上海交通大学硕士学位论文

面向量子任务调度的云平台设计与优化

姓　　名：刘子涵

学　　号：123260910064

导　　师：马汝辉

院　　系：巴黎卓越工程师学院

学 科/专 业：电子信息

申 请 学 位：硕士专业学位

2026 年 5 月 15 日

A Dissertation Submitted to

Shanghai Jiao Tong University for the Degree of Master

DESIGN AND OPTIMIZATION OF A CLOUD PLATFORM FOR QUANTUM TASK SCHEDULING

Author: Liu Zihan

Supervisor: Prof. Ma Ruhui

Shanghai Jiao Tong University

Shanghai, P.R. China

May 15th, 2026

摘 要

随着量子计算技术的快速发展，含噪声中等规模量子（Noisy Intermediate-Scale Quantum, NISQ）计算已成为当前量子硬件的主流形态。NISQ 时代的量子计算机拥有数十至数百个量子比特，但受限于有限的量子比特数量、不完美的门操作以及极短的相干时间，量子系统的实际可用计算能力受到严重制约。如何在受限的量子资源下通过高效调度实现可靠性、吞吐量与公平性的最大化，已成为学术界与工业界共同关注的焦点。针对上述挑战，本文围绕含噪声中等规模量子计算环境下的量子任务调度与量子比特映射问题开展了系统性的研究工作。

本文的主要研究工作包括两个方面：

其一是构建了面向量子任务调度的强化学习仿真平台与性能对比系统。针对现有研究缺乏统一评测基准的问题，本文设计并实现了一个功能完备的量子任务调度仿真平台。该平台对分布式量子计算环境进行了系统建模，引入了系统级与节点级的双层拓扑结构以准确表述量子计算中的连通性约束，并结合量子硬件的异构特性在量子比特层面建模了相干时间、门错误率、读出保真度、工作频率等关键属性。在此基础上，本文将量子任务调度问题形式化为带有多约束条件的优化问题，并进一步转化为马尔可夫决策过程，设计了融合执行时间、截止时间约束、资源冲突以及拓扑映射开销等多维指标的多目标奖励函数。基于该统一建模框架，本文分别构建并实现了基于 DQN、A2C、A3C 及 PPO 等多种经典强化学习智能体，并在所搭建的仿真平台上对其进行了系统性的性能对比分析，从而验证了不同强化学习策略在复杂量子调度场景中的适用性与优劣，为后续研究提供了可靠的评测基准。

其二是探索了量子-经典融合的强化学习调度方法，进一步拓展了强化学习在量子计算领域的应用边界。本文引入变分量子电路作为特征变换与函数逼近模块，构建了混合量子-经典强化学习模型。该方法通过将经典状态特征映射至量子态空间，并利用参数化量子线路对其进行变换处理，从而增强模型在高维希尔伯特空间中的特征表达能力。本文基于此理念设计并实现了 VQC-DQN、VQC-A2C、VQC-A3C 及 VQC-PPO 多种量子增强型智能体，探索了量子计算与强化学习深度融合的可行性。

实验结果表明，本文所设计的仿真平台能够有效对比各类强化学习智能体在不同资源配置与奖励函数配置下的性能表现。其中，DQN 成功率较高，PPO 在复杂约束下收敛路径更受控但平均运行时间较长，而 A2C 与 A3C 则普遍面临成功率瓶颈。

此外，引入变分量子层在某些场景下显著增强了模型的高维特征映射能力。实验结果显示，VQC-PPO 架构比经典 PPO 框架的调度耗时缩减了约 60%，在维持高成功率的同时大幅压缩了执行开销，有力验证了量子增强机制在特定策略梯度框架下的能效优势。本文的研究成果为量子计算环境下的智能资源调度提供了理论支撑与技术路径，对推动量子计算与强化学习的深度融合具有重要的参考价值，同时为未来在实际量子硬件上部署智能调度系统奠定了坚实的研究基础。

关键词：强化学习，量子计算，量子资源调度，变分量子电路

Abstract

With the rapid development of quantum computing technology, Noisy Intermediate-Scale Quantum (NISQ) computing has become the mainstream form of current quantum hardware. NISQ-era quantum computers possess dozens to hundreds of qubits, but their actual available computing capacity is severely limited by finite qubit numbers, imperfect gate operations, and extremely short coherence times. How to achieve maximum reliability, throughput, and fairness through efficient scheduling under constrained quantum resources has become a focal point of common concern in both academia and industry. Addressing these challenges, this dissertation conducts systematic research on quantum task scheduling and qubit mapping in NISQ computing environments.

The main research work of this dissertation includes two aspects:

The first aspect is the construction of a reinforcement learning simulation platform and performance comparison system for quantum task scheduling. To address the lack of unified evaluation benchmarks in existing research, this dissertation designs and implements a fully functional quantum task scheduling simulation platform. This platform provides a systematic modeling of the distributed quantum computing environment, introducing a two-level topology structure at the system level and node level to accurately represent the connectivity constraints in quantum computing. Combined with the heterogeneous characteristics of quantum hardware, key attributes such as coherence time, gate error rates, readout fidelity, and operating frequency are modeled at the qubit level. Based on this, the quantum task scheduling problem is formalized as an optimization problem with multiple constraints, and further transformed into a Markov Decision Process (MDP). A multi-objective reward function integrating execution time, deadline constraints, resource conflicts, and topological mapping overhead is designed. Based on this unified modeling framework, various classical reinforcement learning agents based on DQN (Deep Q-Network), A2C (Advantage Actor-Critic), A3C (Asynchronous Advantage Actor-Critic), and PPO (Proximal Policy Optimization) are constructed and implemented. Systematic performance comparison and analysis are conducted on the built simulation platform, verifying the applicability and characteristics of different reinforcement learning strategies in complex quantum scheduling scenarios, and

providing a reliable evaluation benchmark for subsequent research.

The second aspect explores quantum-classical hybrid reinforcement learning scheduling methods, further expanding the application boundary of reinforcement learning in quantum computing domains. This dissertation introduces Variational Quantum Circuits (VQCs) as feature transformation and function approximation modules to construct a hybrid quantum-classical reinforcement learning model. By mapping classical state features to quantum state space and transforming them using parameterized quantum circuits, the model's feature expression capability in high-dimensional Hilbert space is enhanced. Based on this concept, this dissertation designs and implements various quantum-enhanced agents including VQC-DQN, VQC-A2C, VQC-A3C, and VQC-PPO, exploring the feasibility of deep integration between quantum computing and reinforcement learning.

Experimental results demonstrate that the simulation platform designed in this paper can effectively compare the performance of various reinforcement learning agents under different resource and reward function configurations. Specifically, DQN achieves a high success rate, and PPO exhibits a more controlled convergence path under complex constraints despite a longer average execution time, whereas A2C and A3C generally face significant bottlenecks in success rates. Furthermore, the integration of variational quantum layers significantly enhances the model's high-dimensional feature mapping capabilities in certain scenarios. Experimental data shows that the VQC-PPO architecture achieves a reduction in scheduling latency of approximately 60% compared to the classic PPO framework. By substantially compressing execution overhead while maintaining a high success rate, this validates the energy-efficiency advantages of quantum enhancement mechanisms within specific policy gradient frameworks. The research findings of this dissertation provide theoretical support and technical pathways for intelligent resource scheduling in quantum computing environments, offering significant reference value for promoting the deep integration of quantum computing and reinforcement learning, while laying a solid foundation for the future deployment of intelligent scheduling systems on actual quantum hardware.

Key words: Reinforcement Learning, Quantum Computing, Quantum Resource Scheduling, Variational Quantum Circuit

目 录

插 图

表 格

第 1 章 绪论

1.1 研究背景及意义

随着计算科学技术的飞速发展，量子计算（Quantum Computing）已经成为了推动自然科学和其他相关前沿领域快速发展的重要力量。量子计算能够凭借其独特优势在数值模拟等关键任务中提供显著的、甚至是指数级的加速，因此高能物理学界也正积极利用量子算力对经典技术难以实现的理论模型进行处理，并以此支撑如大型强子对撞机升级等实验所带来的海量数据挑战[1]。随着变分量子算法（Variational Quantum Algorithms, VQAs）[2]和量子机器学习（Quantum Machine Learning, QML）[3-4]相关技术的不断发展，量子计算在现实世界中可能会被进一步广泛应用。尽管具有广阔的前景，现有的量子计算相关设备仍普遍面临着硬件规模受限与稳定性不足的严峻挑战。受限于有限的量子比特数量、不完美的门操作以及极短的相干时间，当前技术仍处于含噪声的中等规模量子（Noisy Intermediate-Scale Quantum, NISQ）阶段[5]。这种硬件层面的局限性不仅限制了系统的整体吞吐量，也对大规模、无错误计算的实现提出了更高要求的调度与资源管理策略[6-7]。

在此背景下，如何在受限的量子资源下通过高效调度实现可靠性、吞吐量与公平性的最大化已成为学术界共同关注的焦点[8-10]。与传统计算集群不同，量子计算设备具有硬件异质性特点（量子比特的物理特性并非完全均等）[11]，其连接性、错误率及相干时间的波动直接决定了计算任务的可行性。与之相矛盾的是，现有的大部分研究仍倾向于将硬件属性抽象化并将量子比特视为同质单元，从而忽略了读出错误率、串扰以及特定拓扑连接等关键约束条件。这种抽象模拟与真实硬件执行之间的失配亟需一种能够深度集成量子比特级与系统级异质性的调度框架[12]进行解决。

针对上述挑战，强化学习（Reinforcement Learning, RL）能够凭借其在处理复杂动态环境中的良好表现[13]为量子计算系统的自适应资源管理方式提供了新的问题解决思路。相较于传统的启发式调度策略，RL 智能体能够通过与量子环境的持续交互在即时执行的可行性与长期系统的整体效率之间找到最佳平衡[14]。随着深度强化学习（Deep Reinforcement Learning, DRL）及多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）[15]等技术的发展逐渐成熟，利用这些技术在不确定性下的自适应能力来优化量子资源配置已成为未来技术发展的必然趋势。为此，本文提出了 QRAP（Quantum Resource Allocation Platform For Adaptive Scheduling Under

Topology Constraints），一个专为分布式 NISQ 系统设计的拓扑感知且硬件敏感的调度框架。本文探索了量子-经典融合的强化学习调度方法，并引入变分量子电路（Variational Quantum Circuit, VQC）作为特征变换与函数逼近模块，构建了混合量子-经典强化学习模型。

与以往的相互独立的研究模式不同，本研究提出的 QRAP 将量子调度从单一算法设计提升到了系统级平台的高度进行研究。QRAP 框架不仅统一了硬件感知建模与任务级抽象，使得调度决策能够显式地权衡量子比特可用性、连接拓扑及噪声特性。更将强化学习作为系统的原生组件，通过将调度问题严谨地映射为马尔可夫决策过程，从而实现了对动态任务到达和时变资源状态的持续优化。

1.2　国内外研究现状

本节对量子计算、强化学习及其在复杂系统资源调度领域的交叉应用相关的国内外研究现状进行系统性介绍。首先，第1.2.1节对量子计算已有研究进行分析，旨在阐明 NISQ 阶段底层硬件的物理约束、量子加速的理论基础以及变分量子电路等核心混合范式的发展态势。其次，第1.2.2节对强化学习在资源调度领域的已有研究进行介绍，剖析强化学习在处理不确定环境序贯决策时的理论优势及经典调度架构的应用演进。最后，针对传统深度强化学习在应对高维状态空间时面临的算力瓶颈与维度灾难，第1.2.3节重点对量子强化学习在资源调度方面的应用相关工作进行介绍，探讨其如何利用量子叠加与纠缠特性突破经典算力限制，实现复杂架构下的高效资源管理与动态寻优。

1.2.1　量子计算已有研究分析

量子计算现已正式步入 NISQ 阶段，尽管一部分拥有数十个量子比特的量子计算机或许已具备执行超越当今经典数字计算机完成任务的能力，但量子门中普遍存在的噪声依然极大程度上限制着能够可靠运行的量子电路的规模。令人不可忽略的是，受限于相干时间不足、错误率较高以及稀疏的比特连通性[5]等问题，目前的设备仍通过超导电路、离子阱、中性原子阵列以及光子架构等多种硬件途径进行实验改进[16-19]。量子计算在物理本质上的核心挑战在于无法在不引起失控干扰的前提下观察量子系统。为了能够可靠地处理信息，必须将系统与外界隔绝，但与此同时为了计算性能又需要比特间的强力相互作用。为研究此类问题，魏世杰等人提出了开放量子系统的对偶量子模拟算法[20]。在此背景下，VQC 已经成为实现量子优势的关键

混合范式，通过在量子处理器上执行参数化逻辑门，同时利用经典优化器迭代更新，VQC 具备了能够在较浅的电路深度下处理更复杂的计算任务的能力，展现出对硬件噪声的独特韧性[2,21]。此外，系统开发者还必须能够从外部精准操控，并最终通过读取系统内部量子比特的状态来获取计算结果。

与传统的经典高性能计算不同，一些特定的任务在量子节点上的处理难度远低于经典计算节点[22-23]。然而，量子计算并非是完美无缺的，例如它无法涵盖经典计算所面临的所有问题。目前的量子计算算法在组合优化领域，例如量子近似优化算法（Quantum Approximate Optimization Algorithm, QAOA）通过将优化任务映射到量子态测量，为经典算法难以处理的问题提供了新的解决路径[24-25]。比如，针对 NP-hard 属性的车间调度问题（Job Shop Scheduling Problem, JSSP），Kurowski 等研究者利用特定的成本哈密顿量寻找最优调度策略[26]。Prasad 与 Masthan 等研究者进一步提出的混合量子调度优化算法（Quantum Scheduling Optimization Algorithm, QSOA）通过约束感知损失函数改进了 QAOA，在航司排班等任务中实现了显著的成本削减[27]。对于更复杂的柔性车间调度问题（Flexible Job Shop Scheduling Problem, FJSSP），基于量子退火的求解算法（Quantum Annealing-based Scheduling Algorithm, QASA）在优化完工时间以及负载平衡等冲突目标上具有良好的表现[28]。尽管如此，对于此类组合搜索问题，目前的研究仍难以获得远超详尽搜索的突破性结果。量子计算机虽然可以加速详尽搜索[29]，但这种提升程度是相对有限的[30]。

针对量子加速争议问题，近期的理论突破成果表明在近似大型矩阵的特征向量时，量子算法具有可观的多项式级加速优势。这意味着即使在噪声干扰下，量子计算在基础线性代数任务中依然保有稳固的优势地位[31]。QML 及与其密切相关的量子强化学习领域正经历飞速发展阶段，VQC 也被视为经典神经网络的量子对应产物，其利用量子特征映射方法将数据投影至高维希伯特空间以增强表征能力[32-33]。尽管量子核方法在一些特定任务下可能面临比数据重上传等模型更高的样本复杂度要求[34]，但该方法在回归分析与微分方程任务中的应用成果已证明了其最优收敛的潜力[35-36]。

为了克服大型 VQC 训练中常见的“贫瘠高原”现象（Barren Plateaus, BP）并提升可训练性[37-38]，目前的研究重点已转向参数初始化策略与对称性保持的电路设计[39-40]。为了提升学习效率，引入量子电路经验回放（Quantum Circuit Experience Replay, QCPR）机制的混合架构能够通过评估经验的重要性来进一步优化决策质量[41]。此外，多智能体量子强化学习（Multi-Agent Quantum Reinforcement Learning, MAQRL）利用量子纠缠和耦合注意力机制提升了智能体间的协作效率[42]。最新的 ARDNS-FN-

Quantum 框架相关成果则表明通过奖励方差和好奇心驱动的自适应策略，量子增强模型在动态环境中能够实现更快的收敛速度和更强的探索稳定性[43]。

1.2.2　强化学习资源调度已有研究分析

强化学习的特征为在不确定环境下的序贯决策过程提供了严谨的理论框架。强化学习, 尤其是深度强化学习, 在系统资源调度与分配领域展现出了突破性的应用潜力。在边缘计算与物联网场景下，资源调度任务主要涉及计算卸载决策与虚拟机 (Virtual Machine, VM) 分配，早期 Mao 等人[44]的经典工作首次证明了直接将调度问题建模为马尔可夫决策过程 (Markov Decision Process, MDP) 并使用策略梯度算法优化的可行性。近期 Wang 等人的研究提出了一种基于 DRL 的云边协同计算资源调度框架[45]，显著提升了任务处理效率并有效控制了任务迁移延迟。针对大规模物联网设备产生的海量碎片化任务，Li 等人[46]利用深度 Q 网络（Deep Q-Network, DQN）探讨了异构节点间的计算与缓存资源联合分配问题。该模型能够根据边缘节点的实时负载状况，动态调整任务的执行队列。

深度强化学习领域的经典进展为复杂环境下的调度难题提供了强大的分析工具，特别是 DQN[47]与近端策略优化（Proximal Policy Optimization, PPO）[48]的应用。DQN 依托基于价值的方法对调度与动作映射进行离散化，并结合经验回放和目标网络机制显著提升了学习效率与系统稳定性。PPO 则采用策略梯度方法在高维随机环境中维持了策略更新的平稳性。

除 DQN 与 PPO 等主流框架外，近期的研究亦开始将强化学习的其他高级架构引入资源管理中。例如在分布式调度场景下，Actor-Critic 架构被用于提升并行处理能力与样本效率[49]。面对高维连续动作空间，深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）[50]或软 Actor-Critic（Soft Actor-Critic，SAC）等方法被部署以学习细粒度的量子比特映射策略[51]。与此同时，Zhang 等人的研究工作将 MARL[52]引入任务调度系统，旨在破解跨多节点的协同分配难题。上述现有研究进展充分表明了 RL 不仅能够胜任静态调度场景下的优化任务，更在分布式与多租户环境中展现出了较强的可扩展性。

1.2.3　量子强化学习资源调度已有研究分析

近年来，量子强化学习 (Quantum Reinforcement Learning, QRL) 作为一种融合量子计算与深度强化学习的前沿技术，在解决复杂资源调度问题中展现出突破性潜力。针对高维状态空间带来的维度灾难，QRL 通过量子叠加与纠缠特性有效降低了模型

复杂度并加速收敛。

在移动边缘计算与无线通信网络方面，研究者重点探索了混合量子架构。Wei 等人[53]提出了一种基于 VQC 的混合模型，专门用于边缘计算中的任务卸载与联合资源分配任务，大幅减少了神经网络的参数量。针对未来日益复杂的通信网络（如 6G 网络），Astiti 与 Lee 等人的综述[54]系统总结了 QRL 在功率控制与信道分配中的应用范式，指出了量子框架在克服组合优化瓶颈上的算力优势。此外，Nguyen 等人设计了变分量子彩虹深度 Q 网络 (Variational Quantum Rainbow Deep Q-Network，VQR-DQN)[55]，在处理 NP 难级别的调度分配任务时，有效提升了决策效率并显著缩短了完工时间。

在云计算平台与底层量子网络管理中 QRL 同样取得了实质性进展。例如，Dai 等人将 QRL 创新性地应用于云端实时作业调度，利用量子特征提取显著提升了系统在高负载突发情况下的作业执行成功率[56]。

1.3 主要研究内容与研究思路

本文围绕含噪声 NISQ 计算环境下量子资源受限、硬件异构性显著以及任务调度复杂等问题，针对量子任务调度与量子比特映射的高效实现展开研究。本文以强化学习为核心方法，结合量子计算系统的物理约束与拓扑结构特性构建统一的调度建模框架，并在此基础上开展经典强化学习方法对比分析及量子增强调度方法探索。整体研究内容可归纳为以下两个方面。

首先，本文基于经典强化学习方法构建多种量子任务调度智能体，并设计实现了仿真平台以对不同算法的调度性能进行系统性评估。在系统建模方面，本文对分布式量子计算环境进行了抽象，并通过引入系统级与节点级的双层拓扑结构用以表述量子计算中的连通性约束。同时，结合量子硬件的异构特性，在量子比特层面建模了相干时间、门错误率、读出错误率等关键属性。在此基础上，本文将量子任务调度问题形式化为带有多约束条件的优化问题，并进一步转化为马尔可夫决策过程，设计了多目标奖励函数。基于该统一建模框架，分别构建并实现了 DQN、A2C/A3C 以及 PPO 等多种强化学习智能体，并在所搭建的仿真平台上对其进行了性能对比分析，从而验证不同强化学习策略在复杂量子调度场景中的适用性与优劣特性。

其次，本文进一步探索量子-经典融合的强化学习调度方法，引入 VQC 作为特征变换与函数逼近模块，构建了一系列量子-经典混合强化学习模型。具体而言，本研究通过角度嵌入技术将经典的系统状态特征映射至高维的量子希尔伯特空间，并利

用多层强纠缠量子线路对其进行深度的非线性特征提取，最后通过 Pauli-Z 基下的期望测量将量子特征反馈至经典决策层。基于此量子-经典混合架构，本文分别设计并实现了 VQC-DQN、VQC-A2C、VQC-A3C 以及 VQC-PPO 等量子增强型强化学习智能体。此外，本文在上述所实现的相同的仿真平台中对量子-经典混合强化学习智能体开展了多维度的性能对比研究，全面评估了量子变分层在提升策略收敛稳定性与增强状态表征能力方面的作用，并深入剖析了量子增强机制对于不同量子-经典混合强化学习智能体在不同资源密度配置下对系统整体任务成功率与执行效率所带来的实质性影响。

1.4　文章结构

本文分为七个章节，具体结构如下：

- **第1章**对本文的研究现状及相关研究工作进行了详细介绍，分析了当前存在的问题与挑战，并据此提出了本文的研究内容和目标。
- **第2章**介绍了量子物理特性、变分量子电路及强化学习等核心理论，为后续章节构建智能调度框架提供了数学依据与理论支撑。
- **第3章**围绕量子任务调度云平台建模与智能优化方法展开研究。首先，探讨了分布式量子计算环境的系统建模方法，通过建立异构算力单元的分层拓扑模型及电路执行时间估算模型，将量子任务调度与量子比特映射问题转化为带有复杂约束条件的数学优化问题。其次，构建了面向量子任务调度的强化学习框架，详细论述了状态空间映射、动作空间设计、多目标奖励函数构建及 DQN、A2C、A3C 和 PPO 等经典强化学习智能体的建模方案。在此基础上，本章进一步引入量子-经典融合的混合智能调度算法架构，论述了 VQC 增强型智能体的设计思路，通过角度嵌入与参数化量子线路实现高维调度状态的特征映射，并构建了 VQC-DQN、VQC-A2C、VQC-A3C 和 VQC-PPO 等混合智能体，从而完成任务调度与量子比特映射的联合优化。最后，本章从计算复杂度、参数规模与潜在性能增益等维度分析了混合模型在处理高维量子调度问题时的适用性与优势。
- **第4章**介绍了量子调度仿真平台的工程构建方案。通过在 Gymnasium 框架中集成异构量子资源的分层拓扑模型，本章系统性地实现了包括经典算法与 VQC 混合算法在内的多类智能体调度逻辑。
- **第5章**对本文所需的对比实验进行了系统性设计，并详述了实验配置与参数基准。

通过对经典强化学习模型在多类奖励导向与资源密度下的自适应表现进行分析，确立了基准模型性能。在此基础上，建立了量子-经典混合智能体与经典算法的对比试验机制。

- **第6章**从收敛性、任务处理成功率及运行效率等维度评估了四类经典算法及其 VQC 增强型智能体在不同奖惩配置下的表现。
- **第7章**对全文的研究工作和成果进行了总结，并对未来的改进方向与工作做了展望。

第 2 章 技术背景及理论基础

本章将对本文涉及的核心理论与建模框架进行系统性梳理，主要涵盖量子计算基础理论与物理特性、任务调度云平台建模基础、变分量子电路及其机理，以及强化学习理论与马尔可夫决策过程四个方面。首先，本章从量子比特、态演化及测量机制出发，说明当前 NISQ 设备在有限相干时间、门错误率和读出误差等方面所受到的物理约束。其次，针对量子任务调度场景，本章阐述了由任务层、硬件层、调度层及评估层构成的建模基础，分析了硬件拓扑约束对逻辑比特映射开销的影响，并论证了该问题的序贯决策本质。随后，本章详细介绍变分量子电路的基本结构与量子-经典混合优化流程，阐明其通过参数化量子线路实现高维特征表达的基本原理。最后，本章对强化学习的交互机制和马尔可夫决策过程进行说明，为后续将量子任务调度问题转化为智能体与环境交互的序贯决策问题奠定理论基础。

2.1 量子计算基础理论与物理特性

量子计算科学建立在量子力学公理之上，与经典系统不同，量子计算的基本信息由量子比特作为载体。在纯态 (Pure State) 假设下，封闭量子系统的状态可用复希尔伯特空间 (Hilbert Space) 中的单位向量描述。单量子比特的状态 $|\psi\rangle$ 是计算基矢 $|0\rangle$ 与 $|1\rangle$ 的线性叠加：

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{2.1}$$

其中 α 和 β 为复数概率幅，且满足归一化条件 $|\alpha|^2 + |\beta|^2 = 1$。当系统向多比特扩展时，其状态空间的维度呈指数级增长，这也同时为处理高复杂度计算任务提供了基础的物理容量。

封闭量子系统的状态演化遵循薛定谔方程，在离散时间计算模型下表现为一系列酉变换 (Unitary Transformation)。对于给定的初始状态 $|\psi_0\rangle$ 与包含 d 个量子门操作的逻辑电路，其理想状态演化为：

$$|\psi_d\rangle = U_d U_{d-1} \cdots U_1 |\psi_0\rangle = \left(\prod_{k=1}^{d} U_k \right) |\psi_0\rangle, \tag{2.2}$$

这种酉演化不仅保证了计算过程的理论可逆性，同时也通过受控操作在多比特间生成非定域的量子纠缠 (Quantum Entanglement) 态。然而，系统的最终观测依赖于量子

测量 (Quantum Measurement)，测量操作会导致波函数依据概率幅的平方随机塌缩至确定性的经典基态。

在实际的物理部署中，底层硬件是一个开放量子系统 (Open Quantum System)。量子态在演化与测量过程中不可避免地会与外界环境发生相互作用，导致退相干 (Decoherence) 与操作失真。退相干本质上是量子系统内部的干涉相位信息向宏观环境的泄露过程，它迫使量子比特从维持并行算力的纯态退化为仅具统计概率的经典混合态 (Mixed State)。衡量这一物理衰减速率的核心指标为相干时间 (Coherence Time)，具体包含衡量能量耗散的纵向弛豫时间 T_1 以及描述相位信息丢失的横向退相干时间 T_2。相干时间构成了量子计算严格的物理截止期，任何量子逻辑门序列的总执行延迟必须远小于比特的相干时间，否则量子态将失去保真度且计算结果会被环境噪声完全吞没。在密度矩阵 (Density Matrix)ρ 的数学表述框架下，这种受到环境噪声干扰的量子信道演化过程由克劳斯算子 (Kraus Operator) 展开式表示为：

$$\mathcal{E}(\rho) = \sum_k E_k \rho E_k^\dagger + \int_0^t \mathcal{L}(\rho(t'))\,\mathrm{d}t', \tag{2.3}$$

其中 E_k 为表示离散错误类型的错误算子且满足完备性关系 $\sum_k E_k^\dagger E_k = I$，而积分项则代表了连续时间的马尔可夫噪声耗散 (主要受时间参数 T_1 与 T_2 主导的指数衰减过程)。这种物理层面的噪声累积直接破坏了量子算法的干涉路径。特别是在多租户并发的硬件调用场景下，相邻物理比特间的串扰效应 (Crosstalk) 与热弛豫过程会进一步急剧压缩实际的相干时间窗口并显著加速密度矩阵的非酉演化。因此，从系统调度的宏观视角来看，底层机制的根本差异使得计算任务的执行不再是简单的计算周期分配，而必须被建模为针对物理节点时变保真度的动态寻优过程。

2.2　任务调度云平台建模基础

鉴于目前 NISQ 设备受到退相干、门错误率、读出误差及拓扑连通限制的共同影响，量子云平台中的任务调度问题已不再是传统云计算环境下对 CPU、GPU 和内存等经典资源的分配。传统云平台通常可以将底层算力抽象为相对稳定且可重复调用的计算单元，而量子云平台中的基础资源则由具有物理噪声、有限相干时间和异构性能特征的量子比特构成。因此，量子任务调度不仅需要考虑任务队列、资源占用和执行延迟等经典调度因素，还必须进一步纳入量子比特质量、量子门操作误差、读出误差以及硬件拓扑连通关系等量子计算特有约束。

