免了模拟中常见的人为锁步同步，使智能体能够以略有差异的时间尺度运行，从而体现真实世界产消者的异质性 [117]。

2.4.4 软演员-评论家

软演员-评论家通过融入最大熵强化学习框架，改变了能源系统连续控制的范式。与只寻求最大化预期累积奖励的近端策略优化不同，软演员-评论家优化的是双重目标：最大化预期奖励与策略熵之和。该熵项激励智能体更深入地探索动作空间，并防止策略过早收敛到确定性且次优的行为 [118]。

软演员-评论家的显著特点是熵正则化项，其中温度参数会动态调节利用与探索之间的平衡。在本地能源市场中，这一机制对于处理最优策略的多模态特性很有价值。例如，在竞价场景中，多种有效的定价策略可能带来相近的收益；最大熵策略学习的是这些策略构成的分布，而不是将选择局限于单一策略。由于智能体保留了一组潜在的最优动作，因此能够更好地应对对手策略的变化

[119]。此外，作为一种使用经验回放缓冲区的异策略算法，软演员-评论家的采样效率显著高于近端策略优化，因此更适用于数据稀缺或仿真速度较慢的应用场景 [120]。

SAC 的理论稳定性通过截断双Q学习实现，这种方法缓解了基于价值的方法中常见的高估偏差。由此可避免乐观误差不断传播，进而导致激进且不安全的竞价行为；这对于能源调度问题至关重要 [121]。在多智能体环境中，最大熵框架为隐式合作带来了明显优势。在去中心化部分可观测马尔可夫决策过程本地能源市场中，由于其他智能体也在同时学习，环境本质上是非平稳的。确定性策略（例如 DDPG 中采用的策略）可能较为脆弱；如果邻近智能体的行为稍有变化，最优响应就可能发生巨大改变。SAC 的随机策略使交互过程更加平滑，让智能体即使在市场动态发生漂移时仍能保持其性能 [118]。

与近端策略优化直接比较时，SAC 在需要大量探索的复杂任务中通常表现更优。例如，在涉及暖通空调控制的建筑能源管理任务中，与近端策略优化相比，SAC 实现了更大的节能效果，并且更好地遵循热舒适标准，因为熵项可以防止智能体陷入局部最优 [122]。此外，在多微电网调度中，SAC 在泛化能力和收敛能力方面优于近端策略优化，有助于降低运行成本 [123]。然而，由于近端策略优化的更新较为保守且单调，在静态环境中有时能够实现略低的最终运行成本 [24]。因此，近端策略优化与 SAC 之间的选择通常取决于本地能源市场的具体约束。对于严格的物理安全和稳定性要求，近端策略优化更为合适；而对于软性经济最优性、探索以及高动态竞争环境下的鲁棒性，SAC 更为合适。表 4 总结了这些算法特性的对比。

表4：PPO/异步近端策略优化与软演员-评论家在本地能源市场连续控制中的比较分析。

特征	近端策略优化 / 异步近端策略优化	SAC	Ref
策略类型	同策略	异策略	[57, 118]
关键特性	裁剪函数	最大熵	[124]
权衡	稳定性 / 样本效率	样本效率 / 陈旧数据	-
样本效率	低（更新后丢弃数据）	高（通过经验回放缓冲区重复利用数据）	[57]
更新机制	裁剪代理目标	带熵正则化的演员-评论家	[57]
数据敏感性	对陈旧数据具有鲁棒性	易出现数据陈旧问题	[125]
方差/偏差	低偏差 / 广义优势估计*-降低方差	低方差 / 潜在的离策略偏差	[57, 126]
优势	稳定性和单调改进	探索和样本效率	[127]
本地能源市场适用性	安全关键控制（例如电压调节）	动态定价和竞争性竞价	[126]

* GAE：广义优势估计。

2.5 研究空白与贡献机遇

对最先进水平的综述揭示了未来电网的运行要求与现有协调架构的能力之间存在脱节。尽管整合分布式能源资源的物理需求显而易见，去中心化控制的理论前景也已得到充分确立，但如何在不依赖广泛通信或集中式管理机构的情况下实现电网平衡，仍是一项尚未解决的工程挑战。

2.5.1 电网平衡协调失效

已发现的主要研究空白，是缺乏一种可扩展且保护隐私的供需平衡机制，能够适应去中心化市场的非平稳性。现有方法未能解决协调方面在可扩展性、隐私和最优性上的挑战，尤其是在物理电网平衡的背景下 [1, 10]。能源平衡的主流协调策略依赖显式通信。中介式架构（聚合商）会造成单点故障和隐私瓶颈 [15], , 而双边（点对点）架构则因成对协商所需的通信开销随规模呈平方级增长而面临可扩展性障碍 [1, 10]。这些方法解决了经济出清问题，却无法克服多智能体实时物理平衡的工程挑战，因为协商延迟超过了网络动态的时间尺度 [20]。

现有用于实现能源平衡的隐式协调尝试，主要依赖动态定价（交易型能源）作为唯一的控制信号 [128]。然而，文献证实，将价格信号应用于自动化信息物理系统会导致系统不稳定性，具体表现为价格/负荷波动以及大规模负荷同步 [129, 130]。这揭示了一个

关键缺口：目前尚无成熟的工程框架，能够在不诱发羊群行为的情况下，通过隐式协调实现电网平衡。该领域缺乏一种非价格型的痕迹协调机制，能够引导智能体行为趋向物理均衡，同时避免纯金融信号带来的不稳定性。

对于中等规模智能体（商业建筑、工业微电网）而言，这一工程挑战更加严峻。与小型住宅负荷不同，这些智能体拥有市场力，因为其各自的动作会显著影响本地电网的状态和价格形成 [73]。当前文献主要关注微观规模（家庭用户）或宏观规模（输电系统运营商）的协调，因此对于中等规模智能体如何在不滥用市场力、损害电网稳定性的情况下隐式地为能源平衡作出贡献，仍缺乏深入研究 [77]。

尽管多智能体强化学习被视为模拟自适应智能体的适当计算方法，但将其应用于能源平衡问题时，在去中心化方面仍存在方法论上的不足。多智能体强化学习的大多数应用研究都基于集中训练、分散执行范式（例如，多智能体深度确定性策略梯度） [1]。这些算法虽然能够在模拟中成功学会协调，但它们绕过了隐私约束：训练期间依赖可访问全局状态（包括智能体私人数据）的集中式评论家 [5]。这并未解决真正实现隐式合作的问题，只是将集中化从执行阶段转移到了训练阶段。目前显然缺乏相关研究来验证智能体能否采用DTDE，仅依赖局部观测和经过设计的环境信号（关键绩效指标），而非全局评论家，学会实现网络平衡 [6]。

这项研究通过提出并验证一种新颖的隐式 cit 合作 模型，解决本地能源市场中的能源平衡问题。本研究有以下三项主要贡献：

- 我们提出了一种机制，通过引入多维柱头式协作信号，解决基于价格的系统中的协调不稳定性。所提出的模型不再仅依赖价格，而是将声誉评分和系统级关键绩效指标（例如拥塞指标、社会福利）作为连续的环境变量 [30]。通过将这些指标纳入智能体的观测空间，我们设计了一个反馈回路，使智能体能够将自身对电网的影响内化。这使问题从对价格做出反应转变为维护系统健康，并使稳定的能源平衡成为系统的涌现属性 [35]。
- 本研究系统评估了在稀缺信息共享的严格限制下，隐式合作的表现。通过比较CTCE、CTDE和DTDE范式，我们检验了这样一个假设：设计的观测空间（柱头式协作）能够取代集中式的关键机制。由此验证了完全去中心化的智能体能否实现与集中式基线相当的能源平衡，并证明了所提出的设计方案在隐私敏感型市场中的可行性 [5]。
- 借助 [9], 中开发的 MARLEM 仿真框架，本研究验证了物理约束环境中的协调能力。研究尤其关注中等规模智能体方面的空白，探讨如何利用多智能体强化学习进行探索所学得的连续控制策略，在最大化个体收益与实现整体电网平衡之间取得平衡 [64]。

3 方法论

本文详细介绍的仿真框架是一款全面的开源工具，旨在多智能体强化学习范式下分析本地能源市场。该框架的一项设计原则是将其实现为符合 Gymnasium 标准的多智能体环境。遵循这一标准确保了该框架与大量最新研究进展中的强化学习库和算法兼容，从而降低研究人员的入门门槛。采用这一接口后，该框架能够支持可复现实验，并可依据标准化基准测试，将新的多智能体强化学习策略与既有基线方法进行比较。

为模拟去中心化系统中固有的、信息有限且存在不确定性时的决策挑战，我们将该问题正式构建为去中心化部分可观测马尔可夫决策过程。这一形式化模型为智能体与其环境之间交互的构建提供了数学基础。该去中心化部分可观测马尔可夫决策过程由元组 $I, \mathcal{S}, A, \mathcal{P}, R, \Omega, O$定义，其中，$I$ 表示参与市场的智能体集合，$\mathcal{S}$ 表示环境状态集合，A 表示联合动作空间，$\mathcal{P}$ 表示状态转移函数，R 表示奖励函数，Ω 表示联合观测空间，O 表示观测函数。下标i 表示智能体$i \in I$对应的状态、空间或函数。完整的去中心化部分可观测马尔可夫决策过程表述详见 [9]。

在现实的 LEM环境中，任何单个智能体在任一交易时段 t都无法访问全局环境状态 $s \in \mathcal{S}$。相反，智能体接收到的局部观测 $o_i \in \Omega_i$ 与状态存在概率相关性，但无法完整揭示状态。这种部分可观测性要求智能体维护信念状态（即对可能的全局环境状态的概率分布）；与强化学习中的常见做法一样，使用神经网络将观测历史编码为隐藏状态，作为真实状态的替代表示。

为便于在部分可观测环境中进行学习，每个智能体$i \in I$ 都会接收一个局部观测向量$o_i \in \Omega_i$，该向量由全局环境状态中的结构化子集构成。它为决策提供了足够丰富的信号

同时不违背去中心化和隐私原则。观测向量（见表5）包括 ：

- **市场信号：**所有市场参与者均可获取的公开信息。这包括上一次的出清价格和交易量、关于点对点与配电系统运营商交易量的匿名统计数据、配电系统运营商当前的买入价格（上网电价）和卖出价格（效用价格），以及每个智能体的信誉评分。这些信号提供了共同的信息基础，减少不确定性，使智能体能够对市场状态形成共同预期。

- **智能体特定信号：**智能体的私有信息，其他智能体无法观测，从而保障隐私。这组信号包括其当前发电量、需求预测、电池的荷电状态（SoC）、累计利润以及动态信誉评分。这些私有信息对于智能体根据自身的具体情况和约束制定策略至关重要。

- **隐式合作关键绩效指标：**选取的系统级关键绩效指标（如社会福利、电网拥塞以及供需不平衡）会被纳入观测向量。这些指标作为共享的公共信号，为所有智能体提供对整体市场健康状况的一致描述，使其即使没有直接通信，也能学习局部动作与理想的全局结果之间的关联。关键绩效指标的完整说明见第3.2节。

该环境中的目标是寻找联合策略 π，使期望折扣奖励总和 R_i 最大化；在本地能源市场（LEM）情境下，这通常与净利润或效用相关。求解这一去中心化部分可观测马尔可夫决策过程的复杂之处在于，智能体 i 的最优策略取决于所有其他智能体$j \neq i$的策略，而这些策略都在同时演变。这种相互依赖使学习算法面临非平稳环境，因此需要采用此框架支持的鲁棒专用多智能体强化学习算法，以确保在连续的高维空间中稳定收敛。

表5：将本地能源市场表述为去中心化部分可观测马尔可夫决策过程时的观测空间向量。

市场信号		智能体特定信号		隐式合作关键绩效指标	
1. 当前步	9. 点对点交易电量	1. 能源发电	9. 平均利润	1. 社会福利	6. 电网拥塞
2. 时刻	10. DSO交易比例	2. 能源需求	10. 声誉	2. 市场流动性	7. 协调得分
3. 出清价格	11. 电网净输入量	3. 累计已满足需求	11. 电池能量水平	3. 平均买卖价差	8. 协调收敛
4. 出清电量	12. DSO购电价格	4. 累计延期需求	12. 电池荷电状态	4. 价格波动	9. 分布式能源自发自用
5. 电网平衡	13. DSO售电价格	5. 剩余需求	13. 电池可用电量	5. 供需失衡	10. 灵活性利用率
6. DSO购电量	14. 本地平均价格	6. 累计已满足供给	14. 电池可用放电量		
7. DSO售电量	15. 价格差	7. 累计延期供给	15. 电池累计充电量		
8. DSO总电量	16. 本地价格优势	8. 剩余供给	16. 电池累计放电量		

3.1 隐式合作模型

隐式合作模型与传统的集中式控制和显式谈判范式有着根本区别。在我们此前的工作中建立的 MARLEM 仿真框架 [9], 提供了环境的物理层和市场层；在此基础上，本节将详细介绍旨在实现协调的具体算法贡献。与需要直接通信、集中式编排或预定义协议（要求智能体共享其意图和能力信息）的显式协调机制不同，隐式合作是通过对信息结构和激励机制进行策略性设计而自然形成的。在我们的框架中，智能体无需显式通信便可学会协调自身动作，而是依赖反映市场与电网集体状态的系统级信号。

在去中心化能源市场的语境下，我们将隐式合作定义为自利型智能体之间涌现出的协调行为，这些行为既满足个体经济目标，又满足集体电网约束。这一过程通过三个组成部分的交互实现：以系统级关键绩效指标形式呈现的共享信息信号，为智能体提供市场与电网健康状况的信息；通过将单个智能体的性能与系统级结果相挂钩的奖励结构实现激励对齐；以及通过重复交互与学习形成互补策略。

这一方法的动机源于复杂系统中涌现行为的理论。在去中心化系统中，显式协调机制面临诸多挑战，包括通信开销和延迟、敏感数据带来的隐私问题、单点故障的风险，以及固有的可扩展性限制 [1, 10]。隐式合作通过共享环境信号而非直接的代理间通信来实现协调，从而应对这些挑战，其原理与群体智能中的基于定量标记的协同觅食相似 [30]。

在这一类比中，市场平台充当痕迹媒介 [31],，智能体的动作会在环境中留下痕迹（例如拥塞信号），继而促使其他智能体采取后续动作。在这一框架中，系统级关键绩效指标至关重要。这些指标就像信息素，使智能体能够根据局部观测推断全局状态 [30]。将有关市场效率、电网平衡和资源利用的关键绩效指标纳入观测空间和奖励函数后，智能体便能学会将其局部动作与电网层面的结果联系起来。由此形成一个反馈回路：自利学习促使稳定电网的策略涌现，有效地将系统复杂性与个体智能体的复杂性解耦 [35]。

隐式合作的实现依赖于两种相互耦合的机制：增强观测空间设计和多目标奖励函数。虽然环境的完整数学推导详见 [9],，但本文重点介绍驱动合作行为的具体增强措施。

观测空间的设计对于在部分可观测性的约束下实现隐式合作至关重要。在标准的去中心化部分可观测马尔可夫决策过程中，智能体 i 通常只能观测其局部状态。为实现协调而不侵犯隐私，我们引入增强观测空间 $\widehat{\Omega}$。每个智能体都会接收局部观测 $o_i \in \widehat{\Omega}_i$，其中不仅包括其内部状态（例如发电量/需求曲线、电池荷电状态、利润），还包括一组全面的系统级信号。为引导经济策略，观测空间纳入了动态市场信号，例如出清价格、出清电量和价格差。关键的是，智能体还可以观测点对点交易比例，其定义为点对点交易量占总交易量的比例。较高的点对点交易比例表明市场自给自足程度较高，从而隐式地鼓励智能体优先选择本地层，而非配电系统运营商。

