分类号：_____ 密　　级：_____

U D C：_____ 学　　号：_____

江西理工大學

硕 士 学 位 论 文

面向动态柔性作业车间调度的多智能体强化学习方法研究

Multi-Agent Reinforcement Learning for Dynamic Flexible Job Shop Scheduling

学 位 类 别：_____

作 者 姓 名：夏 乐

学 科、专 业：_____

研 究 方 向：_____

指 导 教 师：王碧/讲师

年　　月　　日

学位论文独创性声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含已获得江西理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

研究生签名：　　时间：　　年　　月　　日

学位论文版权使用授权书

本人完全了解江西理工大学关于收集、保存、使用学位论文的规定：即学校有权保存按要求提交的学位论文印刷本和电子版本，学校有权将学位论文的全部或者部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版。本人允许本学位论文被查阅和借阅，同意学校向国家有关部门或机构送交论文的复印件和电子版，并通过网络向社会公众提供信息服务。

保密的学位论文在解密后适用本授权书

学位论文作者签名（手写）：　　导师签名（手写）：

签字日期：　　年　月　日　　签字日期：　　年　月　日

摘　要

多智能体强化学习（MARL）因其分散式决策能力和良好适应性，已成为解决动态柔性作业车间调度问题（DFJSP）的有前景方法。然而，MARL 在 DFJSP 中仍面临由多智能体交互引发的环境非稳态、奖励稀疏以及策略收敛缓慢等挑战。为此，本文提出了一种面向机器对齐的 MARL 框架，通过利用优先级派工规则（Priority Dispatching Rules, PDRs）构建抽象化的状态—动作表示，将不同类型机器的决策逻辑统一，从而使单一策略即可协调多智能体的调度行为，有效缓解因智能体交互产生的非稳态问题。在此基础上，本文引入在线专家机制，通过实时生成专家动作对智能体策略进行引导，使其在训练早期即可形成优良的决策倾向，加快收敛并提升策略稳定性。同时，结合好奇心机制构建内在奖励，使智能体在奖励稀疏或延迟的情况下仍能主动探索关键调度状态，从而增强策略的泛化能力与环境适应性。实验结果表明，所提方法在多智能体 DFJSP 环境下能够显著降低作业延迟，提高调度效率，并表现出良好的策略协调性与学习稳定性。

关键字: 关键字 1，关键字 2，关键字 3

Abstract

dd

Keywords: keyword1, keyword2, keyword3

目　录

第一章 引言

1.1 研究背景

现代制造业生产环境日益复杂，具有高柔性、高动态性的特点。动态柔性作业车间调度问题 (DFJSP) 是其中最具挑战性的优化范式之一。调度通过对生产资源的合理安排, 以缩短生产时间、提高资源利用率、降低生产成本, 在生产系统中作用显著。DFJSP 不仅需要决定工序的机床选择（路由）和加工顺序（排序），还必须实时应对一系列动态事件，如随机工件到达、紧急插单、机床故障以及加工时间不确定性等。这些不确定性使得生产环境处于随机性和实时性的持续变化中。传统的调度方法，包括启发式规则（如 FIFO、SPT、EDD 等）和元启发式算法（如遗传算法、粒子群优化等），主要依赖于离线优化。它们在面对 DFJSP 的复杂和动态环境时，表现出适应性差、计算开销大和难以实时决策的局限性[1-2]。

为克服这些限制，以深度强化学习 (DRL) 为代表的数据驱动框架应运而生。DRL 结合了深度学习强大的感知能力和强化学习的自适应决策能力，为车间调度提供了实时、高质量的解决方案。在 DFJSP 中，由于存在多台机器和多个工件需要协调，多智能体强化学习 (MARL) 通过结合 RL 的自适应决策与多智能体系统 (MAS) 的分散式协调，被认为是一种特别适合解决这一复杂、高维、非线性问题的强大工具。

1.2 多智能体强化学习在 DFJSP 中的局限

虽然多智能体强化学习（MARL）为动态柔性作业车间调度问题（DFJSP）提供了新的研究思路，但其在现有研究中的应用仍面临若干关键挑战。这些挑战主要源于 DFJSP 的复杂性以及现有 MARL 框架在异构系统中固有的局限性。