从系统建模角度来看，量子任务调度云平台可以被抽象为由任务层、量子硬件层、调度层和评估层共同构成的动态交互系统。其中，任务层负责描述用户提交的量子计算任务，包括所需量子比特数量、量子门复杂度、任务优先级、截止时间约束和并发执行标志等属性。量子硬件层由多个量子节点及其内部量子比特组成，每个节点可被视为具有有限量子比特容量、特定拓扑结构和硬件性能参数的异构算力单元。调度层根据当前系统状态完成任务选择、节点分配和量子比特映射等任务。评估层则依据任务完成情况、资源利用率、执行延迟和调度失败情况等指标对调度策略进行反馈。

在上述模型中，拓扑结构是影响量子任务可执行性的重要约束。由于当前量子硬件通常只允许特定物理量子比特之间直接执行双量子比特门操作，逻辑量子比特之间的交互关系必须被映射到满足硬件连通条件的物理比特集合上。当任务电路结构与底层硬件拓扑不匹配时，系统往往需要引入额外的交换门或跨节点通信操作，从而增加电路深度、执行时间和累积错误概率。因此，量子任务调度不能仅依据可用量子比特数量进行决策，而需要同时考虑节点内部拓扑、节点间全局拓扑以及由此产生的映射开销。

2.3 变分量子电路及其机理

鉴于上述开放量子系统面临的退相干与克劳斯噪声限制，传统包含大量逻辑门的深层量子算法在当前硬件下难以维持有效保真度。这促使 VQC 成为应对上述物理约束的核心混合计算范式。VQC 通过参数化的浅层量子电路与经典优化器的协同工作，展现出对非酉演化噪声的独特韧性。

VQC 的完整运行机理建立在量子状态制备、哈密顿量测量与经典参数优化的闭环之上。首先，在处理具体的学习或数据驱动任务时，需将经典信息 (如输入特征向量 $\boldsymbol{x}$) 编码至量子态中。这一过程通常通过特征映射模块 (Data Encoding) 完成，即初始态准备为 $|\psi_0\rangle = S(\boldsymbol{x})|0\rangle^{\otimes n}$，其中 $|0\rangle^{\otimes n}$ 为系统基态，$S(\boldsymbol{x})$ 为固定的编码电路 (如振幅编码或角度编码)。

随后，该状态进入 VQC 的核心，一个依赖于经典参数向量 $\boldsymbol{\theta}$ 的参数化量子操作。VQC 的演化算子 $U(\boldsymbol{\theta})$ 由一系列固定的纠缠门与可调的单比特旋转门交替拼接而成。参数化制备的量子态表示为：

$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_0\rangle = \left(\prod_{l=1}^{L} U_l(\theta_l)W_l\right)|\psi_0\rangle, \tag{2.4}$$

其中 L 为电路的层数，W_l 代表不含参数的固定拓扑纠缠操作 (如受控非门 CNOT 集合)，$U_l(\theta_l)$ 代表由参数 θ_l 控制的旋转操作 (通常为 R_x, R_y, R_z 门)。

在此结构中，可调旋转门主要负责探索单比特的希尔伯特子空间，而固定拓扑纠缠操作则利用量子纠缠特性在比特间传递信息，从而使整个电路具备在高维复杂空间中拟合目标函数的强大表达能力 (Expressibility)。根据设计逻辑，变分拟设 (Ansatz) 大致可分为两类:一类是针对特定物理或组合优化问题定制的启发式拟设 (如量子化学中的幺正耦合簇拟设 UCC)；另一类则是以硬件高效拟设 (Hardware-Efficient Ansatz,HEA) 为代表的架构，其直接利用底层硬件天然支持的门集与拓扑连通性，最大程度压缩电路深度以抑制退相干噪声。典型的变分量子电路拓扑结构如图2.1所示，其中单比特参数化门与双比特纠缠门交替排列形成循环层结构。

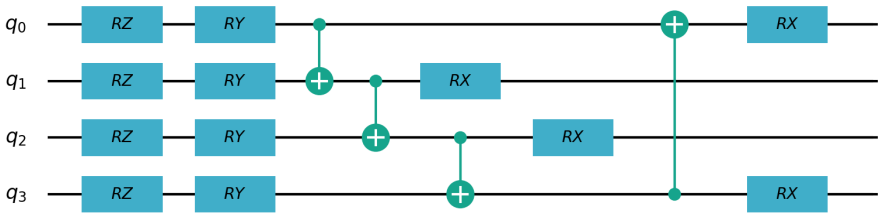

图 2.1　量子电路示意图
Figure 2.1　Quantum Circuit Diagram

在 VQC 的框架下，待求解的复杂计算任务被映射为一个特定的可观测量，即厄米特 (Hermitian) 哈密顿量 H。算法的最终目标是通过寻找最优参数组合 $\boldsymbol{\theta}^*$，最小化该哈密顿量在参数化量子态下的期望值。根据量子力学测量假设，目标损失函数定义为:

$$C(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle = \langle\psi_0|U^\dagger(\boldsymbol{\theta})HU(\boldsymbol{\theta})|\psi_0\rangle. \tag{2.5}$$

在实际计算中，复杂的哈密顿量 H 通常无法被直接测量，而是被分解为张量积形式的泡利字符串 (Pauli Strings) 的线性组合，即 $H = \sum_k c_k P_k$，其中 $P_k \in \{I, X, Y, Z\}^{\otimes n}$。由于量子态的塌缩特性，$C(\boldsymbol{\theta})$ 无法通过单次测量获取，量子硬件需要对每个泡利项进行大量重复制备与基态测量 (即采样 Shot)，以估算其统计平均值，最后在经典计算机中加权求和得出最终的成本函数值。

这完整刻画了 VQC 的量子-经典混合工作流 (Hybrid Loop)：量子协处理器负责高计算复杂度的状态演化与期望评估，而经典主节点则利用这些结果计算梯度并更新参数。为了利用经典优化器 (如梯度下降算法) 迭代更新参数 $\boldsymbol{\theta}$，必须获取损失函数

关于各参数的梯度。由于量子硬件黑盒的特性无法直接利用经典反向传播，学术界提出并严格证明了参数平移规则 (Parameter-Shift Rule) 定理。对于由泡利算符生成的旋转门 $U(\theta) = \exp(-\mathrm{i}\theta P/2)$，其精确梯度的计算展开式表示为：

$$\begin{aligned}\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_i} =& \frac{1}{2}\left[C\left(\boldsymbol{\theta}+\frac{\pi}{2}\boldsymbol{e}_i\right)-C\left(\boldsymbol{\theta}-\frac{\pi}{2}\boldsymbol{e}_i\right)\right] \\ =& \frac{1}{2}\left(\langle H\rangle_+ - \langle H\rangle_-\right),\end{aligned} \tag{2.6}$$

其中 $\boldsymbol{e}_i$ 为第 i 个参数方向的单位向量。该定理从数学层面上证明了量子电路的精确解析梯度可以通过将目标参数正反平移一个固定宏观角度后获得的两次量子设备评估结果的差值来获取，从而解决了量子与经典混合优化中的梯度求解难题。

尽管 VQC 具有极高的工程应用价值，但其在理论上受到 BP(Barren Plateaus，贫瘠高原) 定理的严格制约。BP 现象指出，当变分电路的规模足够大，使得参数化电路构成一个近似的酉 2-设计 (Unitary 2-design) 时，损失函数的梯度期望趋于零，且其方差随系统规模 (量子比特数 n) 呈指数级衰减。包含期望与方差项的梯度渐进衰减公式表示为：

$$\mathrm{Var}\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_i}\right] = \mathbb{E}\left[\left(\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right] - \left(\mathbb{E}\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \theta_i}\right]\right)^2 \approx O\left(\frac{1}{2^{cn}}\right), \tag{2.7}$$

其中 c 为与电路拓扑相关的正常数。BP 现象与前文提及的底层硬件串扰噪声共同作用使得在大规模异构系统上的 VQC 任务极易陷入训练停滞的困境。因此，如何通过智能调度策略缓解硬件噪声以规避梯度消失，进一步构成了量子-经典混合系统资源调度的重要理论动机。

2.4 强化学习理论与马尔可夫决策过程

强化学习主要研究智能体如何在与环境的交互过程中通过试错学习策略以最大化累积奖励信号。与监督学习不同之处在于强化学习不依赖于带有标签的静态数据集，而是通过环境反馈的奖励来指导动作选择的优化。在异构量子计算系统的资源调度场景中，调度器即为智能体，而量子硬件状态与计算任务队列共同构成了极其复杂的动态环境。为进一步直观展示智能体与环境之间的闭环交互机制，其交互逻辑示意图如图2.2 所示。

强化学习算法的数学理论基石是马尔可夫决策过程，通常由一个五元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ 来严谨定义。其中 $\mathcal{S}$ 表示状态空间，包含了系统可能处于的所有状态特征。$\mathcal{A}$ 表示

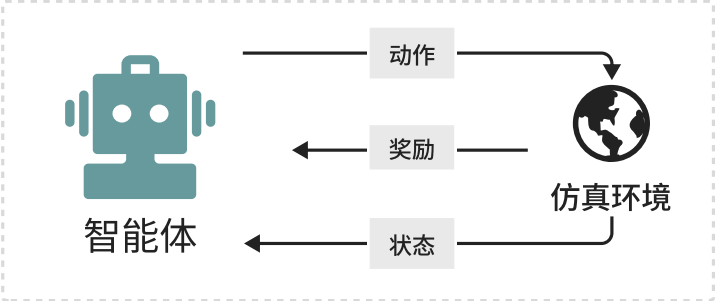

图 2.2　强化学习交互逻辑示意图
Figure 2.2　Reinforcement Learning Diagram

动作空间，代表智能体在任一状态下可执行的调度决策集合。$\mathcal{P}$ 为状态转移概率，决定了环境在受控状态下的动态演化规律。$\mathcal{R}$ 为系统反馈的奖励函数。$\gamma \in [0,1]$ 为折扣因子 (Discount Factor)，用于平衡即期奖励与远期奖励对当前决策的影响。当智能体在时间步 t 观测到状态 $s_t \in \mathcal{S}$ 并严格执行动作 $a_t \in \mathcal{A}$ 时，环境接收该调度动作并向下一个状态 s_{t+1} 进行转移，其条件概率转移矩阵表示为：

$$\begin{aligned}\mathcal{P}_{ss'}^{a} =& \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a] \\ =& \mathbb{P}[S_{t+1} = s' \mid S_1, A_1, \ldots, S_t = s, A_t = a].\end{aligned} \tag{2.8}$$

该等式不仅定义了状态演化的动态学特征，同时也直观体现了马尔可夫性，即系统的未来状态演化仅依赖于当前时刻的状态与动作，而与历史交互轨迹完全无关。伴随状态的每一次转移，环境都会向智能体反馈一个奖励信号 $r_t = \mathcal{R}(s_t, a_t)$。

为了严谨评估某一策略 π（即在给定状态下选择各项动作的概率分布）的长期表现，强化学习引入了状态价值函数 $V_\pi(s)$。它被严格定义为智能体从状态 s 出发并始终遵循既定策略 π 所能获得的期望折扣累积回报：

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right]. \tag{2.9}$$

该价值函数精确反映了特定状态的长期寻优潜能。进一步，状态价值函数可以通过贝尔曼期望方程 (Bellman Expectation Equation) 展开为递归推导形式。对于任意状态 s，其贝尔曼方程表示为：

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} V_\pi(s') \right). \tag{2.10}$$

在异构量子任务调度问题中，由于多维任务队列的连续到达以及底层硬件噪声参数的时变特性，系统建模后的状态空间 $\mathcal{S}$ 与动作空间 $\mathcal{A}$ 往往呈现出典型的高维连

续特征。传统的表格型强化学习求解方法 (如 Q-Learning 算法) 在处理此类高复杂度问题时会面临严重的维度灾难。因此，引入深度神经网络作为策略或价值函数高维近似器的深度强化学习架构打破了经典状态空间的表征瓶颈，正在成为解决混合量子系统复杂资源调度问题的核心理论工具。

2.5 本章小结

本章对本文研究所涉及的核心理论基础进行了系统性的梳理与阐述。首先，本章介绍了量子计算的基础物理特性，指出在现有开放量子系统下退相干与克劳斯噪声对量子态保真度的影响，从而阐明了传统吞吐量优先任务调度策略在量子架构中的失效机理。随后，本章阐述了量子任务调度云平台的建模场景与基础，分析了多层级架构下的动态交互机制及硬件拓扑约束。其次，本章详细介绍了变分量子电路作为量子-经典混合计算范式的数学演化与梯度求解机理，并进一步确立了通过智能调度缓解硬件噪声的研究动机与可行性。最后，本章还系统性地构建了强化学习与马尔可夫决策过程的数学框架，解释了状态价值函数与贝尔曼方程的推导逻辑。本章的理论探讨为后续章节中框架与系统的设计以及算法的实现提供了坚实的理论支撑与数学依据。

第 3 章 量子任务调度云平台建模与智能优化方法

基于第2章对于技术背景及理论，本章从理论建模到算法构建的角度，系统介绍量子计算环境下的任务调度与量子比特映射问题。针对 NISQ 阶段量子硬件资源有限、量子比特异构性显著、节点连通结构受限以及任务执行易受噪声影响等特点，本章首先对量子任务调度云平台的整体架构进行建模，明确任务层、量子硬件层、调度层与评估层之间的交互关系。在此基础上，本章进一步刻画量子节点、量子比特和量子任务等核心对象的属性特征，并通过分层拓扑结构描述节点间与节点内部量子比特之间的连接约束，从而为后续调度决策提供更加贴近真实硬件环境的系统状态表征。

随后，本章将量子任务调度与量子比特映射过程形式化为带有多重约束条件的优化问题，并围绕任务执行时间、资源占用情况、硬件兼容性、拓扑映射开销以及截止时间要求等因素建立数学描述。为了处理动态任务流和时变资源状态下的序贯决策问题，本章进一步引入强化学习框架，将调度器视为智能体，将量子云平台视为交互环境，构建包含状态空间、动作空间和奖励函数的马尔可夫决策过程模型。在此基础上，本章分别设计经典强化学习智能体与量子-经典混合强化学习智能体，系统说明 DQN、A2C、A3C、PPO 及其 VQC 增强型模型在量子任务调度场景中的建模方式与优化思路，为后续系统实现和实验分析奠定理论基础。

3.1 系统概览

所提出的系统对分布式量子计算环境进行了建模，如图3.1所示，构成整体流程框架的四个层中包含五个核心组件：量子比特、量子节点、量子任务、算法库以及评估模块。

这些模块共同为 NISQ 设备中的资源感知调度与性能优化提供了一个全面的框架。其中量子比特与量子节点集合在量子硬件层中，而算法库则由调度层中的多类强化学习智能体体现。

量子比特作为计算系统的基础计算单元。每个量子比特特征包含相干时间、门保真度、读出错误率以及工作频率等硬件级属性。这些特性捕捉了当前量子硬件的异构与含噪本质并直接影响着电路的可执行性、错误率以及整体计算的可靠性。量子节点代表容纳有限数量量子比特的量子处理单元。每个节点同时维护着一个内部拓

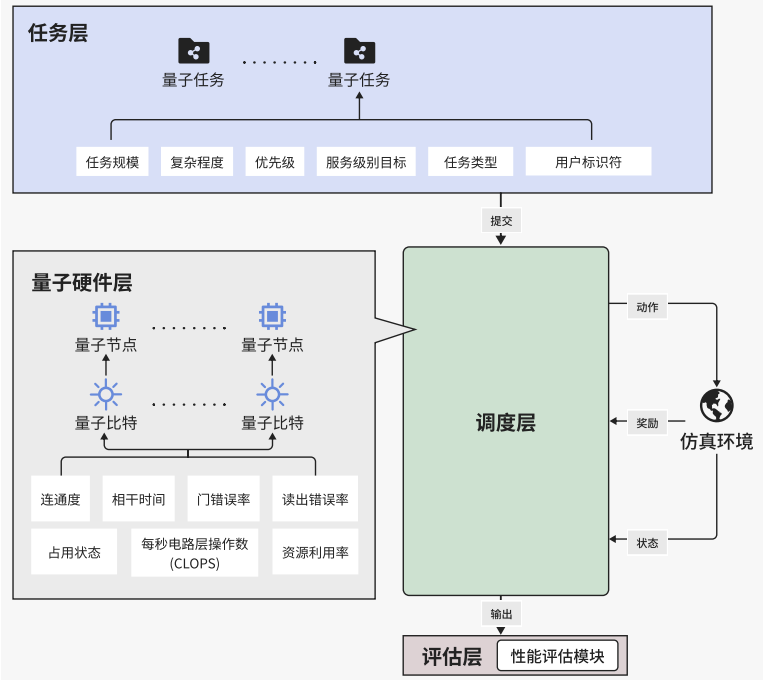

图 3.1　量子任务调度云平台框架及量子-经典融合任务调度算法优化

Figure 3.1　Framework of Cloud-based Quantum Task Scheduling and Optimization for Quantum-Classical Hybrid Scheduling Algorithms

扑 (Internal topology)，用于定义量子节点内部量子比特的连通性，以及一个全局拓扑 (Global topology)，用于定义分布式环境中量子节点间的连通性。

量子节点负责在系统运行中分配与释放量子比特、管理任务执行以及维护利用率与保真度等系统状态。量子节点的双层拓扑结构建模使得系统能够捕捉真实的架构约束，并支持同个节点内与跨节点的任务调度。

量子任务代表提交至系统的计算作业，其中每个任务由一个六元组来描述，包含所需量子比特数、门复杂度、优先级、截止时间、耦合密度以及指示任务是否必须独占运行的并发标志位。量子任务不仅编码了量子电路结构，还结合了时间与资源约束。在调度期间，系统会评估任务的紧迫性、复杂度以及硬件兼容性，从而决定合适的节点分配方案。

该调度框架在算法库中引入了多种强化学习范式。DQN 智能体采用基于价值的方法来学习分配决策的预期累积奖励，从而在动态环境中实现自适应调度。针对更复杂的场景，系统采用了 Advantage Actor-Critic（A2C）及其变体异步 Advantage Actor-Critic（Asynchronous Advantage Actor-Critic，A3C），以利用并行工作节点来加速训练并改善探索过程。在这些策略梯度方法的基础上，PPO 智能体通过稳定且经过裁剪

的更新机制来优化任务分配，在处理高维且具有随机性的调度任务时显著增强了鲁棒性与样本效率。其量子-经典融合任务调度算法优化模块示意图如图3.2所示。

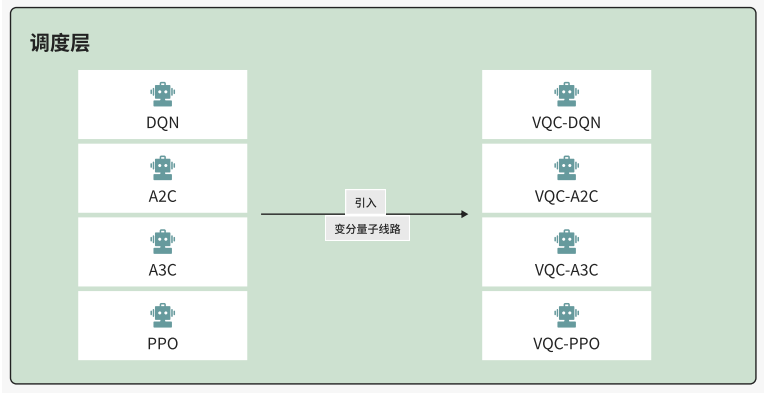

图 3.2　量子-经典融合任务调度算法优化

Figure 3.2　Optimization for Quantum-Classical Hybrid Scheduling Algorithms

仿真环境基于 Gymnasium 框架构建。该仿真环境为强化学习智能体的训练与评估提供了一个标准化的交互平台。该环境充当了底层量子硬件架构与上层调度策略之间的桥梁，负责动态生成随机量子任务流、管理量子节点的状态转移，并精确追踪量子任务整个生命周期。在与智能体的交互过程中，仿真环境会构建并输出一个多维度的字典型观测空间，深度融合了节点特征，任务特征以及反映系统整体积压负载的全局特征，为调度决策提供全面的状态感知。

基于接收到的系统观测状态，智能体在动作空间内输出具体的分配指令，决定将特定任务映射至目标量子节点或执行拒绝操作。仿真环境随即根据节点硬件兼容性与当前资源约束评估该动作的合法性，执行系统状态转移，并通过内置的结合正向激励与惩罚机制的奖励函数向智能体反馈优化梯度的标量奖励。

该奖励函数涵盖了多个维度，包括任务完成度、资源利用率、公平性以及及时性。如发生任务拒绝、分配失败、错过截止时间以及空闲动作等情况将进行惩罚。评估指标的汇总如表3.1所示。

3.2　优化问题与数学表述

基于上一节中已经定义的量子计算环境与任务模型集合 $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$，本节将任务调度与量子比特映射问题转为表述为一个混合整数优化问题，其中使用到的符号含义如表3.2中所示：

表 3.1　奖励函数中使用的评估指标描述

指标	描述	效果
完成度 ($w_{\text{completion}}$)	任务成功执行后赋予的基础奖励。	提高系统任务吞吐量。
效率 ($w_{\text{efficiency}}$)	基于执行时间与资源消耗比率的奖励。	优化任务处理的时效性。
公平性 (w_{fairness})	反映跨节点负载均衡程度的奖励。	防止资源垄断，均衡负载。
截止时间惩罚 (w_{deadline})	任务错过预设截止时间的惩罚。	确保任务处理的及时性。
资源利用率 ($w_{\text{utilization}}$)	与系统节点整体利用情况相关的奖励。	提高底层硬件使用效率。
任务优先级 (w_{priority})	根据任务权重或重要程度缩放的奖励。	促进紧急或核心任务的优先执行。
分配失败惩罚 (w_{failure})	针对不成功的映射或调度尝试的惩罚。	引导更鲁棒的资源分配策略。
拒绝惩罚 ($w_{\text{rejection}}$)	拒绝已提交至队列的任务所产生的惩罚。	减少无效拒绝，提升服务质量。

表 3.2　优化问题中的数学符号说明

符号	定义与物理含义	备注
n	量子任务总数	$\mathcal{T} = \{t_1, \ldots, t_n\}$
t_i	第 i 个量子任务	$t_i \in \mathcal{T}$
D_i	任务 t_i 的预估执行持续时间	取决于门复杂度与硬件吞吐量
S_i	任务 t_i 的实际调度开始时间	决策变量，$S_i \in \mathbb{R}_{\geq 0}$
d_i	任务 t_i 的截止日期 (Deadline)	SLO 硬性指标
y_i	任务接受决策变量	$y_i \in \{0, 1\}$
q_i	任务 t_i 所需的逻辑量子比特数量	任务资源需求
V	物理量子比特集合	硬件节点资源
$M_{i,k}$	量子比特映射决策变量	$M_{i,k} \in \{0, 1\}$
M	大于零的常数 (Big-M)	用于放宽拒绝任务的约束
$\mathbf{1}[\cdot]$	指示函数	条件成立为 1，否则为 0
p_i	并发执行标志位	控制任务隔离或并行
Active(t)	时刻 t 活跃的任务集合	用于冲突检测
E	硬件拓扑边集	描述比特间的物理耦合

$$\text{minimize} \quad \sum_{i=1}^{n} D_i \cdot y_i, \tag{O1}$$

$$\text{subject to}$$

$$s_i + D_i \le d_i + M(1 - y_i), \quad \forall i,\ t_i \in \mathcal{T}, \tag{C1}$$

$$\sum_{k \in V} M_{i,k} = q_i, \quad \forall i,\ t_i \in \mathcal{T}, \tag{C2}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{q_i} M_{i,k} \cdot \mathbf{1}[S_i \le t < S_i + D_i] \le 1, \quad \forall v_k \in V,\ \forall t, \tag{C3}$$

$$p_i = 0 \Rightarrow |\text{Active}(t)| \le 1, \quad \forall t \in [S_i, S_i + D_i], \tag{C4}$$

$$M_{i,u} \cdot M_{i,v} \le \mathbf{1}[(u, v) \in E], \quad \forall u, v \in V \text{ interacting via two-qubit gates in } t_i, \forall t_i \in \mathcal{T} \tag{C5}$$

$$y_i \in \{0, 1\}, \quad M_{i,k} \in \{0, 1\}, \quad S_i \in \mathbb{R}_{\ge 0}, \quad \forall i, k. \tag{C6}$$

目标 (O1) 旨在最小化系统中所有已接受任务的累积执行时间。对于每个量子任务 t_i，其预估执行时间 D_i 由其门复杂度 g_i 与有效硬件吞吐量 (例如 CLOPSγ_i) 得出。$y_i \in \{0, 1\}$ 表示一个二进制决策变量，用于指示任务 t_i 是否被接受并进行调度。通过最小化总和 $\sum D_i \cdot y_i$，调度器可以有效提升系统响应度、缩短平均周转时间并增加整体吞吐量。这对于相干时间窗口较为有限且硬件约束严格的 NISQ 时代量子处理器尤为重要，因为过长的执行时间会显著增加退相干与错误传播的风险。

约束 (C1) 即 SLO 满足约束，通过引入大 M 项 (Big-M) 来放宽对被拒绝任务的限制，强制要求所有已接受的任务 ($y_i = 1$) 必须在预设的截止日期 d_i 前处理完成，从而在严苛的截止时间需求下确保时间调度的可行性。

约束 (C2) 为量子比特映射约束，其强制每个任务将 q_i 个逻辑量子比特精确映射至互不相同的物理量子比特上，以保障电路嵌入的完整性与无冲突性。

在此基础上，约束 (C3) 即互斥约束，进一步确保物理量子比特在任意时刻均不被多个任务同时占用，有效防止了时间层面的重叠并维护了硬件执行的排他性。

针对需要隔离执行的特定任务，约束 (C4) 即并发策略约束规定在指定执行窗口内最多仅允许一个任务处于活跃状态。

约束 (C5) 作为拓扑感知映射约束，确保所有双量子比特门均映射至具备物理连接的量子比特对上，旨在最大限度减少 SWAP 开销并维持门操作的保真度。

最后，约束 (C6) 即变量定义域约束规定了所有决策变量的有效取值范围，以确保数学模型的一致性并杜绝非物理性的赋值。

3.3　基于强化学习的量子任务调度与量子比特映射建模

与先前定义的调度模型保持一致，本部分将量子任务调度与量子比特映射问题形式化为马尔可夫决策过程，并引入强化学习框架来学习最优策略，并由以下元组定义：

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle. \tag{3.1}$$

$\mathcal{S}$ 表示状态空间，其中的每个状态 $s_t \in \mathcal{S}$ 描述了系统当前的调度与硬件资源状态，包括任务队列的状态 (例如等待、执行、完成)、物理量子比特的占用向量 $\mathbf{s} = [s_1, s_2, \ldots, s_{|V|}]$(其中 $s_j \in \{0, 1\}$ 指示量子比特 v_j 是否被占用)、每个物理量子比特的硬件属性、已接受任务的剩余 SLO 时间，以及当前时间步 t。状态可表示为：

$$s_t = (\text{TaskQueue}_t, \text{QubitOccupancy}_t, \text{HardwareAttr}_t, \text{Remaining SLO}_t). \tag{3.2}$$

$\mathcal{A}$ 表示动作空间，其中的每个动作 $a_t \in \mathcal{A}$ 代表针对某个任务的调度与量子比特映射决策，定义为：

$$a_t = (t_i, \{v_{j_1}, \ldots, v_{j_{q_i}}\}, \text{StartTime}, y_i, p_i), \tag{3.3}$$

其中 t_i 表示所选任务。$\{v_j\} \subseteq V$ 表示分配的物理量子比特集合。StartTime 表示执行开始时间。$y_i \in \{0, 1\}$ 表示任务接受指示符。$p_i \in \{0, 1\}$ 表示并发标志位 (是否允许并发执行)。

$\mathcal{P}(s_{t+1}|s_t, a_t)$ 表示转移函数。环境根据智能体的动作以顺序方式更新系统状态。具体而言，每个任务的状态会被更新，以反映其被接受、拒绝或完成的情况。同时，物理量子比特的占用情况会被修改，以捕捉新分配或释放的资源。当前全局时间向前推进，并且活跃任务的剩余执行时间会相应减少。

$R(s_t, a_t)$ 表示奖励函数。该奖励函数旨在最小化执行时间的同时满足服务约束。它由几个子组件构成。执行时间惩罚定义为

$$R_{\text{exec}} = -D_i \cdot y_i, \tag{3.4}$$

用于惩罚较长的任务持续时间。SLO 惩罚考虑是否满足截止时间：

$$R_{\text{SLO}} = \begin{cases} 0, & \text{if } S_i + D_i \leq d_i, \\ -\lambda(S_i + D_i - d_i). & \text{otherwise}, \end{cases} \tag{3.5}$$