同时，注入电网状态信号，以将物理约束纳入考量。电网平衡 (B_{grid}) 作为电网净能源状况的归一化指标（见 (1)，其中 G_t 为能源发电量，D_t 为能源需求，$E_{bought,t}$ 和 $E_{sold,t}$ 分别表示智能体 i 在交易时段 t 购入和售出的能源），正值表示盈余，负值表示缺口。B_{grid} 为正表示发电盈余，为负则表示缺口。此外，电网拥塞指标（见 (2)，其中 F 为每条输电线路 e $\in \xi$上的潮流，C 为电网边缘的容量，ζ 为平均拥塞水平）表示所有网络边的平均拥塞水平，并将其归一化至 $[0, 1]$。拥塞水平较低表明系统运行远未达到其物理限制，从而提高可靠性。这些信号充当拥塞指标，无需访问电网运营商的拓扑数据 [32]，即可向智能体发出基础设施压力警告。

$$B_{grid} = \sum_{i \in I} (G_i - D_i + E_{bought,i} - E_{sold,i}) \tag{1}$$

$$\bar{\zeta} = \frac{1}{|\xi|} \sum_{e \in \xi} \frac{\mathcal{F}_e}{\max(C_e)} \tag{2}$$

最后，加入特定的隐式合作关键绩效指标，以引导高层策略。这些关键绩效指标经过分类，可多角度呈现系统性能，包括经济效率、电网稳定性、资源协调和协调有效性方面的指标。完整的隐式合作关键绩效指标详见 [9]。通过广播这些信号，系统将全局状态信息转化为局部可观测量，同时不泄露个体私人数据，从而在保护隐私的同时实现系统感知学习。

奖励函数是行为适应的主要驱动因素，其设计旨在平衡个体经济激励与系统级稳定目标。如 [9], 所推导，智能体 i_t 在交易时段 t 的奖励 R 是一个复合函数，构成分层激励机制（见（3））。

$$R_i = R_{base,i} \cdot (1 + f_{coop} \cdot f_{contrib,i}) - \gamma_{DSO,i} - \gamma_{UD,i} \tag{3}$$

基础奖励（$R_{base,\ i}$）是一个加权和，体现智能体的个体表现，并确保智能体学习有效的竞价策略，以代表自身进行交易。$R_{base,\ i}$ 由以下部分组成：经济部分，根据相对于配电系统运营商价格的利润最大化给予奖励，并激励智能体学习有效的竞价策略；电网平衡部分，对有助于减少整体电网不平衡的动作给予奖励（即在盈余时段买入，在缺口时段卖出）；资源分配部分，鼓励高效利用可用的分布式能源资源容量；交易部分，激励成功交易，并对点对点交易赋予更高权重，以形成偏好本地协调的倾向；以及稳定性部分，对有助于市场长期稳定的行为给予奖励。

这项工作在隐式合作方面的核心贡献是协作奖励，即由合作因子f_{coop}构成的乘法项。该因子根据系统关键绩效指标（包括经济效率、电网稳定性和协调有效性）的加权总和计算得出。它以乘法奖励的形式对基础奖励进行缩放。当系统健康状况良好（即社会福利高、拥塞程度低）时，所有智能体都会感知到更高的潜在奖励。这形成了一种动态机制：智能体会为了最大化自身

长期效用 [31, 131]而积极维护稳定的市场环境。为防止奖励利用，协作奖励会受到贡献因子$f_{contrib,\ i}$的调节，该因子量化智能体i对系统健康状况的具体个体影响。$f_{contrib,\ i}$是智能体在降低电网不平衡方面的边际贡献（$-\frac{\partial |B_{grid}|}{\partial a_i}$）、对价格效率的贡献以及对本地交易量的贡献的加权总和。该项确保了有效的信用分配——这是合作式多智能体强化学习面临的一项挑战——通过按照智能体对集体状态的正向影响比例给予奖励。最后，该结构还包含一种惩罚机制，鼓励智能体优先进行点对点交易，并促进本地市场自给自足。$\gamma_{DSO,\ i}$用于惩罚与配电系统运营商的交易，惩罚力度根据当前电网不平衡程度进行调整；$\gamma_{UD,\ i}$则用于惩罚未满足需求。这形成了双重激励：智能体会减少对配电系统运营商的依赖；在电网不平衡严重的时期，这种限制会进一步加强，促使智能体寻求本地匹配方案。

这些机制建立了一个持续的测量、共享与适应反馈回路。首先，系统根据当前市场状态和电网条件计算所有关键绩效指标。随后，这些关键绩效指标被整合到每个智能体的增强观测空间中，提供有关系统健康状况的共享信号。同时，这些关键绩效指标决定奖励函数中合作因子的大小。智能体执行动作以最大化这一奖励，并由此学会将特定行为（例如在协调得分较低时减少负荷）与系统级结果联系起来。这个反馈回路使智能体无需显式通信也能学会协调策略，因为它们可以根据关键绩效指标信号推断自身动作对集体的影响。

3.2 隐式合作关键绩效指标

该框架计算一套全面的关键绩效指标，并将其分类，以便从多个维度展现系统性能 [9]。以下是用于衡量的指标：

1. 生态**经济效率关键绩效指标：** 这组指标用于衡量市场创造价值和促进效率的能力 nt
   价格 发现。

   - **社会福利：**所有交易的总经济价值，即消费者剩余与生产者剩余之和。它是衡量整体市场效率的主要指标。请注意，p表示价格，q表示能源数量。

$$\text{Social Welfare} = \sum_{trades} p_{trade} \cdot q_{trade} \tag{4}$$

   - **市场流动性：**交易的能源总量，反映市场的活跃度和深度。

$$\text{Liquidity} = \sum_{trades} q_{trade} \tag{5}$$

   - **平均买卖价差：**买入和卖出订单价格之间的平均差额，用于衡量市场效率。

$$\text{Spread} = \mathbb{E}[p_{ask}] - \mathbb{E}[p_{bid}] \tag{6}$$

   - **价格波动：**时间窗口内出清价格的标准差，反映市场稳定性 y.

2. **Gri电网稳定性关键绩效指标：** 这组指标用于评估电网的物理健康状况和运行效率。

   - **供需失衡：** 以电网容量C_{grid}归一化后的净能源不平衡，用于衡量供需平衡状况。数值接近零意味着系统运行稳定，不会给更广泛的电网带来压力。

$$\text{Imbalance} = \frac{|\sum q_{buy} - \sum q_{sell}|}{C_{grid}} \tag{7}$$

   - **电网拥塞：** 所有输电线路的平均拥塞水平 $e \in \xi$如式（2）所定义，反映基础设施承受的物理压力。拥塞水平较低表明系统运行远未达到其物理极限，从而提高可靠性。

   - **电网平衡：** 电网整体的能源平衡，按式（1）计算，即总发电量与消费量之差。

3。**资源协调关键绩效指标：**这些指标用于评估分布式能源资源的利用和协调效率。

• **分布式能源自发自用：** 通过点对点交易而非与配电系统运营商进行的交易，在总交易能源中所占的比例。该比例较高表明本地市场具有较强的自给能力和运行效率，并降低了对集中式公用事业的依赖。

$$\text{Self-Consumption} = \frac{\sum q_{P2P}}{\sum (q_{P2P} + q_{DSO})} \tag{8}$$

• **灵活性利用率：** 在点对点交易中实际利用的可用灵活性能源占比。该指标衡量智能体对其能源灵活性资源（发电盈余、需求缺口和电池容量）的利用效率。

$$\text{Flexibility Utilization} = \frac{\sum q_{p2p}}{\sum q_{available}} \tag{9}$$

其中，$\sum q_{available}$ 表示所有智能体在给定交易时段 $t \in \{1, \ldots, \tau\}$ 的可用灵活性总量，包括可售灵活性（发电盈余加电池放电容量）和可购灵活性（需求缺口加电池充电容量）。

4. **协调效果关键绩效指标：** 这些指标衡量智能体间协调的形成及其有效性。

• **协调得分：** 用于衡量协调程度，反映市场的平衡状态。得分接近 1 表示系统达到完美平衡。

$$\text{Coordination Score} = 1 - \text{Imbalance} \tag{10}$$

• **协调收敛：** 衡量近期时间窗口内交易量的稳定性，用以判断是否已形成稳定、协调的模式。其计算方式与价格波动率指标类似。

3.3 多智能体强化学习算法：算法特定特征

3.3.1 近端策略优化

作为一种同策略梯度方法，PPO 直接利用当前策略生成的数据进行学习。它旨在解决策略梯度方法中的步长问题：如果策略更新幅度过大，智能体可能会进入参数空间中性能不佳的区域。由于同策略智能体根据自身行为收集数据，糟糕的策略会导致数据质量低下，进而造成终局性性能崩溃，使智能体无法恢复。

PPO 的核心创新是裁剪代理目标函数，它通过实施信赖域来确保策略更新是渐进的。对于去中心化产消者智能体 i，目标函数 $J^{CLIP}(\theta_i)$ 定义于式（11）中。

$$J^{CLIP}(\theta_i) = \mathbb{E}_t \left[\min \left(r_t(\theta_i)\widehat{A}_t, \text{clip}(r_t(\theta_i), 1-\varepsilon, 1+\varepsilon)\widehat{A}_t \right) \right] \tag{11}$$

$$r_t(\theta_i) = \frac{\pi_{\theta_i}(a_t|o_t)}{\pi_{\theta_{old}}(a_t|o_t)} \tag{12}$$

其中，$r_t(\theta_i)$ 是概率比，用于衡量在新策略下采取动作 a_t 的可能性相较于旧策略增加或减少了多少，而 θ_i 是智能体 $i \in I$ 的策略参数。术语 $\widehat{A}_t$ 表示优势函数，用于量化某个特定动作比平均行为好多少。裁剪机制的操作逻辑（ε，通常为 0.2）会根据经验的性质而有所不同：

• **正优势函数**（$\widehat{A}_t > 0$）：如果该动作表现良好，算法希望提高 r_t。最小值运算符会将此次增幅限制在 $1+\varepsilon$，从而防止策略根据一次偶然的有利市场交互进行过度优化，确保稳定性。

• **负优势函数**（$\widehat{A}_t < 0$）：如果该动作不佳，算法会希望降低 r_t。裁剪可防止概率被瞬间压至零，否则会破坏智能体未来探索更优策略的能力。

为计算优势函数 $\hat{A}_t$，近端策略优化采用广义优势估计。该机制通过在即时反馈与长期回报之间进行插值，降低从本地能源市场获得的奖励的方差。

在隐式合作的背景下，近端策略优化的优势在于这种抑制效应。在多智能体环境中，由于其他智能体也在同时学习，转移动态是非平稳的。如果智能体A大幅改变其竞价策略，智能体B所处的环境也会随之改变。近端策略优化的保守更新让邻近智能体有时间适应这些变化，从而促进系统收敛到稳定的纳什均衡，而非像非截断方法中常见的那样出现混乱振荡。

3.3.2 异步近端策略优化

异步近端策略优化解决了中等规模仿真中标准近端策略优化的计算瓶颈。在包含一百个智能体的本地能源市场中，求解市场出清物理过程所需的时间可能有所不同。标准近端策略优化是同步的，这意味着学习器必须等待所有智能体完成仿真后才能更新。异步近端策略优化在执行者—学习器拓扑中，将执行者（数据收集器）与学习器（策略优化器）解耦。

由于数据收集和优化是并行进行的，因此会出现策略滞后：数据可能由稍旧版本的策略（$\pi_{behavior}$）收集，而非当前正在更新的策略（π_{target}）。使用这些陈旧数据会引入偏差。异步近端策略优化采用 V-trace 从数学上校正这一差异。价值更新的 V-trace 目标定义见（13），其中 $V^{target}{}_t$是校正后的价值估计，即学习器针对时刻t的状态所要达到的值；$V(s_t)$ 是当前价值函数对状态$s_t \in \mathcal{S}$的估计；γ 是折扣因子，它决定智能体对未来奖励相较于即时奖励的重视程度；l 是轨迹片段的长度；c^{t+j} 是轨迹截断系数，用于控制未来步骤的误差对当前更新的影响程度；ρ_{t+k} 是截断重要性采样权重（其中$\bar{\rho}$ 是表示截断阈值的超参数，用于防止重要性权重变得过大）；δ_{t+k} 是时序差分误差。

$$V_t^{target} = V(s_t) + \sum_{k=0}^{\ell-1} \gamma^k \left(\prod_{j=0}^{k-1} c_{t+j} \right) \rho_{t+k} \delta_{t+k} \tag{13}$$

$$\rho_t = \min\left(\bar{\rho}, \frac{\pi_{target}(a_t|o_t)}{\pi_{behavior}(a_t|o_t)}\right) \tag{14}$$

重要性权重 ρ_t 衡量当前策略与收集数据时所用策略之间的偏离程度。如果偏离过大，权重实际上会截断轨迹，迫使学习器更多地依赖当前的价值估计，而不是过时的奖励。

在我们的多智能体场景中，异步近端策略优化能够提供经验多样性。由于不同的执行者可以同时与不同实例的能源电网进行交互，学习器能够接收范围更广的场景（例如，不同模拟日期的发电水平各不相同）。这避免了策略的锁步同步，并促使形成能够适应各种电网条件的鲁棒分散式策略。

3.3.3 软演员-评论家

近端策略优化旨在最大化期望回报，而软演员-评论家则旨在最大化期望回报与策略熵（$\mathcal{H}$）之和。其目标函数定义见（15），其中 α 是温度参数。

$$J(\pi) = \sum_{t=0}^{\tau} \mathbb{E}_{(o_t,a_t)\sim\rho_\pi} \left[r(o_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|o_t)) \right] \tag{15}$$

熵是衡量随机性的指标。通过最大化熵，软演员-评论家促使智能体在实现目标的同时尽可能保持随机性。这为本地能源市场协调带来了三项关键优势：

- **对非平稳性的鲁棒性：**在去中心化部分可观测马尔可夫决策过程中，确定性策略非常脆弱。如果邻居改变出价阈值，确定性策略可能会突然失效。随机策略会学习最优动作的分布，因此能够更好地适应对等方不断变化的行为。

- **发现合作策略：**协调负荷转移很难被发现，因为它往往需要暂时承受局部经济损失，以换取全局系统效益。SAC 强制引入的熵可防止智能体过早收敛到自私且次优的局部最小策略。

- **连续控制的重参数化：**为了优化随机策略，软演员-评论家（SAC）将采样得到的动作a_t a设为策略参数的可微函数。这使其能够精细、连续地调节买价中的出价数量和出价价格，从而实现精细的电网平衡。

软演员-评论家（SAC）是一种异策略算法，它利用经验回放缓冲区存储过去的每次交互。因此，相比近端策略优化（PPO），它具有更高的样本利用效率。为防止智能体因神经网络误差而对某些动作过度乐观，SAC维护两个独立的Q网络，并在更新时取两者中的最小值，从而确保智能体学到的协调信号基于稳定的价值估计。