在先前关于 DFJSP 的 MARL 研究中，调度问题通常被建模为异构智能体系统，即由具有不同角色的智能体（例如负责路由决策的工件智能体和负责排序决策的机器智能体）组成[3-6]。

非稳态性：这些智能体存在不对称的观测空间、异构动作空间（路由 vs. 排序）以及事件驱动、异步的决策机制，其动作集随着调度状态动态变化。由于智能体策略的更新会直接改变其他智能体的决策环境，这种交互导致系统呈现强烈的非稳态性，使得学习动力学不稳定[7,6]。尽管现有工作广泛采用集中式训练-分散式执行（CTDE）[1]、交替训练[8-9] 等方法缓解非稳态问题，但策略耦合和协调开销仍然不可避免，异构系统中的非稳态问题并未得到根本解决。

稀疏奖励：在 DFJSP 中，关键性能指标（如作业完成时间、总 tardiness）通常只能在作业完成或整个调度过程结束时才能获得评估。这导致智能体在训练早期难以获得有

效反馈，学习信号稀疏，从而显著降低策略探索效率，并增加训练过程中的不确定性。

收敛困难：非稳态环境和稀疏奖励的双重影响，使得 MARL 系统在 DFJSP 中的收敛速度往往较慢。尤其在大规模或高度动态的车间环境下，智能体之间的协调和策略优化难度显著增加，容易陷入局部最优或产生震荡学习行为，进一步加剧了收敛困难。

综上所述，现有 MARL 方法在处理 DFJSP 时仍存在多智能体交互引发的非稳态性、奖励稀疏以及收敛困难等核心问题。这些挑战表明，需要设计针对智能体间交互的协调机制、改进奖励信号以及加速策略收敛的创新方法，从而为后续提出的单链决策、在线专家引导及内在奖励机制提供研究动机

1.3 本文研究内容与创新点

本文针对动态柔性作业车间调度问题（DFJSP）中多智能体强化学习（MARL）面临的非稳态、奖励稀疏与收敛困难问题，提出了一系列研究内容和创新方法，主要包括以下三个方面：

（1）机器对齐的抽象化 MARL 建模框架本文通过优先级派工规则（Priority Dispatching Rules, PDRs）构建抽象化的状态—动作表示，将不同类型机器的决策逻辑在逻辑上统一，实现单一策略对多智能体的协调控制。该方法不仅简化了异构智能体的建模复杂性，也有效缓解了多智能体交互引发的环境非稳态问题，为高效、可扩展的调度策略学习提供基础。

（2）融入模仿学习的专家引导探索策略为加速策略收敛并提升训练稳定性，本文引入在线专家机制，通过实时生成或提供专家示例对智能体策略进行引导，使智能体在训练早期便能够形成优良的决策倾向。该策略能够在复杂、多智能体交互的环境中有效减少盲目探索，提高策略学习效率。

（3）融入内在激励机制以提升稀疏奖励下的探索能力针对 DFJSP 中奖励稀疏或延迟反馈的问题，本文设计了基于好奇心的内在奖励机制，鼓励智能体主动探索潜在关键调度状态。这一机制不仅增强了智能体在稀疏奖励环境下的探索能力，也显著提高了策略的泛化能力与适应性，为动态调度环境中的稳健决策提供支持