其中要么舍弃超出 SLO 要求的任务，要么在奖励函数中施加此惩罚。为了捕捉资源冲突 (如映射重叠或违反独占性)，本部分引入冲突惩罚：

$$R_{\text{conflict}} = -\gamma \cdot \#\text{Conflicts}. \tag{3.6}$$

为了反映量子比特拓扑映射与 SWAP 操作的成本，本部分将拓扑惩罚定义为：

$$R_{\text{mapping}} = -\alpha \cdot \#\text{SWAPs}. \tag{3.7}$$

总奖励整合了所有这些组件：

$$R_t = R_{\text{exec}} + R_{\text{SLO}} + R_{\text{conflict}} + R_{\text{mapping}}. \tag{3.8}$$

γ 表示折扣因子 (用于累积奖励计算)。

其中 $\lambda, \gamma, \beta, \alpha$ 是可调的惩罚系数，用于平衡执行时间、截止时间遵守情况、冲突解决以及映射成本。

最终所有约束都会被严格执行，以确保生成的状态与物理硬件限制及调度要求保持一致。

3.4 量子节点的分层拓扑建模

量子处理节点是执行量子计算任务的核心物理载体。本研究将量子处理节点建模为具有复杂硬件约束和双层拓扑结构的异构算力单元，旨在捕获 NISQ 时代量子硬件的非理想特性以及分布式量子计算中的通信瓶颈。每个量子节点的状态则受任务负载影响，在空闲 (IDLE)、忙碌 (BUSY)、错误 (ERROR) 或维护 (MAINTENANCE) 等互斥模式间流转，构成了调度决策的基础约束。

本研究引入了双层拓扑架构来描述量子信息的交换约束。节点内拓扑定义了节点内部物理量子比特间的连通结构，生成的邻接矩阵 $\mathbf{A}_{int}$ 直接决定了二比特门操作

的可执行性。只有当任务需求的逻辑连接对满足物理连通性时，量子节点方可直接执行对应的量子门，否则将面临无法嵌入或需引入额外通信成本的风险。量子节点间拓扑定义了分布式网络中量子节点之间的互联结构。该拓扑通过注入机制动态绑定，使得量子节点能够识别其直连邻居并计算跨节点的最短路径距离，为估计跨节点通信或路由开销提供依据。

量子节点的可用性取决于其物理量子比特的占用状态。系统通过量子比特分配与释放机制实时追踪资源消耗，并利用评估函数对电路的可执行性进行校验。电路的执行时间估计模型则综合了量子门总数与硬件基准处理开销，通过线性组合公式实现动态估算。

3.5　基于经典强化学习的智能体建模

3.5.1　使用 DQN 智能体的量子任务调度

DQN 提供了一个适合离散动作空间的基于价值的强化学习框架，本研究将其用于量子任务调度与量子比特映射问题。

每个状态 $s_t \in \mathcal{S}$ 编码了当前的系统状态，包括待处理任务的特征向量集合（如规模、需求、截止时间与可并行性）、指示物理量子比特当前被占用情况的二进制向量、每个量子比特的硬件相关特征（包括门错误率、保真度与连通性矩阵）、在任何任务违反其服务级别目标 (SLO) 之前的剩余松弛时间，以及当前的全局时间。整个观测值作为输入提供给 Q 网络，该网络输出所有可行动作的估计 Q 值。

调度与量子比特映射决策被建模为一个离散动作，通过将二维的任务-节点选择展平为单一索引：

$$a_t = (t_i, n_j), \quad a_t \in \mathcal{A}, \quad |\mathcal{A}| = |\mathcal{T}| \times |\mathcal{N}|, \tag{3.9}$$

其中 t_i 表示所选任务的索引，n_j 表示分配的量子比特节点的索引。奖励函数定义为：

$$R_t = -D_i - \lambda \cdot \max(0, S_i + D_i - d_i) - \gamma \cdot \#\text{Conflicts} - \alpha \cdot \#\text{SWAPs}, \tag{3.10}$$

$-D_i$ 为惩罚执行延迟，$-\lambda \cdot \max(0, S_i + D_i - d_i)$ 表示违反 SLO 约束的惩罚，$-\gamma \cdot \#\text{Conflicts}$ 处理并发或排他性违规问题，而 $-\alpha \cdot \#\text{SWAPs}$ 惩罚由量子比特拓扑映射引起的额外成本。

为了稳定训练过程，系统维护了一个独立的目标网络 Q_{θ^-}，并每隔 $f_{target} = 1,000$ 个训练步与主网络同步。为了缓解高估偏差，本部分的实现采用了双重 DQN(Double

DQN) 逻辑：

$$L(\theta) = \mathbb{E}\left[\left(r_t + \gamma Q_{\theta^-}(s_{t+1}, \text{argmax}_{a'} Q_\theta(s_{t+1}, a')) - Q_\theta(s_t, a_t)\right)^2\right]. \tag{3.11}$$

在回放缓冲区收集到 1,000 个状态转移后开始训练。本部分在每次 Adam 优化步（学习率 0.001）之前应用梯度范数裁剪与 $[-1, 1]$ 内的梯度值截断。动作探索采用 ϵ-贪心策略，探索率 ϵ 使用每步 0.995 的指数衰减因子从 $\epsilon_{start} = 1.0$ 衰减至 $\epsilon_{end} = 0.01$，以平衡探索与利用。

这种建模方法使 DQN 智能体能够学习到有效的调度与映射策略，利用经验回放与双网络机制，在硬件约束下优化资源分配并最小化延迟。

3.5.2 使用 Actor-Critic 智能体的量子任务调度

为了进一步提升在动态量子环境下的调度灵活性与策略稳定性，本部分采用 Actor-Critic 强化学习框架来联合优化量子任务调度与量子比特映射。与第3.5.1节中基于价值的 DQN 方法不同，Actor-Critic 架构通过将策略搜索 (Actor) 与价值评估 (Critic) 相结合，能够更有效地处理高维离散动作空间中的非平稳性问题。

本节首先在第3.5.2.1小节中详细介绍了同步 Advantage Actor-Critic(A2C) 建模方案，通过模块化的网络设计，实现了任务选择与节点分配的联合决策，并引入动作掩码机制以确保物理层面的可行性。并在第3.5.2.2小节中提出了异步增强型 Advantage Actor-Critic(A3C) 变体。通过引入并行智能体交互机制、重要性采样比率优化以及增强的熵正则化约束。这种从同步到异步的演进过程，构建了一个从基础调度到可扩展资源管理的完整强化学习的对比机制。

3.5.2.1 同步 A2C 建模

A2C 算法利用基于梯度的同步框架平衡策略与价值学习，通过共享底层（shared-bottom）架构同时优化调度策略与状态价值函数。

每个状态 $s_t \in \mathcal{S}$ 被表示为一个拼接的一维结构化张量。状态表示包含大小为 $N_{nodes} \times 4$ 的节点级特征矩阵（捕捉物理节点本地资源状态）、大小为 $N_{tasks} \times 4$ 的任务级特征矩阵（描述所需量子比特、优先级、紧急度与复杂度），以及包含归一化决策步数、待处理任务数与运行中任务数的系统级全局特征向量。所有组件拼接后作为策略网络与价值网络的统一输入。

动作空间 $\mathcal{A}$ 采用多离散（Multi-Discrete）分布建模，动作定义为 $a_t = (t_{idx}, n_{idx})$。为确保物理可行性，系统引入动作掩码机制，对不可行的任务-节点对在对数几率（Logits）上赋予 $-\infty$ 惩罚，将采样限制在有效硬件配置范围内。

网络架构由共享特征提取层和三个专门头部组成：任务 Actor 头部输出任务选择概率，节点 Actor 头部生成节点分配概率，Critic 头部估计标量状态价值 $V(s_t)$。各组件参数使用三个独立的 Adam 优化器进行细粒度优化。

策略更新基于优势估计 $\hat{A}_t$，利用 n 步回报进行计算：

$$R_t = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n V(s_{t+n}). \tag{3.12}$$

总损失函数定义为：

$$L(\theta) = L_{\text{actor}} + w_v L_{\text{critic}} - w_e H(\pi_\theta). \tag{3.13}$$

其中通过最大化熵 $H(\pi_\theta)$ 维持策略的探索性。该实现通过同步 n 步轨迹管理，在保证样本效率的同时，相比离策略方法维持了更小的内存占用。

3.5.2.2　通过 A3C 智能体的异步增强

A3C 智能体在 A2C 智能体的基础上引入了关键改进。A3C 智能体采用多个并行智能体与独立的物理环境实例交互，通过异步并行机制有效打破了训练数据的时间相关性。

本研究在 A3C 智能体中实现了差异化优化策略，为共享网络与专门头部设置了不同的学习率($lr_{actor} = 0.0003, lr_{critic} = 0.001$),使全局价值函数能更灵敏地响应环境动态。为增强策略稳定性,更新机制引入了重要性采样比 $\rho_t(\theta) = \exp(\log \pi_\theta - \log \pi_{old})$,其 Actor 损失定义为：

$$L_{\text{actor}} = -\mathbb{E}_t \left[\rho_t(\theta) \cdot \hat{A}_t \right]. \tag{3.14}$$

结合梯度裁剪策略（max_grad_norm = 0.5），该设计减轻了异步更新时可能出现的策略偏移风险。

此外，通过将熵系数设置为 0.01 增强了正则化约束，鼓励智能体在量子比特映射与节点分配决策上进行更广泛的探索，防止收敛至次优节点。

尽管异步机制引入了全局参数同步的开销，但该实现确保了其显著高于 A2C 智能体的训练吞吐量。引入策略比率 $\rho_t(\theta)$ 与熵加权损失仅增加了极低的计算开销，对于在复杂量子任务环境下保持训练稳定性至关重要。

3.5.3　使用 PPO 智能体的量子任务调度

PPO 通过裁剪的替代目标防止过度激进的策略更新，特别适用于具有复杂奖励信号与高维状态空间的量子任务调度场景。

每个状态 $s_t \in \mathcal{S}$ 编码了当前的系统配置，包括：待处理任务特征向量（规模、需求、截止时间、可并行性）、指示物理量子比特占用状态的二进制向量、硬件相关属性（门错误率、保真度、连通性矩阵）以及 SLO 剩余松弛时间。完整状态表示为结构化张量，作为策略网络的输入。

动作空间 $\mathcal{A}$ 涵盖了调度与映射的联合决策，定义为：

$$a_t = (t_i, \{v_j\}, y_i, p_i), \tag{3.15}$$

其中 t_i 为任务索引，$\{v_j\}$ 为分配的物理量子比特集，y_i 为预期开始时间，p_i 为并发执行指示符。奖励函数 R_t 综合了执行延迟惩罚 R_{exec}、基于截止时间的 SLO 惩罚 R_{SLO}、资源冲突惩罚 R_{conflict} 与拓扑映射惩罚 R_{mapping}：

$$R_{\text{SLO}} = \begin{cases} 0, & \text{if } S_i + D_i \le d_i, \\ -\lambda(S_i + D_i - d_i), & \text{otherwise.} \end{cases} \tag{3.16}$$

总奖励表示为：

$$R_t = R_{\text{exec}} + R_{\text{SLO}} + R_{\text{conflict}} + R_{\text{mapping}}. \tag{3.17}$$

为确保训练稳定性，PPO 采用裁剪的替代目标函数：

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \cdot \hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \cdot \hat{A}_t \right) \right], \tag{3.18}$$

其中概率比 $r_t(\theta)$ 被限制在 $[1-\epsilon, 1+\epsilon]$ 范围内。这种裁剪机制在遇到高方差奖励信号（如显著的 SLO 违规惩罚）时，能有效防止策略出现过度偏移或崩溃的情况。

该方法使智能体能够学习到高效的资源分配策略，在遵守物理约束的同时优化量子系统资源使用。在推理阶段，PPO 利用共享底层的 Actor-Critic 架构只通过单次前向传播即可高效生成调度决策。

3.6 量子-经典融合任务调度算法优化

3.6.1 使用量子-经典混合 DQN 智能体的量子任务调度

本节进一步将第3.5.1节定义的经典调度模型扩展为量子-经典混合架构。该模型保留了基于价值的强化学习框架，但核心改进在于利用 VQC 替代了传统深度神经网络的部分隐藏层，旨在通过量子希尔伯特空间的强大表征能力优化高维约束下的调度决策与第3.5.1节直接处理平铺向量不同，混合模型采用了更为精细的分层特征提取机制：系统状态 s_t 首先由三组独立的经典多层感知机进行预处理，分别提取节点

硬件特性、任务队列状态及全局参数的潜在嵌入 (Latent Embeddings)。这些嵌入向量经拼接后通过一个经典压缩层，将高维特征映射为维度 $n_{qubits} = 10$ 的特征向量，从而为后续的量子处理准备输入数据。

该模型价值函数近似器的核心是由基于 PennyLane（一个支持可微分量子编程、能够将量子硬件与机器学习库无缝衔接的开源软件库）实现的 VQC 构成的。处理流程涵盖了数据编码与变分变换两个阶段：首先通过角度嵌入 (Angle Embedding) 将经典特征上传至量子寄存器，利用 R_Y 和 R_Z 旋转门将数据映射至布洛赫球。随后，变分组件由 $L = 3$ 个重复层组成，每层包含可训练的单比特旋转门以及用于产生跨寄存器纠缠的 CNOT 门链。最终，系统测量每个量子比特在 Pauli-Z 算子下的期望值 $\langle \sigma_z \rangle$，并由经典输出层将该量子特征向量映射为展平动作空间 $\mathcal{A}$ 的估计 Q 值。在策略执行方面，动作选择延续了第3.5.1节的 ϵ-贪心策略及相同的指数衰减配置，但在优化阶段需同时调整经典 MLP 权重与 VQC 的旋转角度 θ。为了在混合架构下维持计算效率，本节实现采用了标准 DQN 逻辑而非第3.5.1节的双重 DQN 结构，即直接利用目标网络 Q_{θ^-} 获取最大 Q 值来计算均方误差 (Mean Squared Error, MSE) 损失：

$$L(\theta) = \mathbb{E}\left[\left(r_t + \gamma \max_{a'} Q_{\theta^-}(s_{t+1}, a') - Q_\theta(s_t, a_t)\right)^2\right]. \tag{3.19}$$

系统通过每 1,000 步一次的参数同步以及阈值为 1.0 的梯度范数裁剪确保训练稳定性。

从复杂度角度分析，VQC-DQN 智能体的性能由量子仿真与经典计算共同决定。在时间复杂度方面，推理阶段的延迟主要取决于 VQC 前向传播的 L 个层级，而训练阶段的单次更新复杂度为 $O(B \cdot T_{vqc})$。由于标准 DQN 逻辑减少了额外的网络前向传播，其在混合架构下的计算开销相对可控。在空间复杂度上，内存需求仍由容量为 100,000 的经验回放缓冲区主导。双网络架构则要求同时存储经典权重与量子旋转角，总参数量为 $O(2 \cdot P)$。尽管量子参数量相比同等性能的全连接层更为精简，但这种双集存储机制对于维持混合状态空间下的价值函数近似稳定性至关重要。

3.6.2　使用量子-经典混合 Actor-Critic 智能体的量子任务调度

在第3.5.2节定义的经典 Actor-Critic 框架基础上，本部分引入 VQC 作为函数近似器，构建了量子-经典混合 Actor-Critic 模型。该模型旨在利用量子线路在处理具有高维纠缠特性的数据时表现出的强大表征能力，进一步优化量子任务调度中的复杂比特映射决策。与依赖深层经典神经网络不同，混合架构通过将系统状态编码至量子希尔伯特空间，利用量子门的参数化旋转与受控操作实现策略与价值函数的高效映射，

从而在更少的参数规模下捕捉硬件拓扑与任务需求之间的深层非线性关联。

本部分首先在第3.6.2.1小节中探讨了基于 VQC 的 A2C 建模方案。随后,第3.6.2.2小节详细阐述了针对该混合架构的异步增强机制。

3.6.2.1 使用 VQC 增强型 A2C 建模

本小节将 A2C 算法扩展至量子-经典混合框架，以实现策略优化与价值估计的深度融合。该模型保留了 A2C 的同步梯度更新机制，但核心架构从纯经典网络演进为包含 VQC 的混合神经网络。与第3.5.2.1小节类似，系统状态 $s_t \in \mathcal{S}$ 依然由节点级特征 ($N_{nodes} \times 4$)、任务级特征 ($N_{tasks} \times 4$) 以及全局特征向量拼接而成，形成统一的结构化观测张量作为输入。

该混合模型的底层架构采用共享表征设计，实现了从经典预处理到量子变分处理的平滑过渡。输入向量首先通过一个经典压缩层 (input layer)，将特征维度降低至 $n_{qubits} = 10$。随后，利用角度嵌入将压缩后的特征编码至量子寄存器。核心 VQC 由 $L = 3$ 层强纠缠层组成，在量子希尔伯特空间中执行复杂的非线性变换。测量得到的 Pauli-Z 期望值随后分支为三个专门的经典头部：任务 Actor 头部和节点 Actor 头部输出用于任务与节点选择的对数几率，而 Critic 头部则估计标量状态价值 $V(s_t)$。为确保硬件调度的可行性，系统同样对 Actor 输出应用掩码机制，并在通过 Categorical 分布采样前将无效动作对的对数几率设为 $-\infty$。

在优化阶段，本小节沿用了第3.5.2.1小节中的三独立 Adam 优化器策略，分别针对任务 Actor、节点 Actor 与 Critic 进行细粒度学习率调整。策略更新基于优势估计算法，其中回报 R_t 通过折扣奖励计算，优势估计定义为 $\hat{A}_t = R_t - V(s_t)$ 并经过归一化处理以稳定梯度。总损失函数 $L(\theta)$ 整合了动作损失、价值损失与策略熵正则化项：

$$L(\theta) = L_{actor} + w_v L_{critic} - w_e H(\pi_\theta), \tag{3.20}$$

其中 $L_{actor} = -\mathbb{E}[\log \pi_\theta(a_t|s_t)\hat{A}_t]$。为了促进混合训练期间梯度的稳定流动，所有线性层均采用正交初始化 (Orthogonal Initialization)。

从复杂度角度分析，VQC-A2C 智能体在推理阶段的延迟主要取决于 VQC 前向传播的 T_{vqc} 开销。在训练阶段，单次更新的复杂度为 $O(n \cdot T_{vqc})$，其中 n 为同步轨迹步数。由于压缩层 input layer 与 quantum layer 在三个优化器间共享，梯度更新引入了 $O(3P_{shared} + P_{heads})$ 的额外开销。在空间复杂度方面，内存需求仍由 n 步轨迹缓冲区主导，复杂度为 $O(n \cdot D_{obs})$。VQC-A2C 智能体在每次策略更新后清空缓冲区，无需庞大的经验回放机制，因此保持了更小的内存占用。模型存储需求涵盖了经典权值与

VQC 中 $3 \times n_{qubits} \times 3$ 个可训练旋转角，总复杂度为 $O(P_{classical} + P_{vqc})$，这为在资源受限环境下维持高维价值函数近似提供了保障。

3.6.2.2　使用 VQC 增强型 A3C 建模

在第3.6.2.1小节提到的 VQC-A2C 智能体基础上，VQC-A3C 变体通过引入异步并行机制，进一步应对量子环境的高维与非平稳特征。该模型采用多个并行工作线程 (Worker) 与独立的环境实例进行交互，这种解耦机制有效地打破了训练数据的时间相关性。在量子调度场景中，由于特定的物理节点瓶颈可能导致经验轨迹产生偏差，这种异步并行机制展现出显著的优化优势。

A3C-VQC 在实现细节上继续采用了统一 Adam 优化器下的差异化优化策略。不同于 A2C 基准中相互独立的优化器架构，A3C-VQC 智能体通过为共享的经典压缩层 Input Layer、量子变分层 Quantum Layer 以及专门的输出头部分配不同的学习率，实现了对整个混合网络的统一优化。具体而言，策略相关参数的学习率设为 $lr_{actor} = 0.0003$，而价值估计头部的学习率设为 $lr_{critic} = 0.001$。这种设计使得全局价值函数能够比策略函数更灵敏地响应环境动态的变化。

为了进一步增强混合架构下的训练稳定性，A3C-VQC 智能体引入了重要性采样比 $\rho_t(\theta) = \exp(\log \pi_\theta - \log \pi_{old})$ 来计算 Actor 损失：

$$L_{actor} = -\mathbb{E}_t \left[\rho_t(\theta) \cdot \hat{A}_t \right]. \tag{3.21}$$

结合 0.5 的梯度范数裁剪约束与权重的正交初始化，该公式减轻了多个线程异步更新全局参数时可能出现的灾难性策略偏移风险。此外，通过将熵系数设置为 0.01，模型引入了更强的正则化约束，鼓励智能体在量子比特映射决策上进行更广泛的探索，从而防止过早收敛到表现出临时高保真度的次优物理节点。

从异步开销的角度分析，其时间复杂度随并行工作线程数量 N_w 进行缩放。虽然局部更新复杂度保持为 $O(n \cdot T_{vqc})$，但全局网络同步为参数拉取与梯度推送引入了 $O(N_w \cdot P)$ 的开销，其中 P 涵盖了经典权重与 VQC 旋转角。这种异步机制确保了训练吞吐量显著高于 A2C。不同于 A2C 中的三独立优化器结构，A3C-VQC 通过分组学习率使用单个 Adam 优化器，将单步梯度更新效率优化至 $O(P)$。此外，包含策略比率 $\rho_t(\theta)$ 与 0.01 熵加权损失在计算阶段增加了一个可忽略的常数时间复杂度，而 0.5 的梯度裁剪则确保了跨线程全局参数更新的稳定性。在空间复杂度方面，系统需要在共享内存中额外分配 $O(P)$ 的空间来存储全局模型副本，以实现工作线程间的同步。

3.6.3 使用量子-经典混合 PPO 智能体的量子任务调度

本小节利用变分量子线路增强对状态-动作空间的表征能力，同时利用 PPO-Clip 目标函数确保在复杂量子环境动态下的训练稳定性。与第3.5.3小节所述的经典 PPO 不同，VQC-PPO 通过量子层处理高维约束，能够更精确地捕捉硬件拓扑带来的非线性影响。

系统状态空间 $\mathcal{S}$ 被表示为从结构化特征字典中导出的展平一维张量，整合了节点级物理比特属性、任务级特征以及全局系统指标。在进行量子处理之前，这些高维特征首先通过一个包含线性变换与 ReLU 激活函数的经典压缩层 Input Layer 压缩至 $n_{qubits} = 10$ 维。这种预处理机制为后续进入量子希尔伯特空间的数据编码提供了标准化的输入。

策略与价值近似的核心是一个 Actor-Critic VQC 网络。压缩后的经典特征通过角度嵌入并利用 R_Y 旋转门编码为量子态。该线路采用 $L = 3$ 层强纠缠层，由可训练的单比特旋转门与 CNOT 纠缠器组成。网络架构随后分支为三个专门的头部：用于选择待处理量子任务的任务 Actor 头部、用于分配目标物理节点的节点 Actor 头部，以及用于估计标量状态价值 $V(s_t)$ 的 Critic 头部。动作空间 $\mathcal{A}$ 被建模为多离散分布，其中动作 $a_t = (t_{idx}, n_{idx})$ 同时决定任务选择与映射节点。为强制执行硬件约束，系统在从 Categorical 分布采样前对 Actor 对数几率应用掩码机制，将无效任务-节点对的权重设为 $-\infty$。

为了确保混合训练的稳定性，智能体在多个轮次内最小化复合损失函数：

$$L_{total} = L^{CLIP}(\theta) + c_1 \cdot L_{VF}(\theta) - c_2 \cdot S[\pi_\theta](s_t), \tag{3.22}$$

其中 L^{CLIP} 是裁剪范围 $\epsilon = 0.2$ 的裁剪代理目标，L_{VF} 是系数为 $v_f = 0.5$ 的价值损失，而 S 表示系数为 $e_{ent} = 0.01$ 的熵奖励。优势估计 $\hat{A}_t$ 通过广义优势估计计算，参数设为 $\lambda = 0.95$ 与 $\gamma = 0.99$。计算复杂度由大小为 $n_{steps} = 2048$ 的 On-policy Rollout 缓冲区决定，并使用学习率为 0.0003 的单个 Adam 优化器进行迭代优化，同时应用阈值为 0.5 的梯度范数裁剪。

从复杂度角度分析，VQC-PPO 智能体在推理阶段表现出低延迟特性，每次动作选择的时间复杂度为 $O(T_{vqc})$，该值由经典预处理层与 VQC 的 $L = 3$ 层纠缠层的前向传播决定。在优化阶段，每个更新周期的复杂度为 $O(E \cdot \frac{N_{steps}}{B} \cdot T_{vqc})$，其中轮数 $E = 10$，小批量大小 $B = 64$。计算成本 T_{vqc} 涵盖了通过经典线性头部与变分量子线路的混合反向传播。在空间复杂度方面，内存需求主要由 PPO Rollout Buffer 主导，其

复杂度为 $O(N_{steps} \cdot D_{obs})$，其中 D_{obs} 为展平观测向量的维度 $(4N_{nodes} + 4N_{tasks} + 3)$。作为一种同策略方法，智能体在每个更新周期后清空其缓冲区，确保内存占用不会随时间累积。模型参数空间由经典权重与 VQC 层中 $3 \times n_{qubits} \times 3$ 个可训练旋转角组成，总复杂度为 $O(P_{classical} + P_{vqc})$，在固定量子拓扑下保持恒定。

3.7　本章小结

本章针对量子计算环境下的任务调度与量子比特映射问题，系统性地开展了从底层环境建模、优化问题刻画到智能调度算法构建的研究工作，为后续系统实现与实验验证奠定了理论基础与方法支撑。

首先，在量子计算环境建模方面，本章对分布式量子处理节点进行了分层拓扑建模，将量子节点定义为具有复杂硬件约束和双层连通结构的异构算力单元。通过引入系统级拓扑与节点级拓扑，本章进一步刻画了量子节点之间以及节点内部量子比特之间的连接关系。同时，结合量子比特的相干时间、门错误率、读出错误率和占用状态等硬件属性，建立了节点状态动态流转规则、底层资源占用追踪机制以及综合硬件基准与量子门总数的电路执行时间估算模型，从而使调度过程能够更加贴近真实 NISQ 量子硬件环境。

其次，在问题建模方面，本章将任务调度与量子比特映射的耦合过程转化为一个带有复杂约束条件的混合整数优化问题，明确了以最小化系统累计执行时间为核心的优化目标。在此基础上，本章进一步引入马尔可夫决策过程框架，针对量子任务流动态到达、资源状态实时变化以及硬件约束多维耦合等特点，设计了由任务队列、量子比特占用状态、节点硬件属性和系统全局负载共同构成的结构化状态空间，并构建了融合执行延迟、资源冲突、任务完成度、截止时间约束、资源利用率及拓扑映射开销等因素的多维度复合奖励函数，从而引导智能体在吞吐量、效率、公平性与可靠性之间进行综合权衡。

再次，在智能调度算法构建方面，本章基于上述统一建模框架，分别设计并分析了 DQN、A2C、A3C 和 PPO 等多种经典强化学习智能体在量子任务调度场景中的应用方式。其中，DQN 通过价值函数近似实现离散调度动作的决策学习，Actor-Critic 类方法通过策略网络与价值网络协同优化提升决策效率，PPO 则通过裁剪策略更新机制增强训练过程的稳定性。上述模型共同构成了面向量子任务调度问题的经典强化学习算法基础。

此外，本章进一步探讨了量子-经典融合架构在任务调度优化中的应用，系统性

地构建了基于变分量子电路的混合智能调度方法。在架构设计上，本章确立了由经典多层感知机进行特征预处理、VQC 执行核心特征映射与参数化变换的协同模式。通过角度嵌入技术，调度状态特征被映射至量子希尔伯特空间，并借助量子叠加与纠缠特性增强模型对复杂硬件约束和高维状态空间的表征能力。在算法层面，本章实现了 VQC-DQN、VQC-A2C、VQC-A3C 以及 VQC-PPO 等多种量子-经典混合智能体，并围绕任务选择、节点分配与量子比特映射完成了多目标联合优化。

最后，本章通过对经典强化学习智能体与 VQC 增强型智能体的建模逻辑、结构特征和计算复杂度进行分析，阐明了量子-经典融合方法在参数规模、特征表达能力和推理效率方面的潜在优势。总体而言，本章完成了从量子任务调度问题定义到经典与混合智能体算法设计的完整理论构建，为后续量子调度仿真平台的工程实现、实验配置设计以及不同算法性能对比提供了核心理论依据与技术路径。