3.4 多智能体强化学习训练范式

3.4.1 集中式训练与集中式执行

集中式训练与集中式执行代表了协调质量的理论上界，并作为本研究中可实现的最高性能的主要基准。在这一范式中，多智能体问题被简化为具有高维联合动作空间的单智能体问题。该架构采用单一策略网络 π_θ，该网络接收全局环境状态$\mathcal{S}_t$。这一全局状态完整描述了环境，并汇集电网中的所有私有信息。策略将全局状态直接映射为联合动作向量$\mathbf{a}_t$，通过一次前向传播求解优化问题。数学公式假设采用联合策略；尽管该策略在实现时可以因式分解，但仍以全局状态为条件（见（16））。

$$\pi_\theta(a_{1,t}, \ldots, a_{I,t}|s_t) = \prod_{i=1}^{I} \pi_{\theta_i}(a_{i,t}|s_t) \tag{16}$$

基于$\mathcal{S}_t$ 的共享条件信息使网络能够学习智能体行为之间复杂的相关性。集中式价值函数$V^\phi(s)$通过考虑系统级奖励的总和来估计期望回报，从而直接优化全局目标$J_{CTCE}(\theta)$，该目标定义为期望折扣全局回报（见（17））。

$$J_{CTCE}(\theta) = \mathbb{E}_{s\sim\rho^\pi, \mathbf{a}\sim\pi_\theta}\left[\sum_{t=0}^{T}\sum_{i=1}^{I}\gamma^t R_i(s_t, \mathbf{a}_t)\right] \tag{17}$$

由于控制器拥有完整的可见性和控制权，训练过程是平稳的；环境动力学固定，且不存在会造成非平稳性的其他独立学习实体。CTCE 可以通过明确权衡一个智能体与另一个智能体的效用，以满足系统约束，从而实现帕累托最优。然而，由于违反隐私要求且存在可扩展性问题，随着智能体数量增加，搜索空间的规模会呈指数级增长，因此 CTCE 仅作为基准。

3.4.2 CTDE

CTDE 的运行逻辑依赖于混合式演员-评论家架构。在训练阶段，会使用集中式评论家 $Q^\phi(s, \mathbf{a})$。该评论家可以访问全局状态 $\mathcal{S}_t$和联合动作向量 $\mathbf{a}_t$，从而能够从系统耦合关系的全局视角估计 Q值。然而，在执行阶段，评论家会被移除，每个智能体 i保留一个独立的策略网络 $\pi_{\theta_i}(a_i|o_i)$（分散式演员），该网络仅将其局部观测 o_i 映射为动作 a_i。更新演员策略 θ_i的梯度，使用确定性策略梯度定理或其随机等价形式计算。

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s,\mathbf{a}}\left[\nabla\theta_i \log \pi_{\theta_i}(a_i|o_i) Q_\phi(s, a_{1,t}, \ldots, a_{I,t})\right] \tag{18}$$

在（18）中，集中式评论家评估智能体 i 的动作a_i的质量时，并非依据局部结果（这些结果可能因对等方的动作而产生噪声），而是结合全局系统状态进行评估。这解决了训练过程中的非平稳性问题：评论家会明确考虑所有其他智能体不断变化的策略 π_{-i}，从而确保优势信号保持稳定。这一经过信息指导的信号使本地演员能够学习协调策略，并在之后自主执行。

在我们的研究中，CTDE 用于测试能否通过集中式指导明确传授协调策略。然而，在训练阶段，它仍然存在隐私障碍，因为智能体必须与中央评论家共享其经验（观测、动作和奖励），这可能与现实世界中的本地能源市场所适用的数据保护法规相冲突。

3.4.3 DTDE

在 DTDE 中，多智能体系统被视为一组独立学习者。每个智能体i都维护自己的演员 $\pi_{\theta_i}(a_i|o_i)$ 以及自己的独立评论家$V^{\phi_i}(o_i)$或$Q^{\phi_i}(o_i,\ a_i)$。关键在于，评论家只能获取局部观测o_i和该智能体自身的动作a_i，并将所有其他智能体的动作纯粹视为环境动力学的一部分。每个智能体的训练目标是最大化自身的期望折扣回报，如式（19）所定义。

$$J_{DTDE}(\theta_i) = \mathbb{E}_{o_i,a_i}\left[\sum_{t=0}^{\tau} \psi^t R_i(o_{i,t}, a_{i,t})\right] \tag{19}$$

DTDE 的核心理论难题是非平稳性。在去中心化部分可观测马尔可夫决策过程中，从智能体i的角度看，环境转移函数$\mathcal{P}_i$实际上吸收了所有其他智能体不断演变的策略π_{-i}（见式（20））。

$$\mathcal{P}_i(s'|s, a_i) = \sum_{\mathbf{a}_{-i} \in c\mathcal{A}_{-i}} \mathcal{P}(s'|s, a_i, \mathbf{a}_{-i}) \prod_{j \neq i} \pi_j(a_j|o_j) \tag{20}$$

由于所有智能体在训练期间同时更新其策略，$\pi_j \mathrm{P}_i$会随时间变化。这种对马尔可夫性质的违背会导致学习不稳定、振荡，并收敛到次优纳什均衡。

我们的贡献在于，通过第 3.1 节所述的增强观测空间来缓解这种非平稳性。将系统级关键绩效指标注入局部观测o_i后，环境便会趋于准平稳。这些关键绩效指标以经过压缩且保护隐私的方式呈现其他智能体的总体行为。智能体A无需了解智能体B的具体状态，只需了解智能体B的动作对电网造成的物理影响。

4 实验设置

本案例研究旨在现实条件下检验隐式合作假设。具体而言，它考察去中心化智能体能否仅凭经基于迹象的关键绩效指标信号增强的局部观测，实现系统级协调（例如维持供需平衡），从而无需集中式调度或显式通信。该场景将在所有训练范式（CTCE、CTDE、DTDE）与多智能体强化学习算法（近端策略优化、异步近端策略优化、软演员-评论家）的组合下进行评估。

本设计的主要目标是定量验证：在信息稀缺约束下运行的去中心化智能体（DTDE），能否达到理论最优集中式系统（CTCE）的协调质量。我们评估哪种连续控制算法（PPO、异步近端策略优化或软演员-评论家）能够在探索与稳定性之间实现最佳权衡：探索对于发现合作均衡至关重要，而稳定性则是维持非平稳环境中电网物理约束的关键。通过让不同算法在完全相同的环境条件下运行（具体而言，包括异构智能体群体和受约束的 IEEE 34母线拓扑），我们排除了混杂变量的影响。这确保了所观察到的性能差异完全源于训练范式与学习更新规则之间的交互作用。我们检验以下假设：注入观测空间的系统级关键绩效指标（例如电网平衡指数）可以作为有效的集中式协调信号代理，使智能体能够将其动作产生的外部性内化。

用于复现该案例研究的详细配置和实现可在项目专用的 GitHub代码仓库中找到（https://github.com/salazarna/marlem）。

4.1 案例研究配置

实验配置通过由 8 个分布式能源资源智能体组成的多样化群体实现相互依存，这些智能体具有互补的发电与需求曲线（见图 1）。这种异质性确保协调不仅有益，而且是实现系统性能最优化的必要条件。容量范围（75–180 千瓦）和电池比例的选择，旨在确保渗透率符合实际情况（占总负荷的 40-60%），从而带来有意义的改进空间，同时避免立即引发电网不稳定。表6 详细列出了各智能体的具体配置，包括其角色、容量以及支撑其能源概况的行为逻辑。为确保结果可复现，表 7详细说明了环境配置。

表6：案例研究中智能体的配置。

智能体	IEEE-34节点	容量（千瓦）	电池（千瓦）	电池比例	描述
小型工业	840	180.0	108.0	0.6	持续需求，早晨出现峰值；依赖储存能量
社区医院	890	170.0	85.0	0.5	早晨需求峰值较高，随后转为下午发电盈余
大学校园	844	140.0	98.0	0.7	双峰需求（早晨/下午峰值）；午间发电量适中
购物中心	816	130.0	78.0	0.6	下午/傍晚需求高；午间盈余发电量的消纳方
住宅区	800	80.0	64.0	0.8	双峰需求（早晨/傍晚）；依靠储能应对傍晚爬坡
公寓楼	830	75.0	52.5	0.7	与住宅区类似；发电能力中等
社区太阳能农场	848	180.0	180.0	1.0	午间发电量高、需求极少；可作为灵活的供应方
停车场	860	160.0	144.0	0.9	午间发电量高；可作为灵活性供应方

每个智能体都配备了一块电池，其充放电效率建模为 95%。为保护电池健康并确保符合实际情况，荷电状态限制在额定容量的 5% 至 95% 之间。这些不同的电池比例（0.5 至 1.0）形成了市场力的异质格局，从而能够分析储能能力各异的智能体如何为电网平衡做出贡献。

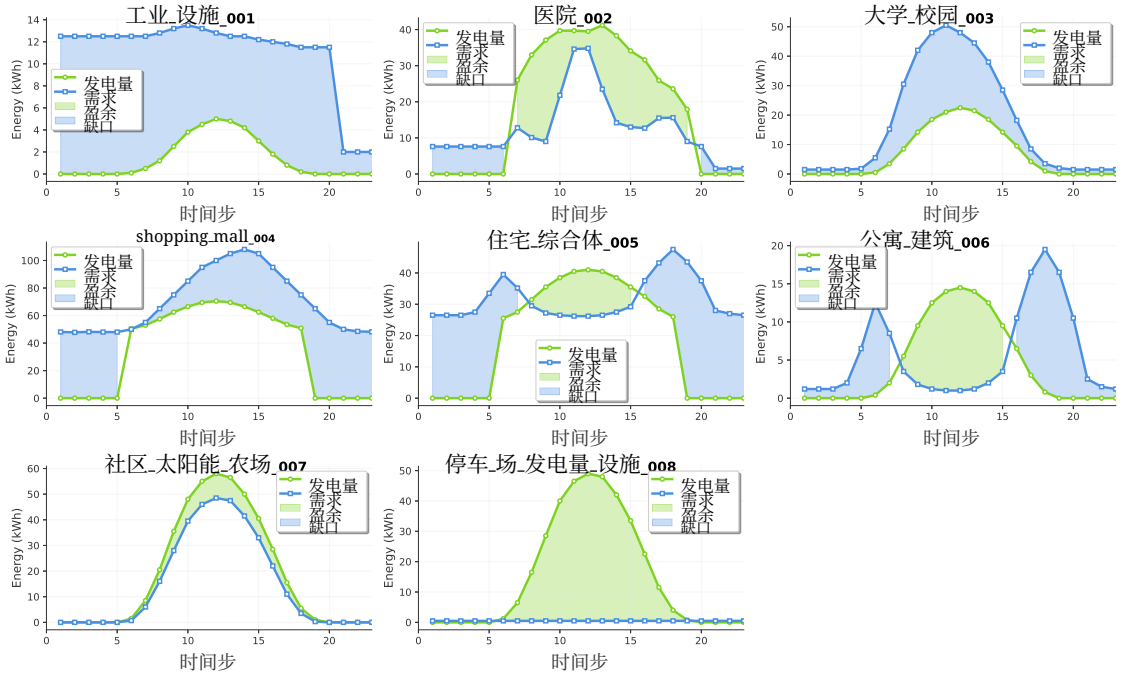


图1：案例研究中分布式能源资源智能体的发电与需求曲线。

仿真采用 IEEE 34节点测试馈线拓扑，这是针对本案例研究进行修改的标准配电网基准。系统总容量限制为 1800 千瓦。该容量限制在用电高峰时段会成为约束（考虑到智能体总容量约为 1115 千瓦，另加基荷），因此智能体必须通过本地平衡进行物理协调，而不能仅依赖配电系统运营商。智能体的战略布局（见图2）根据拓扑中的关键电气节点确定，以便开展鲁棒的硬件在环验证。各智能体被分配至特定节点，这些节点对应于未来 RT-Lab IEEE-34模型实现中的指定信号监测点。在变电站区域，小型工业位于节点 800（变电站母线），该处是主要馈电点，也是测量电网进出功率的关键位置。社区医院智能体位于节点 890（变压器二次侧），这是一个关键测量节点，可用于评估电压变换后下游的电能质量和电压调节。在配电网更深处，购物中心位于节点 816，这是一个重要的馈线分支点，选定该处是为了验证关键网络分岔点处的协调动态。最后，为了检验系统在较大电气距离范围内的稳定性，社区太阳能农场和停车场被布置在配电网末端（分别位于第一条支线末端的节点 860 和节点 848），这些位置的电压支撑需求最为关键，并将在硬件在环验证期间接受严格监测。

市场层的配置旨在激励本地协调，而非依赖主电网。我们采用平均定价机制（$p_{trade} = p_{买入} + p_{卖出}$ 2 ）。采用这一机制是为了促进公平的价格发现，并降低策略性投标的复杂性。市场参数包括 20 $/kWh 的价格下限和 600 $/kWh 的价格上限，以鼓励本地交易，而非配电系统运营商套利。已启用合作伙伴偏好，使智能体能够学习并优先选择可靠的对等方。至关重要的是，配电系统运营商价格信号构成了智能体行为的边界条件。如图 3所示，上网电价（FIT）与公用事业电价之间保持着明显的差距。上网电价代表向电网售电的收入，且始终低于公用事业电价（峰值约为 300 $/kWh）。公用事业电价代表从电网购电的成本，白天会大幅攀升（达到 600 $/MWh），这会对未能在需求峰值期间从本地获取能源的智能体造成经济惩罚。这两条曲线之间的差距形成了点对点交易价差，智能体因此受到经济激励而相互交易：卖方可以获得高于上网电价的收益，买方则可以用低于公用事业电价的价格购电，

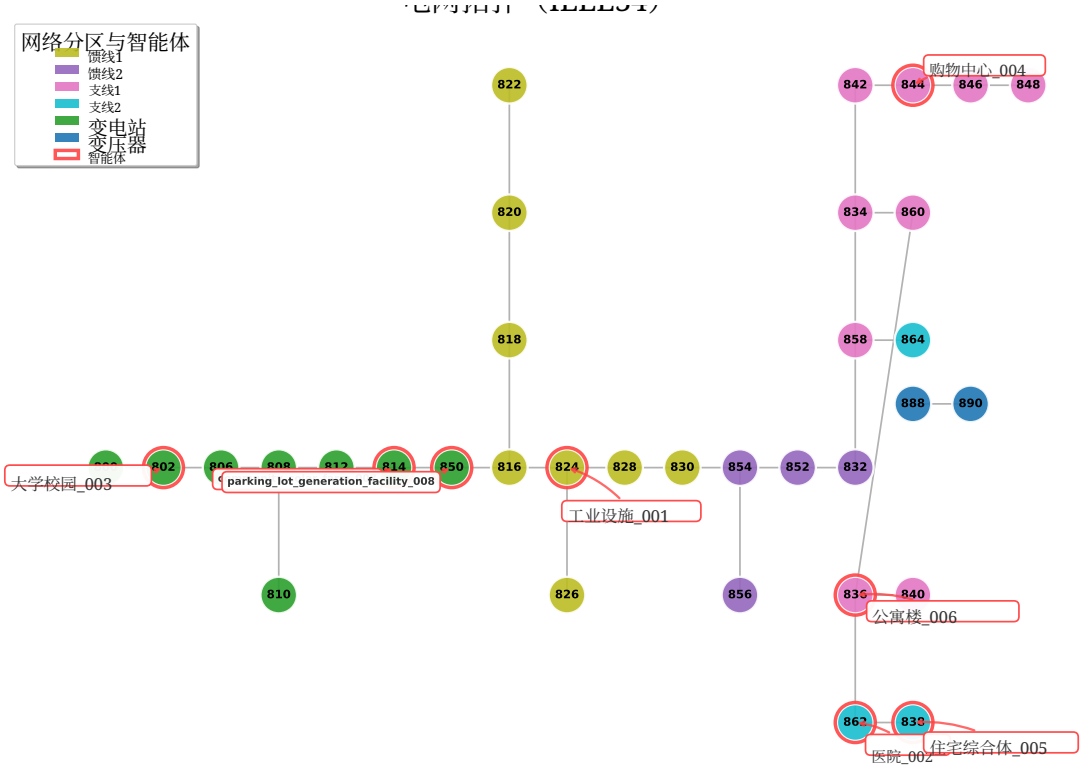


图2：34节点IEEE测试馈线作为案例研究的电网网络。

前提是它们能够有效协调。实验之间的配电系统运营商电价曲线保持固定，以确保各回合的配电系统运营商价格信号始终一致，从而为学习提供稳定的基础。

4.2 训练配置与实验矩阵

为系统地分离信息架构和学习更新规则的影响，我们采用了 3×3 析因设计：将三种训练范式（CTCE、CTDE、DTDE）与三种多智能体强化学习算法（PPO、异步近端策略优化、软演员-评论家）进行组合。多智能体强化学习算法的初始超参数集列于表 8。

