多智能体强化学习（MARL）在多智能体系统（MAS）中的应用：理论、算法、应用与展望

2025年4月26日　17 分钟阅读

#AI #多智能体 #强化学习 #MARL #MAS #论文 #技术

> *本文介绍了多智能体强化学习（MARL）在多智能体系统（MAS）中的应用：理论、算法、应用与展望。*

摘要

多智能体系统（Multi-Agent Systems, MAS）在机器人协作、自动驾驶、智能电网、金融市场等众多领域广泛存在。为了在这些系统中实现智能体的有效协调与决策，多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）已成为一项关键的使能技术。MARL 扩展了单智能体强化学习的范式，专注于解决多个学习智能体在共享环境中交互时的复杂动态问题。然而，MARL 面临着独特的理论和实践挑战，主要包括非平稳性、可扩展性、部分可观察性以及信用分配难题。为了应对这些挑战，研究界提出了多种算法范式，其中中心化训练与去中心化执行（Centralized Training with Decentralized Execution, CTDE）因其在平衡学习效率与执行约束方面的优势而成为主流。在此框架下，值分解方法（如 VDN, QMIX）和多智能体 Actor-Critic 方法（如 MADDPG, MAPPO）等代表性算法被广泛研究和应用。同时，MARL 在机器人集群控制、交通信号优化、策略游戏 AI、网络路由、算法交易等场景展现出巨大潜力。PettingZoo、RLlib、OpenSpiel、EPyMARL、MARLlib、JaxMARL 等开源框架为 MARL 的研究与开发提供了有力支持。尽管取得了显著进展，MARL 在理论完备性、算法鲁棒性、大规模应用的可行性以及安全性、可解释性等方面仍有待突破。未来的研究将聚焦于提升算法的可扩展性与泛化能力、设计更有效的协作与通信机制、深化理论基础，并探索与大语言模型等其他 AI 技术的融合，以推动 MARL 在解决现实世界复杂问题中发挥更大作用。

引言

MARL 的兴起与重要性

多智能体系统（MAS）已渗透到我们日常生活的方方面面，从协同工作的机器人集群到自动驾驶车队，再到智能电网的调度系统，多个决策实体在共享环境中交互以实现各自或共同的目标 1。这些系统的复杂性和动态性对智能体的决策能力提出了极高要求，需要它们不仅能理解环境，还能适应其他智能体的行为，进行有效的协调与合作 1。

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #多智能体 #MAS #论文 #技

在这样的背景下，多智能体强化学习（MARL）应运而生，并迅速成为人工智能领域的一个关键研究方向 1。MARL 建立在单智能体强化学习（RL）的基础之上，其目标是让智能体通过与环境的交互学习最优策略以最大化累积回报 1。然而，MARL 并非简单的智能体数量叠加，它引入了单智能体场景所不具备的独特复杂性 2。多个智能体的并发学习和交互使得环境从单个智能体的角度看是动态变化的（非平稳性），并且智能体之间可能存在合作、竞争或混合关系，需要进行复杂的策略协调。

从单智能体 RL 到 MARL 的转变，不仅仅是数量上的增加，更是性质上的根本改变。单智能体 RL 通常面对一个（假定）固定的环境进行优化，而 MARL 中的每个智能体在优化的同时，其他智能体也在进行优化并改变其策略 2。这种相互影响使得任何单个智能体所处的环境都变得非平稳，这要求我们超越传统的马尔可夫决策过程（MDP）框架，引入博弈论等工具进行战略性推理 1。因此，MARL 的研究不仅涉及机器学习，还与博弈论、控制论、社会学等多个学科交叉融合 1。

报告目标与结构概览

本报告旨在对 MARL 在 MAS 中的应用进行一次全面、深入、专家级的调研与分析。报告将系统梳理 MARL 的理论基础，剖析其核心概念与挑战；详细介绍主流的 MARL 算法范式及其代表性算法；广泛调研 MARL 在机器人、自动驾驶、游戏、资源管理、金融等领域的实际应用案例；评估当前流行的 MARL 开源研究工具；深度剖析 MARL 面临的核心挑战与局限性；并基于现有研究，评估 MARL 技术的发展水平，展望未来的研究方向和趋势。报告结构安排如下：首先阐述 MARL 的理论基础，随后详细介绍主流算法，接着展示广泛的应用场景，然后评估相关开源工具，之后深入分析面临的挑战，最后评估技术成熟度并展望未来。

MARL 理论基础

从单智能体 RL 到 MARL：关键差异

单智能体强化学习（Single-Agent RL）通常基于马尔可夫决策过程（Markov Decision Process, MDP）进行建模 15。一个 MDP 由状态空间 S、动作空间 A、状态转移概率 $P(s'|s, a)$ 和奖励函数 $R(s, a, s')$ 构成。智能体的目标是学习一个策略 $\pi(a|s)$，即在状态 s 下选择动作 a 的概率分布，以最大化长期累积折扣奖励 1。智能体通过与环境的交互循环（观察状态、执行动作、获得奖励、转移到新状态）来学习 1。

相比之下，MARL 处理的是多个智能体在共享环境中交互的场景 2。关键差异在于，一个智能体的动作不仅影响自身的状态和奖励，还会改变环境状态，进而影响其他智能体接收到的观察、奖励以及它们的后续决策 5。这种相互依赖性是 MARL 的核心特征。

如前所述，这种智能体间的相互影响导致了 MARL 相较于单智能体 RL 的一个根本性转变：环境的非平稳性（Non-stationarity）2。当所有智能体同时学习和调整策略时，从任何一个智能体的角度来看，环境的动态特性（状态转移概率和奖励函数）都在不断变化，因为其他智能体的策略是环境动态的一部分 2。这直接违反了标准 RL 算法所依赖的马尔可夫性质（即未来只依赖于当前状态和动作）2，使得学习过程变得不稳定，收敛性难以保证。这个问题被称为“移动目标问题”（moving-target problem）2。因此，MARL 需要更复杂的模型和算法来处理这种动态交互和策略适应过程。

核心概念解析

为了更精确地描述和分析 MARL 问题，需要引入一些核心概念：

随机博弈 / 马尔可夫博弈 (Stochastic Games / Markov Games): 这是 MARL 的标准形式化框架，是对 MDP 的多智能体扩展 13。一个马尔可夫博弈（MG）或随机博弈（SG）通常定义为一个元组，包含：

一组有限的智能体 N={1,...,n}。

一个状态空间 S，描述环境的全局状态。

每个智能体 i 的动作空间 Ai，联合动作空间为 A=A1×...×An。

状态转移概率函数 $P: S \times A \times S \rightarrow $，定义了在状态 s 下采取联合动作 a=(a1,...,an) 后转移到状态 s′ 的概率。

每个智能体 i 的奖励函数 Ri:S×A×S→R，定义了智能体 i 在状态转移后获得的即时奖励。

一个折扣因子 γ∈，其中 π−i 表示除智能体 i 外所有其他智能体的联合策略 13。

联合动作空间 (Joint Action Space): 指所有智能体可能采取的动作组合的空间 A=A1×...×An 2。其大小随着智能体数量 n 呈指数级增长，即 ∣A∣=∏i=1n∣Ai∣。这是导致 MARL 算法（尤其是中心化方法）面临"维度灾难"（curse of dimensionality）的主要原因之一，使得搜索最优联合策略变得极其困难 2。

部分可观察性 (Partial Observability - Dec-POMDP): 在许多现实世界的 MAS 中，智能体无法获取完整的环境状态 s，而只能接收到局部的、可能带有噪声的观察 oi∈Ωi 2。这种设定被称为分散式部分可观察马尔可夫决策过程（Decentralized Partially Observable Markov Decision Process, Dec-POMDP）16。Dec-POMDP 通常由一个元组 ⟨N,S,A,P,R,Ω,O,γ⟩ 定义，其中 Ω=Ω1×...×Ωn 是联合观察空间，O:S×A→Δ(Ω) 是观察函数 49。在这种情况下，智能体需要基于其动作-观察历史 τi=(oi0,ai0,...,oit) 来做出决策，即策略形式为 πi:Ti→Δ(Ai)，其中 Ti 是智能体 i 的历史空间 49。部分可观察性极大地增加了学习的难度，因为智能体需要根据不完整的信息推断隐藏状态 33。

非平稳性 (Non-Stationarity): 如前所述，这是 MARL 的核心挑战之一 1。由于所有智能体都在同时学习和适应，每个智能体所面对的"环境"（包括其他智能体）是动态变化的 2。这意味着从单个智能体的角度看，状态转移和奖励函数不再是固定的，违反了标准 RL 算法的平稳性假设 2。这使得基于经验回放（experience replay）的离策略（off-policy）学习方法尤其困难，因为存储的经验可能很快就"过时"了 22。

博弈论视角

MARL 与博弈论（Game Theory）紧密相关 1。博弈论提供了分析多个理性决策者交互的数学工具，而 MARL 则侧重于智能体如何通过学习（通常是试错）来达到某种（可能是博弈论意义上的）均衡策略 13。理解智能体之间的交互性质至关重要：

合作博弈 (Cooperative Games): 所有智能体拥有共同的目标和共享的奖励函数，需要协同工作以最大化团队的整体回报 5。例如，多机器人协作搬运、团队体育竞技模拟。

竞争博弈 (Competitive Games): 智能体的利益完全对立，通常是零和博弈（zero-sum game），一个智能体的收益等于其他智能体的损失 5。例如，棋类游戏（如围棋、国际象棋）、两人对抗游戏（如星际争霸）。

混合博弈 (Mixed Games): 包含合作和竞争元素，智能体可能需要在不同目标或与其他智能体的关系之间进行权衡 5。例如，社会困境（如囚徒困境、懦夫博弈）、谈判、交通网络中的车辆交互。

在 MARL 中，常用的博弈论解概念包括：

纳什均衡 (Nash Equilibrium, NE): 一种策略组合，其中没有任何一个智能体可以通过单方面改变自身策略而获得更好的收益 13。在对抗博弈中，寻找 NE 是一个核心目标 15。然而，MARL 系统可能存在多个 NE，并且可能收敛到次优的 NE（例如，相对过泛化问题）1。