第 4 章 系统实现

基于第3章中的任务调度云平台设计与各类经典强化学习以及量子-经典混合强化学习智能体设计，本章详细阐述了量子资源调度原型系统的工程实现细节。为了验证前文所提的各类经典强化学习智能体与量子-经典混合强化学习智能体在复杂物理约束下的有效性，本文在 Linux 环境下构建了一个功能完整的量子资源调度仿真系统。系统整体采用 Python 作为核心开发语言，累计代码量超过 7000 行。经典强化学习智能体的参数化策略网络基于 PyTorch 深度学习框架构建，量子-经典混合智能体中的 VQC 则利用 PennyLane 库进行算子定义。两者的模型训练、特征反向传播及梯度更新过程均由 PyTorch 的自动求导机制提供底层支撑，实现了量子参数与经典权重的联合优化。

本章的组织结构如下：第4.1节首先概括了量子资源分配仿真环境的整体架构，详细描述了基于 Gymnasium（一个提供标准化 API 的开源 Python 库，用于开发和比较强化学习算法，是 OpenAI Gym 的官方维护分支）的标准化接口设计，以及环境内部如何通过离散事件驱动实现资源池与任务流的深度解耦。第4.2节阐述了量子任务的建模方式，并介绍了基于状态机的任务生命周期管控逻辑。第4.3节深入说明了基于经典强化学习架构的智能体实现，详细分析了各类智能体在处理异构字典观测空间时的网络设计与物理约束掩码机制。第4.4节阐述了基于 VQC 的混合强化学习智能体的实现方式，说明了 VQC 作为非线性特征提取器与经典神经网络层的交互接口，以及端到端的参数更新策略。最后，第4.5节介绍了量子处理节点的工程实现。

4.1 量子资源分配仿真环境实现

系统基于 Gymnasium 框架构建了一套逻辑闭环且具备高度抽象能力的量子资源分配调度仿真环境。在工程设计层面，该环境通过将复杂的物理硬件约束与任务生命周期管理进行深度解耦，实现了一个支持策略训练与算法评测的离散事件仿真系统。环境的核心实现思路围绕资源实体的层次化建模、任务状态机的流转控制以及多维特征空间的标准化构建展开。

在资源实体的架构设计中，底层通过 `QuantumNode` 类将物理量子节点抽象为逻辑资源池，每个节点内部不仅封装了相干时间、门保真度等关键物理参数，还通过邻接矩阵实现了对硬件比特拓扑连接性的软件定义。这种设计允许系统通过配置文

件动态初始化异构的节点拓扑结构，为后续的拓扑匹配校验提供了底层的静态数据支撑。

在任务流转控制逻辑方面，系统设计了一套由离散事件驱动的状态机机制。环境内部维护了等待、运行、完成及失效四个独立的状态队列，精确刻画了量子任务从提交、调度决策到资源回收的全生命周期流转。每当上层算法通过 step 接口下达分配指令时，环境充当“准入控制网关”的角色：系统首先调用 `_can_allocate_task` 接口进行双重校验，即同时验证目标节点是否有足够的可用比特，以及量子节点物理连接性是否满足任务的拓扑硬约束（`connectivity_requirements`）。只有通过校验的任务才会被执行资源锁定并推入运行队列。随后，系统根据任务的预估执行耗时模拟物理时间的推移，利用异步回收机制自动触发物理比特的释放与状态更新，从而实现了对并行任务实例的集中管控。

此外，为了实现各类调度算法对复杂物理状态的精准感知，系统构建了一套分层级的多维特征观测体系。该体系将底层复杂的硬件原始数据抽象为由节点特征、任务特征与全局系统特征构成的结构化观测空间。其中，节点特征涵盖了资源利用率与平均保真度的动态分布，任务特征则通过对队列中待处理任务的紧迫度与计算复杂度进行非线性映射，体现了当前负载的潜在压力。通过标准化的字典型观测空间（Dict Observation Space）设计，系统实现了高维状态数据的解耦表示，显著增强了底层环境对不同强化学习架构的兼容性与特征提取效率。同时，环境内部集成了一套闭环的实时度量衡监控模块，该模块能够自动追踪资源碎片率、跨时段任务违约数等关键性能指标。这些指标不仅为奖励函数的动态计算提供了量化依据，更构成了系统性能评估的数据底座，确保了调度策略优化的可解释性与客观性。

4.2　量子任务实体实现

量子任务作为最基础的资源请求与逻辑执行单元，其属性建模的深度直接决定了调度器对底层物理约束的感知能力。系统通过封装 `QuantumTask` 类，实现了一套面向调度优化的高保真任务描述模型。

任务的生命周期由严谨的状态机进行管控。如表4.1所示，任务状态涵盖了从任务生成后的 `PENDING` 状态，到获得资源并绑定物理节点后运行的 `RUNNING` 状态，直至最终执行成功或因超时导致的终止状态。为了模拟真实的计算负载，系统利用 `generate_random_task` 静态方法实现随机任务流的采样，并强制校验任务的连接性约束，确保生成的任务均包含如 CNOT 或 CZ 等双比特门操作，从而体现量子拓

扑映射的挑战性。在评估任务难度时，系统采用加权公式量化任务复杂度。这种结构化的任务建模方式精确表述了量子计算任务的异构特征，也为强化学习智能体在复杂环境下的决策提供了高质量的观测数据支撑。

表 4.1 量子任务状态说明

状态名称	语义说明
PENDING	任务进入就绪队列，等待分配物理节点及比特资源。
RUNNING	任务已成功映射至物理拓扑，处于实际计算执行阶段。
COMPLETED	任务顺利完成预定逻辑，资源已回收且计算结果可用。
FAILED	任务因硬件故障、分配超时或违反约束而导致执行中断。
CANCELLED	任务由于用户干预或系统抢占机制被强制撤销执行。

4.3 经典强化学习智能体实现

在量子资源调度仿真环境中，智能体的设计直接影响决策的效率与资源分配的合理性。系统实现了一套基于强化学习的经典智能体框架，涵盖了从基于值函数的方法到策略梯度及 Actor-Critic 架构的多种主流算法。

作为经典值函数方法的代表，DQN 智能体的核心框架设计旨在将复杂的量子资源分配转化为稳定、受控的离散空间价值逼近过程。针对由节点状态、任务属性与全局负载构成的异构字典型观测空间，系统在数据感知前端创新性地构建了多分支特征融合网络。该架构通过并行的子感知机模块分别提取各数据维度的局部表征，并在高维潜在空间中进行拼接聚合，实现了对复杂物理细节的深度解码。针对“任务-节点”匹配的二维决策特性，系统建立了一套空间降维映射机制，将分配组合转化为一维索引序列，使神经网络能够无缝输出全局决策空间的期望收益。为了严格遵守量子硬件的物理约束，该框架在动作选择阶段嵌入了动态掩码策略，通过在推理层面对非法动作施加负向极值惩罚，强制限制智能体在有效资源边界内进行 ϵ-贪婪探索。在保障模型收敛的底层逻辑上，框架不仅通过经验池的乱序采样打破了数据的时序耦合，更在工程实现中集成了双重 Q 学习的动作解耦机制。系统利用在线网络筛选下一状态的最优动作，并交由延迟更新的目标网络进行独立估值，从而在算法底层有效克服了传统 DQN 易引发的价值过估计（Over-estimation）的问题。最后，配合双重梯度限制机制（全局模长截断与逐元素截断）以及探索率的指数衰减设计，该架构规避了梯度爆炸与策略震荡问题，为异构量子计算任务的最优调度提供了一套具备硬约束感知能力且高度鲁棒的算法基础。

在 Actor-Critic 架构的实现中，系统构建了 A2C 与 A3C 智能体，旨在通过策略分布与价值估计的协同优化来提升调度效率。在状态感知端，框架首先实施了特征展平协议，将异构的字典型观测空间无缝拼接为一维稠密张量，以适配底层共享网络的特征提取。在决策输出端，针对量子资源分配的二维动作空间，该架构放弃了易引发维度爆炸的全局索引映射方案，而是将 Actor 网络创新性地解耦为“任务选择”与“节点匹配”两个并行的策略分支。为保障物理约束，系统直接在网络输出的对数几率层面施加负向极值掩码，确保非法动作经 Softmax 激活函数激活后的采样概率严格为零。在参数优化的底层逻辑上，两者均依赖显式计算的优势函数（Advantage Function）来降低策略梯度的方差，并融合了熵正则化机制以遏制策略的过早收敛。在具体的梯度更新设计上，两者展现了差异化的工程解耦思路：A2C 智能体在实例级别拆分了任务 Actor、节点 Actor 与 Critic 的优化器，实现了多决策维度梯度流的物理隔离。而 A3C 智能体则构建了带有参数组独立学习率的统一优化架构，并内置了全局梯度裁剪机制，在维系各网络分支差异化学习步长的同时显著提升模型在应对复杂量子并发负载时的训练鲁棒性。

作为理论上系统中致力于解决策略更新失稳问题的核心算法，PPO 智能体构建了一套兼顾采样效率与训练鲁棒性的 On-policy 优化机制。在数据感知前端，代码显式引入了特征展平协议，将由节点状态、任务特征及全局负载构成的异构字典观测空间精确降维拼接为一维稠密向量，以满足定长轨迹回放缓冲区与底层共享网络的张量存储需求。基于此连续交互轨迹，框架集成了广义优势估计算法，通过动态权衡多步自举价值与真实回报，为策略网络提炼出低方差的优势信号。在决策输出端，该架构延续了 A2C 与 A3C 智能体使用的双分支解耦设计，并严格在网络输出的对数几率层面施加负向极值掩码，从底层阻断了非法物理组合的采样概率。在最关键的参数迭代阶段，PPO 智能体采用了多轮次与微批次相交织的随机打乱优化逻辑。其核心的损失函数构建引入了裁剪替代目标，通过提取新旧策略分布的比率，并利用截断（Clamp）算子将其强制束缚在超参数界定的信任域区间内，从逻辑上遏制了单次更新步长过大导致的策略崩溃。最终，系统将受控的策略损失、基于均方误差的价值回归损失以及用以维持探索机制的熵正则化项聚合为复合目标，并辅以全局梯度截断防范梯度爆炸，从而将高维量子约束下的调度问题转化为一个平滑、高度稳定的参数寻优过程。

如表4.2所示，这些经典智能体共同构成了一个多维度的对比基准，为研究量子计算任务的最优调度策略提供了算法支撑。

表 4.2 经典强化学习智能体特征说明

算法名称	实现特征说明
DQN 智能体	基于值函数估计，利用经验回放与 ϵ-greedy 探索平衡调度收益。
A2C 智能体	同步 Actor-Critic 架构，利用优势函数降低策略梯度的方差。
A3C 智能体	异步并行架构思想，通过精细化调整 Actor 与 Critic 学习率提升收敛性。
PPO 智能体	采用近端策略裁剪机制，在保证训练稳定性的同时优化复杂拓扑下的资源配比。

4.4 量子-经典混合强化学习智能体实现

为了进一步探索量子计算在资源调度决策中的潜力，并为之后的算法研究提供参考，系统实现了一套量子-经典混合强化学习框架。该框架利用变分量子线路作为函数逼近器，替代了传统深度神经网络中的部分隐藏层，构建了包括 VQC-DQN、VQC-A2C、VQC-A3C 以及 VQC-PPO 在内的混合智能体体系。

混合智能体的核心组件是基于 PennyLane 框架构建的变分量子层。该层通过角度编码将经过经典预处理后的特征映射为量子比特的旋转角度，并利用强纠缠层 (Strongly Entangling Layers) 作为可训练参数块，通过测量泡利 Z 算符的期望值输出经典特征向量。为了实现高效的梯度下降，量子电路通过 `TorchLayer` 封装，利用反向传播方法与经典神经网络层无缝集成，实现了量子参数与经典权重的联合优化。

在 VQC-DQN 智能体的代码实现设计中，首先通过经典降维网络将由节点、任务与全局负载拼接而成的庞大观测张量压缩为维度严格适配系统可用逻辑量子比特数的紧凑潜在向量，有效解决了宏观状态与有限量子硬件间的维度错配问题。随后，在核心的量子计算层，系统利用角度编码机制将特征作为旋转参数映射至指数级庞大的希尔伯特空间，并接入由多层参数化单比特旋转与 CNOT 门级联的强纠缠变分层以提取量子原生特征，最终通过泡利-Z（Pauli-Z）可观测量测量输出期望值向量。该包含量子关联信息的向量进而交由后置经典输出网络解码为全局动作空间的 Q 值分布。在工程落地上，系统利用 PennyLane 的 `TorchLayer` 接口并显式指定基于反向传播的微分引擎对量子节点进行了深度封装，使 VQC 能够作为标准算子无缝嵌入 PyTorch 动态计算图中。这一设计在保留原有经验回放与动态动作掩码等成熟机制的前提下，实现了混合网络参数的端到端批量化同步更新，为探索量子算法在复杂资源调度中的潜在优势奠定了坚实的工程基础。

在策略梯度的架构延伸下，系统进一步构建了 VQC-A2C 与 VQC-A3C。区别于价值方法的单点输出映射，混合 Actor-Critic 架构在经典输出端实施了深度的结构解耦。具体而言，量子层输出的含参期望值向量被作为系统状态的泛化量子表征，并行

馈送至三个独立的经典多层感知机分支中：分别负责输出“任务选择”与“节点匹配”独立对数几率的双头 Actor 网络，以及负责状态基线估值的 Critic 网络。针对复杂的混合梯度回传，两种智能体展现了差异化的处理逻辑：VQC-A2C 智能体在实例级别完全物理拆分了三个独立的优化器，使得任务策略、节点策略与价值基线的损失能够通过独立的计算图分支，各自向共享的量子层回传梯度流。而 VQC-A3C 智能体则构建了基于字典映射的统一参数组优化架构，通过为 Actor 分支，Critic 分支以及量子共享层精准分配定制化学习率，在单一优化步骤中完成了全局更新。

为进一步提升混合架构在连续决策环境下的训练稳定性，系统构建了 VQC-PPO，该架构依赖定长轨迹回放缓冲区收集同轨交互片段，并内置广义优势估计算法为变分量子线路的梯度更新提供低方差的优势引导。在策略优化的核心阶段，系统在对数几率动作掩码的硬约束下，将完整的缓冲区轨迹拆解为微批次，并进行多轮次的随机打乱迭代。区别于其他策略梯度方法，VQC-PPO 的算法内核在联合损失中深度集成了裁剪替代目标，利用截断算子将新旧策略分布的演化比率严格限制在信任域超参数区间内，从底层逻辑上阻断了量子变分参数在单步更新中因跨度过大而引发的策略崩溃。此外，不同于 VQC-A2C 物理拆分优化器的做法，VQC-PPO 采用了统御全局网络参数的单一优化器架构，它将受控的策略损失、价值网络均方误差损失与双动作分布的熵正则化项聚合为复合标量目标，并辅以全局梯度模长截断机制，最终实现在极度复杂的量子物理约束下的端到端平滑、受控更新。

如表4.3所示，混合强化学习智能体通过将量子线路集成至决策链路，为解决大规模异构资源调度问题提供了全新的研究维度。

表 4.3　量子-经典混合强化学习智能体特征说明

算法名称	实现特征说明
VQC-DQN	利用 VQC 逼近 Q 函数，通过角度编码将调度状态映射至量子 Hilbert 空间。
VQC-A2C	混合 Actor-Critic 架构，量子层输出作为共享特征流指导策略与价值的学习。
VQC-A3C	结合混合量子架构与异步更新思想，针对 Actor 与 Critic 配置差异化学习率。
VQC-PPO	结合广义优势估计与策略裁剪，利用变分量子参数的表达力优化高维约束下的调度策略。

4.5　量子节点实现

在量子资源调度系统中，物理处理节点被抽象为承载计算任务的底层硬件实体。

在底层物理属性建模方面，系统通过引入 `QubitProperties` 数据类，利用均匀分布采样为节点内的每一个物理量子比特初始化了独立的相干时间、门操作保真

度、读出错误率以及工作频率等退相干参数。这种基于区间的随机采样画像为上层调度器评估计算可靠性提供了真实的物理级观测指标。在资源调度层面，系统摒弃了粗粒度的独占式分配，利用布尔型数组 `busy_qubits` 实现了比特级别的细粒度并发管控。节点通过 `allocate_qubits` 与 `release_qubits` 接口对局部可用比特进行首适应分配与回收，并配合严密的状态机枚举与当前任务队列，确保了量子任务在全生命周期内的资源占用一致性与状态原子性。

该组件在工程上的核心设计在于对拓扑结构的双层解耦实现。在微观层面 (节点内)，系统在初始化阶段基于传入的连接类型实例化了 `internal_topology` 对象，以表征芯片内部量子比特间的物理连接网络。在预估电路可执行性时，系统通过 `can_execute_circuit` 接口，不仅校验当前剩余可用比特数量是否达标，更提取电路逻辑中的双比特门连接需求，利用内部邻接矩阵进行严格的硬约束比对，从而在调度前期精准拦截物理拓扑不匹配的非法映射。在宏观层面 (节点间)，系统预留了由全局环境动态注入的 `global_topology` 机制。通过 `set_global_topology` 接口，系统为孤立的量子节点构建了宏观邻接关系图谱。该全局视图突破了单节点的算力孤岛限制，直接支撑了跨节点的直连判定与最短路径计算，为分布式量子计算场景下的通信路由开销估计与跨节点协同调度奠定了完备的代码底层架构支撑。

4.6 本章小结

本章针对量子计算环境下的资源调度与量子比特映射问题，系统性地完成了从仿真环境构建、核心组件建模到多元智能体开发的设计与实现工作。

首先，本章构建了基于 Gymnasium 框架的仿真环境，通过标准化接口实现了异构量子资源在动态调度中的状态演进与回收。在组件建模方面对量子任务与节点进行了精细化描述：任务侧确立了基于拓扑连接性的异构属性与状态机；节点侧创新性地引入了双层拓扑架构，实现了内部物理连通性与全局网络的动态绑定，为算法提供了感知硬件特性的观测空间。在算法实现层面，本章对比了多种强化学习智能体的构建方案。研究不仅实现了基于经典架构的 DQN，A2C，A3C 以及 PPO 算法，还引入了 VQC 作为函数近似器构建了混合量子-经典架构模型。

综上所述，本章构建了一个从底层物理约束到上层智能决策的完整原型系统，验证了分层拓扑建模在复杂资源管理中的必要性，为后续设计并开展大规模实验仿真与性能评估奠定了坚实的工程基础。

第 5 章 实验设置

本章介绍了基于经典强化学习智能体及基于量子-经典混合强化学习智能体的架构设计、训练机制与实验配置方案。为系统评估不同智能体在量子任务调度与量子比特映射问题中的性能表现，本章首先围绕经典强化学习模型展开实验设置，分别对 DQN、A2C、A3C 和 PPO 等智能体的网络结构、状态观测处理方式、动作决策机制、优化器参数及训练过程进行说明。在此基础上，本章进一步结合不同奖励权重配置与资源密度条件，构建面向任务完成度、执行效率、资源利用率、公平性和截止时间约束等多目标因素的实验场景，从而为后续比较不同调度策略的适应能力和优化效果提供统一的实验基准。

与此同时，为验证量子-经典融合机制在复杂调度问题中的应用潜力，本章进一步设计了 VQC-DQN、VQC-A2C、VQC-A3C 和 VQC-PPO 等混合智能体的实验方案。该部分重点介绍了经典特征预处理模块、变分量子电路结构、量子比特数量、纠缠层数、梯度裁剪策略以及训练超参数等关键设置，旨在保证经典模型与量子增强模型之间具有可比性。通过上述实验设置，本章为后续章节中不同智能体在收敛性、任务成功率、平均执行时间和资源调度效率等方面的性能分析奠定了基础。

5.1 基于经典强化学习智能体架构设计与训练机制

为了系统评估各类调度策略的自适应能力，本节对比分析了四种经典强化学习智能体的性能表现。在特征表征层面，所有智能体均遵循一套统一的、基于字典结构的状态观测处理逻辑，借此实现节点状态、任务属性与全局特征的深度融合。

5.1.1 DQN 智能体

本小节围绕奖励塑形（Reward Shaping）与局部资源密度开展了一系列对比实验。在底层架构上，该智能体基于深度 Q 网络实现，并引入了经验回放（Experience Replay）机制以有效打破训练数据的相关性——同时配合独立构建的目标网络，显著提升了模型迭代过程的稳定性。

针对所有实验配置，系统均统一采用 Adam 优化器，学习率被严格设定为 3×10^{-4}，折扣因子 $\gamma = 0.99$，且经验回放缓冲区（Replay Buffer）的容量被配置为 100,000 条状态转移记录。此外，模型在探索阶段执行 ϵ-贪婪策略，其 ϵ 值在最初的 10,000 步交

互中由 1.0 平滑衰减至 0.1，旨在确保早期广泛探索与后期稳定利用之间的动态平衡。

5.1.1.1　奖励函数设计

为了探究不同的优化目标如何引导 DQN 智能体的调度行为，本部分设计了四种截然不同的奖励权重配置。这些配置（即均衡型、严格约束型、吞吐量导向型与公平性导向型）的详细权重分配总结于表5.1中。

通过对比这些设置，可以观察到智能体在硬约束与系统全局效率之间所作出的权衡。均衡配置（配置 1）作为性能基准，通过在所有指标上保持适中的权重，它鼓励智能体在遵守截止时间约束与资源效率的同时实现较高的任务完成率。针对时间敏感型场景，严格约束配置（配置 2）显著增加了完成奖励（$w_{\text{completion}} = 3.0$）并施加了沉重的截止时间惩罚（$w_{\text{deadline}} = -10.0$）。该配置对分配失败与拒绝分配施加了严厉的惩罚，以严格抑制可能导致系统拥塞或资源碎片的风险映射行为。

相比之下，吞吐量导向配置（配置 3）通过放大效率（$w_{\text{efficiency}} = 2.5$）与资源利用率（$w_{\text{utilization}} = 2.5$）的权重，将量子网络的整体处理能力置于首位。在此设置下，智能体受到激励，以最大化量子比特的占用率与任务吞吐量。

最后，公平性导向配置（配置 4）通过赋予公平性指标最高优先级（$w_{\text{fairness}} = 2.5$），将重点放在了长期稳定性与弹性上。通过对过度的负载不均衡进行惩罚，引导智能体在分布式量子网络中弹性地分配工作负载，从而防止因局部过载而导致的性能退化。

表 5.1　用于 DQN 智能体性能分析的奖励权重配置

指标权重	**配置 1** (均衡)	**配置 2** (严格)	**配置 3** (吞吐量)	**配置 4** (公平性)
完成度 ($w_{\text{completion}}$)	2.0	3.0	1.0	1.5
效率 ($w_{\text{efficiency}}$)	1.2	0.5	2.5	1.0
公平性 (w_{fairness})	0.6	0.2	0.2	2.5
截止时间惩罚 (w_{deadline})	-6.0	-10.0	-3.0	-5.0
资源利用率 ($w_{\text{utilization}}$)	1.0	0.5	2.5	1.0
任务优先级 (w_{priority})	1.5	2.5	1.0	0.5
分配失败 (w_{failure})	-1.5	-4.0	-1.0	-2.0
拒绝惩罚 ($w_{\text{rejection}}$)	-1.5	-4.0	-1.0	-3.0

5.1.1.2　资源密度分析

除了为 DQN 智能体设计不同的奖励函数外，本研究还进一步通过动态调整环境中的 `max_qubits_per_node` 属性，深入研究了局部资源密度指标对调度性能的具体影响。在维持基准配置 $N = 5$ 个节点不变的前提下，本实验在三种截然不同

的资源密度水平上对 DQN 智能体进行了全面测试：（一）低密度（2 个量子比特/节点）、（二）中密度（5 个量子比特/节点）以及（三）高密度（10 个量子比特/节点）。此举旨在充分考量 NISQ 设备中相干时间这一严苛物理限制的背景下，精确评估随着单节点可用量子资源的逐步增加，智能体将如何自适应地扩展并优化其调度策略。

5.1.2 A2C 智能体

为了全面评估 A2C 智能体的自适应能力与敏感度，本小节与上一小节中设置相似的对比实验。在底层架构上，该智能体采用同步 Advantage Actor-Critic 框架，其中共享底层网络分支为独立的任务 Actor 与节点 Actor 策略头。为了实现对各自学习率的细粒度控制，系统为任务 Actor、节点 Actor 以及 Critic 组件分别配置了三个独立的 Adam 优化器。针对所有实验配置，Actor 的学习率均被严格设定为 3×10^{-4}，而 Critic 的学习率则设为 1×10^{-3}，旨在优先保障价值基准的迭代稳定性。

5.1.2.1 奖励函数设计

为了探究不同的优化目标如何引导 A2C 智能体的调度行为，本部分设计了四种不同的奖励权重配置。这些配置（即均衡型、严格约束型、吞吐量导向型与公平性导向型）的详细权重分配总结于表5.2中。

通过对比这些设置，可以观察 A2C 智能体在硬约束与系统全局效率之间所作出的权衡。均衡配置（配置 1）作为性能基准，通过在所有指标上保持适中的权重鼓励智能体在遵守截止时间约束与资源效率的同时实现较高的任务完成率。针对时间敏感型场景，严格约束配置（配置 2）显著增加了完成奖励（$w_{\text{completion}} = 3.0$）并施加了沉重的截止时间惩罚（$w_{\text{deadline}} = -6.0$）。该配置对分配失败与拒绝分配施加了严厉的惩罚，以严格抑制可能导致系统拥塞的风险映射行为。

相比之下，吞吐量导向配置（配置 3）通过放大效率（$w_{\text{efficiency}} = 2.0$）与资源利用率（$w_{\text{utilization}} = 2.5$）的权重，将量子网络的整体处理能力置于首位。在此设置下，智能体受到激励以最大化量子比特占用率与任务吞吐量。最后，公平性导向配置（配置 4）通过赋予公平性指标最高优先级（$w_{\text{fairness}} = 1.8$），将重点放在了长期稳定性与弹性上。通过对过度的负载不均衡进行惩罚引导智能体在分布式量子网络中弹性地分配工作负载。

5.1.2.2 资源密度分析

本部分进一步通过动态调整 `max_qubits_per_node` 属性，深入考察了局部资源密度的具体影响。针对 A2C 智能体实验，本研究特别聚焦于中等与高密度水

表 5.2　用于 A2C/A3C 智能体性能分析的奖励权重配置

指标权重	配置 1 (均衡)	配置 2 (严格)	配置 3 (吞吐量)	配置 4 (公平性)
完成度 ($w_{\text{completion}}$)	1.5	3.0	1.0	1.2
效率 ($w_{\text{efficiency}}$)	1.2	0.8	2.0	1.0
公平性 (w_{fairness})	0.6	0.2	0.4	1.8
截止时间惩罚 (w_{deadline})	-3.0	-6.0	-1.5	-3.0
资源利用率 ($w_{\text{utilization}}$)	1.0	0.5	2.5	1.0
任务优先级 (w_{priority})	1.5	2.5	1.0	0.8
分配失败 (w_{failure})	-1.5	-3.5	-1.0	-2.0
拒绝惩罚 ($w_{\text{rejection}}$)	-2.0	-4.0	-1.0	-3.0

平：（一）中密度（5 个量子比特/节点）以及（二）高密度（10 个量子比特/节点），并主动剔除了低密度配置（2 个量子比特/节点）的对比实验。

这一决策主要源于 A2C 算法在极端资源受限环境下容易遭遇的失效机制。首先，在仅有 2 个比特的资源约束下，待分配任务的比特需求极易超过节点剩余容量，导致合法动作空间被极度压缩，智能体难以通过随机探索获取有效的正向学习信号，产生了严重的样本稀疏性问题。其次，在这种“高频失败、极少成功”的环境中，优势函数的计算会出现剧烈抖动，产生高方差的梯度信号，使得神经网络难以从海量的失败惩罚中辨识出微弱的成功路径。此外，由于 A2C 采用同步更新机制，个别并行环境产生的正向梯度极易被大多数环境回传的失败惩罚梯度所淹没，导致策略梯度在更新过程中信噪比极低，进而诱发策略退化现象。在此状态下，智能体往往会陷入持续拒绝任务以规避高额惩罚的“消极陷阱”，无法演化出有效的调度逻辑。因此，仅有 2 个比特的资源约束下的实验结果对强化学习任务处理效果的参考价值较小。通过配置 5 与 10 个量子比特，系统能够提供充足的资源深度以产生高质量的动作轨迹，从而有效评估智能体在复杂映射与节点内调度优化中的综合效能。

5.1.3　A3C 智能体

为了全面评估 A3C 智能体的自适应能力，本小节设计了一系列与 A2C 智能体保持高度一致的对比实验。该 A3C 智能体将 Actor-Critic 架构进一步拓展为异步执行模型，通过部署多个并行的工作节点来与相互独立的环境实例进行实时交互。