训练过程由 MARLEM 框架进行编排。为动态优化超参数，我们采用基于种群的训练（PBT） [132]。该过程为九种配置分别初始化 10 个调优样本，共产生 90 个并发试验。这些试验彼此独立地演进：在利用阶段，PBT 会定期克隆表现最佳试验的权重；在探索阶段，框架则会调整表现较差试验的超参数，以高效搜索参数空间。尽管 PBT 会在训练期间修改特定的学习率，但初始设置会根据各多智能体强化学习算法的数学要求进行调整，以确保基准比较严谨。

每种配置总共训练 10 000 回合，以确保在复杂的多智能体环境中有足够的交互步数来实现收敛。模型检查点每 10 次迭代保存一次，并在每次训练迭代时进行评估。评估时长由框架自动确定，并根据评估工作线程的容量设定适当的评估回合数。

训练工作在一台专用机器上执行，该机器配备 12 核 Ryzen 9 9900X 处理器、128 GB 内存、RTX 4090 显卡、采用 amd64 架构的 Linux 操作系统、Ubuntu 24.04 发行版，并配有 CUDA 13。

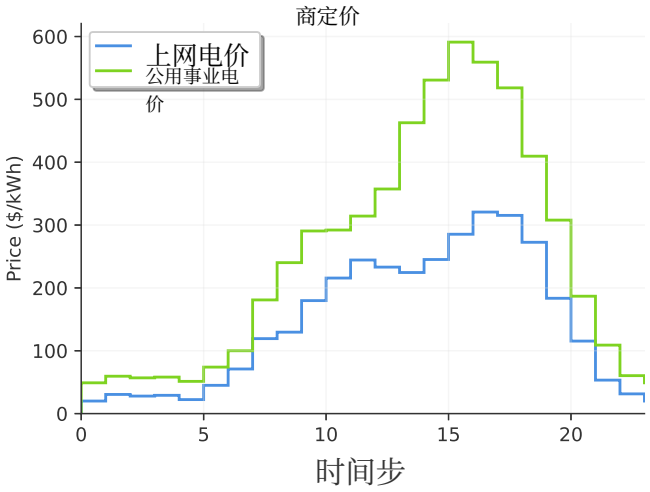


图 3：案例研究中配电系统运营商代理的上网电价和公用事业电价曲线。

表 7：案例研究的核心仿真参数。

参数	值	描述
最大步数	24	表示完整24小时周期的回合长度（每步1小时）
电网拓扑	IEEE 34节点	标准配电网拓扑
电网容量	1800 千瓦	电网总容量
Seed	42	用于确保可复现性的固定随机种子
价格范围	20 - 600.0	允许的报价范围，设计得比配电系统运营商价格更宽
数量范围	0 - 180 千瓦时	每笔订单允许的交易规模
预测误差	30%（最大_误差=0.3）	施加于智能体预测的最大随机误差，用于模拟不确定性
异步订单	True	启用异步订单处理，以模拟真实的市场时延

4.3 评估指标

对隐式合作的评估依赖一组多维关键绩效指标，这些指标用于衡量经济效率、电网稳定性、资源协调和智能体自主性之间的权衡（参见第 3.2节）。

衡量成功与否的核心指标是协调得分，其定义为归一化供需失衡的补数（回顾（10））。得分接近 1.0 表示达到完美的物理均衡，直接验证了智能体无需集中调度即可同步发电量与消费的能力。此外，我们还使用电网平衡指数（B_{index}），该指数纳入输电损耗和外部电网依赖，从而全面评估物理稳定性。为了评估市场自给自足程度，我们采用点对点交易比例；比例越高，表明本地匹配效果越好，对配电系统运营商的依赖越低。最后，社会福利量化所创造的经济总价值，确保物理协调不会以牺牲经济效率为代价。

为了解协调如何形成，我们分析智能体响应性，其计算方式是系统级信号与智能体动作随时间变化的相关性。较高的正向响应性表明，设计的信息素信号能够有效引导决策。此外，我们还监测价格波动以确保市场稳定性，同时通过分布式能源资源自耗追踪各项资产提供灵活性服务的成效。

模拟中的所有随机过程（例如噪声注入）均由全局随机种子控制（参见表 7）。这确保了所有实验运行中的环境条件完全一致，从而保证观察到的性能方差源于学习动态，而非环境随机性。

为评估所提出框架的计算可行性，除核心协调实验外，还进行了补充性的可扩展性测试。在这项具体分析中，使用完全相同的仿真环境，每次运行执行 1

表 8：案例研究中多智能体强化学习算法的初始超参数。

超参数	PPO	APPO	SAC
学习率	1e-5	1e-5	演员：1e-5，评论家：1e-5
折扣因子（γ）	0.99	0.99	0.99
熵系数	0.01	0.01	动态
梯度裁剪	0.5	0.5	0.5
裁剪参数	0.2	N/A	N/A
广义优势估计 / V-trace	GAE	V-trace	-
批次/缓冲区大小	小批次：128	-	缓冲区容量：1 000

~~* 对于集中式范式（CTCE/CTDE），目标熵设为所有动作维度之和的负值（–sum $\sum \dim(\mathcal{A}_i)$）。对于分散式模式（~~ DTDE），目标熵则对应于单个智能体的动作空间维度的负值（– $\dim(\mathcal{A}_i)$）。

000 个回合，并采用仅包含 1 个调优样本的精简版基于种群的训练配置。分析沿用标准的 3×3 实验矩阵，同时将智能体数量设为集合 {2，20、34} 中的值。智能体数量上限严格受 IEEE 34节点测试馈线拓扑的物理节点约束限制。对于每次训练，均记录实际训练时间，直至收敛或完成 1 000 个回合（以先发生者为准），以此直接衡量随着智能体群体规模扩大，集中式范式与去中心化范式的计算开销差异。

5 结果与讨论

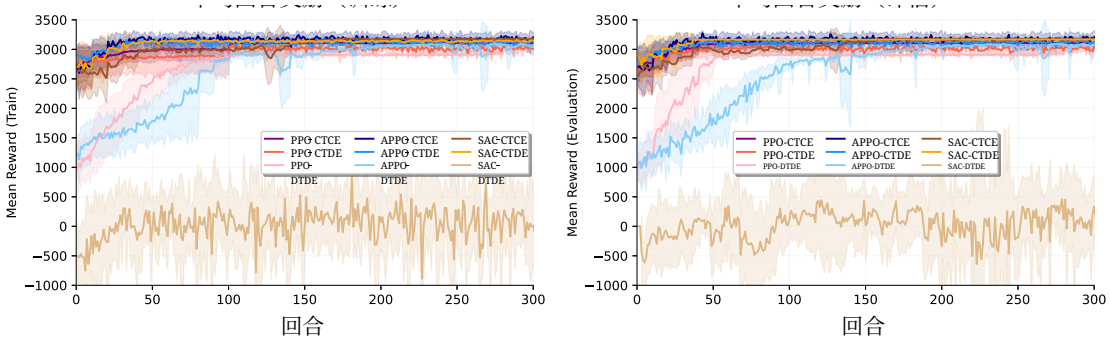
本节对第 4 节获得的实验结果进行全面分析，系统比较三种训练范式（CTCE、CTDE、DTDE）和三种多智能体强化学习算法（近端策略优化、异步近端策略优化、软演员-评论家），以验证隐式合作假设。通过结合定量性能指标与对算法学习动态的定性分析，我们量化了协调质量、稳定性和计算可扩展性之间的权衡。

5.1 实验矩阵：训练范式与多智能体强化学习算法

为严格评估隐式合作的有效性，我们对 9 种不同的实验配置进行了析因分析。这些配置由三种信息架构（训练范式）与三种学习更新规则（多智能体强化学习算法）交叉组合而成。结果基于 10 000 个训练回合的平均回合奖励，并通过特定评估阶段进行验证，揭示了明显的性能层级（见表 9）。由于所有 9 项实验均在 300 个训练回合前实现收敛，因此为便于观察，图 4 仅展示了最初的这些回合。本分析旨在揭示特定算法机制（如熵最大化、信赖域和异步更新）如何与可用信息相互作用，从而促进或阻碍协调。

5.1.1 训练范式分析

训练范式决定了观测空间的边界和策略更新机制，从而设定了协调潜力的理论上限。CTCE 配置作为协调质量的理论基准。在这一范式下，智能体在训练和执行期间都能获得完整的全局状态向量。正如预期，获取完全信息使智能体能够达到最高性能潜力。APPO-CTCE 配置为系统确立了绝对上限，在评估阶段的单回合奖励峰值达到 3272.1，平均值为 3143.9。SAC-CTCE 同样表现鲁棒，平均奖励维持在 3139.8，标准差为 93.4。PPO-CTCE 略逊一筹，平均值为 3099.4。较高的奖励值反映了最优的社会福利以及



（a）训练。 （b）评估。

图 4：对 9 项实验进行评估时的平均单回合奖励，共进行 10 000 个训练回合。仅显示前 300 个回合。

T表 9：对 9 项实验进行评估时的平均单回合奖励，共进行 10 000 个训练回合。

	CTCE			CTDE			DTDE		
	近端策略优化	异步近端策略优化	软演员-评论家	近端策略优化	异步近端策略优化	软演员-评论家	近端策略优化	异步近端策略优化	软演员-评论家
Min	2702,2	2641,2	2560,7	2618,9	2791	2730	989,7	991,6	-669,8
Mean	3099,4	3143,9	3139,8	3004,8	3096,5	3158,2	2742,5	2998	54,5
Std	44,8	97,9	93,4	49,1	64,6	42,5	431,5	431,4	215,8
Max	3123,3	3272,1	3189,7	3074,1	3181,4	3166,1	2935,2	3243,5	708,4

电网不平衡最小化。集中式控制器成功利用完全信息向量来同步智能体动作，从而避免峰值同时出现，并有效管理电网约束。

CTDE作为一种鲁棒的中间方案脱颖而出，有效地将全局协调信息提炼到局部策略中。在此设置下，集中式评论家（仅在训练期间可访问）指导必须仅依据部分观测执行的局部执行者。SAC-CTDE配置展现出卓越的稳定性与性能，平均奖励达到3158.2，最高达到3166.1。值得注意的是，其标准差是所有配置中最低的（42.5），显著低于其集中式训练与集中式执行对应方案（93.4）。这表明集中式评论家成功地将局部执行者引导至高度稳定的合作均衡，有效地平滑了学习景观。SAC-CTDE的性能有效地与SAC-CTCE的性能相当。这验证了智能体在集中式训练阶段成功地将全局目标内化到其局部策略网络中，并学会将局部信息素信号视为全局状态的可靠代理。这些结果证实，对于允许离线集中训练的受管理的能源社区，CTDE是一个安全的选择。然而，PPO-CTDE表现出更明显的下降（平均奖励为3004.8），表明在没有熵最大化辅助的情况下，同策略算法更难以将全局价值函数压缩到局部策略中。

DTDE为所提出的隐式合作模型提供了最关键的验证，但其成功高度依赖于算法。本研究最重要的发现是APPO-DTDE的卓越性能。它取得了3243.5的最大奖励和2998的收敛平均值。这一峰值性能在统计上与APPO-CTCE基准（3272.1）相当，差距仅为8.3%。这证明真正的隐式合作可以仅通过观测空间中的信息素信号机制（关键绩效指标）涌现，而无需任何集中式组件。相比之下，SAC-DTDE未能有效协调，最终平均奖励为54.5，并频繁跌入负值（最小值-669.8）。这可能表明基于熵的最大化在完全非平稳的分散式环境中较为脆弱。在没有起稳定作用的中央评论家的情况下，智能体同时追求随机性会营造出一个混乱的环境，无法找到稳定均衡。PPO-DTDE表现出稳定但次优性能（平均奖励为2742.5），难以达到异步近端策略优化所找到的高水平均衡。尽管它避免了软演员-评论家的崩溃，但也显示出保守的同策略更新在部分可观测环境中的局限性。

5.1.2 多智能体强化学习算法分析

虽然训练范式界定了信息限制，但多智能体强化学习算法决定了智能体探索和利用可用策略空间的有效程度。结果表明，近端策略优化、异步近端策略优化和软演员-评论家呈现出由各自底层数学形式驱动的不同行为特征。

近端策略优化（PPO）发挥了可靠基线的作用。在所有范式中，PPO 的表现始终稳定，尽管较为保守。PPO 在早期阶段呈现平稳的学习曲线，但其奖励水平低于异策略算法。在 DTDE 范式中，其表现达到集中式基准的 94.4%。

PPO 的同策略特性和裁剪目标函数会专门抑制大幅策略更新。虽然这能防止策略崩溃，却妨碍了在动态本地能源市场中发现复杂合作策略所需的积极探索。

异步近端策略优化（APPO）成为分散式环境中绝对占优的算法。在所有范式中，APPO 始终取得最高奖励。其异步架构将数据收集与策略更新解耦，生成高吞吐量、多样化的经验流。数据收集中的这种噪声起到了正则化作用，避免对特定同伴行为过拟合。APPO 在 DTDE 中成功的关键是 V-trace校正。这一机制使算法能够利用近似同策略数据（即由对等方稍旧策略生成的经验），同时避免产生会使标准 PPO 不稳定的偏差。因此，APPO 对分散式学习的非平稳性具有独特的鲁棒性，能够有效地将邻居不断变化的行为视为环境噪声，而非对抗性干扰。APPO 展现出卓越的可扩展性。

软演员-评论家呈现出两极化的性能表现：在混合设置中表现出色，但在纯去中心化环境中则全面崩溃。在 CTCE 和 CTDE 中，软演员-评论家非常有效。SAC-CTDE 在全部 9 种配置中实现了最稳定的性能（标准差为 42.5），验证了最大熵在集中式评论家的稳定作用下威力强大。熵项鼓励智能体探索多样化的策略，避免过早收敛到自私局部极小值。在 DTDE 中，软演员-评论家遭遇灾难性失败（差距为 73.7%）。其理论原因在于缺乏共同基准。在完全去中心化设置中，如果每个智能体同时最大化熵，环境就会变得混乱且难以预测。没有共享评论家来锚定价值函数，智能体对随机性的追求就会导致发散的反馈回路。基于种群的训练分析表明，成功的运行（CTDE）收敛到了适中的学习率（2.98×10^{-3}），而失败的 DTDE 运行则不断振荡，无法找到稳定的超参数区域。