帕累托效率 (Pareto Efficiency): 指一种状态，不可能在不损害至少一个智能体利益的情况下改善任何一个智能体的状况 2。在合作博弈中，通常期望达到帕累托最优的 NE。

MARL 的研究不仅关注如何达到这些均衡，更关注智能体在动态交互和学习过程中如何演化出这些策略，以及如何解决多均衡选择、协调失败等问题 1。

主流 MARL 算法范式与详解

概述

针对 MARL 的独特挑战，研究者们开发了多种算法范式。这些范式试图在去中心化执行的需求（通常由物理限制或通信带宽引起 2）与学习过程的稳定性、效率之间取得平衡。核心目标是为每个智能体 i 学习一个策略 πi，以最大化其（或团队的）期望回报 Vπ 15。

独立学习者 (Independent Learners - IL)

原理: 这是最简单直接的方法，每个智能体将其他智能体视为环境的一部分，并独立地运行一个单智能体 RL 算法来学习自己的策略 22。例如，每个智能体可以独立运行 Q-learning（称为 Independent Q-Learning, IQL）18 或 PPO（称为 Independent PPO, IPPO）59。

优势: 实现简单，计算开销相对较低，具有良好的可扩展性，因为智能体之间没有显式的协调或通信开销 32。

劣势: 主要问题在于严重受到非平稳性挑战的影响。由于其他智能体的策略在不断变化，每个智能体感知的环境动态也在变化，这违反了单智能体 RL 的核心假设，导致学习不稳定，缺乏收敛保证 2。此外，由于忽略了智能体间的交互，难以学习到复杂的协调策略 47。经验表明，其性能在不同任务上表现不一，有时效果尚可，有时则很差 43。

适用场景: 主要用于相对简单的任务，或者智能体间交互不那么密集、协调要求不高的场景。也常作为评估更复杂 MARL 算法性能的基线 43。

中心化训练与去中心化执行 (Centralized Training with Decentralized Execution - CTDE)

原理: CTDE 是目前 MARL 领域最流行和研究最广泛的范式 63。其核心思想是在**训练阶段**利用中心化的信息（例如全局状态、所有智能体的动作、观察、奖励甚至策略参数）来指导和稳定学习过程，从而更好地处理非平稳性和信用分配等问题 2。然而，在**执行阶段**，每个智能体仅根据其自身的局部观察（和历史）来独立决策，不依赖于中心化的信息或与其他智能体的实时通信 63。

优势:

缓解非平稳性: 在训练中获取全局信息有助于稳定学习目标 2。

促进协调: 中心化训练使得学习协调策略成为可能。

保持执行效率: 去中心化执行满足了许多现实应用中通信受限、需要快速响应的要求，并保持了较好的可扩展性 63。

平衡性能与可扩展性: CTDE 在学习性能和系统可扩展性之间提供了一个有效的折衷方案 63。

劣势:

训练要求: 需要一个能够提供额外信息的中心化训练环境（如模拟器），这在某些完全在线或无法进行模拟的场景中可能不可行 63。

设计复杂性: 如何有效利用中心化信息以及如何将其转化为去中心化策略是 CTDE 算法设计的关键难点。

异质性挑战: 处理具有不同状态或动作空间的异质智能体时可能表现不佳 41。

理论争议: 关于中心化评论家（critic）是否真的能促进合作，还是仅仅稳定训练过程，存在一些讨论 75。

CTDE 范式的广泛采用反映了 MARL 领域的一个基本现实：纯粹的去中心化学习（如 IL）往往因非平稳性而失败，而纯粹的中心化学习与执行（CTE）则因联合状态-动作空间的指数级增长而面临严重的可扩展性问题 32。CTDE 通过在训练和执行阶段采用不同的假设，巧妙地规避了这两者的主要缺陷。训练时利用额外信息来稳定学习 41，执行时保持去中心化以保证可扩展性和实用性 63。这种务实的区分使得 CTDE 成为解决许多 MARL 问题的有力武器，并催生了多种具体的算法类别。

值分解方法 (Value Decomposition Methods - VDN, QMIX, etc.)

范式: 这是 CTDE 框架下的一大类方法，主要应用于合作型 MARL 任务 16。其核心思想是学习一个联合动作值函数 Qtot（代表团队整体价值），并通过将其分解为每个智能体的个体效用函数 Qi（或值函数）来实现。

VDN (Value Decomposition Networks):

原理: VDN 是最早的值分解方法之一 73。它做出了一个较强的假设：联合动作值函数 Qtot 可以简单地表示为所有智能体个体动作值函数 Qi 的和，即 $Qtot(\tau,u)=\sum_{i=1}^{n}Qi(\tau i,ui)$，其中 τ 是联合历史，u 是联合动作，τi,ui 分别是智能体 i 的局部历史和动作 48。每个智能体学习自己的 Qi 网络（通常是基于局部历史 τi 的循环神经网络），但在训练更新时（例如计算 TD 误差），会将所有 Qi 的输出求和得到 Qtot 48。

优势: 分解方式简单，易于实现 73。

劣势: 线性求和的假设限制了其表示能力，无法对智能体间复杂的非线性协同效应进行建模 51。此外，它通常不利用训练期间可能获得的额外状态信息 s 51。在部分可观察环境下，可能面临"懒惰智能体"（lazy agent）问题（即某些智能体缺乏学习动力）和虚假奖励（spurious rewards）问题 48。

QMIX:

原理: QMIX 对 VDN 进行了泛化，旨在表示更复杂的联合动作值函数 16。它引入了一个**混合网络**（mixing network），将个体 Qi 值非线性地混合成 Qtot。关键在于，这个混合网络被设计成关于每个 Qi 都是**单调**的，即 $\partial Qi \partial Qtot \geq 0$ 51。这种单调性通常通过限制混合网络的权重为非负来实现，这些权重可以由

依赖于全局状态 s 的超网络（hypernetworks）生成 51。单调性保证了**个体-全局最大化（Individual-Global-Max, IGM）**原则：最大化 Qtot 的联合动作 u∗ 等同于每个智能体独立最大化其 Qi 所得到的动作组合 (argmaxu1Q1 ,...,argmaxunQn) 52。这使得在去中心化执行时，每个智能体只需贪婪地选择最大化自身 Qi 的动作即可。QMIX 通过最小化联合 TD 误差进行端到端训练 60。

优势: 比 VDN 具有更强的表示能力，可以建模更复杂的智能体交互关系 51。通过混合网络（和超网络）有效利用了中心化训练阶段的全局状态信息 s 51。在许多基准测试（尤其是 SMAC）上取得了优异的性能 18。

劣势: 单调性约束虽然保证了 IGM，但也限制了 QMIX 能够表示的 Qtot 函数类别，无法表示某些非单调的协作关系（例如，一个智能体的最优动作依赖于其他智能体选择特定动作）16。其性能可能对超参数和实现技巧（如代码层面的优化）比较敏感 81。通常需要相对密集的奖励信号才能有效学习 59。为了解决表示限制问题，后续研究提出了加权 QMIX（Weighted QMIX, CW-QMIX, OW-QMIX），通过在 QMIX 的投影损失中引入权重来强调更优的联合动作 65。

QTRAN: 试图提出比 QMIX 更通用的值分解方法，放宽单调性约束，但实现和理论分析相对复杂 51。

QPLEX: 另一种先进的值分解方法，采用双工决斗（duplex dueling）结构 52。

值分解方法为合作型 MARL 中的 CTDE 提供了一种优雅的实现方式，它们通过建立个体价值 Qi 与团队价值 Qtot 之间的联系来解决信用分配问题。然而，这些方法普遍面临着表示能力（VDN < QMIX < 更一般的方法）与保证去中心化执行可行性的理论约束（如 IGM 属性）之间的权衡。VDN 的线性假设过于简单，而 QMIX 的单调性假设虽然更灵活且保证了 IGM，但仍限制了其处理某些复杂协调任务的能力。像 Weighted QMIX 这样的后续工作正试图突破这些限制。这反映了在设计基于中心化价值函数的去中心化策略时，表达能力与执行约束之间持续存在的张力。

多智能体策略梯度与 Actor-Critic 方法

范式: 这是 CTDE 框架下的另一大类主流方法 17。这类方法直接学习参数化的策略（Actor），通常会利用一个中心化的评论家（Critic）来估计状态值函数（V）或状态-动作值函数（Q），以提供更稳定和低方差的梯度估计，从而指导策略更新。

MADDPG (Multi-Agent Deep Deterministic Policy Gradient):

原理: MADDPG 是将单智能体 DDPG 算法扩展到多智能体场景的代表作 25。每个智能体 i 学习一个确定性策略（Actor）μi(oi)，该策略仅基于其局部观察 oi。关键在于，它为**每个智能体**学习一个**中心化的 Critic** Qi(x,a1,...,an)，该 Critic 在训练时可以访问全局信息 x（如所有智能体的观察 o1,...,on）以及所有智能体的动作 a1,...,an 25。这个中心化的 Critic 用于评估当前策略的好坏，并指导对应 Actor 的更新 26。MADDPG 可适用于合作、竞争及混合环境 49。

优势: 自然地处理连续动作空间 49。中心化 Critic 考虑了联合动作和全局信息，有助于稳定学习过程 25。原则上可用于非合作性设置 74。

劣势: Critic 的训练需要获取所有智能体的动作（或策略信息），这在某些场景下可能难以实现 26。如果 Actor 更新过于独立，可能导致方差较大和协调问题 49。在一些合作任务基准测试中，性能可能不如值分解方法或 MAPPO，尤其是在离散动作空间 59。可能受到相对过泛化（relative overgeneralization）问

题的影响 49。对于大量同质智能体，需要设计排列不变的 Critic 结构以保证可扩展性 27。

MAPPO (Multi-Agent Proximal Policy Optimization):