这一机制不仅显著提升了系统整体的训练吞吐量，更有力地打破了复杂量子资源状态下所固有的数据相关性。

5.1.3.1 奖励函数设计

由于 A3C 与 A2C 共享着相同的核心优势函数计算逻辑及 Actor-Critic 网络结构，本部分采用了与 A2C 实验完全相同的四种奖励权重配置。

通过保持奖励函数的高度一致性，本部分能够直接对比在处理相同的操作目标（如优先级敏感度或吞吐量优化）时，异步并行更新与同步批量更新对模型收敛速度及策略质量的具体影响。这种一致性有效消除了奖励塑形作为混杂变量的干扰，使本研究得以专注剖析异步机制在降低策略梯度方差方面的核心贡献。

5.1.3.2 资源密度分析

在本部分关于资源密度的研究中，本研究同样聚焦于中密度（5 个量子比特/节点）与高密度（10 个量子比特/节点）水平。需要说明的是，此处剔除 2 个量子比特/节点的低密度配置实验，其原因与前文所述 A2C 算法面临的失效机制一致。相较于同步架构，A3C 得益于其固有的异步更新机制，在上述密度配置下展现出独特的性能优势。随着节点资源密度由 5 跃升至 10 个量子比特，A3C 理论上能够更为高效地利用各个独立工作节点所探索出的多样化任务映射路径。这种高度并行的探索能力有助于缓解模型因全局同步约束而可能引发的策略停滞（Policy Stagnation）难题，从而在更深层次的资源环境中实现更优的策略演化。

5.1.4 PPO 智能体

在底层架构上，该智能体严格遵循 Actor-Critic 架构，其共享网络（Shared Network）负责同时输出动作分布（Action Distribution）与状态价值。

针对参数更新机制，每次迭代均包含 4 个轮次，且微批次大小被设定为 64。此外，系统统一采用 Adam 优化器，学习率被精确配置为 3×10^{-4}，同时折扣因子设为 $\gamma = 0.99$。

5.1.4.1 奖励函数设计

本部分中设计了四种截然不同的奖励权重配置，详细权重分配总结于表5.3中。

通过对比这些设置，可以观察 PPO 智能体在硬约束与系统全局效率之间所作出的权衡。均衡配置（配置 1）作为性能基准，通过在所有指标上保持适中的权重鼓励智能体在遵守截止时间约束与资源效率的同时实现较高的任务完成率。针对时间敏感型场景，严格约束配置（配置 2）显著增加了完成奖励并施加了沉重的截止时间惩罚（$w_{\text{deadline}} = -5.0$）。该配置对分配失败与拒绝分配施加了严厉的惩罚，以严格抑制可能导致系统拥塞的风险映射行为。

相比之下，吞吐量导向配置（配置 3）通过放大效率与资源利用率的权重，将量子网络的整体处理能力置于首位。在此设置下，智能体受到激励以最大化量子比特占用率与任务吞吐量，即使以放宽即时截止时间约束为代价。

最后，公平性导向配置（配置 4）通过赋予公平性指标最高优先级（$w_{\text{fairness}} = 1.8$），将重点放在了长期稳定性与弹性上。通过对过度的负载不均衡进行惩罚，它引导智能体在分布式量子网络中弹性地分配工作负载，从而防止因局部过载而导致的单一节点性能退化。

表 5.3　用于 PPO 智能体性能分析的奖励权重配置

指标权重	**配置 1** (均衡)	**配置 2** (严格)	**配置 3** (吞吐量)	**配置 4** (公平性)
完成度 ($w_{\text{completion}}$)	1.5	2.5	1.0	1.2
效率 ($w_{\text{efficiency}}$)	1.2	0.8	1.8	0.8
公平性 (w_{fairness})	0.6	0.2	0.4	1.8
截止时间惩罚 (w_{deadline})	-3.0	-5.0	-2.0	-2.5
资源利用率 ($w_{\text{utilization}}$)	1.0	0.5	2.0	1.0
任务优先级 (w_{priority})	1.5	2.0	1.0	0.5
分配失败 (w_{failure})	-1.5	-3.0	-1.0	-1.5
拒绝惩罚 ($w_{\text{rejection}}$)	-2.0	-4.0	-1.0	-3.0

5.1.4.2　资源密度分析

本部分进一步通过动态调整环境中的 `max_qubits_per_node` 属性，深入考察了局部资源密度的具体影响。尽管基准配置采用了包含 $N = 5$ 个节点的全连通拓扑结构，本实验仍在三种不同的资源密度水平上对智能体进行了全面测试：（一）低密度（2 个量子比特/节点）、（二）中密度（5 个量子比特/节点）以及（三）高密度（10 个量子比特/节点）。此举旨在充分考量硬件保真度极限的背景下，精确评估随着单节点可用量子资源的逐步增加，PPO 智能体将如何自适应地扩展并优化其调度策略。

5.2　基于量子-经典混合强化学习智能体架构设计与训练机制

在本节在上一节经典强化学习智能体相关实验的基础上，构建了一系列与上一节所设置的实验类似的量子-经典混合强化学习智能体架构的对比实验，并对其与经典强化学习智能体的实验表现进行对比。

5.2.1 VQC-DQN 智能体

VQC-DQN 智能体核心由经典预处理层、变分量子电路层以及经典输出映射层协同构成，旨在通过量子态的高维希尔伯特空间增强策略网络的表达能力。

在模型架构设计上，异构环境状态首先通过由三组独立线性层构成的特征提取网络，分别对节点特征、任务特征与全局特征进行融合。拼接后的高维特征向量经由经典输入层映射至与量子比特数（$n_{\text{qubits}} = 10$）相匹配的低维空间。量子电路部分采用 PennyLane 框架实现，包含数据编码与变分层两个阶段，经典信息通过角度嵌入作用于量子比特的 R_Y 与 R_Z 旋转门。由 3 层（$n_{\text{layers}} = 3$）强纠缠层构成可训练部分，利用单比特旋转门与 CNOT 纠缠门在量子比特间建立复杂的关联特性。最后，系统测量各比特在 Pauli-Z 基下的期望值，将其作为高度非线性的量子特征反馈至经典输出层，映射生成最终的动作 Q 值。

在训练策略与超参数配置方面，VQC-DQN 延续了经典 DQN 的稳定性增强机制。系统统一采用 Adam 优化器，学习率设定为 1×10^{-3}，折扣因子 $\gamma = 0.99$。经验回放机制维持 100,000 条记录的缓冲区容量，单次梯度更新的批次大小配置为 32。为提升训练过程中的数值稳定性，智能体引入了独立的目标网络（Target Network），并以 1,000 步为周期执行状态同步。探索阶段执行 ϵ-贪婪策略，其 ϵ 值由 1.0 起始，按 0.995 软衰减至 0.01，以确保量子参数空间内的广泛搜索与后期策略利用之间的平衡。

为了保证对比试验的公平性，在配置方面，本部分继续采用表5.1中与经典 DQN 智能体相同的奖励函数配置，在维持基准配置 $N = 5$ 个节点不变的前提下，选取中密度（5 个量子比特/节点）对智能体进行测试。此外，针对量子梯度计算的特殊性，系统应用了范数裁剪技术，将梯度阈值严格限制在 1.0 以内，有效防止了深度迭代过程中的梯度爆炸问题。

5.2.2 VQC-A2C 智能体

为了进一步验证量子变分电路在策略梯度类算法中的泛化能力，本部分构建了基于量子-经典混合架构的 VQC-A2C 智能体。该智能体延续了上一节经典强化学习 A2C 的同步更新框架，但在底层特征处理上引入了量子增强机制。

在模型架构方面，VQC-A2C 智能体采用了双头输出结构，通过共享的量子变分层提取高维希尔伯特空间特征。异构状态输入经由经典线性层映射至 10 个量子比特（$n_{\text{qubits}} = 10$）的编码空间。量子层采用 PennyLane 的 TorchLayer 封装实现，其变分电路包含 3 层强纠缠层（$n_{\text{layers}} = 3$），利用 R_Y 与 R_Z 角度嵌入进行数据编码，并配

合 CNOT 门实现比特间的纠缠变换。电路测量的 Pauli-Z 期望值被同步馈送至任务 Actor、节点 Actor 以及 Critic 三个独立分支，分别映射生成动作概率分布与价值基准估计。

在训练策略与超参数配置方面，为确保与经典 A2C 智能体对比的公平性，系统维持基准配置 $N = 5$ 个节点及中密度（5 个量子比特/节点）设置不变。在优化机制上，VQC-A2C 同样为任务 Actor、节点 Actor 与 Critic 组件配置了三个独立的 Adam 优化器，以实现细粒度的策略更新控制。针对量子参数的训练敏感性，Critic 的学习率设定为 3×10^{-4}，而两类 Actor 的学习率通过设定比例系数 actor_2_critic_lr = 0.8，使其最终实际执行的学习率为 2.4×10^{-4}，旨在优先保障复杂量子网络环境下价值估计的收敛稳定性。此外，系统统一采用折扣因子 $\gamma = 0.99$ 与经验范数裁剪技术（阈值为 1.0），并在损失函数中引入熵正则化项（权重为 0.001），以维持量子策略空间在训练后期的探索活性。

在配置方面，继续采用表5.2中与经典 A2C 智能体相同的奖励函数配置。

5.2.3　VQC-A3C 智能体

为了全面评估 A3C 智能体的自适应能力，本部分设计了一系列与其同步对应版本（VQC-A2C）保持高度一致的深度对比实验。该 VQC-A3C 智能体将 Actor-Critic 架构进一步拓展为异步执行模型，通过部署多个并行的工作节点来与相互独立的环境实例进行实时交互。

在底层架构实现上，VQC-A3C 采用量子-经典混合神经网络作为函数近似器。异构环境特征经由经典输入层处理后，进入包含 10 个量子比特（$n_{\text{qubits}} = 10$）的变分量子电路层。该电路层通过角度嵌入实现数据编码，并利用 3 层强纠缠层（$n_{\text{layers}} = 3$）进行非线性特征提取，其测量的 Pauli-Z 期望值同步馈送至任务 Actor、节点 Actor 及 Critic 三个策略头。在训练策略与超参数配置方面，为确保与经典 A3C 智能体对比的公平性，系统维持基准配置 $N = 5$ 个节点及中密度（5 个量子比特/节点）设置不变。

此外，为了在异步更新过程中平衡梯度贡献，系统采用分层优化的 Adam 优化器。其中，Actor 分支的学习率设定为 3×10^{-4}，而 Critic 分支的学习率设定为 1×10^{-3}，旨在强化价值基准对异步经验流的拟合速度。在训练配置方面，VQC-A3C 继续沿用折扣因子 $\gamma = 0.99$ 与范数裁剪技术（阈值为 0.5），并引入了系数为 0.01 的熵正则化项，以确保量子策略空间在异步梯度更新过程中的探索活性。

在配置方面，继续采用表5.2中与经典 A2C/A3C 智能体相同的奖励函数配置。

5.2.4 VQC-PPO 智能体

在底层架构实现上，VQC-PPO 同样引入了包含 10 个量子比特（$n_{\text{qubits}} = 10$）与 3 层强纠缠层（$n_{\text{layers}} = 3$）的变分量子电路作为非线性特征提取器。异构环境特征经由经典预处理层映射至量子空间进行编码与变换，其测量的期望值作为共享特征，分别馈送至任务 Actor、节点 Actor 以及 Critic 策略头，实现了量子层参数在策略更新与价值评估过程中的高效共享。

针对参数更新机制，VQC-PPO 采用了基于广义优势估计的离线策略更新范式。在每次迭代过程中，系统通过部署容量为 2,048 条状态转移记录的采样缓冲区收集经验。更新阶段包含 10 个轮次，且微批次大小被设定为 64。系统统一采用 Adam 优化器，学习率被精确配置为 3×10^{-4}，同时折扣因子设为 $\gamma = 0.99$。为了保证对比试验的公平性，在配置方面，继续采用与经典 PPO 智能体相同的奖励函数配置，并维持基准配置 $N = 5$ 个节点及中密度（5 个量子比特/节点）设置不变。此外，为抑制量子策略更新过程中的剧烈波动，系统引入了裁剪范围为 0.2 的近端策略损失函数，并配合系数为 0.01 的熵正则化项与 0.5 的最大梯度裁剪阈值，以确保智能体能够在维持策略连贯性的前提下，实现对量子网络资源的高效调度。

在配置方面，继续采用表5.3中与经典 PPO 智能体相同的奖励函数配置。

5.3 本章小结

本章通过构建多层次的实验架构，系统性地阐述了面向分布式量子网络调度的智能体训练机制。首先，本章详细设计了基于经典强化学习（DQN、A2C、A3C、PPO）的基准模型，通过引入四种差异化的奖励函数配置（均衡、严格约束、吞吐量导向及公平性导向）与三种量级的局部资源密度（低、中、高密度），全面刻画了经典智能体在异构约束下的自适应表现。

随后，本章将 VQC-DQN、VQC-A2C、VQC-A3C 及 VQC-PPO 等量子-经典混合强化学习智能体架构，通过在维持物理环境参数与节点拓扑一致性的前提下，与经典强化学习的基准模型设置对比试验。本章所定义的实验配置与参数基准，为后续评估量子增强机制在策略收敛稳定性、任务完成成功率以及运行时间的性能提供了系统性的实验平台。

第 6 章 实验结果与分析

本节将对前文提出的各强化学习智能体实验进行全面的效果评估，并从三个维度深入分析算法的性能表现。首先，通过分析训练过程中的平均奖励值，评估模型在学习过程中的收敛速度与策略优化的稳定性。其次，重点考量平均任务执行成功率随训练轮次的变化趋势，以验证智能体在任务调度任务中的可靠性提升过程。最后，结合平均执行时间随训练轮次的变化曲线，衡量调度算法对决策效率的优化能力。通过这三个指标的综合评估，旨在全方位揭示智能体在动态环境下的学习演化轨迹与最终调度效能。

6.1 经典强化学习智能体在多种策略配置下的性能分析研究

为了确保公平的定性比较，本部分评估了 DQN，PPO，A2C 与 A3C 智能体在四种预定义的策略配置下的表现：均衡、严格、吞吐量与公平性配置。在这组实验中，本部分将资源密度固定为 `max_qubits_per_node=5`。

6.1.1 均衡配置 (配置 1) 下的性能表现

如图6.1左上子图所示，DQN 智能体在配置 1 环境下展现出高度平稳的学习行为。在整个训练过程中，原始回合奖励表现出明显的方差，但其 100 回合移动平均线在整个训练周期内未表现出显著的阶段性增长或降低。这种长期处于平台期的演化特征揭示了 DQN 智能体在均衡资源配置下存在学习瓶颈，即在当前探索机制与网络参数规模下，智能体的策略性能迅速进入并长期处于饱和状态，难以通过增加训练回合数来实现收益中枢的进一步突破。

图6.2上面两幅子图展示了 DQN 智能体在配置 1 下的任务成功率与平均执行时间演变过程。由左上子图可见，任务成功率在训练初期迅速爬升，在经历约 150 回合的小幅波动后趋于平稳，并最终维持在约 99% 的高位水平。这一演化特征表明，智能体在均衡资源配置下能够学习到具有高度一致性的调度逻辑，保障了绝大部分量子任务的成功执行。

在平均执行时间方面（右上子图），DQN 智能体展现出开局阶段迅速跌落并迅速收敛的特征趋势。基于此实验现象可以推断，智能体在经历初期的快速试错后，能够迅速锁定一种统计意义上较为可靠的任务调度模式。这种长期处于平台期的演变过

程揭示了 DQN 智能体在当前物理环境下的性能边界，即在满足高成功率约束的前提下，受限于量子网络资源分布或通信时延等刚性瓶颈，DQN 智能体难以通过增加训练回合数进一步压缩整体执行耗时。

与基于价值的 DQN 智能体不同，A2C 智能体实现了策略与状态价值函数的联合优化。如图6.1右上子图所示，在均衡配置下，其 100 回合移动平均线轨迹在训练全周期内表现得较为平缓，仅有较小幅度的提升，这表明在 Critic 的价值估计引导下，策略虽有所改进，但并未出现大规模提升。与 DQN 智能体相比，尽管原始奖励的瞬时波动依然频繁，但其整体趋势受约束程度较高。由此可以推断，A2C 智能体在面对随机任务到达与动态资源可用性时，展现出了一种相对稳健的行为特征，能够维持策略的统计一致性，从而在不确定的环境变化中保持基础任务调度的可靠性。

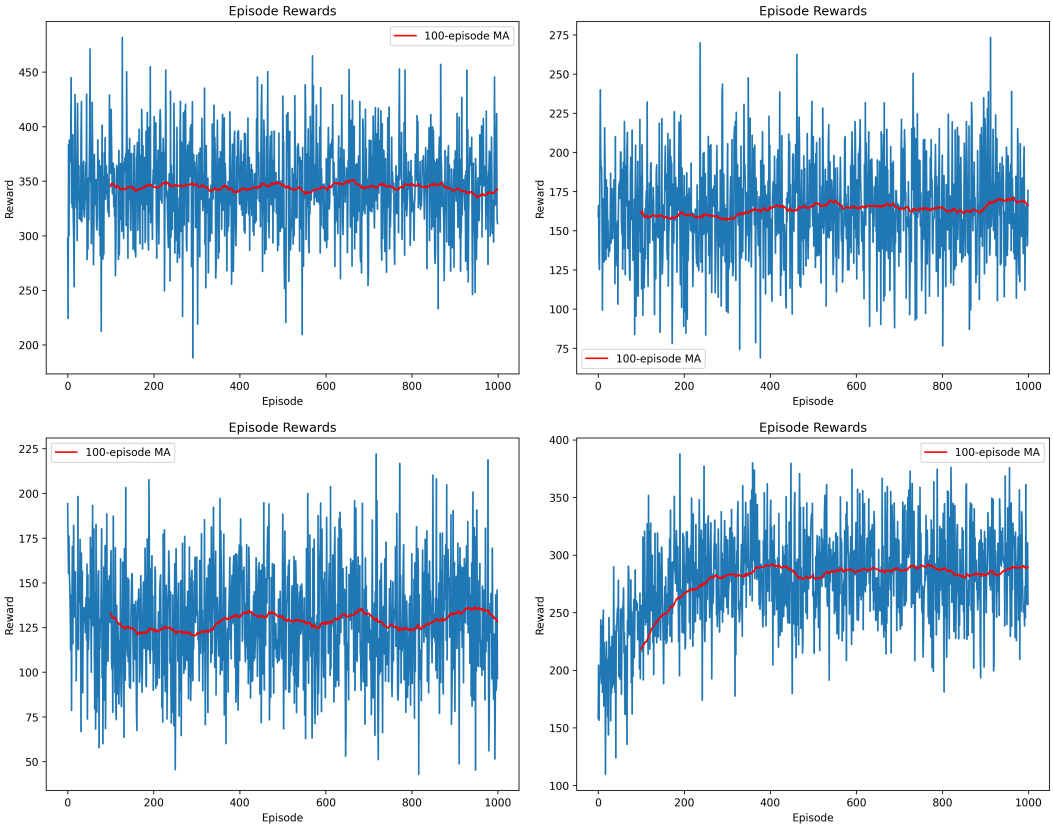

图 6.1　DQN，A2C，A3C 与 PPO 智能体在配置 1 中密度指标下的训练奖励曲线

Figure 6.1　Training reward curve of the DQN,A2C,A3C and PPO agent under medium density in Configuration 1

图6.2下面两幅子图展示了 A2C 智能体在配置 1 中密度指标下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练初期出现急速上升后迅速

回落，并最终维持在约 71% 的水平，并未展现出绝对的收敛优势。

在平均执行时间方面（右下子图），训练初始阶段经历了一次与 DQN 智能体类似的剧烈上升与回落，随后并未立即进入平稳期，而是伴随着较长时间的宽幅震荡。进入训练中后期，得益于策略与价值函数的联合优化，平均执行时间呈现出缓慢下降的趋势，但仍观察到明显的波动。这说明 A2C 智能体经历了更为复杂的探索与动态调整过程，尽管其最终试图通过压低执行耗时来优化性能，但受限于环境约束或算法自身的随机策略特性，其任务成功率与执行效率的平衡上仍表现出一定的不稳定性，未能完全平稳收敛。

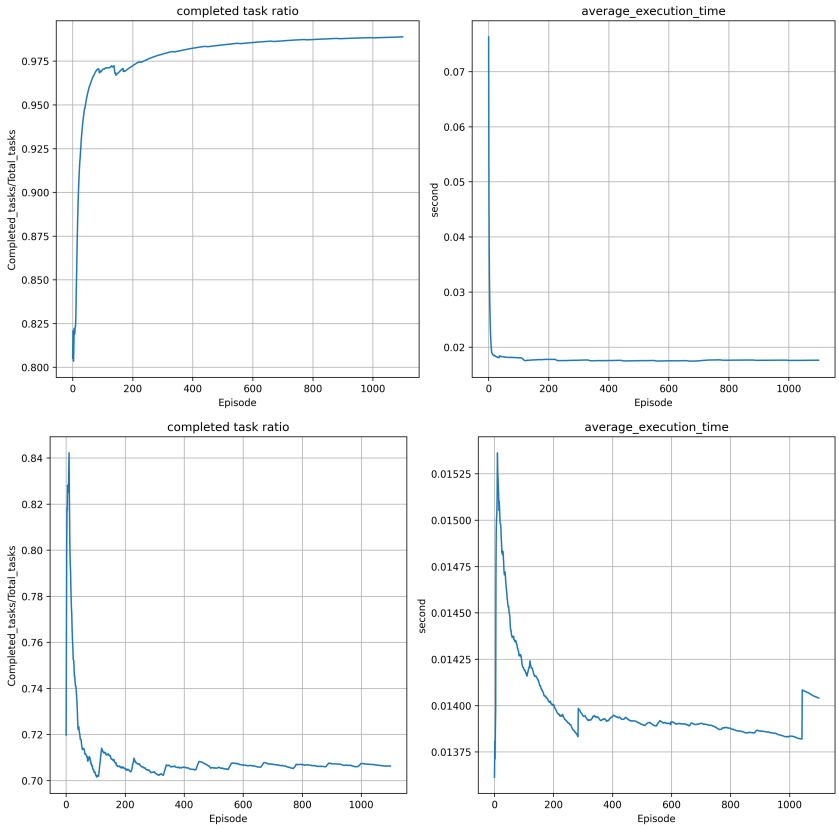

图 6.2 DQN 与 A2C 智能体在配置 1 中密度指标下的任务成功率与平均用时情况

Figure 6.2 Task success rate and execution time of the DQN and A2C agent under medium density in Configuration 1

与 A2C 所采用的同步更新机制不同，A3C 智能体引入了异步训练范式，通过多个独立工作线程与环境交互并更新共享的全局模型。如图6.1左下子图所示，在均衡配置环境下，100 回合移动平均线在训练初期即表现出相对平稳的特征，即便在面临任务到达随机性的挑战时，A3C 智能体仍能通过全局模型的参数整合，在统计意义

上保持策略的稳定性，使奖励收益能够在动态环境下趋于一个相对平衡的稳态。

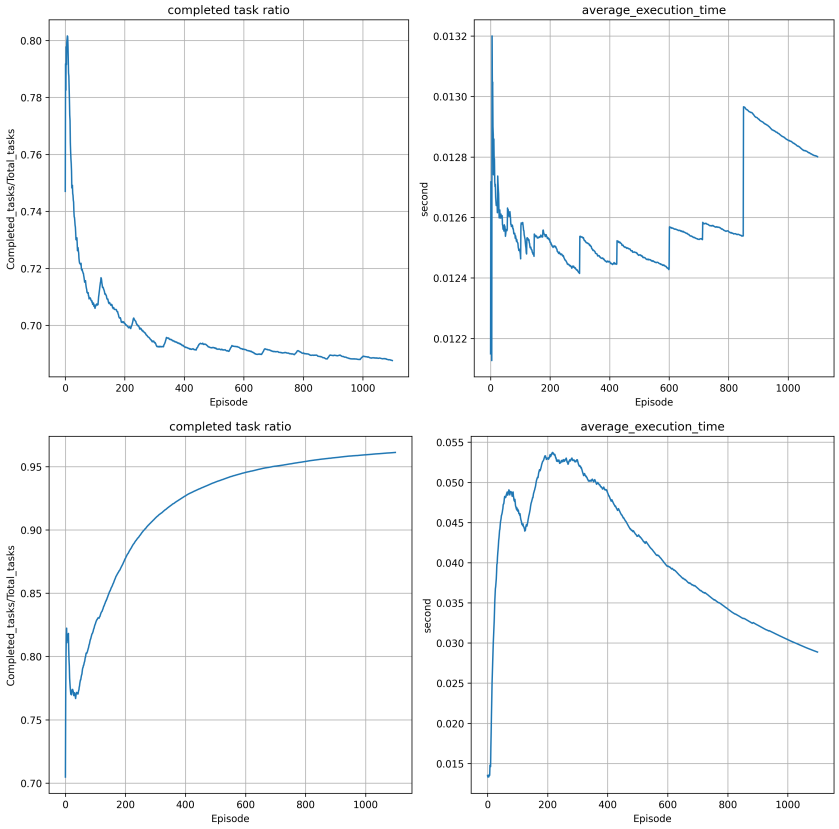

图 6.3　A3C 与 PPO 智能体在配置 1 中密度指标下的任务成功率与平均用时情况

Figure 6.3　Task success rate and execution time of the A3C and PPO agent under medium density in Configuration 1

图6.3上面两幅子图展示了 A3C 智能体在配置 1 下的任务成功率与平均执行时间演变过程。由左上子图可见，任务成功率在训练初期短暂升至约 80% 后迅速下滑，并最终长期收敛至 69% 的水平。这一结果表明，在当前资源配置下，异步并行探索虽能提供多样化的样本，但难以在保障任务成功率方面取得突破。

在平均执行时间方面（右上子图），A3C 展现出了由异步更新机制带来的特有轨迹特征。指标在训练伊始经历了一次剧烈的上升及下降，随后表现出多次显著的阶梯式反弹与锯齿状波动，这些波动直观地反映了多个工作线程在独立环境探索时产生的策略分歧，以及全局模型在同步更新过程中的动态拉扯。这种演化过程揭示了 A3C 凭借异步并发的探索能力，虽能不断尝试优化调度路径，但在处理高度动态的任务流时，策略梯度的稳定性仍受到并行样本差异性的显著影响。

与 DQN、A2C 和 A3C 方法不同，PPO 智能体通过裁剪的替代目标（Clipped Sur-

rogate Objective）来约束策略更新幅度，以此提升训练过程的鲁棒性。如图6.1右下子图所示，在训练初期，其 100 回合移动平均线展现出显著且稳步的上升趋势，反映出策略在优势函数的引导下得到了持续调整。

随着训练的进程 PPO 智能体的奖励均值趋于平稳。基于此实验现象可以推断，PPO 的限制性更新机制在当前训练环境下能够有效抑制策略失效的风险，使智能体在应对动态调度需求时能够逐步收敛至一种相对稳定的行为策略。这种演化特征表明 PPO 在处理该配置下的奖励结构时，展现出了更为受控的收敛路径。

图6.3下面两幅子图展示了 PPO 智能体在配置 1 下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练初期经历小幅波动后稳步攀升，并最终稳定在约 96% 的水平，并仍有较明显的上升趋势，可以推断如果增加其训练轮次，其成功率将有望进一步的提高。

在平均执行时间方面（右下子图），PPO 展现出与其他三种智能体显著不同的演化特征。该指标在训练初期经历了较大幅度的提升，这反映了智能体在初期尝试通过扩大探索空间来寻找可行的调度方案。随后平均执行时间呈现出一条相对平滑且近似单调下降的曲线。基于此实验现象可以推断，PPO 通过裁剪替代目标机制，在允许初期进行必要探索的同时有效限制了策略更新的幅度，从而规避了训练过程中的剧烈震荡。直到训练后期平均执行时间指标仍保持着缓慢递减的态势，表明智能体仍在对调度策略进行微调优化，并有可能在接下来的训练过程中将平均执行时间进一步降低。这种演化过程揭示了 PPO 在处理动态任务调度时，具有更为受控且连贯的策略改进路径，但可能收敛时间较长。

6.1.2 严格配置 (配置 2) 下的性能表现

如图6.4左上子图所示，鉴于对违反截止时间与分配失败的严厉惩罚 ($w_{\text{deadline}} = -10.0$)，通过观察其 100 回合移动平均线可以发现尽管瞬时奖励值起伏显著，但其均值几乎维持在同一水平线，这并非意味着算法已达到最优，而是反映了在强负强化约束下智能体被迫形成了一种倾向于风险规避的调度逻辑，以牺牲部分理论最大化收益为代价，换取对惩罚机制的规避，从而使奖励基线在统计意义上保持在一个相对稳定的高位。