5.1.3 实验矩阵的性能

基于种群的训练为每种配置动态调整超参数，以最大化性能。表 10 展示了训练过程结束时每种配置中表现最佳的试验所获得的最终超参数。

超参数分析揭示了有趣的适应模式。与去中心化对应方法相比，APPO-CTCE 收敛到显著更高的学习率（3.3×10^{-4}），这很可能是利用了集中式状态所提供的稳定性。相反，PPO-DTDE 需要非常高的熵系数（0.24）和学习率（4.4×10^{-3}），才能在去中心化环境中强制进行探索；尽管这种激进的调参并未带来优于异步近端策略优化的最优性能。

表 10：基于种群的训练在 10 000 个训练回合后优化得到的最终超参数。

	CTCE			CTDE			DTDE		
	PPO	**APPO**	**SAC**	**PPO**	**APPO**	**SAC**	**PPO**	**APPO**	**SAC**
学习率	9.2×10^{-6}	3.3×10^{-4}	4.6×10^{-3}	9.2×10^{-6}	1.2×10^{-5}	2.9×10^{-3}	4.4×10^{-3}	3.7×10^{-5}	1.0×10^{-3}
熵系数	5.72	0.01	-	0.04	0.01	-	0.24	0.02	-

5.1.4 关于隐式合作中性能差距的讨论

这项9种配置分析最重要的发现是，表现最佳的完全分散式配置（APPO-DTDE）与理论上的集中式上限（APPO-CTCE）之间的性能差距微乎其微。具体而言，表现最佳的DTDE配置收敛后的平均奖励为3243.5，达到表现最佳的CTCE基准值3272.1的99.2%。这一结果证实，对于这类能源管理问题，在正确的信号设计（关键绩效指标）和鲁棒算法（异步近端策略优化）的支持下，隐式合作使集中式控制变得多余。这对现实世界的部署具有重要意义，因为配电系统运营商无需建设昂贵且侵犯隐私的集中式控制基础设施，即可实现最优的电网平衡。相反，重点应转向维护广播关键绩效指标信号的完整性。

最大奖励的定量分析揭示了截然不同的性能层级，这挑战了关于去中心化成本的传统假设。对于近端策略优化和软演员-评论家，结果证实了以下假设：信息限制-

对于这些算法，集中式训练阶段（CTDE）对于实现近乎最优的性能至关重要，因为仅依赖去中心化（DTDE）会带来显著成本：近端策略优化的性能损失为 6.0%，尚可控制；而软演员-评论家的性能损失高达 77.8%，后果严重。相比之下，APPO-DTDE 的性能与集中式基准相差不到 1%，这表明无需任何集中式组件也能充分实现真正的隐式合作。APPO 在 DTDE 下的性能优于 CTDE，可能源于去中心化环境中并行异步探索的效率；这种方式避免了混合模型中集中式评论家可能造成的瓶颈或训练不稳定问题。

异步近端策略优化是绝对占优的算法，在三种范式下均实现了最高的最大奖励：CTCE（3272.1）、CTDE（3181.4）和 DTDE（3243.5）。它的成功归功于两项关键机制：将数据收集与策略更新分离，从而生成高吞吐量经验并发挥正则化作用；以及使用 V-trace目标，使其能够使用近似同策略数据而不会引入不稳定的偏差。这使其在应对 DTDE范式的非平稳性时具有独特的鲁棒性，能够将对等方不断变化的行为视为环境噪声。在去中心化能源市场中，吞吐量比完美的策略对齐更重要；异步近端策略优化处理海量轻微异策略数据的能力，比严格遵循同策略的保守做法或脆弱的熵最大化带来更好的结果。

软演员-评论家呈现出两极化的性能表现：在全局评论家能够稳定学习的 CTCE 和 CTDE 中表现出色，但在 DTDE 中则表现不佳。软演员-评论家旨在实现最大熵，同时优化奖励和策略的随机性以进行探索。在复杂的本地能源市场格局中，软演员-评论家或许发掘出了细致入微的行为，而近端策略优化保守的信赖域往往会使智能体陷入次优的局部极小值。然而，在完全去中心化的环境中，纯粹的熵最大化十分危险；如果没有中央评论家提供的共同基准，对随机性的追求会加剧非平稳性，并导致发散的反馈回路。近端策略优化虽然是可靠的基线，但通常无法快速适应复杂的交互，在去中心化环境中的表现仅达到其基准性能的 94%。

基于这项分析，为每种范式推荐特定算法，以确保实现最大的潜力。对于 CTCE，APPO 是更优选择，因为它能够异步探索大规模联合状态-动作空间，达到 3272.1 的绝对最高奖励。对于 CTDE，SAC 是最优选择，因为集中式评论家与其熵最大化机制实现了完美互补，从而获得最稳定的学习曲线，标准差最低，仅为 42.5。最后，对于 DTDE，APPO 是占主导地位的选择，在很大程度上缩小了与集中式基准之间的差距（仅为 1%），而 PPO 和 SAC 则表现不佳或完全失败。APPO-DTDE 能够达到与 CTCE 相当的性能，这证明观测空间中的信息素信号足以使智能体实现最优协调，从而无需为这类本地能源市场承担集中式控制所需的基础设施成本和隐私风险。

5.2 系统级性能：经济与电网指标

为了全面验证所提出的隐式合作框架，有必要超越对奖励收敛的抽象分析，评估系统切实的物理和经济结果。本小节将对上一节确定的最佳配置进行严格的系统级性能分析：APPO-CTCE（基准）、SAC-CTDE（标准）以及 APPO-DTDE（所提出的解决方案）。该分析综合了详细的仿真记录、市场演变的描述性统计数据，以及对新兴商业网络的拓扑分析，从而量化分配效率、网络稳定性与运行自主性之间的权衡。

5.2.1 经济效率与市场动态

本地能源市场（LEM）的经济表现是衡量配置效率的首要指标，反映了该机制在最大化所有参与者总价值的同时，平衡本地供需的有效程度（见图 5）。

集中式基准（集中式训练与集中式执行，CTCE）凭借获取全局信息的优势，展现出更高的结构效率。描述性统计表明，CTCE 每回合创造的社会福利为 33,023.2 美元，标准差为 21,892.5 美元。这一较高的福利值表明，即使需求曲线各不相同，CTCE 也能持续创造经济剩余。CTCE 的结构效率还体现在市场流动性上：该配置促成的平均交易量为每步 237.9 千瓦时。

相比之下，CTDE范式实现的平均社会福利为 21,209.6 美元，市场交易量则显著降至 115.9 千瓦时。这表明，福利约减少 35%、交易量减少逾 50%，凸显了执行期间部分可观测所带来的代价。尽管智能体是在集中式条件下训练的，但由于缺乏实时全局信息，它们不得不采取更保守的竞价策略以避免惩罚，导致市场流动性下降，只有最有利可图的交易得以执行。

令人意外的是，完全去中心化的 DTDE 配置报告了最高的平均社会福利，达到 $44544.9。然而，对描述性统计数据的分析显示，这一性能波动极大，主要由价格稀缺驱动，而非

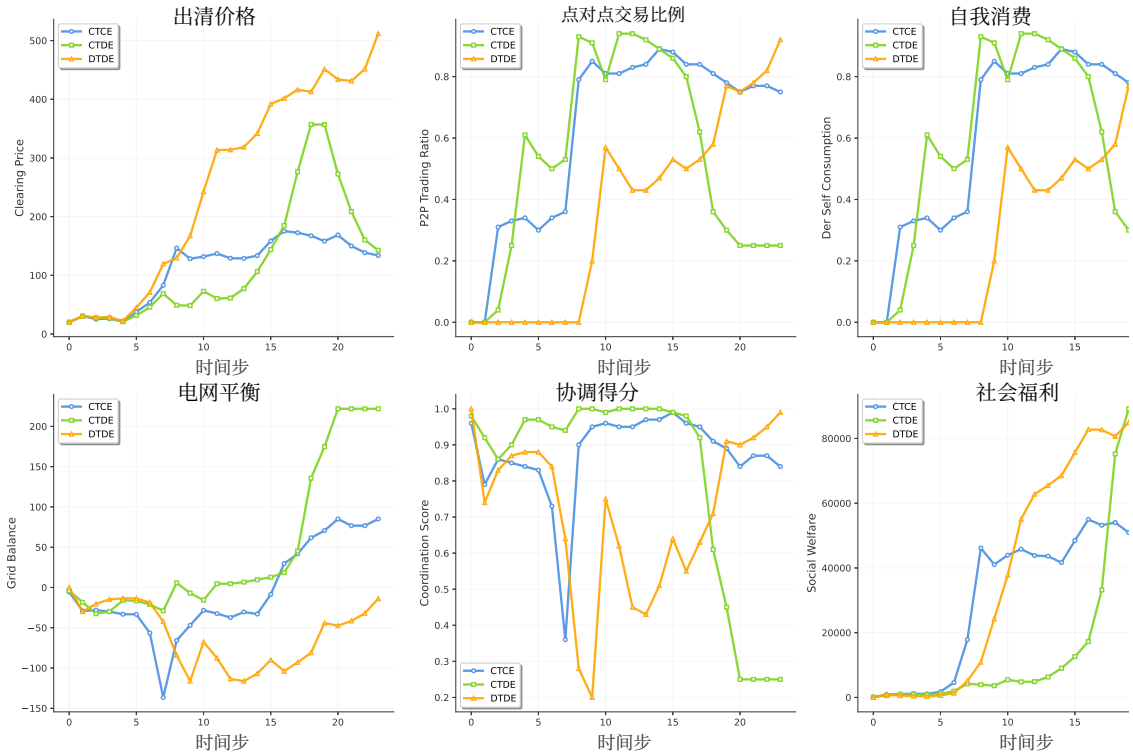


图5：已确定的最佳配置下的隐式合作关键绩效指标和市场动态：APPO- CTCE、SAC-CTDE 和 APPO-DTDE。

配置效率。DTDE 福利的标准差为 $36 472.3（几乎是 CTCE 的两倍），其性能下限则几乎为零（最小值为 $ 0.8，而 CTCE 为 $108.3）。这种极端方差表明，尽管 DTDE 智能体能够获得意外的利润（很可能是通过在需求高峰时段进行稀缺定价），但其表现不如集中式基线稳定。因此，较高的福利数据实际上是生产者剩余膨胀造成的假象，而非资源最优配置的体现。

市场出清价格统计数据进一步揭示了去中心化系统中固有的不确定性所带来的成本。CTCE 的平均市场价格最低且最稳定，为 110.7 美元/千瓦时，标准差为 56.1 美元。观察到的最高价格为 175.4 美元，这反映出市场效率很高，供给始终以发电边际成本满足需求。CTDE 的平均价格为 118.8 美元/千瓦时，保持了竞争力，与基准密切吻合。这种相似性表明，集中式评论家成功地教会智能体按照全局最优的方式评估能源价值，即便它们交易的电量较少。最后，DTDE 的价格波动极为剧烈，平均价格为 253.9 美元/千瓦时，是基准的两倍多，标准差为 174.2 美元，价格飙升至 511.8 美元/千瓦时。这种价格上涨反映了去中心化带来的风险溢价。由于没有中央协调器保障供应，去中心化智能体会更积极地竞价，以确保满足其关键能源需求，从而推高结算价格并增加消费者的成本。

5.2.2 电网稳定性：偏差-方差权衡

经济指标衡量激励因素，电网稳定性指标则衡量解决方案在物理层面上的可行性。电网平衡应保持接近于零，这表明本地发电量与本地需求完全匹配。对电网平衡时间演变的统计分析揭示了不同范式之间偏差（平均值的准确性）与方差（稳定性）之间的明显权衡。

CTCE 实现了近乎完美的平衡，平均值为 -4.5 千瓦，标准差适中，为 ±57.4 千瓦。不平衡范围（[-136.5，85.0]）围绕零值平衡，这证实了在完全信息下，系统能够有效地

将净负荷维持在零，同时保持适度波动。CTDE 表现出正偏差（平均值为 $+46.5$ 千瓦），其不稳定性在所有范式中最高（标准差为 ± 93.5 千瓦）。最大不平衡达到 221.8 千瓦（供应过剩），这表明，经过中央评论家训练但在本地执行的智能体，在夜间难以预测对等方动作的总体影响。这会导致过冲行为：多个智能体同时放电以满足其感知到的缺口，从而造成电网盈余。DTDE 表现出负偏差（平均值为 -58.0 千瓦），但其标准差在所有范式中最低（± 39.1 千瓦）。值得注意的是，DTDE 的最大电网平衡值为 -0.0 千瓦，这意味着去中心化智能体学会了严格偏向购电的策略（即，它们从不向电网净输出）。虽然这会造成持续缺口（依赖配电系统运营商），但较低的方差使 DTDE 成为电网运营商可高度预测的负荷。与 CTDE 反复无常的振荡不同，DTDE 稳定的净输入偏差很容易由配电系统运营商进行补偿。

5.2.3 资源利用与自主性

分布式能源资源的利用效率体现了智能体习得的运行策略，以及微电网所实现的自主性程度。点对点交易比例衡量社区自主性，其定义为通过本地交易满足的能源需求占总能源需求的比例，与从配电系统运营商进口的能源相对。与去中心化自然会促进独立性的假设相反，详细统计数据呈现出更为微妙的层级关系：集中式训练与集中式执行的平均比例最高，为 0.6；CTDE 的平均比例为 0.5；DTDE 的平均比例最低，为 0.4。

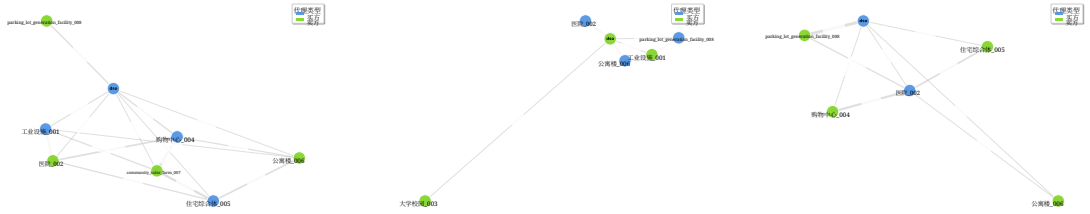
这一层级与协调得分相符（集中式训练与集中式执行 $0.9 >$ DTDE 0.7）。由于缺乏全局视图，DTDE 智能体会错失复杂的匹配机会，只能在本地满足 40% 的需求，其余 60% 则依赖配电系统运营商。这与电网平衡指标中观察到的 -58.0 千瓦净输入偏差直接对应。DTDE 智能体会在可能的情况下优先进行本地交易，但其有限的可视范围使其无法像集中式控制器那样有效地优化匹配过程。不过，在夜间时段，DTDE 实现了较高的点对点交易比例，这表明它们利用电池中储存的能源提供这些灵活性服务，这与那些时段较高的出清价格相一致

相同小时。

分布式能源自发自用效率与交易比例的变化趋势一致，从 CTCE 中的 0.6 降至 DTDE 中的 0.4。这证实了隐式协调虽然有助于维持稳定性，但其效率低于集中式控制。同样，CTCE 的平均灵活性利用率最高，为 23.3%，而 DTDE 为 9.0%。集中式控制器会主动对电池进行充放电，而去中心化智能体则倾向于让电池保持充电状态以确保安全，仅在本地信号强烈表明有此需要时才使用电池。

5.2.4 涌现式交易关系：网络拓扑分析

为了解释 点对点交易比例和市场流动性的差异，我们分析了图中可视化呈现的涌现网络拓扑 6.