原理: MAPPO 是将流行的单智能体 PPO 算法应用于 MARL 的扩展，通常结合参数共享（parameter sharing）用于同质智能体 18。它采用去中心化的 Actor πi(oi)（基于局部观察）和一个**中心化的 Critic** V(s)（基于全局状态 s）59。中心化的 Critic 学习状态值函数，用于计算优势函数（advantage function），进而通过 PPO 的截断替代目标（clipped surrogate objective）来更新 Actor 的策略 70。MAPPO 主要设计用于合作环境 70。其对应的独立学习版本是 IPPO，使用去中心化的 Critic 59。

优势: 在各种合作型 MARL 基准测试（如 MPE, SMAC, GRF, Hanabi）中展现出非常强劲的经验性能，常常能达到甚至超越先进的离策略方法 59。算法结构相对简单，是对 PPO 的直接扩展 70。相比于传统观念，其样本效率在 MARL 中表现得相当有竞争力 70。对稀疏奖励环境具有较好的鲁棒性 59。

劣势: 作为一种在线策略（on-policy）算法，可能需要大量的并行环境或较大的批次大小（batch size）来收集足够的样本，相比离策略方法数据利用率可能较低 70。中心化 Critic 要求在训练阶段能够访问全局状态 70。性能可能依赖于仔细的超参数调整和实现细节 70。在某些博弈中可能仍然会遇到相对过泛化问题，收敛到次优均衡 65。

COMA (Counterfactual Multi-Agent Policy Gradients): 一种 Actor-Critic 方法，使用中心化 Critic，并通过计算一个反事实基线（counterfactual baseline）来解决信用分配问题 49。在某些基准测试中，性能似乎低于 MAPPO 或值分解方法 59。

FACMAC: 结合了 MADDPG 和 QMIX 思想（因子分解的 Critic）的 Actor-Critic 方法，声称通过对联合动作空间的中心化梯度估计能实现更好的协调 49。

混合方法

除了纯粹基于值或策略的方法，也存在一些混合方法。例如，MAVEN 算法使用了一个隐变量空间，价值型智能体的行为以该隐变量为条件，而这个隐变量由一个层级策略控制，从而融合了价值和策略方法 22。

表 1: 关键 MARL 算法对比

算法 (Algorithm)	类别 (Category)	核心思想 (Core Idea)	挑战处理 (Challenge Handling)	优势 (Pros)	劣势 (Cons)	典型应用/动作空间 (Typical Application/Action Space)
IQL (Independent Q-Learning)	IL (独立学习)	每个智能体独立学习 Q 函数，将其他智能体视为环境一部分 22。	**非平稳性:** 无法有效处理；**信用分配:** 不适用（无联合奖励概念）；**可扩展性:** 好；**部分可观察性:** 差。	简单，可扩展性好 32。	学习不稳定，无收敛保证，协调能力差 32。	简单任务，弱交互场景，基线 43。离散动作。
VDN (Value Decomposition Networks)	VD (值分解) / CTDE	联合 Q 值是所有个体 Q 值的和: $Q_{tot} = \sum Q_i$ 48。	**非平稳性:** 通过 CTDE 缓解；**信用分配:** 通过值分解隐式处理；**可扩展性:** 中等；**部分可观察性:** 通过 RNN 处理局部历史，但分解假设可能受影响。	实现简单，保证 IGM 60。	表示能力有限（线性），忽略状态信息 51。	合作任务，相对简单的协调 59。离散动作。
QMIX	VD / CTDE	联合 Q 值是个体 Q 值的单调混合: $Q_{tot} = mix(\{Q_i\}, s)$, $\partial Q_{tot}/\partial Q_i \geq 0$ 60。	**非平稳性:** 通过 CTDE 缓解；**信用分配:** 通过值分解和混合网络处理；**可扩展性:** 中等；**部分可观察性:** 通过 RNN 处理局部历史，混合网络可利用全局状态。	比 VDN 表示能力强，利用状态信息，经验性能好 51。保证 IGM 60。	单调性仍是限制，无法表示非单调关系 82。对实现和奖励密度敏感 59。	合作任务，复杂协调（如 SMAC） 59。离散动作。
MADDPG	AC (Actor-Critic) / CTDE	每个智能体学习确定性 Actor $\mu_i(o_i)$ 和中心化 Critic $Q_i(x, a_1..a_n)$ 25。	**非平稳性:** 中心化 Critic 考虑联合动作缓解；**信用分配:** Critic 评估联合动作；**可扩展性:** Critic 输入维度随 N 增长；**部分可观察性:** Actor 基于局部观察，Critic 基于全局信息。	处理连续动作，适用于混合/竞争环境 49。	Critic 需访问所有动作，协调性可能不足，离散动作需技巧 49。	连续控制，混合/竞争博弈 49。连续/离散动作。

算法 (Algorithm)	类别 (Category)	核心思想 (Core Idea)	挑战处理 (Challenge Handling)	优势 (Pros)	劣势 (Cons)	典型应用/动作空间 (Typical Application/Action Space)
MAPPO	AC / CTDE	去中心化 Actor πi(oi)，中心化 Critic V(s)，使用 PPO 更新 59。	**非平稳性:** 中心化 Critic 稳定价值估计；**信用分配:** 通过优势函数估计；**可扩展性:** Actor 去中心化，Critic 输入为全局状态；**部分可观察性:** Actor 基于局部观察，Critic 基于全局状态。	经验性能强，样本效率相对较高（对 on-policy 而言），对稀疏奖励鲁棒 59。	On-policy 数据收集可能需并行化，中心化 Critic 需全局状态 70。	合作任务（MPE, SMAC, GRF, Hanabi 等）59。离散/连续动作。

MARL 应用场景与前沿案例

概述

MARL 作为一种强大的建模和解决涉及多个交互实体的复杂决策问题的工具，已经在众多领域找到了应用 1。这些应用场景的多样性凸显了 MARL 框架的普适性。

(a) 机器人协作

集群控制与编队: MARL 被用于协调大量简单机器人组成的集群（swarm），以完成单个机器人难以完成的任务 19。这种方法的优势在于其去中心化的特性，使得系统对单个机器人的故障具有鲁棒性，并且易于扩展 53。应用包括大规模搜索与救援、环境监测、自动化仓库管理等 53。通常结合群体智能（Swarm Intelligence, SI）的原理 88。一个挑战是如何在部分可观察和动态变化的环境中实现有效协调 53。最近的研究如 MARLIN 尝试结合大型语言模型（LLM）进行协商，以指导 MARL 训练，加速策略部署 66。

多机器人导航与探索: 这是 MARL 在机器人领域的经典应用，涉及多个机器人在共享空间中的路径规划、避碰以及协作达到目标点或探索未知区域 11。研究工作还包括处理异构机器人团队（具有不同感知、运动或执行能力的机器人）19 以及应对现实世界中普遍存在的异步决策问题（机器人并非同时做出决策）19。MARL 也被用于更高层次的任务规划 19。

协作操纵与装配: 对于需要多个机器人手臂或移动平台物理协作来搬运大型物体或进行复杂装配的任务，MARL 可以学习必要的协调策略 36。

(b) 自动驾驶

车辆协调与决策: MARL 为互联自动驾驶车辆（Connected and Autonomous Vehicles, CAVs）在复杂交通场景（如高速公路合流区、无信号交叉口）中的协同决策提供了解决方案 4。例如，通过 MARL 学习车辆间的意图共享策略（如 MAPPO-PIS）可以提高通行效率和安全性 86。此外，MARL 也被用于优化自动驾驶车队的路径选择，尤其是在人类驾驶员和自动驾驶车辆混合存在的交通流中 89。

交通信号控制 (TSC): 将交叉口的交通信号灯视为智能体，利用 MARL 来学习协调控制策略，以优化整个城市交通网络的流量，减少拥堵和延误 10。研究比较了中心化（如 QMIX）和去中心化（如 IQL）方法在不同交通模拟环境（如 SUMO, CityFlow）中的表现，发现中心化方法在需要更强协调的动态路由环境中通常表现更优 18。像 PyTSC 这样的专用框架旨在简化 MARL 在 TSC 领域的研究 18。

(c) 游戏 AI

多人策略游戏: 游戏是 MARL 算法发展和测试的重要试验场。MARL 在许多复杂的多人策略游戏中取得了突破性进展，达到了超越人类顶尖玩家的水平，例如星际争霸 II (StarCraft II) 10、扑克 (Poker) 57、外交 (Diplomacy) 10 以及合作类游戏 Hanabi 10。这些游戏环境通常具有巨大的状态-动作空间、部分可观察性以及复杂的长期策略需求，为 MARL 算法提供了极具挑战性的基准 10。通过自我对弈（self-play）等技术，MARL 能够在这些环境中涌现出复杂的、有时甚至是反直觉的策略和高水平的协调或对抗行为 1。自我对弈过程中策略的不断进化形成了所谓的“自课程”（autocurricula）13。

团队竞技模拟: MARL 也被用于模拟团队体育项目，如谷歌研究足球（Google Research Football, GRF）环境 69。这类任务需要智能体学习团队协作、角色分配和战术执行。

(d) 资源管理

智能电网调度: MARL 被应用于优化智能电网的运行，包括电力负荷平衡、需求响应、以及可再生能源（Renewable Energy Sources, RES）的整合 12。可以将分布式能源（DERs）、储能系统、需求侧单元等建模为智能体，通过 MARL 学习实时的、去中心化的调度决策，以提高电网效率、可靠性，降低成本，并减少碳排放 20。MARL 的自适应性使其特别适合处理 RES（如太阳能、风能）的间歇性和不确定性 20。研究表明，相比传统的基于规则或数学优化的方法以及单智能体 RL，MARL 在动态和复杂的电网环境中具有优势 20。挑战主要在于大规模电网的可扩展性和智能体间的通信协调 20。