图6.5上面两幅子图展示了 DQN 智能体在配置 2 中密度指标下的任务成功率与平均执行时间演变过程。由左上子图可见，任务完成率在训练初期迅速攀升并最终稳定在 97% 以上的较高水平。

在平均执行时间方面（右上子图）可以看出在训练初期，由于 DQN 智能体处于

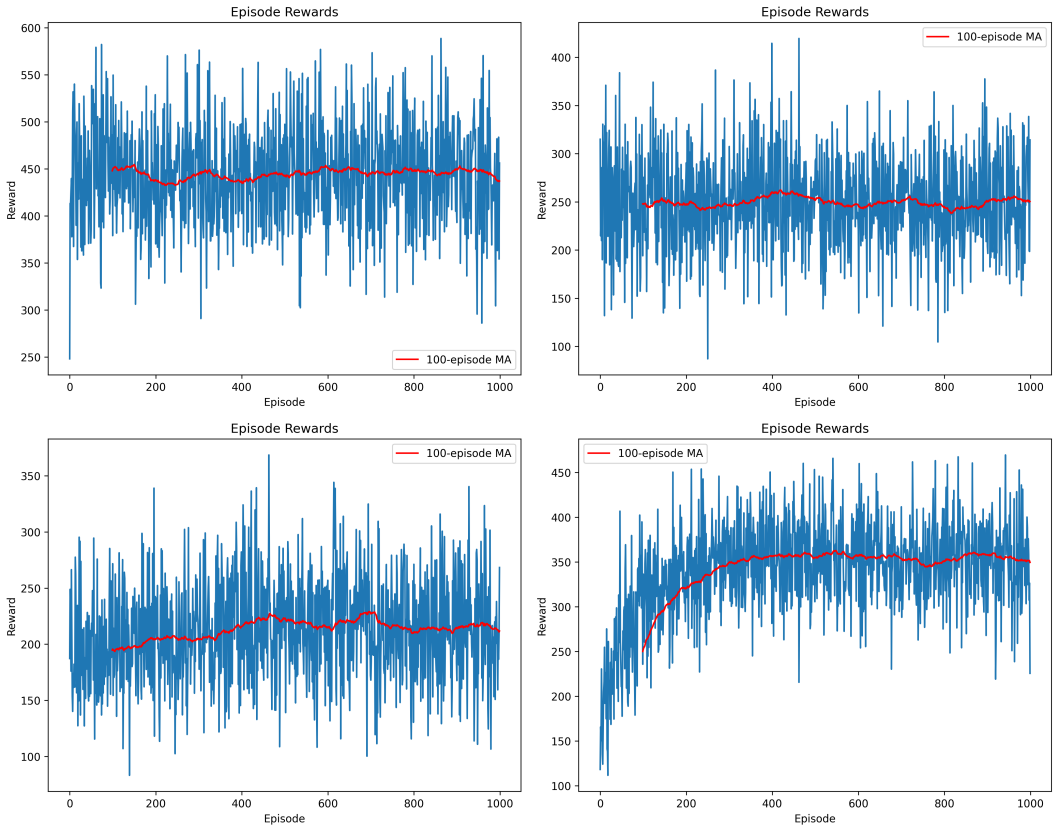

图 6.4　DQN，A2C，A3C 与 PPO 智能体在配置 2 中密度指标下的训练奖励曲线

Figure 6.4　Training reward curve of the DQN,A2C,A3C and PPO agent under medium density in Configuration 2

无序探索阶段，执行时间处于峰值。然而在极短的几十个回合内平均运行时间曲线便急速下降并迅速上升，之后趋于平稳。这种现象反映了 DQN 智能体在面对严苛的负强化约束时为了规避违规惩罚迅速放弃了低效的随机探索。由此可以推断出这种学习特征表明尽管该策略通过牺牲部分探索空间换取了执行效率的快速稳定，但也可能难以进一步挖掘潜在的更优解。

在配置 2 环境下，A2C 智能体在适应严苛惩罚机制的同时，展现了典型的 On-policy Actor-Critic 学习特征。如图6.4右上子图所示，原始回合奖励在整个训练周期内呈现出持续且不断震荡的模式，反映出随机策略在动态调度环境中的探索不确定性。

其 100 回合移动平均线在经历初始波动后，表现出相对平稳的轨迹，这说明 A2C 基于优势函数的更新机制在一定程度上能够将策略引导至可行调度区域，并通过牺牲部分潜在的高收益探索换取对超出截止时间违规风险的规避，从而使策略仍能保持统计意义上的行为一致性，并未出现明显的性能溃缩。

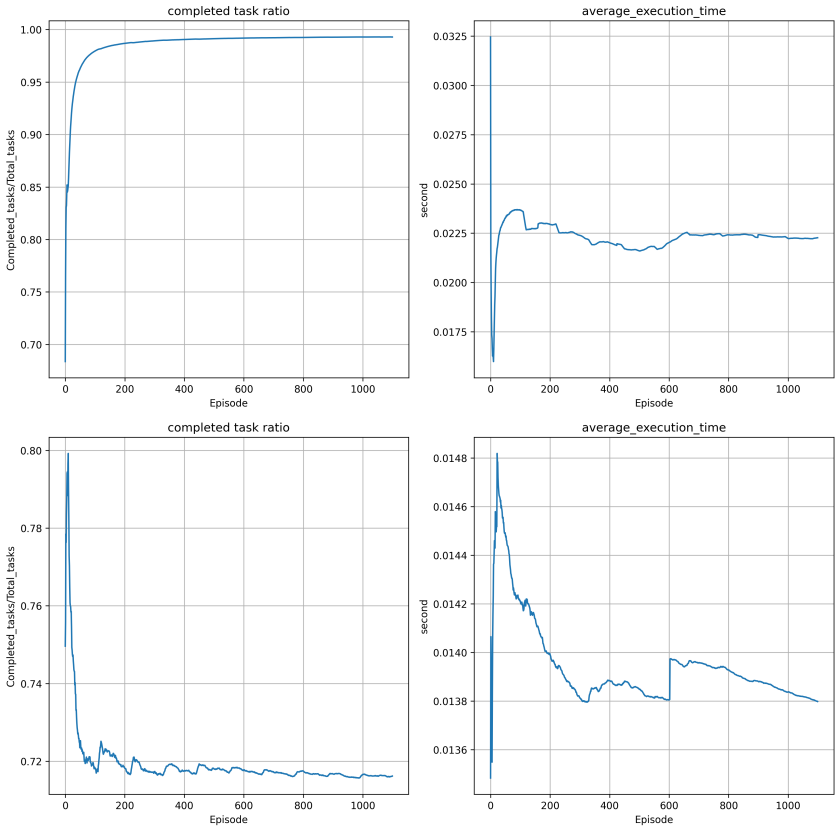

图 6.5 DQN 与 A2C 智能体在配置 2 中密度指标下的任务成功率与平均用时情况

Figure 6.5 Task success rate and execution time of the DQN and A2C agent under medium density in Configuration 2

图6.5下面两幅子图展示了 A2C 智能体在配置 2 中密度指标下的任务成功率与平均执行时间演变过程。在左下子图中任务成功率在训练开始阶段急速上升并下降，并最终维持在约 71.5% 的水平。这一现象反映出在配置 2 的严格约束下，智能体在初步尝试后难以维持高成功率的调度输出。

在平均执行时间方面（右下子图），A2C 智能体展现出先迅速上升后波动下行的学习轨迹。在面对严苛惩罚机制的初期，策略探索引发了调度耗时的短暂激增。但在跨越该峰值后，平均执行时间呈现出明显的衰减趋势，表明基于优势的更新机制正在引导智能体脱离极度低效的动作空间。与 DQN 智能体那种快速趋于平稳的特征不同，A2C 智能体在训练中后期表现出多次阶梯式反弹，但平均执行耗时最终仍收敛并稳定在约 0.0138s 的较低水平。这说明 A2C 智能体虽能通过持续的策略微调来压缩执行时间，但在逼近潜在的优选方案的过程中仍有一定程度的不稳定性。

A3C 智能体在配置 2 环境下展现出特有的学习轨迹。如图6.4左下子图所示，通

过观察其 100 回合移动平均线可以发现，其在训练前 200 回合内呈现出缓慢的上升态势，随后进入平台期在微弱波动中逐渐出现收敛趋势。

即便在引入严厉惩罚机制（$w_{\text{deadline}} = -10.0$）的条件下，奖励均值并未出现剧烈的持续性回撤。基于此实验现象可以推断，由多个工作线程异步收集的并行经验在一定程度上增强了策略的鲁棒性，使全局模型能够逐步识别满足严格系统约束的调度方案。虽然在统计意义上奖励轨迹并未表现出显著的单调上升潜力，但这种相对平稳的演化特征揭示了 A3C 在处理高方差负强化信号时具有一定的适应力。异步更新机制通过整合不同线程的探索反馈，在一定程度上缓解了单一线程可能陷入的局部崩溃风险，使调度策略在严苛约束下趋于一种统计平衡。

图6.6上面两幅子图展示了 A3C 智能体在配置 2 下的任务成功率与平均执行时间演变过程。由左上子图可见，任务成功率在训练初期冲高至约 71% 后迅速跌落至约 63%，随后虽表现出阶梯式回升趋势，但最终仅稳定在约 65.5% 的水平。这一现象表明，在严苛的任务截止时间约束下，异步并行探索机制难以在保障全局任务成功率方面取得显著突破。

在平均执行时间方面（右上子图），A3C 前期展现出了和之前类似的变化趋势。但在训练中后期出现了多次明显的大幅度阶梯式提升。这种独特的波动模式直观地体现了 A3C 异步更新机制在面对高方差惩罚时的动态拉扯，工作线程的局部探索不时打破全局模型的现有平衡，并未实现收敛。

如图6.4右下子图所示，PPO 智能体在配置 2 环境下展现出与配置 1 类似的训练趋势。在训练的前 300 回合内，100 回合移动平均线呈现出显著的上升趋势，随后奖励均值逐渐趋于平稳。

即便在引入严苛惩罚机制（$w_{\text{deadline}} = -10.0$）的条件下，PPO 智能体的原始回合奖励虽然存在瞬时方差，但整体波动范围相对受控。基于此实验现象可以推断，PPO 智能体所采用的裁剪替代目标在处理高负强化信号时，能够有效约束策略更新的幅度。这种机制在一定程度上缓解了智能体在触发严重违规惩罚时可能出现的策略崩溃风险，使其能够在应对严格约束的同时，维持一条相对连贯的策略改进路径，反映出 PPO 智能体在处理复杂非线性约束时具有较好的行为统计一致性。

图6.6下面两幅子图展示了 PPO 智能体在配置 2 中密度指标下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练初期经历波动后表现出稳步上升的态势，并最终在训练末期达到约 96% 的水平，并有明显继续上升的趋势，有望达到非常高的成功率数据。

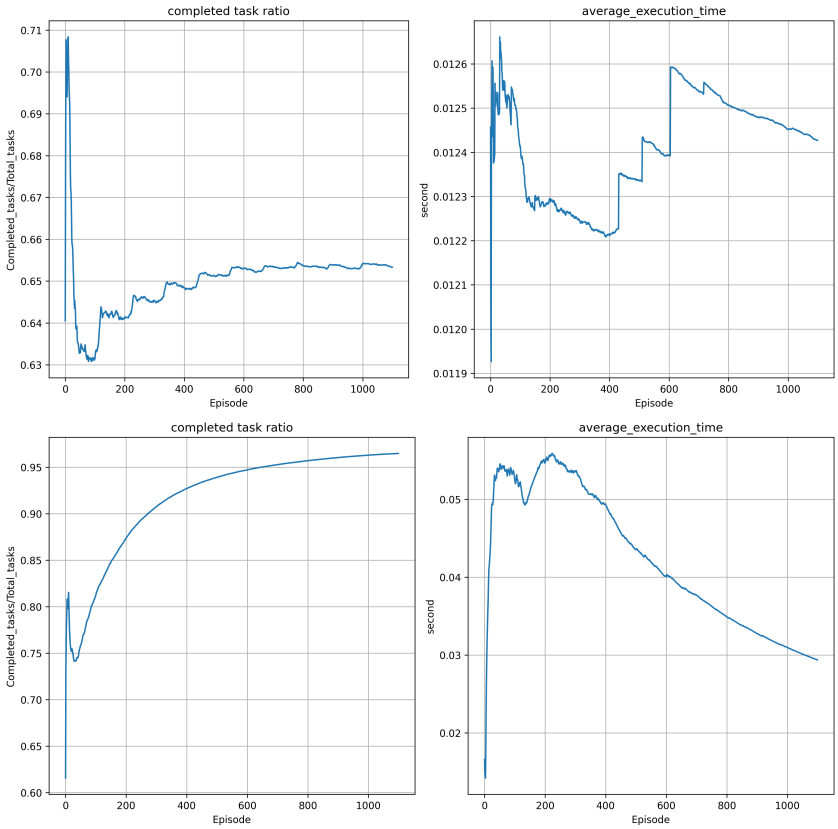

图 6.6 A3C 与 PPO 智能体在配置 2 中密度指标下的任务成功率与平均用时情况

Figure 6.6 Task success rate and execution time of the A3C and PPO agent under medium density in Configuration 2

在平均执行时间方面（右下子图），与此前配置 1 下的演化过程类似，PPO 再次展现出明显的初始阶段起伏后平滑下降的特征。在训练初始阶段出现了一次剧烈的上升，这表明智能体在面对严苛惩罚机制时，仍通过必要的大范围探索来尝试规避局部最优。随后平均执行时间并未出现如其他智能体中常见的剧烈震荡或阶梯式跳跃，而是呈现出一条相对平滑的单调下行曲线。基于此实验现象可以推断，即使在惩罚方差较高的配置 2 环境下，PPO 凭借其裁剪替代目标机制，能够有效抵御大幅梯度波动带来的干扰。这种演化轨迹揭示了智能体在完成初步探索后，能够维持较为受控且连贯的策略优化路径，逐步向低耗时的调度模式收敛。

6.1.3 吞吐量配置 (配置 3) 下的性能表现

如图6.7左上子图所示，配置 3 下的 DQN 智能体回合奖励表现出更为显著的非平稳性。由其 100 回合移动平均线可以得出奖励均值并未在初始预热后迅速稳定，而是在约 200 至 300 的范围内呈现出大幅度的周期性摆动。

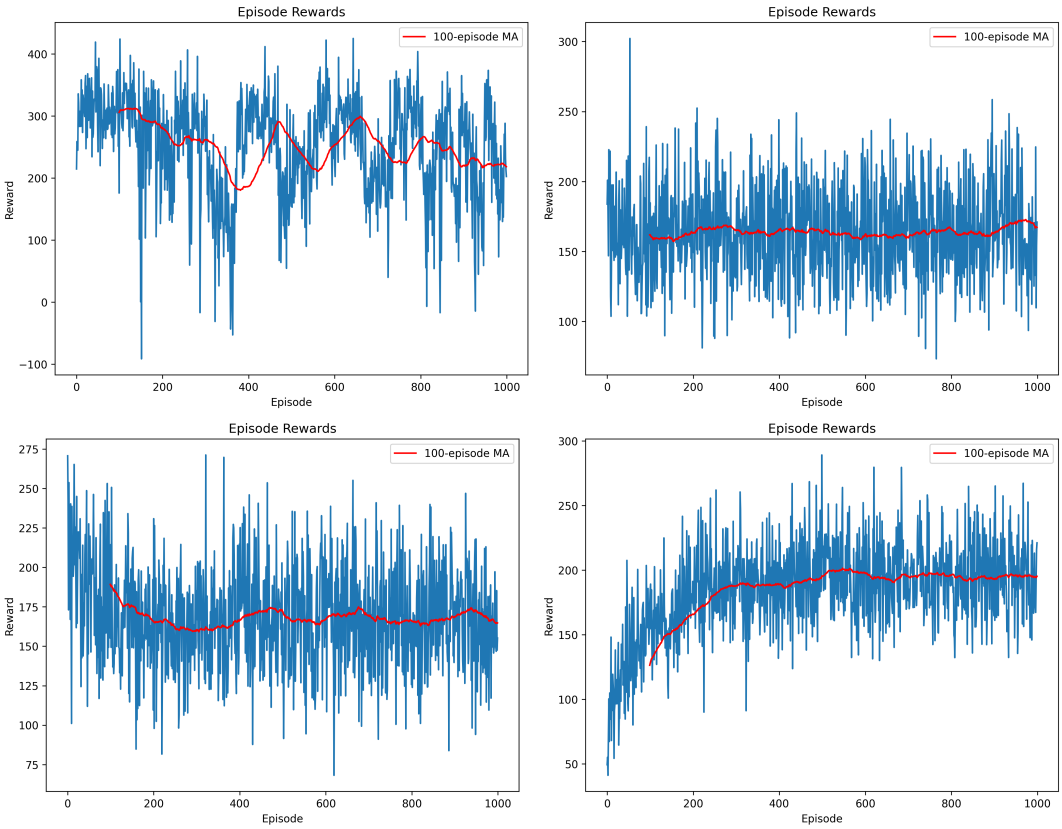

图 6.7　DQN，A2C，A3C 与 PPO 智能体在配置 3 中密度指标下的训练奖励曲线

Figure 6.7　Training reward curve of the DQN,A2C,A3C and PPO agent under medium density in Configuration 3

基于此实验现象可以推断，尽管 DQN 智能体在尝试适应环境的变化，但由于配置 3 中可能存在的复杂约束或资源竞争，其调度策略在探索过程中不断进行大幅度调整，从而导致长期奖励收益出现明显的起伏。这表明 DQN 智能体虽然能够在统计意义上维持基础的奖励水平，但其策略的鲁棒性仍受到环境波动的显著影响，未能实现绝对平稳收敛。

图6.8左上子图展示了 DQN 智能体在配置 3 下的任务成功率演变过程。其任务成功率在训练初期虽一度逼近 98%，但随后遭遇了显著的性能滑坡，最终仅能在 87% 左右的水平维持震荡。

在平均执行时间方面（右上子图），指标呈现出先急剧下降后快速上升并陷入宽幅波动的轨迹。训练前期平均执行时间快速下降的表现反映出智能体曾尝试过一种耗时较低的激进调度策略，但随之而来的成功率大幅波动使智能体对动作选择做出修正。进入训练中后期，平均执行时间伴随着细密的锯齿状特征缓慢抬升并逐渐趋于

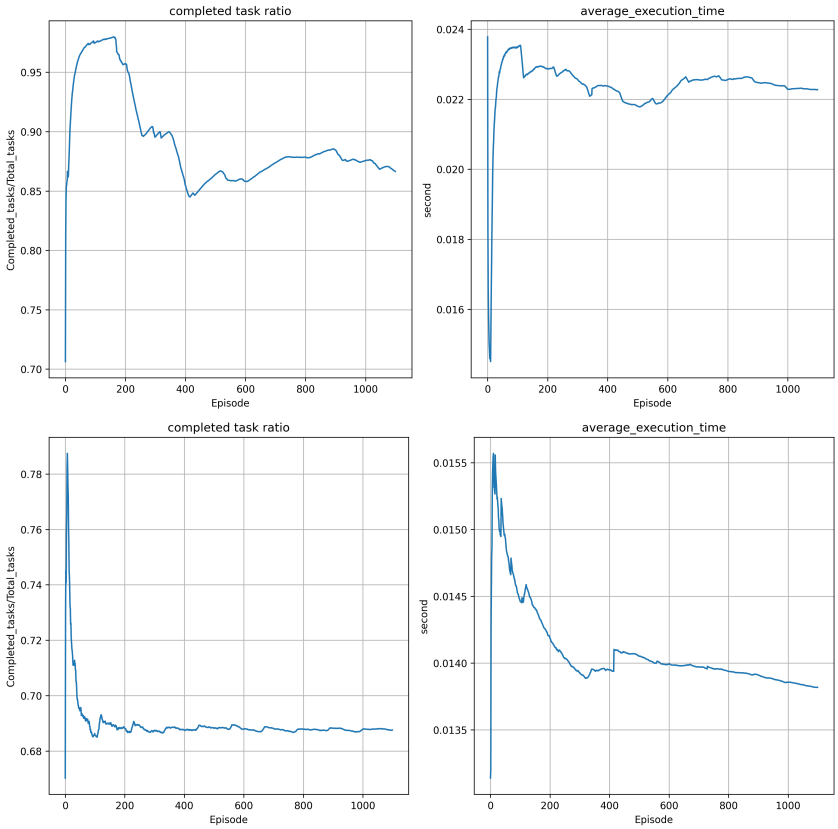

图 6.8 DQN 与 A2C 智能体在配置 3 中密度指标下的任务成功率与平均用时情况

Figure 6.8 Task success rate and execution time of the DQN and A2C agent under medium density in Configuration 3

约 0.022s 以上的区间。可以看出智能体最终不得不放弃极端激进的调度方案，转而锁定于一种执行耗时略长但能勉强平衡各方指标的策略。

如图6.7右上子图所示，在配置 3 环境下，A2C 智能体的原始回合奖励在整个训练周期内呈现出一直震荡的特征。但其 100 回合移动平均线学习轨迹自始至终保持着一条近乎水平的平稳线，数值波动范围极小。

这种现象表明，A2C 智能体在训练初期便锁定了一种相当固定的任务调度模式。基于此实验现象可以推断，其基于优势函数的优化机制在当前环境下并未引发大规模的重新探索，而是使策略能够稳定地维持在一种固有的表现基准上，以应对持续到达的动态任务。这种演化特征说明了该智能体在面对特定约束时，避免了显著的策略偏移，但也在某种程度上也限制了其进一步挖掘潜在更优解的探索空间。

图6.8下面两幅子图展示了 A2C 智能体在配置 3 中密度指标下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练出现了几乎垂直上升并

快速下降的情况，随后最终长期维持在约 69% 的较低水平。这一现象表明，在配置 3 特定的环境约束下，智能体难以维持高成功率的任务调度输出。

右下子图所展示的平均执行时间呈现出开局急剧攀升后波动下行的演化轨迹。在跨越初期探索的峰值后，平均执行时间开始逐步下降，但在约 400 回合处观察到了明显的阶梯式反弹。进入训练中后期，尽管该指标步入了一段相对平缓的缩减期，但这种执行耗时的压低是以牺牲显著的任务成功率为代价的。基于此实验现象可以推断，A2C 智能体在配置 3 下 1000 episodes 内并未真正锁定一种高效且可靠的任务处理模式。

A3C 智能体在配置 3 环境下的学习轨迹展现出显著的波动性。如图6.7左下子图所示，观察 100 回合移动平均线可以发现，指标在训练初期出现了明显回撤，随后进入长期的平台期，并逐渐有收敛趋势。

这种演化特征表明，在配置 3 特定的系统约束下，异步架构提供的并行探索机制并未转化为持续的策略改进动力。这种相对平稳但缺乏上升梯度的轨迹揭示了 A3C 在当前配置下的学习瓶颈，即智能体虽能通过异步更新维持基础的调度表现，但难以在 1000 个回合内突破当前的局部最优平衡点，总体优化效果表现有限。

图6.9上面两幅子图展示了 A3C 智能体在配置 3 下的任务成功率与平均执行时间演变过程。由左上子图可见，任务成功率在训练初期出现短暂冲高后遭遇显著下滑，并最终下降至约 66% 的低位。这一现象表明，在当前资源密度与任务负载下，A3C 智能体难以保障全局任务的成功率，表现出明显的性能瓶颈。

在平均执行时间方面（右上子图），A3C 再次展现出由异步更新机制引发的典型锯齿状震荡特征。指标在训练初期经历剧烈起伏并达到峰值后，表现出一种不连贯的下行趋势。基于此实验现象可以推断，每次突发性反弹反映了异步工作线程在独立探索过程中产生的策略偏差。这种演化特征揭示了 A3C 在高并发任务环境下，其策略梯度更新容易受到局部非最优样本的干扰，导致执行效率的优化过程陷入反复拉扯的平台期，难以实现彻底的平滑收敛。

如图6.7右下子图所示，与前述配置下的表现类似，PPO 智能体在训练初期展现出较为清晰的上升趋势。观察红色的 100 回合移动平均线可以发现，奖励均值在训练中期稳步增长，随后进入平稳期并有收敛趋势。

在优先考虑吞吐量目标的配置 3 下，虽然受随机任务到达与资源利用策略的影响，原始回合奖励仍表现出一定的瞬时方差，但其整体收敛轨迹表现得相对连贯。基于此实验现象可以推断，PPO 通过裁剪替代目标有效约束了策略更新的步长，使其

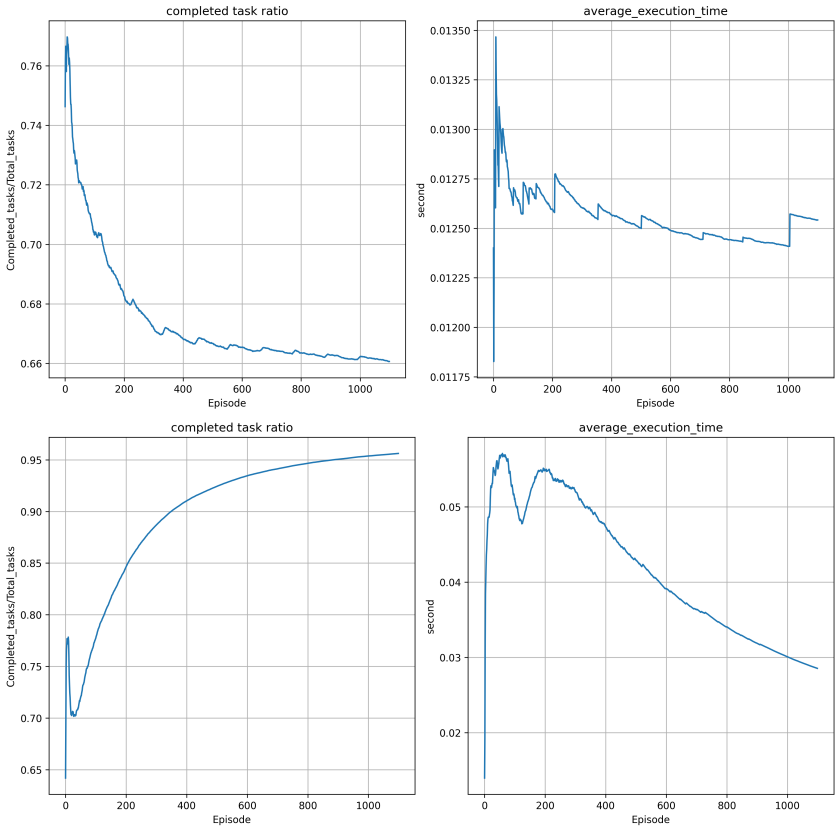

图 6.9 A3C 与 PPO 智能体在配置 3 中密度指标下的任务成功率与平均用时情况

Figure 6.9 Task success rate and execution time of the A3C and PPO agent under medium density in Configuration 3

在追求激进资源利用的同时，能够抑制由大幅梯度更新引发的性能剧烈波动。

图6.9下面两幅子图展示了 PPO 智能体在配置 3 中密度指标下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练初期经历大幅波动后呈现稳步攀升态势，并最终趋于约 96% 的水平，并有继续上升的趋势。

在平均执行时间方面（右下子图），与配置 1 及配置 2 下的演化过程类似，PPO 智能体再次展现出显著的单峰起伏后平滑下降的演化特征。这种现象反映了智能体在优先考虑吞吐量目标的复杂环境下通过必要的大范围探索来寻找可行的调度方案。随后呈现出一条相对平滑的单调下行曲线。这种演化趋势说明了 PPO 智能体在完成初步探索后，能够维持较为受控且连贯的策略优化路径，逐步向更低耗时的调度模式收敛。