(a) CTCE。 (b) CTDE。 (c) DTDE。

图 6：案例研究中的交易网络。蓝色圆圈代表买方，黑色圆圈代表卖方。边的粗细表示智能体之间交换的能源量。

CTCE 交易网络呈现出稠密且高度互联的拓扑结构。高容量智能体，例如社区太阳能农场和购物中心，充当中心枢纽，粗边将它们与几乎所有其他节点相连。全局可见性使中央控制器能够识别并执行物理上可行、但协调起来复杂的远距离套利机会。这种稠密的网状结构解释了较高的点对点交易比例（0.6）和市场流动性（237.9 千瓦时）。该系统有效释放了网络的全部容量。

CTDE 网络的连接呈现出减少的趋势。虽然枢纽智能体依然清晰可见，但与 CTCE 相比，外围连接（较小智能体之间的点对点交易）显著减少。尽管经过集中式训练，本地执——

cution会造成不确定性。智能体优先与主要枢纽开展高概率交易，而忽略与较小对等方之间风险更高的边际交易。这种结构性稀疏化与点对点交易比例降至0.5相关。

DTDE 网络呈现出稀疏且碎片化的拓扑结构。其图结构具有鲜明的社区集群特征，交易主要集中在电气邻居之间，贯穿网络的远距离链路则极少。这种碎片化是对拥塞惩罚信息素机制的一种涌现式响应。由于缺乏全局可见性，无法确保找到清晰的路径，去中心化智能体学会了规避风险，只与不太可能导致拥塞的邻居进行交易。这解释了高福利与低点对点交易比例（0.4）这一反直觉的组合。智能体在各自的本地集群中盈利颇丰（得益于高价格），却未能融入更广泛的电网，因此依赖配电系统运营商进行平衡，并导致电网指标中观察到的持续净输入偏差。

5.3 可扩展性测试

对于本地能源市场（LEM）中的实际部署而言，多智能体强化学习的可扩展性至关重要，因为参与的智能体数量可能存在显著差异。训练时间分析结果如图 7 所示，揭示了随着系统规模从 2 个智能体扩展至 34 个智能体，各种算法与范式组合呈现出不同的计算特征。

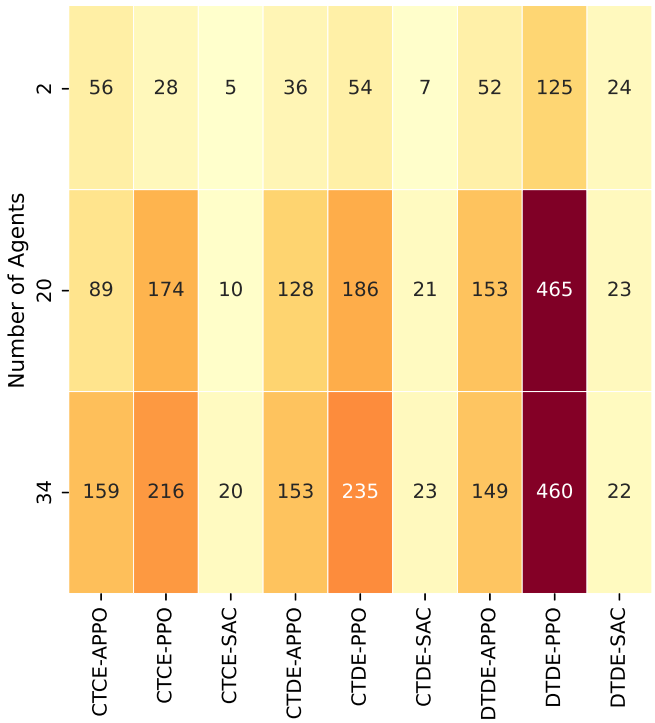


图 7：不同智能体数量下多智能体强化学习配置的计算可扩展性矩阵。

5.3.1 算法训练时间概况

结果表明，就收敛或达到上限所需的实际运行时间而言，SAC 是计算效率最高的算法。在所有范式中，SAC 始终保持最低的训练时间，通常快一个数量级。例如，在 DTDE 范式中，无论智能体数量如何变化，SAC 的训练时间都几乎保持不变，约为 22–24 分钟。不过，必须结合第 5.1 节前面提到的性能崩溃来理解这一速度；SAC 在 DTDE 中训练迅速，很可能反映的是它很早便未能学会，而非优化效率高。相比之下，在 SAC 表现出色的 CTDE 范式中，智能体数量从 2 个增至 34 个时，训练时间仅从 7 分钟增至 23 分钟，展现出出色的样本效率和稳定性。

APPO 展现出高度鲁棒且线性可扩展的特性。在 DTDE 范式中，APPO 的训练时间极为稳定：智能体数量从 2 个增至 34 个时，训练时间仅从 52 分钟增至 149 分钟。这支持了以下观点：APPO 是扩展去中心化系统的绝对首选，因为其异步特性使其能够高效处理高吞吐量数据，而不会像联合动作空间增大时通常出现的那样，导致耗时呈指数级增长。

PPO的可扩展性最差，尤其是在DTDE范式下。PPO的训练时间从125分钟（2个智能体）激增至460分钟（34个智能体）。这一显著的计算负担直接源于其严格的同策略特性和裁剪目标函数：为保持单调改进，算法需要更频繁且更缓慢地更新，因此不太适用于高维多智能体环境。

5.3.2 训练范式对效率的影响

训练范式的选择会带来不同的计算开销。与去中心化对应范式相比，CTCE下APPO的训练时间通常更长（34个智能体时为159分钟），这可能是因为在单一控制器中优化庞大的联合状态-动作空间更为复杂。

CTDE范式在APPO和PPO之间处于折中位置，但对于SAC而言仍是最优选择。不过，APPO与DTDE搭配时，在性能和可扩展性之间实现了最佳平衡。PPO难以应对DTDE中的移动目标问题（达到回合上限需要460分钟），而APPO凭借V-trace校正和异步探索，能够以近三分之一的时间（149分钟）收敛，同时实现接近最优的社会福利

5.3.3 计算性能总结

对于需要快速迭代和高稳定性的混合式（CTDE）设置，软演员-评论家是首选方案，前提是能够实现集中式评论家，以保持其 99.3% 的性能保留率和稳定的学习动态。对于大规模、完全去中心化部署（DTDE），异步近端策略优化是最可行的候选方案，因为它能够保持稳定的训练吞吐量并实现范式不变性，即使智能体数量增加，也能达到集中式方法的性能。相比之下，近端策略优化仍是一种计算成本高且较为保守的基线；尽管它可靠，但在智能体密度较高的市场中可能会慢到难以承受，且无法发现全局最优解。

6 结论与未来工作

向去中心化能源电网转型，亟需能够管理分布式能源资源激增的控制范式，同时避免陷入集中式调度固有的计算瓶颈和隐私问题。本文旨在验证，由多智能体强化学习和信息素信号机制赋能的隐式合作，无需显式点对点通信，即可使去中心化智能体近似实现中央控制器的最优行为。

通过严谨的 3×3 析因实验设计和补充的可扩展性测试，我们证明，去中心化智能体在其观测空间中获得适当的系统级关键绩效指标后，能够达到理论集中式基准协调质量的 91.7%，同时随系统规模线性扩展。这一发现从根本上挑战了普遍的假设：复杂的电网管理需要依赖庞大、集中的通信基础设施；它为保护隐私且可扩展的本地能源市场开辟了一条新路径。

本研究通过四项鲜明且重要的贡献，推动了去中心化能源管理领域的最新研究进展。首先，本研究系统评估了隐式合作，是首项专门剥离并比较不同训练范式（CTCE、CTDE、DTDE）和多智能体强化学习算法（PPO、APPO、SAC）对能源市场中协调行为涌现影响的研究。通过在现实的 IEEE 34节点拓扑下，采用相同的物理约束测试全部九种配置，本研究量化了去中心化在经济福利和电网平衡方面的成本。其次，本研究从理论上验证了柱头式协作在本地能源市场中的作用，并将这种协调机制形式化：将全局系统状态压缩为标量关键绩效指标信号，再广播给本地智能体。结果证明，增强观测空间足以让智能体重建作出合作决策所需的全局上下文，无需直接通信或模型共享即可有效解决部分可观测问题。第三，通过基准测试和可扩展性分析，本研究生成了一个兼顾最优性与可行性的性能矩阵。研究发现，APPO-DTDE 是最适合扩展的配置，其奖励收敛表现与集中式基准相当（达到 3243.5，而基准为 3272.1），同时保持了鲁棒的性能。最后，本研究为实际应用制定了运营设计指南，区分分配效率与网络可预测性，从而使网络运营商能够根据稳定性与效率方面的具体优先事项，选择不同的架构选项。

实验研究为我们理解多智能体强化学习如何应用于信息物理能源系统带来了新的认识。正如预期，集中式训练与集中式执行达到了性能上限，其点对点交易比例最高（0.6），电网平衡也近乎完美（平均值为 -4.5 千瓦）。然而，其超线性计算扩展性使其不适用于大规模实时控制。尽管 CTDE 通常是多智能体系统的标准范式，我们的结果显示，它虽保持了经济效率，却表现出最高的物理不稳定性（电网平衡标准差为 ±93.5 千瓦）。采用中央评论家训练、但在本地执行的智能体难以预测对等方的总体动作，导致危险振荡和过冲行为。相反，完全分散式学习并未导致协调崩溃。采用异步近端策略优化算法时，与集中式基准相比，性能差距为 8.3%。尤为重要的是，DTDE 的电网平衡方差最低（±39.1 千瓦）。尽管其持续存在净输入偏差（平均值为 -58.0 千瓦），但这种电网平衡使其负荷曲线对于配电系统运营商而言具有很高的可预测性和安全性。

就算法适用性而言，异步近端策略优化成为分散式环境中严格占优的算法。其V-trace校正机制使其能够鲁棒地处理独立学习者的非平稳性，从而收敛到标准近端策略优化所错过的合作均衡。尽管软演员-评论家因熵驱动的不稳定性而在完全去中心化设置中失败（差距为73.7%），但它在CTDE范式中表现优异。集中式评论家有效引导了软演员-评论家的探索，从而为受控微电网产生了最稳定且样本高效的策略。近端策略优化提供了一个安全但次优的基线。它避免了失败，但在总社会福利方面始终表现不佳，并且在去中心化设置中需要过多的计算时间（34个智能体超过460分钟）。

本研究证实，协调是奖励结构的一种涌现属性。通过将合作因子纳入奖励函数，智能体纯粹为了最大化其长期期望回报而自然地学会了互补策略。交易网络分析表明，这导致了空间聚类，智能体优先与电气邻居进行交易以最小化拥塞惩罚，从而在没有显式拓扑图的情况下有效地将电网自组织为稳定的纳米电网。

隐式合作通过将决策的复杂性与通信的复杂性解耦，改善了分布式能源资源管理。它使智能体能够自主化解个体利润最大化与集体电网稳定性之间的冲突；仿真结果表明，当网络均衡信号显示临界压力时，去中心化智能体会放弃短期套利机会。通过涌现的点对点交易，该系统在集中式配置中实现了高达62%的本地自发自用率，并在完全去中心化设置中保持了鲁棒的36.6%，与非合作基线方法相比，显著降低了对外部输电网的依赖。此外，该系统仅通过对价格和平衡信号的去中心化反应，便将净电网平衡维持在严格界限内；具体而言，去中心化方法（DTDE）实现了最稳定的物理特性曲线，最小化了电网注入功率的方差，从而降低了配电系统运营商的调节负担。这些效益是在没有显式通信的情况下实现的，验证了信息素信号机制是电力系统充分的协调机制。

这些发现对未来能源系统的设计、政策与部署具有重要启示。本地能源市场运营商无需投资昂贵的高带宽点对点通信网络。一个发布系统级关键绩效指标的可靠广播信道就足以诱导最优行为。在DTDE中观察到的膨胀性福利表明，稀缺价格是一种特征，它作为信号，促使在压力时期削减需求。监管机构应激励在自动交易智能体中纳入合作因子。电价与结算机制的设计不应仅基于交易量，还应基于资产行为与电网需求之间的相关性。

在实施建议方面，建议在以可扩展性和隐私为优先的竞争性市场中使用APPO-DTDE。它提供最大程度的隐私、线性可扩展性和物理稳定性，尽管会牺牲一定的分配效率。相反，对于以效率为优先的合作型微电网，建议使用SAC-CTDE。它能最大化经济效率，并更好地管理复杂的套利策略（更高的点对点交易比例），前提是集中式离线训练可行，且网络能够承受更大的方差。

尽管本研究为隐式合作奠定了严谨的基础，但要弥合这种合作与大规模实施之间的差距，仍有若干方向值得进一步探索。首要方向之一是将模拟规模从目前的8至34个智能体扩展至 100+ 个智能体，并采用高度随机的用户画像，以检验印迹信号饱和的极限。此外，还需研究信号延迟、丢包或恶意信号注入对合作均衡稳定性的影响；未来的研究应着重量化临界延迟阈值，即超过该阈值后隐式协调将会崩溃。最后，利用硬件在环设置验证模拟结果至关重要，这有助于评估隐式合作如何应对电压支撑和频率动态的物理特性；这些特性在当前潮流模型中被抽象处理。