网络路由优化: 在通信网络中，MARL 可用于动态优化数据包的路由策略和进行流量工程（Traffic Engineering, TE），以减轻网络拥塞，提高服务质量（QoS）10。路由器可以被视为智能体，通过学习来调整路由决策 46。MARL 方法旨在克服传统路由协议（如 OSPF）对网络动态（如拥塞）不敏感以及简单 RL 方法（中心化或独立学习）在可扩展性或非平稳性方面的局限 46。研究探索了结合图神经网络（GNN）95、涌现通信（emergent communication）46 等技术来增强 MARL 路由算法的性能和适应性，特别是应对网络拓扑动态变化（如链路故障、节点增删）的能力 46。RouteRL 框架则专注于使用 MARL 解决自动驾驶车辆的集体路径选择问题 89。

(e) 金融市场

多主体交易策略: MARL 被用于设计和优化算法交易策略，特别是在高频交易（High-Frequency Trading, HFT）领域 6。可以将市场中的不同参与者（如做市商、套利者）或不同的交易算法实例建模为智能体 21。MARL 算法（如 VDN, MAPPO）可以学习适应动态的市场环境、预测其他市场参与者的行为，并制定最优的买卖决策 21。研究表明，基于 MARL 的策略在夏普比率、年化回报等方面可能优于单智能体 RL 或传统算法交易策略 79。MARL 也被用于增强经典的投资组合保险策略，如 CPPI 和 TIPP 96。

风险管理: 通过 MARL 模拟多个交易策略在市场中的交互，可以更好地评估系统性风险和策略鲁棒性。

这些广泛的应用案例揭示了一个重要现象：尽管各个领域（物理系统如机器人、电网，虚拟系统如游戏，抽象系统如金融、网络）的具体细节千差万别，但它们都存在需要协调多个决策者的去中心化优化问题 1。MARL 提供了一种通用的**方法论**来应对这类问题。虽然核心挑战——如协调的需求、非平稳性、可扩展性——在不同应用中普遍存在 20，但这些挑战的具体表现形式以及最有效的解决方案往往需要根据特定领域的特点进行调整和优化。这表明 MARL 具有强大的跨领域适用潜力，但也要求研究者和实践者深入理解具体问题背景，进行针对性的算法设计和实现。

开源 MARL 研究框架与平台评估

概述

标准化的开源框架和基准测试环境对于推动 MARL 研究的进展至关重要，它们有助于提高实验的可复现性、促进算法之间的公平比较，并降低研究门槛 10。目前，已涌现出多个流行的 MARL 开源库，各自具有不同的侧重点和功能。

PettingZoo

描述: PettingZoo 定位为“MARL 领域的 Gymnasium”，旨在为多智能体环境提供一个标准、简洁、符合 Python 风格的 API 98。

功能: 提供了两种主要的 API：AEC (Agent Environment Cycle) API 用于处理智能体按顺序行动的环境（如棋盘游戏），Parallel API 用于处理智能体同时行动的环境 100。包含了大量多样的参考环境，涵盖合作、竞争、混合模式，如图形化协调任务 (Butterfly)、经典游戏 (Classic)、多玩家 Atari 游戏、粒子环境 (MPE)、机器人仿真 (SISL) 等 100。其核心在于环境接口的标准化，使得用户可以像使用 Gymnasium 一样与多智能体环境交互 101。它还与 SuperSuit 等库集成，方便使用常见的环境包装器（如帧堆叠、观察归一化）100。

易用性/社区: 拥有完善的文档网站 (pettingzoo.farama.org) 100。由 Farama Foundation 积极维护，社区活跃，并设有 Discord 服务器 100。由于其标准化的接口，被许多其他 MARL 库（如 RLlib, MARLlib）用作环境后端 89。

RLlib (Multi-Agent)

描述: RLlib 是一个基于 Ray 构建的可扩展、工业级的强化学习库，对 MARL 提供了强大的原生支持 18。

功能: 定义了自己的 MultiAgentEnv API，用于描述多智能体环境的交互逻辑（返回字典形式的观察、奖励等），同时提供了对 PettingZoo 和 DeepMind OpenSpiel API 的封装器 103。支持灵活的智能体到策略的映射（一个策略可以控制多个智能体）103。能够处理同时行动、轮流行动或任意混合模式的环境 103。利用 Ray 的分布式计算能力，可以通过配置 EnvRunner 和 Learner 的数量轻松实现大规模并行采样和训练 105。内置了多种常用的单智能体和多智能体 RL 算法 106。具备良好的容错性 106。

易用性/社区: 文档详尽 103。依托 Ray 拥有庞大的用户群体和活跃的社区。对于不熟悉 Ray 的用户来说，学习曲线可能相对陡峭。

OpenSpiel

描述: 由 DeepMind 开发，专注于**博弈**中的强化学习和搜索/规划研究的框架 90。

功能: 支持广泛的博弈类型：n 人博弈、零和/合作/一般和博弈、完全/不完全信息博弈、序贯/同时移动博弈等 108。核心库用 C++ 实现以提高效率，并提供 Python 接口 (pybind11) 108。包含了众多经典和现代博弈的实现（如棋类、牌类游戏）108。内置了来自 RL 和计算博弈论的多种算法（如 CFR, DQN, 策略梯度）108。提供分析学习动态（如可利用度计算）的工具 108。设计哲学强调代码简洁性和减少外部依赖 108。

易用性/社区: 提供了良好的文档、教程和示例 108。主要面向博弈论和相关 AI 研究社区。由 Google DeepMind 维护 109。

EPyMARL

描述: EPyMARL 是 PyMARL（最初专注于 SMAC 环境）的扩展版本，旨在提供一个用于在合作型 MARL 任务中进行算法基准测试的框架 18。

功能: 基于 PyTorch 实现，包含了多种流行的 MARL 算法，如 IQL, VDN, QMIX, 以及多种 Actor-Critic 变体（IA2C, IPPO, MADDPG, MAA2C, MAPPO）78。支持多种基准环境，包括 SMAC (v1/v2/Lite), Level-Based Foraging (LBF), Multi-Robot Warehouse (RWARE), Multi-agent Particle Environment (MPE), PettingZoo 环境, VMAS, 矩阵博弈等 104。提供了灵活的配置选项，如是否共享参数、硬/软更新、奖励标准化等 59。支持超参数搜索、模型保存与加载、以及使用 Weights & Biases (W&B) 进行日志记录 104。还支持个体奖励设置 104。

易用性/社区: 代码库在 GitHub 上开源 104。被用于发表 MARL 算法的基准比较研究 59。

MARLlib

描述: MARLlib 是一个基于 Ray/RLlib 构建的库，其目标是**统一**不同的 MARL 算法和环境，简化研究流程 69。

功能: 核心设计在于其**智能体级别的分布式数据流**（agent-level distributed dataflow）78。通过标准化的环境封装器处理不同环境（如 PettingZoo, SMAC, MPE, GRF, MAMuJoCo 等）在数据提供（如全局状态、动作掩码）和奖励结构（团队奖励 vs 个体奖励）上的差异 99。实现了包括独立学习、中心化 Critic、值分解等多种范式下的数十种算法（如 IQL, VDN, QMIX, MADDPG, MAPPO 等）69。采用灵活的策略映射机制来自动处理任务与算法之间的兼容性问题 99。支持合作、竞争、混合等所有任务模式 107。

易用性/社区: 提供在线文档和快速入门指南 72。旨在通过处理兼容性问题来降低 MARL 研究的复杂性 99。在基准测试中与 EPyMARL 进行了比较 78。代码在 GitHub 开源 99。

JaxMARL

描述: JaxMARL 是一个完全使用 JAX 实现的 MARL 库，旨在利用 JAX 的即时编译（JIT）和硬件加速（GPU/TPU）能力，实现极高的训练效率 77。

功能: 提供了多种常用 MARL 环境的 JAX 原生实现，包括 Coin Game, Hanabi, MPE, Overcooked, SMAX（SMAC 的 JAX 版本，无需 SC2 引擎）等 77。实现了流行的基线算法的 JAX 版本，如 IPPO, MAPPO, Q-learning 变体 77。其核心优势在于**速度**：通过端到端的 JIT 编译，训练速度据称比传统基于 CPU 环境的库快 14 倍到数万倍 77。采用了类似 PettingZoo 的并行环境 API 114。

易用性/社区: 拥有专门的文档网站 (jaxmarl.foersterlab.com) 114。主要面向追求高性能和熟悉 JAX 生态的研究者 114。是 JAX RL 生态系统的重要组成部分 115。开发活跃，鼓励社区贡献 114。

当前 MARL 开源生态系统的发展呈现出**成熟与碎片化并存**的特点。一方面，多个成熟库的出现（PettingZoo 专注于环境 API 标准化 98，RLlib 提供强大的可扩展性 106，OpenSpiel 深耕博弈论应用 108，EPyMARL 和 MARLlib 致力于算法基准测试和统一 78，JaxMARL 则追求极致的运行速度 114），标志着该领域工具链的日益完善。这些库各自在特定方面表现出色，为研究者提供了丰富的选择。

然而，另一方面，这种专业化分工也导致了生态系统的碎片化。没有一个单一的库能够完美覆盖所有用户的全部需求。研究者可能需要根据具体任务（环境类型、智能体数量）、资源限制（计算能力）、算法需求或技术偏好（如 JAX vs PyTorch）来仔细选择，甚至组合使用不同的工具。例如，需要多样化环境的研究者可能会选择 PettingZoo 作为起点 102，而需要处理大规模智能体或进行分布式训练的用户可能会倾向于 RLlib 106。博弈论研究者可能首选 OpenSpiel 109，进行算法横向比较的研究者可能会使用 EPyMARL 或 MARLlib 78，而对训练速度有极高要求的研究者则会考虑 JaxMARL 114。这种现状凸显了像 MARLlib 这样试图整合和统一不同组件的库的重要价值 99，它们旨在降低研究者在工具选择和集成上花费的精力。