1.4 论文结构安排

第二章 相关研究综述

2.1 FJSP 与 DFJSP】

经典调度模型

动态调度与再调度策略

研究进展（近 2-5 年重点）

2.2 启发式与元启发式方法

常见 PDR 介绍

PDR 在动态环境中的优势与局限

2.3 强化学习在 DFJSP 中的应用

单智能体 RL

动态调度 RL 方法综述

现有 RL 的扩展性瓶颈

2.4 多智能体强化学习在 DFJSP 中应用

训练范式（集中式训练、分散式执行）

局限：非稳态性、信用分配、异质性

2.5 模仿学习

行为克隆 BC

专家策略选择

混合专家 / 多规则专家

2.6 内在奖励

RND、ICM 等机制

稀疏奖励中的探索优势

2.7 本章小结

第三章 动态柔性作业车间调度问题建模

3.1 问题定义

3.2 动态事件模型

3.3 性能评价指标

3.4 本章小结

第四章 机器对齐的多智能体强化学习方法

4.1 机器对齐（Machine Alignment）与 MDP 建模

4.2 状态空间

4.3 动作空间

4.4 模仿学习的引入

4.5 奖励设计与内在奖励机制的引入

4.6 多智能体策略优化框架

4.7 本章小结

第五章 实验设计与结果分析

5.1 实验环境

5.2 训练细节

5.3 实验结果

5.4 消融实验

5.5 本章小结

第六章 总结与展望

结论是一篇学位论文的收尾部分，是以研究成果为前提，经过严密的逻辑推理和论证所得出最终的、总体的结论。换句话说，结论应是整篇论文的结局，而不是某一局部问题或某一分支问题的结论。结论应体现学生更深层的认识，且从全篇论文的全部材料出发，经过推理、判断、归纳等逻辑分析过程而得到的新的学术总观念、总见解。

结论是论文主要成果的总结，客观反映了论文或研究成果的价值。论文结论与问题相呼应，同摘要一样可为读者和二次文献作者提供依据。结论的内容不是对研究结果的简单重复，而是对研究结果更深人一步的认识‘是从正文部分的全部内容出发，并涉及引言的部分内容，经过判断、归纳、推理等过程而得到的新的总观点。毕业论文的研究结论通常由三部分构成：研究结论、不足之处、后续研究或建议。

第一，毕业论文的结论主要是由研究的背景与问题、文献综述、研究方法、案例资料分析与整理等研究得到的，其中核心的结论是正文部分的资料分析与研究的结果得出的结论和观点，即论文的基本结论。本研究结论说明了什么问题，得出了什么规律性的东西，解决了什么实际问题。研究结论必须清楚地表明本论文的观点，有什么理论背景的支持，对实践有什么指导意义等，若用数字来说明则效果嫌佳，说服力最强。不能模棱两可，含糊其辞。避免使人有似是而非的感觉，从而怀疑论文的真正价值。

第二，研究的不足，表明本论文的局限性所在，包括研究假设、资料收集、研究方法方面的不足之处，可以为后来的研究在该领域进一步完善指明方向。对于一篇学位论文的结论，上述基本结论是必需的，而不足之处和研究建议则视论文的具体内容可以多论述或少论述。论文的结论部分具有相对的独立性，应提供明确、具体的定性和定量信息。可读性要强。

参考文献

[1] Liu R, Piplani R, Toro C. A deep multi-agent reinforcement learning approach to solve dynamic job shop scheduling problem[J]. Computers & Operations Research, 2023, 159: 106294.

[2] Zhao Z, Zhou M, Liu S. Iterated greedy algorithms for flow-shop scheduling problems: A tutorial[J/OL]. IEEE Transactions on Automation Science and Engineering, 2022, 19(3): 1941-1959. DOI: 10.1109/TASE.2021.3062994.

[3] Zhang L, Yan Y, Yang C, et al. Dynamic flexible job-shop scheduling by multi-agent reinforcement learning with reward-shaping[J]. Advanced Engineering Informatics, 2024, 62: 102872.

[4] Kaven L, Huke P, Göppert A, et al. Multi agent reinforcement learning for online layout planning and scheduling in flexible assembly systems[J]. Journal of Intelligent Manufacturing, 2024, 35(8): 3917-3936.

[5] Pu Y, Li F, Rahimifard S. Multi-agent reinforcement learning for job shop scheduling in dynamic environments[J]. Sustainability, 2024, 16(8): 3234.

[6] Jing X, Yao X, Liu M, et al. Multi-agent reinforcement learning based on graph convolutional network for flexible job shop scheduling[J]. Journal of Intelligent Manufacturing, 2024, 35(1): 75-93.

[7] Son K, Kim D, Kang W J, et al. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]//International conference on machine learning. PMLR, 2019: 5887-5896.

[8] Liu R, Piplani R, Toro C. Deep reinforcement learning for dynamic scheduling of a flexible job shop[J]. International Journal of Production Research, 2022, 60(13): 4049-4069.

[9] Gergely M I. Multi-agent deep reinforcement learning for collaborative task scheduling. [C]//ICAART (3). 2024: 1076-1083.

致　谢

攻读学位期间的研究成果

已发表论文:

1. XXX，XX. 国内经济相互作用研究 [J]. 有色金属科学与工程，2010，21(3): 70-74.
2. XXX，XX. 风化壳淋积型稀土矿提取除杂技术现状及进展 [J]. 稀土，2011 (已录用).