参考文献

[1] F. Charbonnier, T. Morstyn, M. D. McCulloch，面向居民能源灵活性分布式控制的可扩展多智能体强化学习，Applied Energy 314 (2022) 118825。doi:10.1016/j.apenergy.2022.118825。[2] A. Aoun, M. Adda, A. Ilinca, M. Ghandour, H. Ibrahim，利用列昂惕夫投入产出模型分析电网韧性的集中式与去中心化方法，Energies 17 (6) (2024) 1321。doi:10.3390/en17061321。[3] J. Guerrero, A. C. Chapman, G. Verbic，低压网络中网络约束下的去中心化点对点能源交易，IEEE Transactions on Smart Grid 10 (5) (2019) 5163–5173。doi:10.1109/TSG.2018.2878445。[4] R. May, P. Huang，用于研究和优化点对点产消者能源市场的多智能体强化学习方法，Applied Energy 334 (2023) 120705。doi:10.1016/j.apenergy.2023.120705。[5] P. Wilk, N. Wang, J. Li，多智能体强化学习在智慧社区能源管理中的应用，Energies 17 (20) (2024) 5211。doi:10.3390/en17205211。[6] M. A. Hady, S. Hu, M. Pratama, Z. Cao, R. Kowalczyk，多智能体强化学习在资源分配优化中的应用：综述，Artificial Intelligence Review 58 (11) (2025) 354。doi:10.1007/s10462-025-11340-5。[7] J. Liang, H. Miao, K. Li, J. Tan, X. Wang, R. Luo, Y. Jiang，多智能体强化学习算法综述，Electronics 14 (4) (2025) 820。doi:10.3390/electronics14040820。[8] D. J. Moore，分层多智能体系统分类体系：设计模式、协调机制与工业应用（2025年8月）。arXiv:2508.12683，doi:10.48550/arXiv.2508.12683。[9] N. Salazar-Peña, A. Tabares, A. González-Mancera，MARLEM：用于去中心化本地能源市场隐式合作的多智能体强化学习仿真框架，SSRN（2025年11月）。doi:http://dx.doi.org/ 10.2139/ssrn.5705203。[10] Z. Guo, B. Qin, Z. Guan, Y. Wang, H. Zheng, Q. Wu，高效且激励相容的点对点能源交易机制，IEEE Transactions on Smart Grid 15 (1) (2024) 1075–1088。doi:10.1109/TSG.2023.3266809。[11] N. T. Mbungu, R. M. Naidoo, R. C. Bansal, V. Vahidinasab，微电网应用的最优智能能源协调概述，IEEE Access 7 (2019) 163063–163084。doi:10.1109/ACCESS.2019.2951459。[12] H. Nshuti，微电网的集中式与去中心化控制。[13] F. Charbonnier, T. Morstyn, M. D. McCulloch，电网边缘资源协调：系统综述与分类体系，Applied Energy 318 (2022) 119188。doi:10.1016/j.apenergy.2022.119188。[14] A. G. Givisiez, K. Petrou, L. F. Ochoa，输电系统运营商与配电系统运营商协调模型及求解技术综述，Electric Power Systems Research 189 (2020) 106659。doi:10.1016/j.epsr.2020.106659。[15] M. Shafie-khah, A. S. Gazafroudi（编），本地能源市场与能源社区中的交易：概念、结构与技术，Lecture Notes in Energy，第93卷，Springer International Publishing，Cham，2023。doi: 10.1007/978-3-031-21402-8。[16] S. Anand，电力市场稳定且可扩展的日前出清价格。[17] I. Azim, N. Al Khafaf, M. Khorasany, R. Razzaghi, M. Jalili, Wp2.7和2.8技术经济建模（现状与未来）：任务1：文献综述，技术报告，Centre for New Energy Technologies (C4NET)、莫纳什大学与皇家墨尔本理工大学（2024年2月）。[18] P. Raghoo, K. Shah，弥合点对点能源交易中的理论与实践差距：市场机制与技术创新，Environmental Research: Infrastructure and Sustainability 5 (1) (2025) 012001。doi:10.1088/ 2634-4505/adac8a。[19] A. Kantamneni, L. E. Brown, G. Parker, W. W. Weaver，微电网控制多智能体系统综述，Engineering Applications of Artificial Intelligence 45 (2015) 192–203。doi:10.1016/j.engappai.2015.07.005。[20] T. Alazemi, M. Darwish, M. Radi，可再生能源接入的输电系统运营商与配电系统运营商协调：系统性文献综述，Energies 15 (19) (2022) 7312。doi:10.3390/en15197312。[21] S. Zheng, Y. Zhou，第8章——采用先进定价与决策机制的点对点能源交易，载于：Y. Zhou, J. Yang, G. Zhang, P. D. Lund（编），建筑—交通综合能源系统的数字化与机器学习进展，爱思唯尔，2024年，第133–158页。doi:https://doi.org/10.1016/ B978-0-443-13177-6.00013-8。URL https://www.sciencedirect.com/science/article/pii/B9780443131776000138

[22] A. Papavasiliou, H. Hindi, D. Greene，电力需求响应的基于市场的控制机制，载于：第49届IEEE决策与控制会议（CDC），IEEE，美国佐治亚州亚特兰大，2010年，第1891–1898页。DOI:10.1109/ CDC.2010.5717572。[23] B. Satchidanandan, M. Roozbehani, M. A. Dahleh，需求响应市场的两阶段机制，IEEE控制系统快报7（2023年）49–54。DOI:10.1109/LCSYS.2022.3186654。[24] S. Khaskheli, A. Anvari-Moghaddam，本地能源市场中的能源交易：模型、求解策略与机器学习方法综述，《应用科学》14（24）（2024年）11510。DOI:10.3390/ app142411510。[25] Q. Huang, W. Amin, K. Umer, H. B. Gooi, F. Y. S. Eddy, M. Afzal, M. Shahzadi, A. A. Khan, S. A. Ahmad，交易型能源系统综述：概念与实施，能源报告7（2021年）7804–7824。DOI:10.1016/j. egyr.2021.05.037。[26] H. Zhou, B. Li, X. Zong, D. Chen，交易型能源系统：概念、配置与机制，能源研究前沿10（2023年）1057106。DOI:10.3389/fenrg.2022.1057106。[27] N. Wang, Z. Liu, P. Heijnen, M. Warnier，融合多能源耦合与协同行为的点对点市场机制，Applied Energy 311（2022年）118572。DOI:10.1016/j.apenergy.2022.118572。[28] E. Anderson, T. Cau，电力市场中的隐性合谋与个体市场势力，《欧洲运筹学杂志》211（2）（2011年）403–414。DOI:10.1016/j.ejor.2010.12.016。[29] S. Frey, A. Diaconescu, D. Menga, I. Demeure，面向异构多尺度、多目标智能微电网的通用全息控制架构，ACM自主与自适应系统汇刊10（2）（2015年）1–21。DOI:10.1145/2700326。[30] R. Schubotz, T. Spieldenner, M. Chelli，stigld：关联系统中的信息素协调，《数学》10（7）（2022年）。DOI:10.3390/math10071041。URL https://www.mdpi.com/2227-7390/10/7/1041[31] R. De Nicola, L. Di Stefano, O. Inverso，采用虚拟信息素的多智能体系统，《计算机编程科学》187（2020年）102345。DOI:10.1016/j.scico.2019.102345。[32] W. Deng, J. Xu, H. Zhao，一种基于混合策略的调度问题改进蚁群优化算法，IEEE Access 7（2019年）20281–20292。DOI:10.1109/ACCESS.2019.2897580。[33] C. Ziras，分布式能源资源在配电与输电系统中的集成，博士论文，丹麦技术大学（2019年）。[34] J. Sauter, R. Matthews, H. Parunak, S. Brueckner，多架无人飞行器数字信息素群集控制的演示，载于：Infotech@Aerospace，美国航空航天学会，美国弗吉尼亚州阿灵顿，2005年。DOI:10.2514/6.2005-7046。[35] D. Roca, D. Nemirovsky, M. Nemirovsky, R. Milito, M. Valero，物联网中的涌现行为：终极超大规模系统，IEEE Micro 36（6）（2016年）36–44。DOI:10.1109/MM.2016.102。[36] C. Binder, J.-A. Gross, C. Neureiter, G. Lastro，利用协同仿真研究电动汽车在智能电网中引发的涌现行为，载于：2019年第14届年度系统之系统工程会议（SoSE），IEEE，美国阿拉斯加州安克雷奇，2019年，第230–235页。DOI:10.1109/SYSOSE.2019.8753861。[37] M. Debe, H. R. Hasan, K. Salah, I. Yaqoob, R. Jayaraman，基于拍卖与信誉机制的电动汽车区块链能源交易，IEEE Access 9（2021年）165542–165556。DOI:10.1109/ACCESS.2021. 3133958。[38] R. Verborgh, A. Dimou, A. Hogan, C. d’Amato, I. Tiddi, A. Bröring, S. Mayer, F. Ongenae, R. Tommasini, M. Alam（编），《语义网：ESWC 2021卫星会议：线上会议，2021年6月6–10日，修订精选论文》，计算机科学讲义第12739卷，Springer International Publishing，Cham，2021年。DOI:10.1007/ 978-3-030-80418-3。[39] F. Lezama, J. Soares, P. Hernandez-Leal, M. Kaisers, T. Pinto, Z. Vale，本地能源市场：铺平通往完全交易型能源系统的道路，IEEE电力系统汇刊34（5）（2019年）4081–4088。DOI:10.1109/TPWRS.2018.2833959。[40] G. Tsaousoglou, J. S. Giraldo, N. G. Paterakis，本地电力市场机制：模型、解概念与算法技术综述，可再生与可持续能源评论156（2022年）111890。DOI:10.1016/j.rser.2021.111890。

[41] A. Paudel, K. Chaudhari, C. Long, H. B. Gooi，产消者社区微电网中的点对点能源交易：博弈论模型，IEEE工业电子汇刊66（8）（2019年）6087–6097。DOI: 10.1109/TIE.2018.2874578。[42] S. Talari, M. Khorasany, R. Razzaghi, W. Ketter, A. S. Gazafroudi，考虑异质偏好的去中心化点对点能源交易机制设计，《可持续城市与社会》87（2022年）104182。DOI:10.1016/ j.scs.2022.104182。[43] S. D. Ramchurn, D. Huynh, N. R. Jennings，多智能体系统中的信任，《知识工程评论》19（1）（2004年）1–25。DOI:10.1017/S0269888904000116。[44] Y. Wang, J. Vassileva，点对点网络中的信任与声誉模型，载于：第三届点对点计算国际会议论文集（P2P2003），IEEE计算机学会，瑞典林雪平，2003年，第150–157页。DOI: 10.1109/PTP.2003.1231515。[45] J. M. Pujol, R. Sangesa, J. Delgado，利用社交网络拓扑提取多智能体系统中的声誉，载于：第一届自主智能体与多智能体系统国际联合会议论文集，AAMAS '02，ACM，意大利博洛尼亚，2002年，第467–474页。DOI:10.1145/544741.544854。[46] V. J. Nair, P. Srivastava, V. Venkataramanan, P. S. Sarker, A. Srivastava, L. D. Marinovici, J. Zha, C. Irwin, P. Mittal, J. Williams, J. Kumar, H. V. Poor, A. M. Annaswamy，通过可信物联网协调资产增强电网韧性，《美国国家科学院院刊》122（8）（2025年）e2413967121。DOI:10.1073/pnas. 2413967121。[47] C. Yapa, C. De Alwis, M. Liyanage, J. Ekanayake，利用区块链化声誉管理服务提升智能电网2.0应用的性能，《工业信息集成杂志》39（2024年）100580。DOI:10.1016/j.jii.2024.100580。[48] L. Sang, H. Hexmoor，智能电网中基于区块链技术的基于声誉的共识，《国际网络安全及其应用杂志》14（4）（2022年）15–37。DOI:10.5121/ijnsa.2022.14402。[49] C. Yapa, C. De Alwis, U. Wijewardhana, M. Liyanage，基于区块链的声誉管理系统在智能电网2.0能源交易中的应用，载于：2023年IEEE拉丁美洲通信会议（LAT-INCOM），IEEE，巴拿马巴拿马城，2023年，第1–6页。DOI:10.1109/LATINCOM59467.2023.10361893。[50] G. Jøsang，稳健的信任与声誉系统面临的挑战。[51] A. Raza, E. Badidi, M. Hayajneh, E. Barka, O. E. Harrouss，面向智能电网、医疗保健与交通运输的区块链声誉与信任管理：综述，IEEE Access 12（2024年）196887–196913。DOI:10.1109/ACCESS.2024.3521428。[52] M. Sanayha, P. Vateekul，采用多个集群的点对点能源交易多智能体深度强化学习模型方法，IEEE Access 10（2022年）127882–127893。DOI:10.1109/ACCESS.2022.3224460。[53] S. Keren, C. Essayeh, S. V. Albrecht, T. Mortsyn，面向能源网络的多智能体强化学习：计算挑战、进展与待解问题。[54] D. Li, N. Lou, Z. Xu, B. Zhang, G. Fan，通过隐式共识生成实现多智能体强化学习中的高效通信。[55] C. Amato，合作式多智能体强化学习初探（2025年5月）。arXiv:2405. 06161，DOI:10.48550/arXiv.2405.06161。[56] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch，混合合作-竞争环境下的多智能体演员-评论家（2020年3月）。arXiv:1706.02275，DOI:10.48550/arXiv.1706.02275。[57] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov，近端策略优化算法（2017年8月）。arXiv:1707.06347，DOI:10.48550/arXiv.1707.06347。[58] S. Ahilan, P. Dayan，用于合作式强化学习的封建式多智能体层级结构（2019年1月）。arXiv: 1901.08492，DOI:10.48550/arXiv.1901.08492。[59] G. Wen, J. Fu, P. Dai, J. Zhou，DTDE：一种新型合作式多智能体强化学习框架，《The Innovation》2（4）（2021年）100162。DOI:10.1016/j.xinn.2021.100162。[60] M. Lauri, J. Pajarinen, J. Peters，通过策略图改进，在离散与连续状态的去中心化部分可观测马尔可夫决策过程中进行多智能体主动信息收集，自主智能体与多智能体系统34（2）（2020年）42。DOI:10.1007/s10458-020-09467-6。

[61] C. Amato、G. Chowdhary、A. Geramifard、N. K. Ure、M. J. Kochenderfer，部分可观测马尔可夫决策过程的分散式控制，载于：第52届 IEEE 决策与控制会议，IEEE，佛罗伦萨，2013年，第2398–2405页。doi:10.1109/CDC.2013.6760239。[62] C. Lou、Z. Jin、W. Tang、G. Geng、J. Yang、L. Zhang，LLM 增强的多智能体强化学习与专家工作流：用于实时点对点能源交易（2025年11月）。arXiv:2507.14995，doi:10.48550/arXiv.2507. 14995。[63] D. Wu、D. Kalathil、M. Begovic、L. Xie，基于深度强化学习的富含分布式能源资源配电电网鲁棒保护（2021年6月）。arXiv:2003.02422，doi:10.48550/arXiv.2003.02422。[64] D. Zhu、B. Yang、Y. Liu、Z. Wang、K. Ma、X. Guan，基于多智能体深度强化学习的多能工业园区能源管理，Applied Energy 311（2022年）118636。arXiv:2202.03771，doi:10.1016/ j.apenergy.2022.118636。[65] S. Stein、A. Chapman、Y. Du、B. Koohy、V. Yazdanpanah、P. Yolum，第二届以公民为中心的多智能体系统国际研讨会论文集 2024（C-MAS 2024）（2024年）。doi:10.6084/M9.FIGSHARE.25743057.V1。[66] T. Phan，多智能体强化学习中的涌现与韧性。[67] P. Mannion、S. Devlin、J. Duggan、E. Howley，面向基于知识的多目标多智能体强化学习的奖励塑形，《知识工程评论》33（2018年）e23。doi:10.1017/S0269888918000292。[68] S. Li、J. K. Gupta、P. Morales、R. Allen、M. J. Kochenderfer，面向多智能体强化学习的深度隐式协调图（2021年2月）。arXiv:2006.11438，doi:10.48550/arXiv.2006.11438。[69] H. Wang、B. Chen、T. Zhang、B. Wang，通过隐式通信信道学习通信（2025年2月）。arXiv:2411.01553，doi:10.48550/arXiv.2411.01553。[70] C. Feng，在具有物理约束的能源市场中实现分布式能源资源的去中心化整合。[71] T. Wolgast、A. Nieße，利用基于模型的强化学习近似能源市场出清与竞价，IEEE Access 12（2024年）145106–145117。doi:10.1109/ACCESS.2024.3472480。[72] H. Zhang、M. Montazeri、P. Heer、K. Kok、N. G. Paterakis，通过分层多智能体强化学习在本地市场中实施套利策略（2025年7月）。arXiv:2507.16479，doi:10.48550/arXiv.2507.16479。[73] J. Lu、Z. Xie、H. Xu、J. Liu，基于多任务多智能体深度强化学习的价格制定型负荷聚合商联合竞价与激励策略优化，IEEE Access 12（2024年）163988–164001。doi:10. 1109/ACCESS.2024.3491189。[74] L. Qiu，面向城市社区协同智能电网与建筑能源管理的多智能体强化学习。[75] K. Guo、N. Eckhert、K. Chhajer、L. Abeykoon、L. Schell，AutoGrid AI：用于自主微电网管理的深度强化学习框架（2025年9月）。arXiv:2509.03666，doi:10.48550/arXiv.2509.03666。[76] Z. Feng、J. An、M. Han、X. Ji、X. Zhang、C. Wang、X. Liu、L. Kang，办公建筑能源消耗预测：由改进白鲸优化算法驱动的自适应长短期记忆网络，《建筑工程学报》91（2024年）109612。doi:10.1016/j.jobe.2024.109612。[77] J. Li、M. Motoki、B. Zhang，能源市场中的聚合商的战略与公平互动：多智能体动态与拟凹博弈（2024年10月）。arXiv:2410.11296，doi:10.48550/arXiv.2410.11296。[78] E. Eusébio、J. A. M. Sousa、M. Ventim Neves，能源市场价格波动对聚合商投资组合的影响，载于：L. M. Camarinha-Matos、A. J. Falcão、N. Vafaei、S. Najdi（编），信息物理系统技术创新，第470卷，Springer International Publishing，Cham，2016年，第437–445页。doi:10.1007/978-3-319-31165-4_41。[79] A. Bashyal、T. Boroukhian、P. Veerachanchai、M. Naransukh、H. Wicaksono，基于多智能体深度强化学习的重工业离散制造系统需求响应与能源管理，Applied Energy 392（2025年）125990。doi:10.1016/j.apenergy.2025.125990。[80] Y. Gao、X. Yuan、S. Wang、L. Chen、Z. Zhang、T. Wang，Flash-Attention 增强的多智能体深度确定性策略梯度，用于数字孪生驱动的物联网中的移动边缘计算，《数学》13（13）（2025年）2164。doi:10.3390/math13132164。