表 2: 主流 MARL 开源框架对比

框架 (Framework)	主要侧重 (Primary Focus)	关键特性 (Key Features)	支持算法示例 (Supported Algorithms (Examples))	支持环境示例 (Supported Environments (Examples))	后端 (Backend)	易用性/文档 (Usability/Docs)	
PettingZoo	环境 API 标准化	AEC & Parallel API, 多样化环境库, SuperSuit 集成 100	不直接提供算法实现	Atari, Butterfly, Classic, MPE, SISL 102	Python	良好，文档清晰 101	
RLlib (Multi-Agent)	可扩展性, 工业级应用	基于 Ray, MultiAgentEnv API, 支持 PettingZoo/OpenSpiel, 灵活策略映射, 分布式训练, 容错 103	PPO, DQN, SAC, MADDPG, QMIX 等多种算法 106	Gymnasium, PettingZoo, OpenSpiel, 自定义环境 103	Python, Ray, PyTorch, TF	较好，文档全面，但 Ray 较复杂 103	
OpenSpiel	博弈论研究, 游戏 AI	支持多种博弈类型, C++/Python API, 博弈论/RL 算法, 分析工具 108	CFR, DQN, Policy Gradients, Search 108	棋类, 牌类 (Kuhn/Leduc Poker), Goofspiel, Grid Worlds 108	C++, Python	良好，教程丰富 108	
EPyMARL	合作 MARL 算法基准测试	PyMARL 扩展, 多种 CTDE 算法, 多环境支持, 参数共享选项, W&B 集成 104	IQL, VDN, QMIX, IA2C, IPPO, MADDPG, MAA2C, MAPPO 104	SMAC, LBF, RWARE, MPE, PettingZoo, VMAS, Matrix Games 104	Python, PyTorch	一般，主要通过代码和配置文件理解 112	
MARLlib	算法与环境统一, 兼容性	基于 Ray/RLlib, 智能体级数据流, 标准化环境封装器, 灵活策略映射 78	IQL, VDN, QMIX, MADDPG, MAPPO 等 18+ 算法 69	SMAC, MPE, GRF, MAMuJoCo, PettingZoo 69	Python, Ray, PyTorch	良好，提供文档 72	
JaxMARL	高性能, JAX 生态	完全基于 JAX, 端到端 JIT 编译, 极高训练速度, JAX 原生环境/算法 77	IPPO, MAPPO, Q-Learning 变体 114	Coin Game, Hanabi, MPE, Overcooked, SMAX, STORM 114	Python, JAX	良好，面向 JAX 用户 114	

MARL 的核心挑战与局限性深度剖析

尽管 MARL 取得了显著进展，但在理论和实践中仍面临一系列深刻的挑战和固有的局限性。这些挑战相互关联，共同构成了 MARL 研究的核心难点。

非平稳性 (Non-Stationarity) 与环境动态性

深度分析: 这是 MARL 最根本的挑战之一 1。由于系统中的多个智能体同时进行学习和策略更新，导致从任何单个智能体的视角来看，环境的动态特性（状态转移概率和奖励函数）都在不断变化 2。这种"移动目标问题" 2 直接违背了传统单智能体 RL 方法所依赖的马尔可夫环境假设 2。非平稳性使得智能体难以稳定地学习最优策略，因为过去学习到的经验可能不再适用于当前由其他智能体策略构成的"环境"。这对依赖经验回放的离策略算法（如 DQN 及其变体）尤其具有破坏性，因为样本池中的数据可能很快失效 22。

应对策略: CTDE 范式是应对非平稳性的主要手段，通过在训练中引入全局信息来稳定学习目标 41。其他策略包括对对手或队友进行显式建模 41，采用特定的算法结构（如具有中心化 Critic 的 Actor-Critic 方法 22），调整学习率 39，或使用指纹信息（fingerprints）来标记经验的时效性 44。智能体间的通信，特别是意图或策略的共享，也有助于缓解非平稳性问题 37。

可扩展性 (Scalability) 与维度灾难

深度分析: 随着智能体数量 N 的增加，联合状态空间和联合动作空间的维度通常呈指数级增长 2。这使得完全中心化的方法（如将整个 MAS 视为单个智能体进行学习和控制，即 CTE）在计算上变得不可行 32。即使采用去中心化执行，训练过程（尤其是在 CTDE 中需要处理联合信息时）的计算复杂度和样本复杂度也会急剧上升 17。

应对策略: 采用去中心化执行（如 IL 或 CTDE）是基本要求 32。对于同质智能体，参数共享是一种常用的降低模型复杂度的技术 59。值函数分解方法通过将联合价值函数分解为个体价值函数来控制复杂度 52。注意力机制（Attention Mechanisms）可以帮助智能体关注相关的其他智能体信息 19。对于极大数量的智能体，平均场（Mean-Field）方法将大量智能体的影响近似为平均效应。分层强化学习（Hierarchical RL）可以将复杂的协调问题分解为不同层级的子问题 20。利用图神经网络（GNN）处理智能体间的结构化交互 1。将智能体分组或因子分解协调 57。以及模型压缩技术，如网络稀疏化 55。

信用分配 (Credit Assignment) 难题

深度分析: 在合作型 MARL 任务中，所有智能体通常共享一个团队奖励信号。这使得难以判断每个智能体的具体动作对最终团队成果（奖励）的贡献大小 2。当奖励稀疏或延迟时，这个问题尤为突出 15。如果不能准确地将功劳（或过失）分配给导致它的智能体及其动作，智能体就很难有效地学习到有益的个体行为和协作策略 50。

应对策略: 值分解方法（VDN, QMIX）通过建立个体 Qi 和联合 Qtot 之间的函数关系，隐式地解决了部分信用分配问题 51。更显式的方法包括使用反事实基线（如 COMA），它评估一个智能体采取某个动作相对于采取默认动作所带来的边际贡献 51。差分奖励（Difference Rewards）是类似的思想 51。奖励塑形（Reward Shaping）可以设计更丰富的奖励信号。中心化 Critic 由于可以评估联合动作，也有助于信用分配 49。最新的方法如 Deconfounded Value Decomposition (DVD) 尝试从因果推断的角度来解决信用分配中的混淆偏倚问题 52。

部分可观察性 (Partial Observability) 下的决策挑战

深度分析: 现实世界中，智能体通常只能获得关于环境和彼此的局部信息，而非全局状态 1。这要求智能体必须基于不完整的观察历史来推断环境的潜在状态和其他智能体的状态或意图，极大地增加了策略学习的复杂性 33。部分可观察性可能导致智能体采取次优甚至冲突的行动，并可能引发"懒惰智能体"问题，即某些智能体因为无法观察到协作的全部好处而缺乏学习动力 48。该问题通常用 Dec-POMDP 模型来描述 24。

应对策略: 在智能体的策略网络或价值网络中使用循环神经网络（RNNs, LSTMs, GRUs）来处理历史信息，维持一个内部的信念状态（belief state）32。智能体之间的通信是另一种关键策略，通过交换信息来弥补各自观察的不足 1。CTDE 范式允许在训练阶段利用全局状态信息来辅助学习 63。注意力机制也可以帮助智能体聚焦于来自其他智能体的关键信息。

智能体通信 (Communication) 机制的设计与学习

深度分析: 通信被认为是解决部分可观察性和协调问题的有效途径 1。然而，设计有效的通信机制本身就是一个挑战 1。关键问题包括：智能体应该在何时通信？通信什么内容（原始观察、处理后的特征、意图、策略参数）？如何设计通信协议和网络拓扑（广播、点对点、基于图的通信）？ 1。此外，通信会带来额外的开销（如带宽占用、延迟）2，并且需要学习如何有效地编码和解码信息，甚至是从零开始学习一种"语言"（即涌现通信）1。

应对策略: "学习通信"（Learning to Communicate, L2C）方法让智能体在优化任务目标的同时学习通信策略 1。利用注意力机制动态选择通信伙伴或信息内容。使用图神经网络（GNNs）来建模和利用智能体之间的结构化通信关系 1。设计机制来控制或过滤通信内容，以减少开销，例如 VBC 算法通过控制消息方差来去除噪声 58。建立点对点通信渠道 62。甚至利用大型语言模型进行更高级别的协商式通信（如 MARLIN）66。

学习稳定性与收敛性

深度分析: MARL 的学习过程比单智能体 RL 更容易不稳定。非平稳性是主要原因之一，它破坏了许多 RL 算法的收敛性基础 2。即使所有智能体都收敛到各自的局部最优策略，它们的联合策略也未必是全局最优的，甚至可能不是一个稳定的均衡点 22。MARL 中还存在一些特有的学习病理（learning pathologies），如相对过泛化（relative overgeneralization），即智能体倾向于收敛到一个虽然次优但对其他智能体策略变化更不敏感（更"安全"）的纳什均衡 1；以及协调失败（miscoordination），即当存在多个同样好的均衡时，智能体可能无法就选择哪一个达成一致 1。多智能体环境中的探索（exploration）也更具挑战性，因为一个智能体的探索行为会影响其他智能体的学习环境 1。

应对策略: CTDE 范式通过引入中心化信息来提高稳定性 41。算法设计层面，如 QMIX 的单调性约束有助于保证 IGM 一致性 60，PPO/TRPO 类算法的信任域约束有助于稳定策略更新 65。需要设计针对 MARL 的特定探索策略，鼓励协同探索 1。采用乐观（optimistic）方法来处理不确定性或多均衡问题 44。虽然存在一些理论工作试图提供收敛性保证，但它们通常局限于特定的博弈类别（如零和博弈）或需要较强的假设 15。

这些挑战并非孤立存在，而是**相互交织、相互影响**。例如，部分可观察性使得智能体更难判断奖励的来源，从而加剧了信用分配的难度 48，同时也使得预测其他智能体的行为更加困难，放大了非平稳性的影响 42。可扩展性问题限制了使用完全中心化方法来解决非平稳性和部分可观察性的可行性 32。通信被视为解决部分可

观察性和协调问题的有效手段 37，但通信本身又带来了新的设计复杂性和潜在的带宽瓶颈 1。因此，设计有效的 MARL 算法需要对这些相互关联的挑战进行整体考虑，寻求系统性的解决方案。