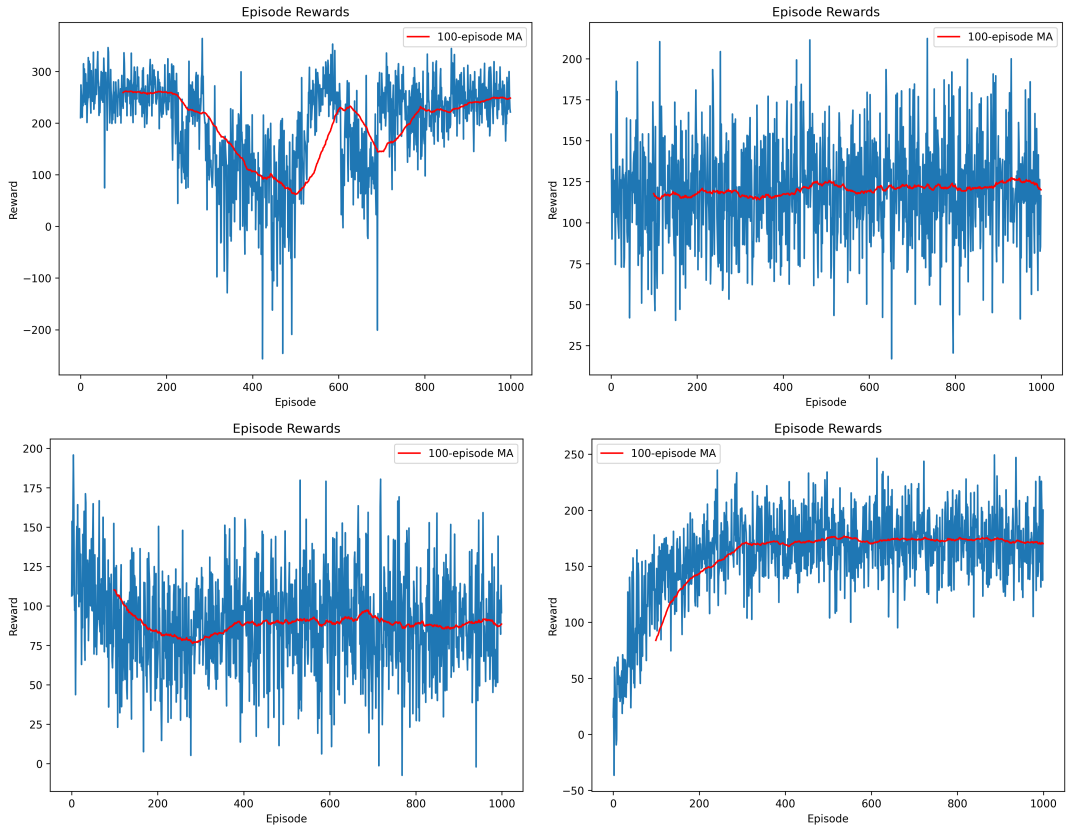

图 6.10　DQN，A2C，A3C 与 PPO 智能体在配置 4 中密度指标下的训练奖励曲线

Figure 6.10　Training reward curve of the DQN,A2C,A3C and PPO agent under medium density in Configuration 4

6.1.4　公平性配置 (配置 4) 下的性能表现

在引入公平指数以平衡节点负载的配置 4 中，如图6.10左上子图所示，DQN 智能体的奖励轨迹呈现出一种剧烈且非线性的演变特征。观察其 100 回合移动平均线可以发现，奖励在训练初期出现大幅度下跌。这表明，在多目标优化的初期阶段，智能体在试图平衡公平约束与调度效率时经历了显著的策略失调。基于此趋势可以推断，智能体在随着训练的推进开始通过调整动作空间以适应新的奖励结构。在表现出一段明显的波动后开始逐渐上升。这反映出公平性约束与系统整体收益之间存在着较强的动态博弈关系，而非简单的线性兼容。

图6.11上面两幅子图展示了 DQN 智能体在配置 4 下的任务成功率与平均执行时间演变过程。由左上子图可见任务成功率在训练初期短暂上升至 95% 后出现了剧烈且长期的下滑趋势，训练末期曲线有所反弹，最终仅能回升至 83% 左右。

在平均执行时间方面（右上子图），指标在经历剧烈波动后表现出处于相对高位

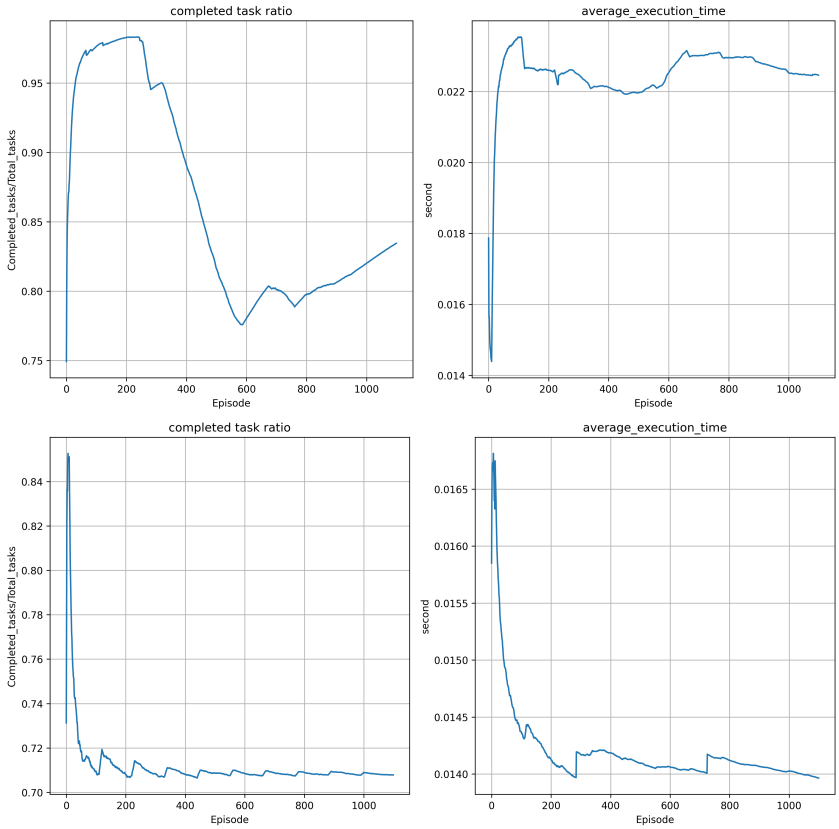

图 6.11 DQN 与 A2C 智能体在配置 4 中密度指标下的任务成功率与平均用时情况

Figure 6.11 Task success rate and execution time of the DQN and A2C agent under medium density in Configuration 4

的演化特征。训练前期，执行时间在极速下探后随即出现急剧反弹，这一轨迹反映出智能体在激进探索极低耗时策略时，因无法兼顾复杂的负载均衡要求而被迫对动作分布进行大幅调整。进入训练中后期，平均执行时间逐渐摆脱了大幅震荡，转而进入一段相对稳定的平稳期并伴有较为缓慢的下行趋势。基于此现象可以推断，在引入公平性约束后，DQN 智能体以牺牲部分任务成功率为代价，降低了执行耗时。

与此前配置下的表现相似，A2C 智能体在应对平衡任务完成与负载分配的多目标挑战时，展现了 Actor-Critic 方法特有的稳定性。如图6.10右上子图所示，尽管奖励函数的复杂性因引入公平性指标而显著增加，但回合奖励的学习趋势并未出现剧烈波动。

通过其 100 回合移动平均线可以发现，其轨迹在训练开始后迅速趋于平稳。基于此现象可以推断，A2C 的同策略更新机制结合优势函数的引导，对公平性指标引入的非线性约束具有较强的适应能力。这一演化特征表明，在多目标博弈环境下，A2C

能够维持策略的统计一致性，虽未实现奖励水平的显著突破，但有效保障了训练过程的平稳性与调度决策的连贯性。

图6.11左下子图展示了 A2C 智能体在配置 4 下的任务成功率在训练初期短暂升至 85% 左右后迅速下滑，最终长期维持在约 71% 的水平。这一结果表明，在引入公平性指标的多目标约束下，智能体难以在保障负载均衡的同时维持较高的任务完成率。

在平均执行时间方面（右下子图），A2C 智能体表现出与此前配置截然不同的剧烈震荡特征。执行时间在开局阶段经历迅速跌落后随即发生数次小范围阶梯状反弹，这反映出智能体在初次尝试平衡公平性约束时，策略探索出现了显著的分歧与重置。基于此实验现象可以推断，尽管整体耗时在统计意义上呈现下降趋势，但频繁的局部波动揭示了同策略更新在处理高度非线性约束时的不稳定性。智能体在试图压低执行耗时以提升效率的过程中，不断受到公平性奖励项的拉扯，导致策略始终处于动态微调的震荡状态。

如图6.10左下子图所示，在配置 4 环境下，A3C 智能体的回合奖励展现出显著的震荡特征。基于此实验现象可以推断，尽管异步并行探索在统计意义上维持了奖励中枢的相对稳定，但由于策略梯度容易在具有潜在冲突的多目标约束中陷入停滞，智能体难以通过目前的并行架构识别出全局最优的调度方案。这种长期处于低位震荡的轨迹揭示了 A3C 在处理复杂公平性约束时的学习瓶颈，反映出其策略网络在局部最优平衡点附近的性能锁定现象，总体优化效果表现平平。

图6.12左上子图展示了 A3C 智能体在配置 4 下的任务成功率演变过程。任务成功率在训练伊始出现短暂的峰值后遭遇显著下滑，并最终长期处于约 66% 的低位区间。

在平均执行时间方面（右上子图），指标呈现出开局阶段极速拉升后震荡下行的轨迹。然而，不同于理想的平滑收敛，A3C 在该环境下表现出频繁的锯齿状波动与显著的阶梯式反弹。基于此实验现象可以推断，每当智能体试图压缩执行耗时，异步工作线程间的策略分歧便会引发全局模型的波动。直至训练末期，执行时间仍未能稳定在低耗时区间，反映出其异步更新机制在导航高度非线性奖励空间时，难以彻底摆脱局部非最优样本的干扰，无法锁定一种兼顾负载均衡与执行效率的稳固调度模式。

如图6.10右下子图所示，PPO 智能体在面对配置 4 的多目标挑战时，展现出较为连贯的学习轨迹。通过观察 100 回合移动平均线可以发现，奖励均值在训练初期迅速攀升，并最终趋于平稳。

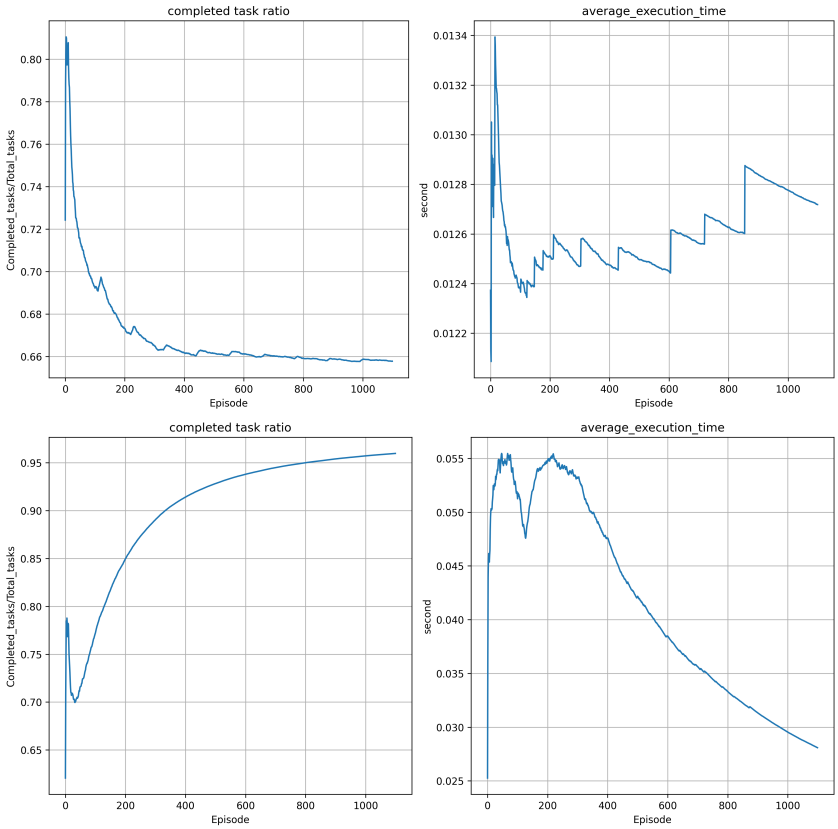

图 6.12 A3C 与 PPO 智能体在配置 4 中密度指标下的任务成功率与平均用时情况

Figure 6.12 Task success rate and execution time of the A3C and PPO agent under medium density in Configuration 4

基于此实验现象可以推断，PPO 通过裁剪替代目标机制，能够有效缓解由面向公平的非线性约束所引发的梯度冲突风险。相较于其他基线方法，PPO 在处理多目标权衡任务时能够更稳健地维持策略的统计一致性。在训练过程中，智能体通过受控的步长更新，逐步在负载均衡与调度收益之间找到了一个相对稳定的平衡点，并未因环境复杂性的增加而导致收敛过程失效。

图6.12左下子图展示了 PPO 智能体在配置 4 下的任务成功率演变过程。任务成功率在训练初期经历波动后表现出稳步攀升的态势，并最终超过 95% 的成功率，并且有继续上升的明显趋势。

在平均执行时间方面（右下子图），与前述配置下的演化特征相似，PPO 的平均运行时间在训练初期经历了显著的提升后呈现出一条相对平滑且近似单调递减的曲线。基于此实验现象可以推断，PPO 通过裁剪替代目标机制，在面对由公平性约束引入的非线性梯度挑战时，能够维持较为连贯的策略优化路径。

6.2　经典强化学习智能体多种资源密度配置下的性能分析研究

为了全面评估各类深度强化学习算法在不同网络拓扑与硬件约束下的泛化能力与鲁棒性，本节将开展多种资源密度配置下的对比研究。在不同资源密度下对 DQN、A2C、A3C 以及 PPO 智能体的表现进行对比，并与前文中研究的中密度下各智能体的表现进行对比分析。为了保证对比实验的一致性，这部分均使用配置 1 进行试验。这一系列对比分析将为未来在不同规模与复杂度的实际量子网络部署中，挑选与适配最佳的智能体架构提供详实的理论与实证依据。

6.2.1　低密度配置下的智能体表现

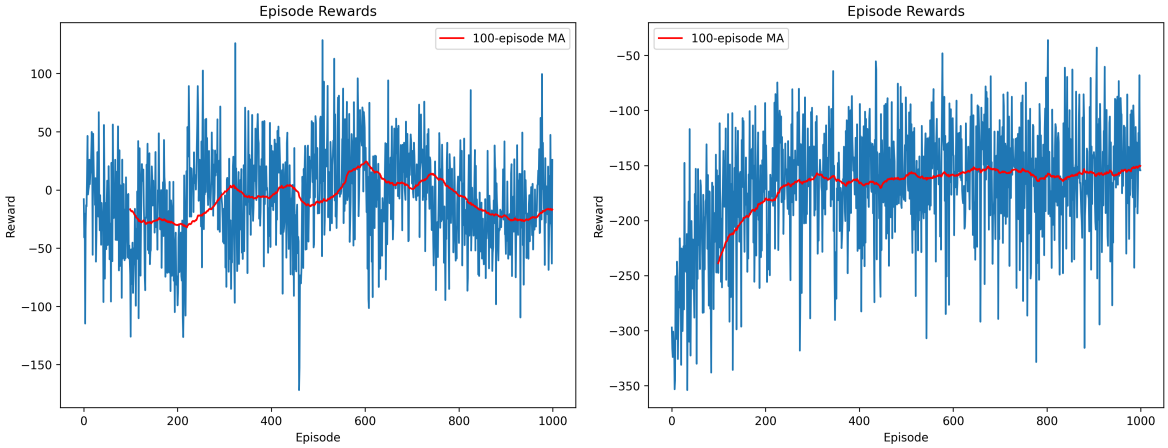

图 6.13　DQN 与 PPO 智能体在低密度指标下的训练奖励曲线
Figure 6.13　Training reward curve of the DQN and PPO agent under low density

6.2.1.1　DQN 智能体

如图6.13左子图所示，在低密度环境下，DQN 智能体的回合奖励表现出显著的非平稳性与较差的收敛特性。通过观察 100 回合移动平均线可以发现奖励均值长期处于-25 至 25 的极低区间内往复震荡，未能实现有效的收益突破。

这种持续波动的奖励轨迹说明了当环境负载密度降低至物理瓶颈以下时，DQN 的贪婪搜索策略容易导致其在非最优解空间内循环往复，从而无法锁定一个统计一致性的调度方案，这表明传统价值学习机制在该特定约束环境下存在明显的适应性局限。

图6.14上面两幅图展示了 DQN 智能体在每个量子节点拥有 2 个量子比特的低密度环境下，任务成功率与平均执行时间的演变过程。由左上子图可见，任务成功率在经历训练初期的剧烈震荡后大幅下滑，最低跌至约 33%，并在训练中后期仅表现出

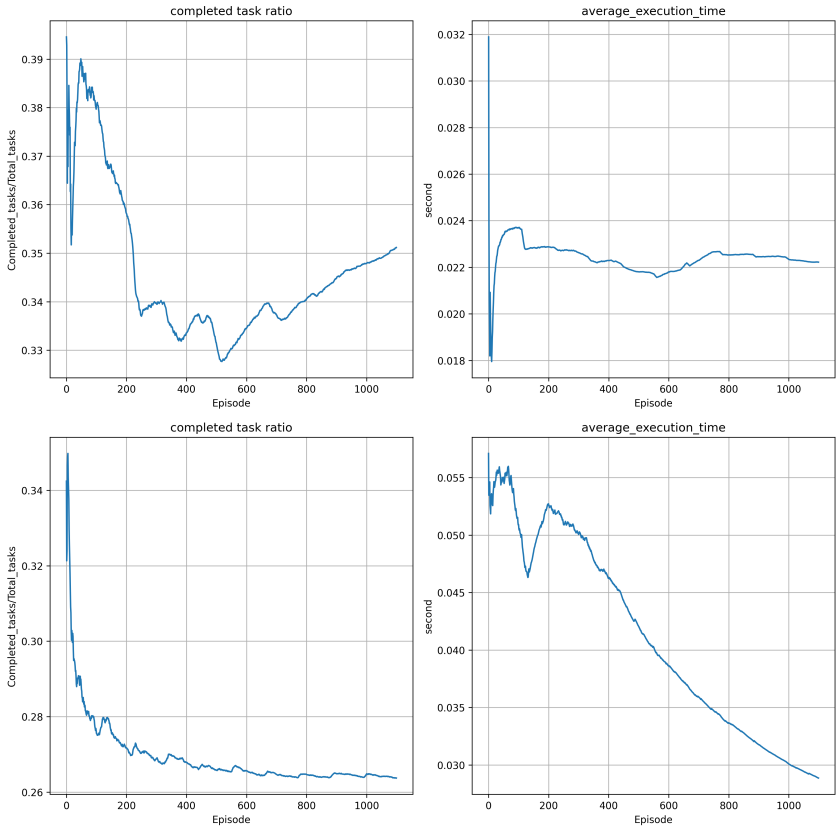

图 6.14 DQN 与 PPO 智能体在低密度指标下的任务成功率与平均用时情况
Figure 6.14 Task success rate and execution time of the DQN and PPO agent under low density

微弱的回升趋势，最终维持在约 35% 的极低水平。

在平均执行时间方面（右上子图），指标呈现出开局快速上升后迅速下降，随后上升并趋于平稳的轨迹。结合此前观察到的低位波动的回合奖励，可以推断出由于可调度的量子资源极其有限，DQN 智能体难以在当前状态空间内通过价值迭代寻找更高效的动作组合，导致其丧失了持续优化的能力。这种演化特征揭示了当环境负载密度降至物理极限附近时，传统的价值学习方法极易陷入局部最优或搜索死点，无法通过增加训练时长来实现性能的有效突破。

6.2.1.2 PPO 智能体

如图6.13右子图所示，在低密度环境下，PPO 智能体的回合奖励展现出持续上升的演化轨迹。

尽管奖励值始终处于负收益区间，但与 DQN 在此环境下的无序震荡不同，PPO 表现出更为连贯的学习趋势。然而，受限于物理资源的刚性瓶颈，智能体在追求更高奖励的过程中面临显著的性能上限。这种演化特征揭示了当环境负载密度降至临界

点以下时，即便是具备鲁棒更新机制的 PPO 算法，也难以通过单纯的策略优化将总收益扭转为正，这反映了底层硬件资源匮乏对调度性能带来的本质性制约。

图6.14左下子图展示了 PPO 智能体在 2 个量子比特的低密度环境下，任务成功率与平均执行时间的演变过程。由左图可见，任务成功率在训练伊始出现短暂波动后遭遇了显著的下滑，并最终长期处于约 26.5% 的极低水平。这一现象表明，在物理计算资源匮乏的约束下，智能体难以保障任务的全局成功率。

在平均执行时间方面（右下子图），PPO 展现出与 DQN 显著不同的演化特征。指标在经历训练初期的剧烈震荡与阶段性反弹后，呈现出一条相对平滑且近似单调下降的轨迹。结合长期较低任务成功率可以发现，这种执行时间在某种程度上反映了智能体在极端物理限制下倾向于通过快速处理极少数可行任务来维持基础的调度表现。这种演化特征揭示了当环境负载密度不足时，算法的优化空间受到了硬件资源匮乏的本质性挤压，难以在确保高效执行的同时提升任务的整体完成质量。

6.2.2　高密度水平下的智能体表现

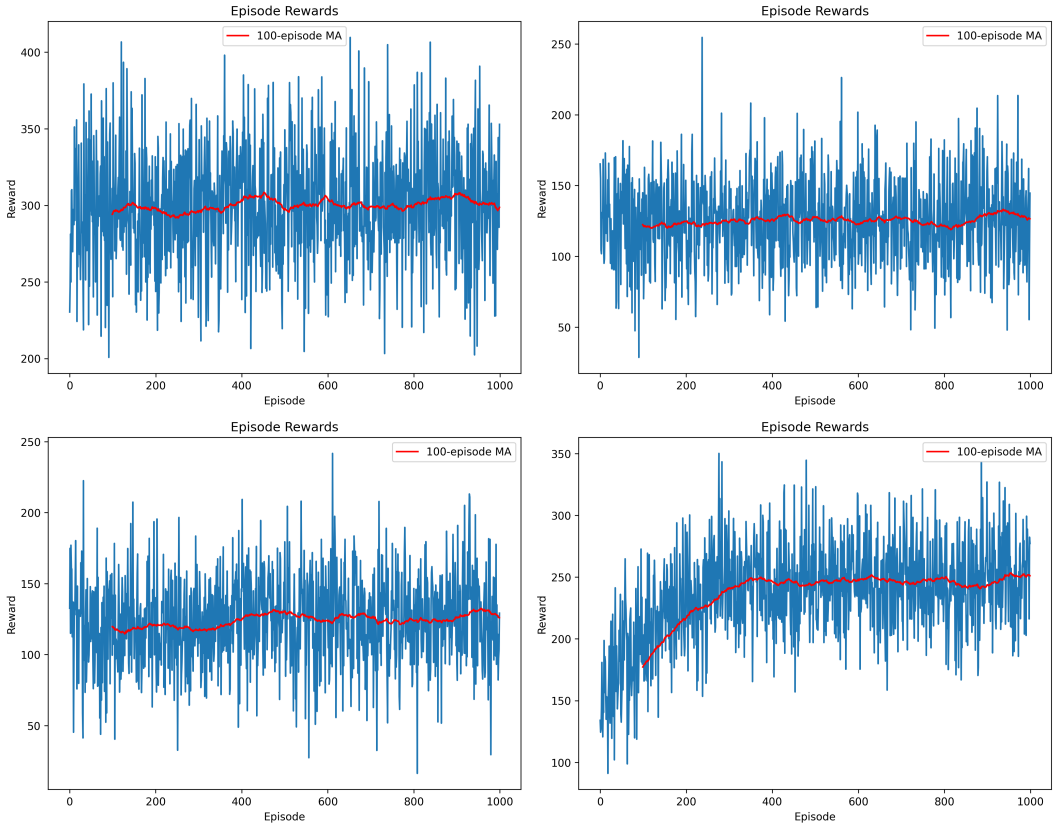

图 6.15　DQN，A2C，A3C 与 PPO 智能体在高密度指标下的训练奖励曲线
Figure 6.15　Training reward curve of the DQN,A2C,A3C and PPO agent under high density

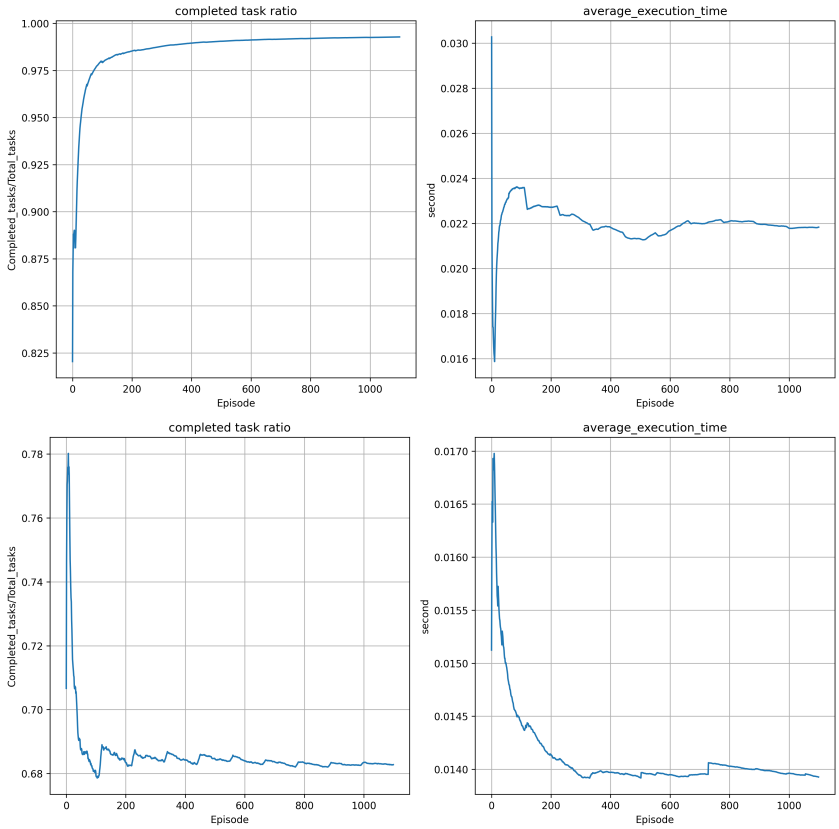

图 6.16 DQN 与 A2C 智能体在高密度指标下的任务成功率与平均用时情况
Figure 6.16 Task success rate and execution time of the DQN and A2C agent under high density

6.2.2.1 DQN 智能体

如图6.15左上子图所示，当网络资源处于高密度水平时，原始回合奖励呈现出剧烈且密集的宽幅高频震荡，反映出高负载环境下状态转移的高度复杂性。观察其 100 回合移动平均线可以发现其并未表现出明显的向上攀升趋势。基于此实验现象可以推断，在高密度任务到达的压力下，DQN 基于价值迭代的机制虽能维持基础的收益水平，但难以通过策略优化进一步突破物理资源的利用率上限。这种演化特征揭示了当任务规模接近系统承载极限时，智能体的探索空间受到严重压缩，导致其调度策略陷入局部最优的停滞状态，无法实现性能的持续迭代提升。

图6.16上面两幅子图展示了 DQN 智能体在高密度指标下的任务成功率与平均执行时间演变过程。由左上子图可见，任务成功率在训练初期经历了显著的爬升阶段，随后轨迹逐渐平缓，并最终稳定在约 99% 的水平。这一结果表明，在高密度任务负载下，DQN 智能体能够通过学习建立起一套高成功率的调度逻辑。

在平均执行时间方面（右上子图），指标在开局阶段经历了极速跌落与随后的快

速反弹，反映了智能体在初期探索过程中的动作调整。跨越波动期后，平均执行时间进入了一段平稳期。基于此实验现象可以推断，在高维状态空间与密集任务到达的压力下，DQN 基于价值迭代的更新机制虽能保障较高的任务完成率，但在进一步压低执行耗时方面表现出明显的停滞特征。这种变化趋势说明当系统资源利用率接近饱和时，DQN 智能体的策略改进空间受到物理环境的限制导致其调度模式锁定于当前的局部平衡点，难以实现执行效率的持续突破。

6.2.2.2　A2C 智能体

如图6.15右上子图所示，在高密度资源环境下，A2C 智能体的回合奖励展现出极高频率且宽幅的震荡特征。其 100 回合移动平均线可以发现，其在整个训练周期内表现得较为平缓，并未表现出任何明显的上升趋势。

基于此实验现象可以推断出在高维状态空间与复杂环境约束的共同作用下，A2C 智能体难以通过当前的探索机制发现更优的调度逻辑，导致其策略性能锁定在当前的局部平衡点，无法实现奖励收益的持续迭代提升。

图6.16下面两幅图展示了 A2C 智能体在高密度指标下的任务成功率与平均执行时间演变过程。由左下子图可见，任务成功率在训练初期短暂攀升后出现剧烈下滑，随后在整个训练中后期几乎稳定于约 68% 的低位水平但伴随有几次跃升情况，说在高并发任务负载的压力下，A2C 智能体难以维持较高的任务完成质量。