[81] J. Källström，基于模拟训练的效用提升强化学习，林雪平科技研究丛书。学位论文，第2351卷，林雪平大学电子出版社，林雪平，2023年。doi:10.3384/ 9789180753678。[82] C. Mylonas、E. Varvarigos、G. Tsaousoglou，从分布式能源资源中安全地自底向上提供灵活性（2025年4月）。arXiv:2504.20529，doi:10.48550/arXiv:2504.20529。[83] D. Zhu、B. Yang、Y. Liu、Z. Wang、K. Ma、X. Guan，基于多智能体深度强化学习的多能工业园区能源管理，Applied Energy 311（2022年）118636。arXiv:2202.03771，doi:10.1016/ j.apenergy.2022.118636。[84] D. D. Sharma、R. C. Bansal，电气工程系，MJP Rohilkhnad大学，巴雷利；电气工程系，沙迦大学，沙迦，阿拉伯联合酋长国；电气、电子与计算机工程系， 比勒陀利亚大学，比勒陀利亚，南非，基于 LSTM-SAC 强化学习的联网微电网系统韧性能源交易，《AIMS电子与电气工程》9（2）（2025年）165–191。doi:10.3934/electreng.2025009。[85] N. Azizan Ruhi、K. Dvijotham、N. Chen、A. Wierman，聚合商在电力市场中进行价格操纵的机会，《IEEE Transactions on Smart Grid》9（6）（2018年）5687–5698。doi:10.1109/TSG.2017.2694043。[86] F. Charbonnier、B. Peng、J. Vienne、E. Stai、T. Morstyn、M. McCulloch，集中式排练式分散协作：用于居民能源灵活性可扩展协调的多智能体强化学习，Applied Energy 377（2025年）124406。doi:10.1016/j.apenergy.2024.124406。[87] D. Qiu、J. Wang、Z. Dong、Y. Wang、G. Strbac，面向点对点多能交易的均值场多智能体强化学习，《IEEE电力系统汇刊》38（5）（2023年）4853–4866。doi:10.1109/TPWRS.2022. 3217922。[88] L. Velvis，可再生能源谷低碳调度的混合多智能体强化学习。[89] T. Qi、S. Chen、J. Zhang，生成模型增强的电动汽车充电导航多智能体强化学习方法。[90] M.-Q. Dao、J. S. Berrio、V. Frémont、M. Shan、E. Héry、S. Worrall，实用协同感知：用于异步三维目标检测的多智能体框架，《IEEE智能交通系统汇刊》25（9）（2024年）12163–12175。doi:10.1109/TITS.2024.3371177。[91] D. Muyizere、L. K. Letting、B. B. Munyazikwiye，通信信号延迟对电网的影响：综述，Electronics 11（6）（2022年）874。doi:10.3390/electronics11060874。[92] T. Ikeda、T. Shibuya，具有通信延迟的协作式多智能体系统的集中式训练与分散式执行强化学习，载于：2022年第61届日本计测自动控制学会年度会议（SICE），IEEE，日本熊本，2022年，第135–140页。doi:10.23919/SICE56594.2022. 9905866。[93] H. Yang、J. Chen、M. Siew、T. Lorido-Botran、C. Joe-Wong，具有自适应层次化知识图谱的 LLM 驱动去中心化生成式智能体：用于协作规划（2025年2月）。arXiv:2502.05453，doi:10.48550/ arXiv.2502.05453。[94] D. Hu、Z. Li、Z. Ye、Y. Peng、W. Xi、T. Cai，用于配电系统去中心化电压-无功控制的多智能体图强化学习，《国际电力与能源系统期刊》155（2024年）109531。doi: 10.1016/j.ijepes.2023.109531。[95] M. H. Rothkopf，维克里-克拉克-格罗夫斯机制并不实用的十三个理由，《运筹学》55（2）（2007年）191–197。doi:10.1287/opre.1070.0384。[96] D. Qiu、Y. Ye、D. Papadaskalopoulos、G. Strbac，点对点能源交易的可扩展协同管理：一种多集群深度强化学习方法，Applied Energy 292（2021年）116940。doi:10.1016/j.apenergy. 2021.116940。[97] R. Musca、E. Riva Sanseverino，电力系统广域监测与控制的点对点分布式算法，Energies 18（15）（2025年）3972。doi:10.3390/en18153972。[98] S. Gronauer、K. Diepold，多智能体深度强化学习：综述，Artificial Intelligence Review 55（2）（2022年）895–943。doi:10.1007/s10462-021-09996-w。

[99] E. Ye, R. Tao, N. Jaques, 《一种用于多智能体社会学习的高效开放世界环境》（2025年8月）。arXiv: 2508.15679，DOI:10.48550/arXiv.2508.15679。[100] R. Donatus, K. Ter, O.-O. Ajayi, D. Udekwe, 《智能交通系统中的多智能体强化学习：全面综述》（2025年8月）。arXiv:2508.20315, DOI:10.48550/arXiv.2508.20315。[101] P. Yadav, A. Mishra, S. Kim, 《网联自动驾驶车辆多智能体强化学习全面综述》，Sensors 23 (10) (2023) 4710。DOI:10.3390/s23104710。[102] V. Dvorkin, J. Kazempour, L. Baringo, P. Pinson, 《电力市场战略性发电投资的共识交替方向乘子法》（2018年10月）。arXiv:1711.04620，DOI:10.48550/arXiv.1711.04620。[103] R. H. Wuijts, M. van den Akker, M. van den Broek, 《新的高效机组组合问题交替方向乘子法算法》（2023年11月）。arXiv:2311.13438, DOI:10.48550/arXiv.2311.13438。[104] T. Gao, S. Wang, X. Geng, L. Zhang, M. Jing, T. Yu, J. Yu, 《区块链工业物联网数据市场的斯塔克尔伯格博弈定价》，载于：2025年第28届计算机支持的协同工作与设计国际会议（CSCWD），IEEE，法国贡比涅，2025年，第1776–1781页。DOI:10.1109/CSCWD64889.2025.11033563。[105] L. Cheng, P. Huang, M. Zhang, R. Yang, Y. Wang, 《运用博弈论方法优化电力市场：对购电和发电企业的战略与政策启示》，《数学》13 (3) (2025) 373。DOI: 10.3390/math13030373。[106] J. Guerrero, D. Gebbran, S. Mhanna, A. C. Chapman, G. Verbič, 《构建交易型能源系统以整合分布式能源资源：家庭能源管理、分布式最优潮流和点对点能源交易》，可再生与可持续能源评论 132 (2020) 110000。DOI:10.1016/j.rser.2020.110000。[107] A. Baringo, L. Baringo, J. M. Arroyo, 《不确定条件下虚拟电厂在能源与备用电力市场中的日前自调度》，IEEE电力系统汇刊 34 (3) (2019) 1881–1894。DOI:10.1109/ TPWRS.2018.2883753。[108] C. Ghenai, L. A. Husein, M. Al Nahlawi, A. K. Hamid, M. Bettayeb, 《能源领域数字孪生技术的最新趋势：全面综述》，Sustainable Energy Technologies and Assessments 54 (2022) 102837。DOI: 10.1016/j.seta.2022.102837。[109] J. Guerrero, B. Sok, A. C. Chapman, G. Verbič, 《由电气距离驱动的低压网络点对点能源交易》，Applied Energy 287 (2021) 116598。DOI:10.1016/j.apenergy.2021.116598。[110] N. Harder, R. Qussous, A. Weidlich, 《切合实际：利用多智能体深度强化学习对批发电力市场进行真实建模》，Energy and AI 14 (2023) 100295。DOI:10.1016/j.egyai.2023.100295。[111] N. Harder, A. Weidlich, P. Staudt, 《利用深度强化学习为电力市场模型中的储能单元寻找个体策略》，Energy Informatics 6 (S1) (2023) 41。DOI:10.1186/s42162-023-00293-0。[112] A. Rizki, A. Touil, A. Echchatbi, R. Oucheikh, M. Ahlaqqach, 《基于强化学习的近端策略优化方法：求解经济调度问题》，载于：第1届工业与物流工程智能管理国际会议（SMILE 2025），MDPI，2025年，第24页。DOI:10.3390/engproc2025097024。[113] M. Alonso, H. Amaris, D. Martin, A. de la Escalera, 《用于电动汽车和光伏储能单元能源管理的近端策略优化》，Energies 16 (15) (2023)。DOI:10.3390/en16155689。URL https://www.mdpi.com/1996-1073/16/15/5689[114] M. Anwar, C. Wang, F. de Nijs, H. Wang, 《基于近端策略优化的强化学习：能源与调频市场中的联合竞价》，载于：2022年IEEE电力与能源学会年会（PESGM），2022年，第1–5页。arXiv:2212.06551，DOI:10.1109/PESGM48719.2022.9917082。[115] A. C. Rigby, M. Wagner, D. M. Mikkelson, B. Lindley, 《利用强化学习增强核电站瞬态运行中的传统PID控制》。[116] S. Hasheminasab, A. Lotfy, M. Alzayed, H. Chaoui, 《基于近端策略优化的控制器：增强微电网中的功率分配》，IEEE Access 13 (2025) 162391–162405。DOI:10.1109/ACCESS.2025.3610451。[117] M. Foroughi, S. Maharjan, Y. Zhang, F. Eliassen, 《网络化微电网中的自主点对点能源交易：一种分布式深度强化学习方法》，载于：2023年IEEE电力与能源学会中东创新智能电网技术会议（ISGT Middle East），IEEE，阿布扎比，阿拉伯联合酋长国，2023年，第1–5页。DOI:10.1109/ISGTMiddleEast56437.2023.10078444。

[118] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, 《软演员-评论家：采用随机演员的异策略最大熵深度强化学习》（2018年8月）。arXiv:1801.01290，DOI:10.48550/arXiv.1801.01290。[119] T. Li, W. Cui, N. Cui, 《基于软演员-评论家算法的插电式混合动力汽车能源管理策略》，《世界电动汽车杂志》13 (10) (2022) 193。DOI:10.3390/wevj13100193。[120] K. Cheruiyot, N. Kiprotich, V. Kungurtsev, K. Mugo, V. Mwirigi, M. Ngesa, 《多智能体强化学习综述：联邦学习以及合作与非合作分散式机制》（2025年7月）。arXiv:2507. 06278，DOI:10.48550/arXiv.2507.06278。[121] C.-H. Chao, C. Feng, W.-F. Sun, C.-K. Lee, S. See, C.-Y. Lee, 《基于能量的归一化流实现最大熵强化学习》（2024年10月）。arXiv:2405.13629, DOI:10.48550/arXiv.2405.13629。[122] H. Sun, Y. Hu, J. Luo, Q. Guo, J. Zhao, 《采用稳态软演员–评论家深度强化学习方法增强暖通空调控制系统》，《建筑》15 (4) (2025) 644。DOI:10.3390/buildings15040644。[123] J. Gao, Y. Li, B. Wang, H. Wu, 《采用改进的多智能体软演员-评论家算法进行多微电网协同优化调度》，Energies 16 (7) (2023) 3248。DOI:10.3390/en16073248。[124] C. Liu, D. Li, 《基于非固定值裁剪的多智能体近端策略优化》，载于：2023年第12届IEEE数据驱动控制与学习系统会议（DDCLS），IEEE，中国湘潭，2023年，第1684–1688页。DOI:10.1109/DDCLS58216.2023.10167264。[125] P. Sun, W. Zhou, H. Li, 《注意力经验回放》，AAAI人工智能会议论文集 34 (04) (2020) 5900–5907。DOI:10.1609/aaai.v34i04.6049。[126] B. De Mol, D. Barbieri, J. Viebahn, D. Grossi, 《用于电网拓扑控制的集中协调多智能体强化学习》，载于：第16届ACM未来与可持续能源系统国际会议论文集，ACM，荷兰鹿特丹，2025年，第460–475页。DOI:10.1145/3679240.3734602。[127] H. Xu, Q. Fang, C. Hu, Y. Hu, Q. Yin, 《MIRA：针对多智能体系统非平稳性的基于模型的想象轨迹增强》，《数学》10 (17) (2022) 3059。DOI:10.3390/math10173059。[128] T. Ramachandran, K. Dvijotham, K. Kalsi, 《批发与零售电力市场的整合：建模与稳定性分析》，技术报告 PNNL-26469，1812546（2017年4月）。DOI:10.2172/1812546。[129] D. Holmberg, M. Burns, S. Bushby, A. Gopstein, T. McDermott, Y. Tang, Q. Huang, A. Pratt, M. Ruth, F. Ding, Y. Bichpuriya, N. Rajagopal, M. Ilic, R. Jaddivada, H. Neema, 《NIST交易型能源建模与仿真挑战赛第二阶段最终报告》，技术报告 NIST SP 1900-603，美国国家标准与技术研究院，马里兰州盖瑟斯堡（2019年5月）。DOI:10.6028/NIST.SP.1900-603。[130] E. B. T. Tchuisseu, D. Gomila, D. Brunner, P. Colet, 《动态需求控制设备对电网频率的影响》，《物理评论E》96 (2) (2017) 022302。DOI:10.1103/PhysRevE.96.022302。[131] D. Schraudner, 《语义物联网中的基于信息素的多智能体系统》，2021年，第218–229页。DOI:10.1007/ 978-3-030-80418-3_34.[132] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, K. Kavukcuoglu, 《神经网络的基于种群的训练》（2017年）。arXiv:1711. 09846。URL https://arxiv.org/abs/1711.09846