MARL 技术成熟度评估与未来展望

当前技术成熟度与局限性总结

成就: 近年来，尤其是在深度学习的推动下，MARL 领域取得了显著进展 15。在多个复杂的基准测试环境（特别是策略游戏如星际争霸、围棋、扑克、外交等）中取得了令人瞩目的成就，展示了超越人类水平的能力 10。CTDE 范式为实际应用提供了一个相对成熟和有效的框架 63。涌现出了一批性能优异的算法（如 QMIX, MAPPO）59。应用领域不断扩展，相关的开源工具和库也日益丰富 19。

局限性: 尽管成就斐然，但 MARL 仍面临诸多瓶颈。前述的核心挑战（非平稳性、可扩展性、信用分配、部分可观察性）在实际部署中仍然是巨大的障碍 2。算法的样本效率仍有提升空间，尤其是在复杂环境中 70。严格的理论收敛保证往往缺失或依赖于较强的假设 43。评估标准和实验的可复现性仍需加强，存在"评估危机"的担忧 10。模拟环境中的优异表现与现实世界中的鲁棒性之间仍存在差距 11。处理真正开放、动态、不可预测的环境仍然困难 11。

前沿研究方向

为了克服现有局限并推动 MARL 的发展，以下研究方向备受关注：

可扩展性增强: 开发能够处理成百上千甚至更大规模智能体群体的算法，例如基于平均场理论的 MARL、更有效的因子分解或值分解方法、分层 MARL 架构 20。利用 JAX 等框架进行高效的并行和分布式训练，大幅缩短实验周期 77。

鲁棒性与泛化性: 提高 MARL 策略对环境变化、智能体故障、噪声干扰甚至恶意攻击的鲁棒性 15。研究如何使学习到的策略能够泛化到不同的任务、环境配置或智能体数量 19。探索在开放环境（open environment）下的 MARL，即环境中的关键因素（如智能体数量、目标、规则）可能发生变化的场景 11。

信用分配与协作机制: 发展更精细化的信用分配技术，例如引入因果推断 52。深入理解复杂协作行为的涌现机制，并设计能够有效促进合作的算法 1。研究更高效、自适应、甚至可学习的通信协议 1。探索基于协商、社会学习规范的 MARL 35。

安全性、公平性与可解释性: 在 MARL 中引入安全约束，确保智能体在学习和执行过程中的行为符合安全规范 87。研究多智能体系统中的公平性问题，如资源分配的公平性、奖励分配的公平性。提高 MARL 策略的可解释性，理解智能体做出决策的原因以及群体行为的模式 15。利用 MARL 进行 AI 对齐（AI alignment）研究，探索如何使 AI 系统（或多 AI 系统）的目标与人类价值观保持一致 13。

理论基础深化: 寻求更强的 MARL 算法收敛性理论保证，尤其是在部分可观察和非平稳环境下的收敛性。加深对多智能体系统中均衡选择问题的理解。进一步融合博弈论和深度强化学习的理论成果 1。

与其他 AI 领域融合: 探索 MARL 与大型语言模型（LLMs）的结合，利用 LLM 进行高级规划、任务分解、生成通信内容或解释智能体行为 34。结合模仿学习（Imitation Learning）和逆强化学习（Inverse RL）从专家演示中学习协调策略 15。发展更强大的基于模型的 MARL（Model-based MARL）方法 1。研究

多智能体环境下的迁移学习（Transfer Learning）2 和元学习（Meta-learning）14。

未来发展趋势与潜在突破

展望未来，MARL 技术有望在以下方面取得突破并产生深远影响：

更广泛的现实世界应用: 随着算法鲁棒性和可扩展性的提升，MARL 将在机器人、自动驾驶、智能制造、物流优化、能源管理、金融科技等领域得到更广泛和深入的应用 4。

通用多智能体智能: 出现更通用的 MARL 智能体，能够适应不同的社会交互场景（合作、竞争、混合），并展现出更强的泛化能力。

大规模群体智能: 在处理包含数千乃至数百万智能体的超大规模系统方面取得突破，可能借鉴物理学（如平均场理论）或生物学（如群体智能）的原理。

人机协同: 实现 MARL 智能体与人类用户或其他（非学习型）系统的无缝、高效协同。

社会科学洞见: 通过构建更真实的 MARL 社会模拟，为理解经济学、社会学、生态学等领域的复杂集体行为提供新的计算工具和洞见。

当前和未来的研究趋势共同指向了一个明确的方向：MARL 正在从主要关注在受控环境（如游戏）中展示能力的阶段，**转向致力于解决现实世界应用中的复杂性、鲁棒性、安全性和可扩展性问题。**虽然早期成功多集中于游戏基准 13，但现在的应用研究已广泛涉足机器人、自动驾驶、智能电网、金融等实际领域 19。同时，研究焦点也越来越多地放在克服鲁棒性、安全性、大规模可扩展性、开放环境适应性等瓶颈上，并将 MARL 与 LLM 等其他前沿 AI 技术相融合 11。这表明 MARL 领域正在走向成熟，其目标不再仅仅是模拟智能，而是要在真实、动态、充满不确定性的世界中可靠地部署智能。

结论

核心观点与洞见回顾

多智能体强化学习（MARL）是驱动多智能体系统（MAS）实现智能化协调与决策的关键技术。它虽然源于单智能体 RL，但引入了非平稳性、指数级增长的联合动作空间、部分可观察性以及信用分配等独特的、相互关联的挑战。为了应对这些挑战，中心化训练与去中心化执行（CTDE）已成为最实用且研究最深入的范式，它在训练效率和执行约束之间取得了有效平衡。在 CTDE 框架下，值分解方法（如 VDN、QMIX）和基于中心化 Critic 的 Actor-Critic 方法（如 MADDPG、MAPPO）是两类主要的算法途径，各自在表示能力、适用场景和理论保证方面存在权衡。MARL 已在机器人协作、自动驾驶、游戏 AI、资源管理和金融市场等多个领域展示了其解决复杂去中心化问题的潜力。同时，一个日益丰富的开源工具生态系统（包括 PettingZoo, RLlib, OpenSpiel, EPyMARL, MARLlib, JaxMARL 等）正在为 MARL 的研究和开发提供支持，尽管这个生态系统也呈现出一定的碎片化特征。

MARL 在推动复杂系统智能化中的价值与展望

尽管面临诸多挑战，MARL 仍然是理解和构建能够在复杂动态环境中进行协作、竞争和适应的智能系统的最有前途的途径之一。它为解决从微观的机器人交互到宏观的社会经济系统模拟等一系列重要问题提供了强大的计算框架。随着算法的不断进步、理论基础的日益完善以及与 LLM 等其他 AI 技术的深度融合，MARL 有望在未

来解锁更高级别的群体智能，推动自动驾驶、智能制造、智慧城市、科学发现等领域实现变革性的发展 1。克服当前在可扩展性、鲁棒性、安全性和样本效率方面的挑战，将是释放 MARL 全部潜力的关键。

参考文献

引用的著作

Multi-agent Reinforcement Learning: A Comprehensive Survey - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/html/2312.10256v1

arxiv.org, 访问时间为 四月 25, 2025， https://arxiv.org/pdf/2312.10256

[2312.10256] Multi-agent Reinforcement Learning: A Comprehensive Survey - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/abs/2312.10256

[2203.07676] An Introduction to Multi-Agent Reinforcement Learning and Review of its Application to Autonomous Mobility - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/abs/2203.07676

All You Need to Know About Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025， https://adasci.org/all-you-need-to-know-about-multi-agent-reinforcement-learning/

Single-Agent vs Multi-Agent AI Comparison - saasguru, 访问时间为 四月 25, 2025， https://www.saasguru.co/single-agent-vs-multi-agent-ai-comparison/

What is the difference between single-agent and multi-agent systems? - Milvus, 访问时间为 四月 25, 2025， https://milvus.io/ai-quick-reference/what-is-the-difference-between-singleagent-and-multiagent-systems

Single-Agent vs Multi-Agent Systems: Two Paths for the Future of AI | DigitalOcean, 访问时间为 四月 25, 2025， https://www.digitalocean.com/resources/articles/single-agent-vs-multi-agent

多智能体强化学习综述 - ASC实验室, 访问时间为 四月 25, 2025， https://asc.xmu.edu.cn/storage/file/202111/45c2e243172d3ca62987922e77496221.pdf

arXiv:2312.08463v2 [cs.AI] 26 Jan 2024, 访问时间为 四月 25, 2025， https://arxiv.org/pdf/2312.08463

[2312.01058] A Survey of Progress on Cooperative Multi-agent Reinforcement Learning in Open Environment - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/abs/2312.01058

What is Multi-Agent Reinforcement Learning (MARL)? — Klu, 访问时间为 四月 25, 2025， https://klu.ai/glossary/multi-agent-reinforcement-learning

Multi-agent reinforcement learning - Wikipedia, 访问时间为 四月 25, 2025， https://en.wikipedia.org/wiki/Multi-agent_reinforcement_learning

基于学习机制的多智能体强化学习综述 - 工程科学学报, 访问时间为 四月 25, 2025， https://cje.ustb.edu.cn/article/doi/10.13374/j.issn2095-9389.2023.08.08.003

基于多智能体强化学习的博弈综述 - 自动化学报, 访问时间为 四月 25, 2025， http://www.aas.net.cn/cn/article/doi/10.16383/j.aas.c240478?viewType=HTML

多智能体深度强化学习研究进展 - 计算机学报, 访问时间为 四月 25, 2025, http://cjc.ict.ac.cn/online/onlinepaper/dsf-202479161153.pdf

A Review of Multi-Agent Reinforcement Learning Algorithms - MDPI, 访问时间为 四月 25, 2025, https://www.mdpi.com/2079-9292/14/4/820

PyTSC: A Unified Platform for Multi-Agent Reinforcement Learning in Traffic Signal Control, 访问时间为 四月 25, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11902778/

Cooperative and Asynchronous Transformer-based Mission Planning for Heterogeneous Teams of Mobile Robots - arXiv, 访问时间为 四月 25, 2025, https://arxiv.org/html/2410.06372v2

Optimizing Renewable Energy Integration in Smart Grids Using Multi-Agent Reinforcement Learning - ResearchGate, 访问时间为 四月 25, 2025, https://www.researchgate.net/publication/387088909_Optimizing_Renewable_Energy_Integration_in_Smar Agent_Reinforcement_Learning

The Impact of Multi-Agent Reinforcement Learning (MARL) - Rapid Innovation, 访问时间为 四月 25, 2025, https://www.rapidinnovation.io/post/multi-agent-reinforcement-learning-marl-and-its-impact

Survey of Recent Multi-Agent Reinforcement Learning Algorithms Utilizing Centralized Training, 访问时间为 四月 25, 2025, https://qiniu.pattern.swarma.org/pdf/arxiv/2107.14316.pdf

开放环境下的协作多智能体强化学习进展综述 - LAMDA - 南京大学, 访问时间为 四月 25, 2025, https://www.lamda.nju.edu.cn/lilh/file/openmarl.pdf

基于通信的多智能体强化学习进展综述, 访问时间为 四月 25, 2025, http://scis.scichina.com/cn/2022/SSI-2020-0180.pdf

arXiv:2208.14447v1 [cs.LG] 30 Aug 2022, 访问时间为 四月 25, 2025, https://arxiv.org/pdf/2208.14447

A Learning Multi-Agent Communication through Structured Attentive Reasoning: Appendix, 访问时间为 四月 25, 2025, https://proceedings.neurips.cc/paper/2020/file/72ab54f9b8c11fae5b923d7f854ef06a-Supplemental.pdf

PIC: Permutation Invariant Critic for Multi-Agent Deep Reinforcement Learning, 访问时间为 四月 25, 2025, http://proceedings.mlr.press/v100/liu20a/liu20a.pdf

single agent vs multiple agent reinforcement learning - Computer Science Stack Exchange, 访问时间为 四月 25, 2025, https://cs.stackexchange.com/questions/48174/single-agent-vs-multiple-agent-reinforcement-learning

Multi Agent RL : r/reinforcementlearning - Reddit, 访问时间为 四月 25, 2025, https://www.reddit.com/r/reinforcementlearning/comments/14fcpad/multi_agent_rl/

From Single-Agent to Multi-Agent Reinforcement Learning: Foundational Concepts and Methods - cs.utah.edu, 访问时间为 四月 25, 2025, https://users.cs.utah.edu/~tch/CS6380/resources/Neto-2005-RL-MAS-Tutorial.pdf

Game Theory and Multi-Agent Reinforcement Learning : From Nash Equilibria to Evolutionary Dynamics - arXiv, 访问时间为 四月 25, 2025, https://arxiv.org/html/2412.20523

Multi-Agent Reinforcement Learning: A Review of Challenges and ..., 访问时间为 四月 25, 2025， https://www.mdpi.com/2076-3417/11/11/4948

CICC科普栏目｜多代理强化学习综述：原理、算法与挑战- 中国指挥 ..., 访问时间为 四月 25, 2025， http://www.c2.org.cn/h-nd-1320.html

LantaoYu/MARL-Papers: Paper list of multi-agent reinforcement learning (MARL) - GitHub, 访问时间为 四月 25, 2025， https://github.com/LantaoYu/MARL-Papers

RaghuHemadri/Multi-Agent-Reinforcement-Learning-Survey-Papers - GitHub, 访问时间为 四月 25, 2025， https://github.com/RaghuHemadri/Multi-Agent-Reinforcement-Learning-Survey-Papers

Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents - MIT Media Lab, 访问时间为 四月 25, 2025， https://web.media.mit.edu/~cynthiab/Readings/tan-MAS-reinfLearn.pdf

Communication Learning for True Cooperation in Multi-Agent Systems - MARMot Lab, 访问时间为 四月 25, 2025， https://www.marmotlab.org/projects/comms_learning.html

Progress on cooperative multi-agent reinforcement learningin open environment - SciEngine, 访问时间为 四月 25, 2025， https://www.sciengine.com/doi/10.1360/SSI-2023-0335

多智能体强化学习控制与决策研究综述, 访问时间为 四月 25, 2025， http://www.aas.net.cn/cn/article/doi/10.16383/j.aas.c240392?viewType=HTML

多智能体深度强化学习研究进展 - 计算机学报, 访问时间为 四月 25, 2025， http://cjc.ict.ac.cn/online/onlinepaper/dsf-2024715101353.pdf

开放环境下的协作多智能体强化学习进展综述 - LAMDA - 南京大学, 访问时间为 四月 25, 2025， https://www.lamda.nju.edu.cn/zhangzq/zzq_index_files/openmarl.pdf

基于奖励滤波信用分配的多智能体深度强化学习算法 - 计算机学报, 访问时间为 四月 25, 2025， http://cjc.ict.ac.cn/online/onlinepaper/xc-2022116221914.pdf

Investigation of independent reinforcement learning algorithms in multi-agent environments - PMC - PubMed Central, 访问时间为 四月 25, 2025， https://pmc.ncbi.nlm.nih.gov/articles/PMC9530713/

I2Q: A Fully Decentralized Q-Learning Algorithm, 访问时间为 四月 25, 2025， https://proceedings.neurips.cc/paper_files/paper/2022/file/8078e8c3055303a884ffae2d3ea00338-Paper-Conference.pdf

META PROXIMAL POLICY OPTIMIZATION FOR COOPERATIVE MULTI-AGENT CONTINUOUS CONTROL, 访问时间为 四月 25, 2025， https://legacy.cs.indiana.edu/ftp/techreports/TR745.pdf

DRAMA: A Dynamic Packet Routing Algorithm using Multi-Agent Reinforcement Learning with Emergent Communication - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/html/2504.04438v1

Multi-agent Reinforcement Learning with Deep Networks for Diverse Q-Vectors - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/html/2406.07848v1

(PDF) Value-Decomposition Networks For Cooperative Multi-Agent Learning, 访问时间为 四月 25, 2025，

https://www.researchgate.net/publication/317649800_Value-Decomposition_Networks_For_Cooperative_Multi-Agent_Learning

papers.nips.cc, 访问时间为 四月 25, 2025，
https://papers.nips.cc/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper.pdf

Conditionally Optimistic Exploration for Cooperative Deep Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://proceedings.mlr.press/v216/zhao23b/zhao23b.pdf

Value-Decomposition Multi-Agent Actor-Critics, 访问时间为 四月 25, 2025，
https://ojs.aaai.org/index.php/AAAI/article/view/17353/17160

Deconfounded Value Decomposition for Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025， https://proceedings.mlr.press/v162/li22l/li22l.pdf

Multi-Agent Reinforcement Learning for Swarms in Uncertain Environments - OpenReview, 访问时间为 四月 25, 2025，
https://openreview.net/pdf/e0e1e12914f8ad52ec3112741af686145281976b.pdf

Multi Agent RL - A Survey, 访问时间为 四月 25, 2025，
https://ppriyank.github.io/MARL/final.html

Value-Based Deep Multi-Agent Reinforcement Learning with Dynamic Sparse Training - NIPS papers, 访问时间为 四月 25, 2025，
https://proceedings.neurips.cc/paper_files/paper/2024/file/31888563b194f9bb33ce1aebc7e1551c-Paper-Conference.pdf

Value-Based Deep Multi-Agent Reinforcement Learning with Dynamic Sparse Training, 访问时间为 四月 25, 2025，
https://neurips.cc/virtual/2024/poster/95865

MARL-LNS: Cooperative Multi-agent Reinforcement Learning via Large Neighborhoods Search - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/html/2404.03101v1

Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based Control, 访问时间为 四月 25, 2025，
http://papers.neurips.cc/paper/8586-efficient-communication-in-multi-agent-reinforcement-learning-via-variance-based-control.pdf

datasets-benchmarks-proceedings.neurips.cc, 访问时间为 四月 25, 2025，
https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/a8baa56554f96369ab93e4f3bb068c22-Paper-round1.pdf

<www.jmlr.org>, 访问时间为 四月 25, 2025，
https://www.jmlr.org/papers/volume21/20-081/20-081.pdf

NeurIPS Poster IMP-MARL: a Suite of Environments for Large-scale Infrastructure Management Planning via MARL, 访问时间为 四月 25, 2025，
https://neurips.cc/virtual/2023/poster/73459

Privacy-Engineered Value Decomposition Networks for Cooperative Multi-Agent Reinforcement Learning - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/pdf/2311.06255

(PDF) An Introduction to Centralized Training for Decentralized Execution in Cooperative Multi-Agent Reinforcement Learning - ResearchGate, 访问时间为 四

月 25, 2025，
https://www.researchgate.net/publication/383791849_An_Introduction_to_Centralized_Training_for_Decent Agent_Reinforcement_Learning

An Introduction to Centralized Training for Decentralized Execution in Cooperative Multi-Agent Reinforcement Learning - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/html/2409.03052v1

Optimistic Multi-Agent Policy Gradient - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/pdf/2311.01953

MARLIN: Multi-Agent Reinforcement Learning Guided by Language-Based Inter-Robot Negotiation - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/html/2410.14383v1

Tutorials – AAMAS 2025 Detroit, 访问时间为 四月 25, 2025，
https://aamas2025.org/index.php/conference/program/tutorials/

A Comprehensive Survey on Multi-Agent Reinforcement Learning for Connected and Automated Vehicles - MDPI, 访问时间为 四月 25, 2025，
https://www.mdpi.com/1424-8220/23/10/4710

jianzhnie/deep-marl-toolkit: MARLToolkit: The Multi-Agent Rainforcement Learning Toolkit. Include implementation of MAPPO, MADDPG, QMIX, VDN, COMA, IPPO, QTRAN, MAT... - GitHub, 访问时间为 四月 25, 2025，
https://github.com/jianzhnie/deep-marl-toolkit

proceedings.neurips.cc, 访问时间为 四月 25, 2025，
https://proceedings.neurips.cc/paper_files/paper/2022/file/9c1535a02f0ce079433344e14d910597-Paper-Datasets_and_Benchmarks.pdf

Proximal Policy Optimization Family — MARLlib v1.0.0 documentation - Read the Docs, 访问时间为 四月 25, 2025，
https://marllib.readthedocs.io/en/latest/algorithm/ppo_family.html

MARLlib: A Multi-agent Reinforcement Learning Library — MARLlib v1.0.0 documentation, 访问时间为 四月 25, 2025， https://marllib.readthedocs.io/

Free Full-Text | Performance Evaluation of Multi-Agent Reinforcement Learning Algorithms, 访问时间为 四月 25, 2025，
https://www.techscience.com/iasc/v39n2/56498/html

MADDPG Explained - Papers With Code, 访问时间为 四月 25, 2025，
https://paperswithcode.com/method/maddpg

Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://archive.illc.uva.nl/AAMAS-2021/pdfs/p844.pdf

(PDF) Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training - ResearchGate, 访问时间为 四月 25, 2025，
https://www.researchgate.net/publication/350821361_Survey_of_recent_multi-agent_reinforcement_learning_algorithms_utilizing_centralized_training

JaxMARL: Multi-Agent RL Environments and Algorithms in JAX - NIPS papers, 访问时间为 四月 25, 2025，
https://papers.nips.cc/paper_files/paper/2024/file/5aee125f052c90e326dcf6f380df94f6-Paper-Datasets_and_Benchmarks_Track.pdf

MARLlib: Extending RLlib for Multi-agent Reinforcement Learning - OpenReview, 访问时间为 四月 25, 2025， https://openreview.net/forum?id=q4qocCgE3uM

Multi-Agent Reinforcement Learning for High-Frequency Trading Strategy Optimization, 访问时间为 四月 25, 2025，
https://www.researchgate.net/publication/386279469_Multi-Agent_Reinforcement_Learning_for_High-Frequency_Trading_Strategy_Optimization

Review for NeurIPS paper: Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://proceedings.neurips.cc/paper/2020/file/73a427badebe0e32caa2e1fc7530b7f3-Review.html

Revisiting the Monotonicity Constraint in Cooperative Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://openreview.net/forum?id=F6S_3RSWFI7

<www.cs.ox.ac.uk>, 访问时间为 四月 25, 2025，
http://www.cs.ox.ac.uk/people/shimon.whiteson/pubs/rashidnips20.pdf

rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/pdf/2102.03479

Off-Policy Correction For Multi-Agent Reinforcement Learning - deepsense.ai, 访问时间为 四月 25, 2025，
https://deepsense.ai/resource/off-policy-correction-for-multi-agent-reinforcement-learning/

Distributional Actor-Critic for Risk-Sensitive Multi-Agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://research.ibm.com/publications/distributional-actor-critic-for-risk-sensitive-multi-agent-reinforcement-learning

[2408.06656] MAPPO-PIS: A Multi-Agent Proximal Policy Optimization Method with Prior Intent Sharing for CAVs' Cooperative Decision-Making - arXiv, 访问时间为 四月 25, 2025， https://arxiv.org/abs/2408.06656

Multi-Agent Constrained Policy Optimisation - OpenReview, 访问时间为 四月 25, 2025， https://openreview.net/forum?id=BlyXYc4wF2-

Swarm Intelligence-Based Multi-Robotics: A Comprehensive Review - ResearchGate, 访问时间为 四月 25, 2025，
https://www.researchgate.net/publication/384580993_Swarm_Intelligence-Based_Multi-Robotics_A_Comprehensive_Review

RouteRL: Multi-agent reinforcement learning framework for urban route choice with autonomous vehicles - arXiv, 访问时间为 四月 25, 2025，
https://arxiv.org/html/2502.20065v1

sjtu-marl/malib: A parallel framework for population-based multi-agent reinforcement learning. - GitHub, 访问时间为 四月 25, 2025，
https://github.com/sjtu-marl/malib

APPLICATION OF DEEP REINFORCEMENT LEARNING FOR REAL-TIME DEMAND RESPONSE IN SMART GRIDS - IRJMETS, 访问时间为 四月 25, 2025，
https://www.irjmets.com/uploadedfiles/paper//issue_3_march_2025/69155/final/fin_irjmets1741852009.pdf

(PDF) Optimizing Energy Efficiency in Smart Grids Using Machine Learning Algorithms: A Case Study in Electrical Engineering - ResearchGate, 访问时间为 四

月 25, 2025，
https://www.researchgate.net/publication/385479518_Optimizing_Energy_Efficiency_in_Smart_Grids_Using_

Energy Aware Load Balancing Framework for Smart Grid Using Cloud and Fog Computing, 访问时间为 四月 25, 2025，
https://pmc.ncbi.nlm.nih.gov/articles/PMC10098693/

Multi-Agent Reinforcement Learning for Traffic Flow Management of Autonomous Vehicles - PMC - PubMed Central, 访问时间为 四月 25, 2025，
https://pmc.ncbi.nlm.nih.gov/articles/PMC10007156/

Is Machine Learning Ready for Traffic Engineering Optimization? - IEEE ICNP 2021, 访问时间为 四月 25, 2025，
https://icnp21.cs.ucr.edu/papers/icnp21camera-paper25.pdf

2303.11959v2 | PDF - Scribd, 访问时间为 四月 25, 2025，
https://www.scribd.com/document/838052909/2303-11959v2

What is Multi-Agent Reinforcement Learning (MARL) - Activeloop, 访问时间为 四月 25, 2025，
https://www.activeloop.ai/resources/glossary/multi-agent-reinforcement-learning-marl/

CITATION.cff - Farama-Foundation/PettingZoo - GitHub, 访问时间为 四月 25, 2025，
https://github.com/Farama-Foundation/PettingZoo/blob/master/CITATION.cff

MARLlib: A Scalable and Efficient Library For Multi-agent Reinforcement Learning, 访问时间为 四月 25, 2025，
https://www.jmlr.org/papers/volume24/23-0378/23-0378.pdf

himanshu-02/PettingZoo-MARL: Gym for multi-agent reinforcement learning - GitHub, 访问时间为 四月 25, 2025，
https://github.com/himanshu-02/PettingZoo-MARL

PettingZoo Documentation, 访问时间为 四月 25, 2025，
https://pettingzoo.farama.org/

Farama-Foundation/PettingZoo: An API standard for multi-agent reinforcement learning environments, with popular reference environments and related utilities - GitHub, 访问时间为 四月 25, 2025，
https://github.com/Farama-Foundation/PettingZoo

Multi-Agent Environments - RLlib - Ray Docs, 访问时间为 四月 25, 2025，
https://docs.ray.io/en/latest/rllib/multi-agent-envs.html

Huey Song/epymarl - Gitee, 访问时间为 四月 25, 2025，
https://gitee.com/hueysong_sxu/epymarl?skip_mobile=true

Environments — Ray 2.44.1 - Ray Docs, 访问时间为 四月 25, 2025，
https://docs.ray.io/en/latest/rllib/rllib-env.html

RLlib: Industry-Grade, Scalable Reinforcement Learning - Ray Docs, 访问时间为 四月 25, 2025， https://docs.ray.io/en/latest/rllib/index.html

Introduction — MARLlib v1.0.0 documentation - Read the Docs, 访问时间为 四月 25, 2025， https://marllib.readthedocs.io/en/latest/handbook/intro.html

[1908.09453] OpenSpiel: A Framework for Reinforcement Learning in Games - ar5iv - arXiv, 访问时间为 四月 25, 2025，
https://ar5iv.labs.arxiv.org/html/1908.09453

open_spiel/docs/intro.md at master - GitHub, 访问时间为 四月 25, 2025，
https://github.com/google-deepmind/open_spiel/blob/master/docs/intro.md

google-deepmind/open_spiel: OpenSpiel is a collection of environments and algorithms for research in general reinforcement learning and search/planning in games. - GitHub, 访问时间为 四月 25, 2025，
https://github.com/google-deepmind/open_spiel

OpenSpielWrapper — torchrl 0.6 documentation - PyTorch, 访问时间为 四月 25, 2025，
https://pytorch.org/rl/0.6/reference/generated/torchrl.envs.OpenSpielWrapper.html

Pi-Star-Lab/epymarl_resco: EPyMARL codebase modified to operate the RESCO benchmark environments - GitHub, 访问时间为 四月 25, 2025，
https://github.com/Pi-Star-Lab/epymarl_resco

oxwhirl/pymarl: Python Multi-Agent Reinforcement Learning framework - GitHub, 访问时间为 四月 25, 2025， https://github.com/oxwhirl/pymarl

JaxMARL Documentation, 访问时间为 四月 25, 2025，
https://jaxmarl.foersterlab.com/

purejaxrl/RESOURCES.md at main - GitHub, 访问时间为 四月 25, 2025，
https://github.com/luchris429/purejaxrl/blob/main/RESOURCES.md

Installation - JaxMARL Documentation, 访问时间为 四月 25, 2025，
https://jaxmarl.foersterlab.com/installation/

kinalmehta/marl-jax: JAX library for MARL research - GitHub, 访问时间为 四月 25, 2025， https://github.com/kinalmehta/marl-jax

MABIM：多智能体强化学习算法的"炼丹炉" - Microsoft Research, 访问时间为 四月 25, 2025， https://www.microsoft.com/en-us/research/articles/mabim/

Part 1 of 3 — Proximal Policy Optimization Implementation - YouTube, 访问时间为 四月 25, 2025，
https://m.youtube.com/watch?v=MEt6rrxH8W4&pp=ygUMI3Bwb3RyYWluaW5n

Chain of Draft 论文解读

本文介绍了 Chain of Draft（CoD）论文，并对其技术...

Test-Time Scaling 相关论文解读

本文介绍了 Test-Time Scaling（测试时扩展）的概念，并对...