在平均执行时间方面（右下子图），指标呈现出开局阶段迅速上升并下降的轨迹，在随后的下行过程中逐渐收敛，但也伴随着数次阶梯式上升。基于此实验现象可以推断，A2C 的更新机制在处理高密度资源调度时，整体耗时在统计意义上呈现出微弱的下降趋势，但频繁的局部震荡揭示了智能体在追求执行效率的过程中受到复杂环境约束的剧烈干扰，难以稳定在一种高效且连贯的调度模式上。

6.2.2.3　A3C 智能体

如图6.15左下图和图6.17上面两幅图所示，在高密度资源环境下，A3C 智能体的 100 回合移动平均线在整个训练周期内表现平缓，并未表现出明显的上升梯度。

在任务成功率方面，指标在初期冲高至约 78% 后迅速回落，最终稳定在约 68% 的区间。然而在平均执行时间方面 A3C 展现出与 A2C 不同的波动形态。指标在经历训练初期的极速跌落与反弹后出现了明显的阶梯式大幅度上涨，但基于此实验现象可以推断，A3C 的异步并行更新机制在一定程度上有助于抑制策略的剧烈退化。相较于同策略单线程版本，其在调度效率上维持了更低的耗时基准，说明了异步架构在处理高密度任务流时，对执行效率优化具有一定的促进作用。

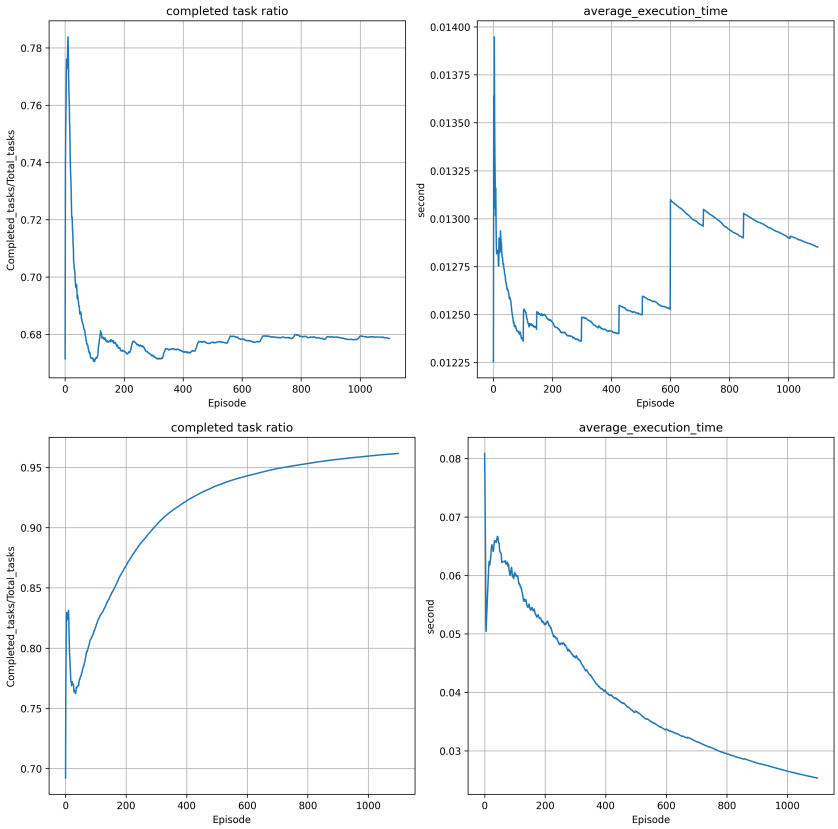

图 6.17 A3C 与 PPO 智能体在高密度指标下的任务成功率与平均用时情况
Figure 6.17 Task success rate and execution time of the A3C and PPO agent under high density

6.2.2.4 PPO 智能体

如图6.15右下子图和图6.17下面两幅图所示，PPO 智能体在高密度环境下的 100 回合移动平均线自训练初期起呈现出显著的上升趋势，并最终奖励水平趋于平稳。基于此实验现象可以推断，PPO 通过裁剪替代目标机制，在一定程度上缓解了高维状态空间带来的探索压力，引导策略向更高收益的区间演进。

奖励的持续增长与后期耗时的平稳下降相统一，揭示了 PPO 智能体在处理复杂量子网络环境时，能够通过限制更新幅度来维持策略的连贯性。这种演化特征表明，即便在任务密度极高且搜索空间巨大的环境下，智能体仍能通过受控的探索逐步优化调度决策。PPO 智能体通过牺牲部分初期的时间效率来确保策略不发生剧烈退化，从而在 1000 个回合的演进中，实现了执行效率与任务成功率的同步改进。

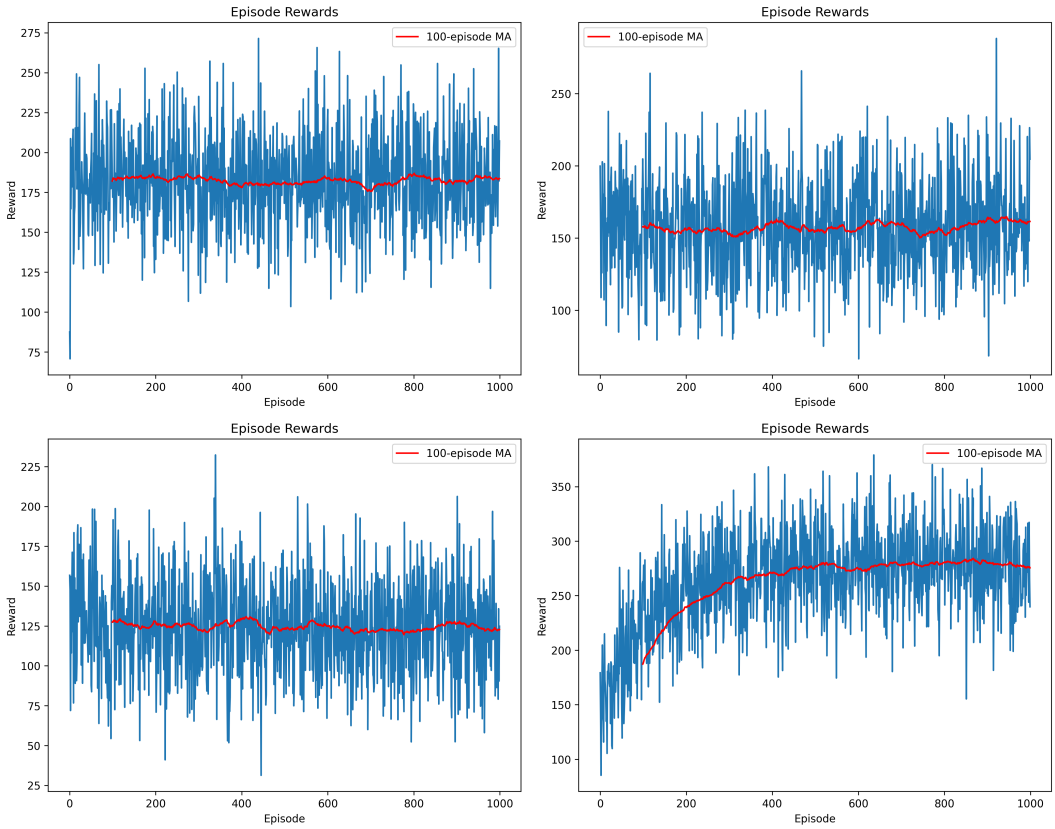

图 6.18　VQC-DQN，VQC-A2C，VQC-A3C 与 VQC-PPO 智能体在中密度指标下的训练奖励曲线

Figure 6.18　Training reward curve of the VQC-DQN,VQC-A2C,VQC-A3C and VQC-PPO agent under medium density

6.3　量子-经典混合强化学习智能体性能分析研究

6.3.1　VQC-DQN 智能体中密度水平下表现

如图6.18左上子图所示，在中密度网络资源环境下，VQC-DQN 智能体的训练奖励表现出独特的收敛特征。其 100 回合移动平均线保持平稳，未出现明显的性能衰减或攀升。

基于此实验现象可以推断，VQC-DQN 智能体在中密度任务压力下展现出了快速收敛的能力。引入 VQC 后的模型能够迅速在庞大的状态空间中锁定最优策略边界。其 100 回合均值线的平稳表现证明了 VQC-DQN 智能体已经稳定地达到了物理资源利用率的平衡点，能够在中密度环境下维持高效且可靠的调度决策水平。

如图6.19上面两幅子图所示，在中密度资源配置下，VQC-DQN 智能体的任务成功率与平均执行时间展现出高度协同的轨迹特征。由左上子图可见，任务成功率在训

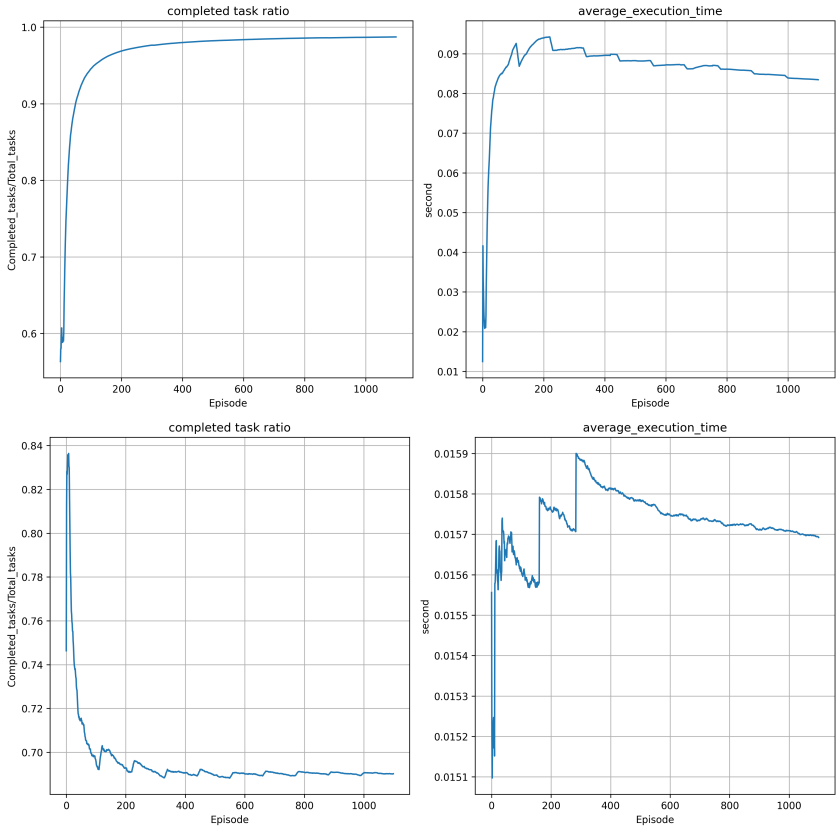

图 6.19 VQC-DQN 与 VQC-A2C 智能体在中密度指标下的任务成功率与平均用时情况
Figure 6.19 Task success rate and execution time of the VQC-DQN and VQC-A2C agent under medium density

练初期经历了迅速的攀升，最终稳定在约 98% 至 99% 的高位水平。这一演化特征表明，引入 VQC 增强了智能体对中等密度任务流的表征能力，使其能够迅速学习到具有高度一致性的调度逻辑，保障了绝大部分量子任务的可靠执行。

在平均执行时间方面（右上子图），VQC-DQN 展现出与成功率爬升相对应的演变特征。在训练起始阶段，平均执行时间因智能体优先保障任务完成质量而出现短暂的激增，随后曲线进入了一个稳健且缓慢的下行收敛区间。基于此实验现象可以推断，在满足高成功率约束的前提下，VQC-DQN 能够通过变分参数的持续微调，在统计意义上不断挖掘任务编排的优化空间，从而进一步压缩整体执行耗时。这说明 VQC-DQN 在中密度物理环境下具备优异的性能调优潜力，能够有效兼顾调度的高成功率与执行的高时效性，但收敛时间较为漫长。

6.3.2　VQC-A2C 智能体中密度水平下表现

如图6.18右上子图所示，在中密度网络资源环境下，VQC-A2C 智能体的训练奖励 100 回合移动平均线在整个训练周期内表现出小幅波动的震荡但微小提升的趋势。

基于此实验现象可以推断，VQC-A2C 在中密度任务压力下展现出了更具探索性的策略迭代过程。引入 VQC 后的 Actor-Critic 框架在训练过程中采用较为稳定的策略平衡，但也在通过对高维状态空间的持续缓慢搜索来挖掘潜在的高收益调度路径。

如图6.19下面两幅子图所示，在中密度资源配置下，VQC-A2C 智能体的任务成功率与平均执行时间展现出较为波动的变化轨迹。由左下子图可见，任务成功率在训练初期经历了短暂的升高，随后出现剧烈下滑并在震荡中趋于平稳，最终维持在约 69% 的低位水平。这一演化特征表明，VQC-A2C 在中密度任务环境下虽然具备一定的快速探索能力，但面对复杂的环境约束与资源竞争时，难以维持长期稳定的高完成质量，导致其调度逻辑在统计意义上陷入了局部最优的瓶颈。

在平均执行时间方面（右下子图），VQC-A2C 展现出频繁震荡后缓慢下行的特征，训练结束时仍有较明显的下降趋势。基于此实验现象可以推断，VQC-A2C 虽然能够通过策略迭代在一定程度上压缩任务执行的冗余开销，使其在统计意义上呈现出微弱的下行趋势，但其整体性能受限于较低的任务成功率。这种低成功率与低平均执行时间的特征揭示了 VQC-A2C 在中密度物理环境下由于过度追求执行效率而牺牲了调度的可靠性，难以实现任务成功率与执行效率的有效兼顾。

6.3.3　VQC-A3C 智能体中密度水平下表现

如图6.18左下子图所示，在中密度网络资源环境下，VQC-A3C 智能体的训练奖励展现出与 VQC-A2C 类似的趋势。其 100 回合移动平均线在整个训练周期内表现出小幅波动的震荡但微小提升的趋势。

如图6.20上面两幅子图所示，VQC-A3C 智能体的任务成功率与平均执行时间演进曲线表现出较强的不稳定性。由左上子图可见，任务成功率在训练初期快速冲高至约 85% 的峰值后，随即陷入剧烈且持续的下滑趋势，最终收敛于约 70.5% 的低位。在平均执行时间方面（右上子图），VQC-A3C 展现出更为剧烈的阶梯式波动与跳变特征。指标在起始阶段迅速下降后，出现了几次阶梯状上升，并最终有收敛趋势。VQC-A3C 在异步更新机制下，虽然其最终平均执行时间处于较低水平，但由于这种低时延是建立在牺牲近 30% 任务成功率的基础之上，说明 VQC-A3C 在中密度环境下难以形成有效且连贯的调度策略。

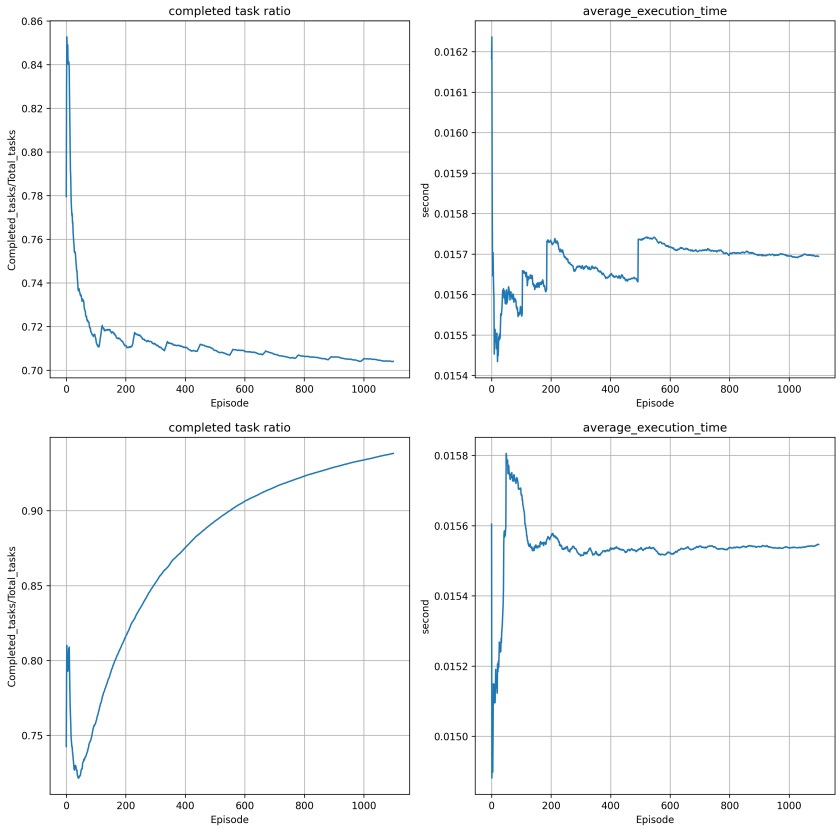

图 6.20 VQC-A3C 与 VQC-PPO 智能体在中密度指标下的任务成功率与平均用时情况
Figure 6.20 Task success rate and execution time of the VQC-A3C and VQC-PPO agent under medium density

6.3.4 VQC-PPO 智能体中密度水平下表现

如图6.18右下子图所示，在中密度网络资源环境下，VQC-PPO 智能体的训练奖励展现出稳健且显著的性能提升趋势。其原始回合奖励虽然伴随有一定程度的震荡，但其 100 回合移动平均线经历了非常平滑且持续的快速攀升。

基于此实验现象可以推断，VQC-PPO 在中密度任务压力下表现出了较强的策略优化能力与训练稳定性。引入 VQC 后的 PPO 框架凭借其特有的概率比率裁剪机制，有效抑制了训练过程中由于奖励信号波动导致的策略大幅偏移。这证明了 VQC-PPO 能够通过稳健的梯度更新实现了性能的深度迭代，最终锁定在一种高效且高收益的调度策略上。

在任务成功率方面（图6.20左下子图），VQC-PPO 智能体任务成功率在训练初期由于剧烈的策略探索而出现大幅波动。然而，随着 PPO 裁剪机制对更新幅度的稳健约束，指标随后开启了单调递增过程，从约 73% 稳步爬升，并在训练末期突破了

93.5%，并在训练结束时仍有较明显的上升趋势。

在右下子图所示的平均执行时间方面，VQC-PPO 智能体展现出一种先波动后逐渐平稳的收敛特征。基于此实验现象可以推断，VQC-PPO 在中密度资源配置下表现出了较强的策略学习能力与鲁棒性，能够有效利用其概率代理目标的优势，规避资源冲突并提升成功率的同时保持在较低的运行时间。

6.4　总体性能对比

如图6.21所示，DQN 系列智能体在不同资源密度与奖励配置下展现出显著的性能差异。在任务成功率维度，除低密度环境下受限于匮乏的量子比特资源外，DQN 在其余中、高密度配置下均表现出较强的稳定性，均能达到 83% 以上的高成功率，甚至在中密度配置 1 和配置 2 以及 VQC-DQN 在中密度配置 1 设置下能够达到 98% 以上的极高成功率。这表明 DQN 算法能够有效学习到量子网络中的资源分配逻辑，在资源充足的情况下展现出较高的任务交付质量。

基于此实验现象可以推断，VQC-DQN 在保持与经典 DQN 相当的较高成功率的同时，其约 33% 的运行时间的增加主要源于量子-经典混合架构中量子态编码与变分层梯度回传所带来的额外计算开销。然而，VQC-DQN 在高收益平稳性上的表现，证明了量子层在提取量子网络状态特征时具有更强的非线性映射能力，虽然牺牲了一定的时间效率，但在应对复杂调度策略时展现出了更具潜力的鲁棒性。

从图6.22和图6.23可以发现，A2C 与 A3C 系列智能体（包括其对应的 VQC 量子增强版本）在处理当前量子任务需求时，普遍表现出成功率瓶颈。在预设的复杂资源约束与任务截止时间要求下，A2C 与 A3C 架构的整体任务完成质量仅能达到约 60% 到 70% 的低成功率，但相较于 DQN 智能体在总体运行时间方面有约 40% 的优势。

引入 VQC 后，VQC-A2C 与 VQC-A3C 在同等配置下的运行时间分别出现了约 10% 与 26% 的增长，达到了对应所有实验组中的最高耗时水平。在实际调度表现中，量子增强层理论上更高维度的特征映射能力的增益被巨大的回传计算开销所掩盖。实验结果揭示了在现有的策略梯度框架下，VQC-A2C 及 VQC-A3C 变种因其较高的执行耗时导致了能效比的进一步恶化，并不适合执行本文所设计的量子任务。

依据图6.24可以得出,PPO 系列智能体表现出极佳的任务交付稳定性。除在计算资源匮乏的低密度环境下受限于物理约束而表现不佳外，在其余中、高密度配置下，PPO 均展现出显著高于同步/异步 Actor-Critic 架构的成功率中枢，但也带来了更大的执行时间。

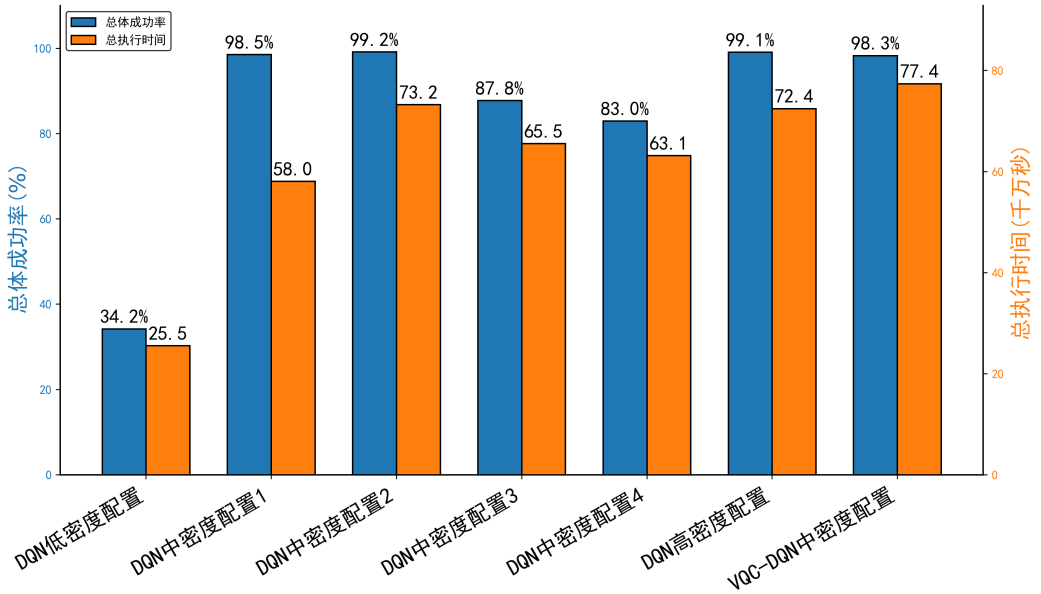

图 6.21 DQN 与 VQC-DQN 智能体配置方案性能对比

Figure 6.21 Comparative Performance Evaluation of DQN and VQC-DQN Agent Configurations

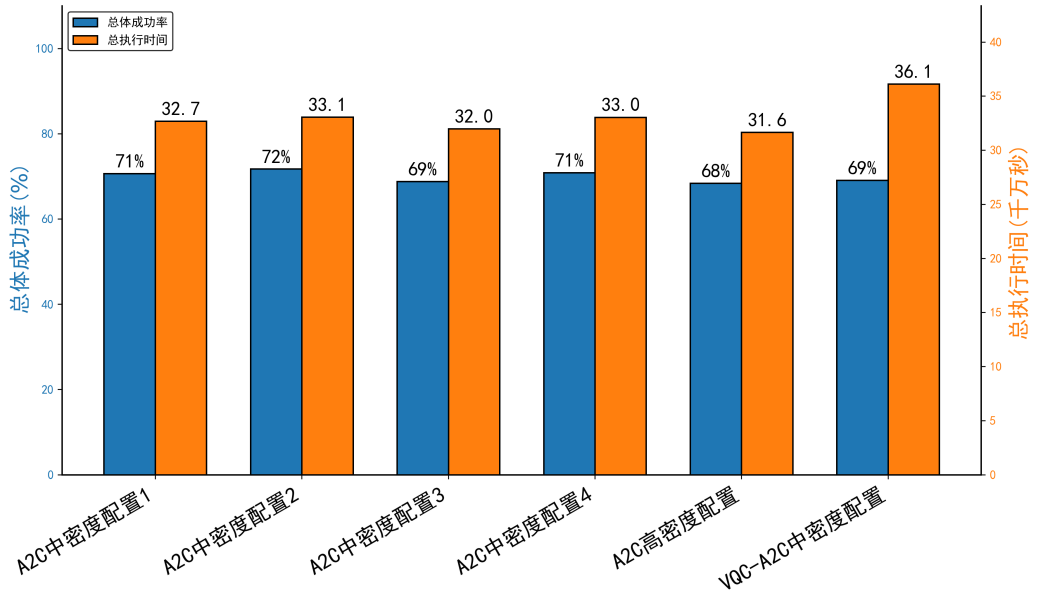

图 6.22 A2C 与 VQC-A2C 智能体配置方案性能对比

Figure 6.22 Comparative Performance Evaluation of A2C and VQC-A2C Agent Configurations

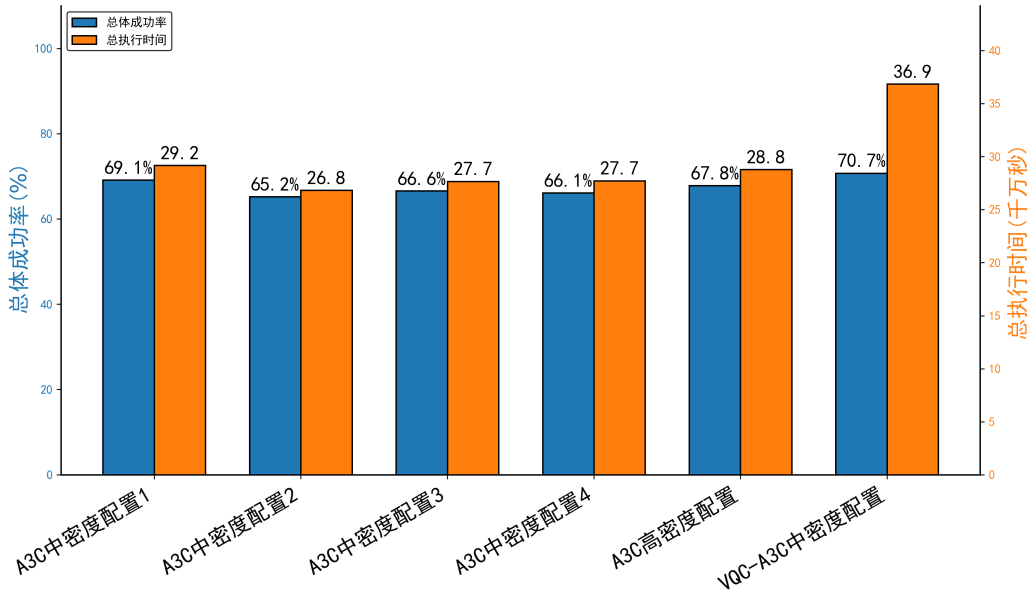

图 6.23　A3C 与 VQC-A3C 智能体配置方案性能对比
Figure 6.23　Comparative Performance Evaluation of A3C and VQC-A3C Agent Configurations

在运行效率方面，该智能体展现出了与其他 VQC 变种智能体截然不同的性能特征。尤为显著的是，引入变分量子电路层后的 VQC-PPO 智能体，在维持与经典版本相近的约 90% 以上的高任务完成率高成功率水平下，其总执行时间较经典 PPO 算法出现了约 60% 的大幅度缩减，对未来处理本文所设置的类似类型的量子任务具有巨大发展潜力及研究价值。

6.5　本章小结

本章通过构建仿真实验体系评估并对比了四类经典深度强化学习智能体（DQN、A2C、A3C、PPO）及其变分量子电路增强版本（VQC 变体）在量子调度任务中的性能表现。实验涵盖了均衡、严格、吞吐量及公平性四种差异化奖励配置，并深入探讨了不同资源密度环境下算法的自适应能力与鲁棒性。

实验结果表明，DQN 智能体在中高密度环境下展现出较高的任务交付稳定性与成功率，但总运行时间较长。而 PPO 智能体凭借裁剪替代目标机制，在复杂非线性约束下实现了受控的收敛路径与较高的任务成功率。通过引入变分量子层，VQC-DQN 与 VQC-PPO 表现出更强的高维特征映射能力。尤其需要注意的是，VQC-PPO 架构在维持高成功率的同时，大幅压缩了调度执行耗时，验证了量子增强机制在特定策略梯度框架下的能效优势。相比之下，A2C 与 A3C 架构在处理当前量子任务时面临明

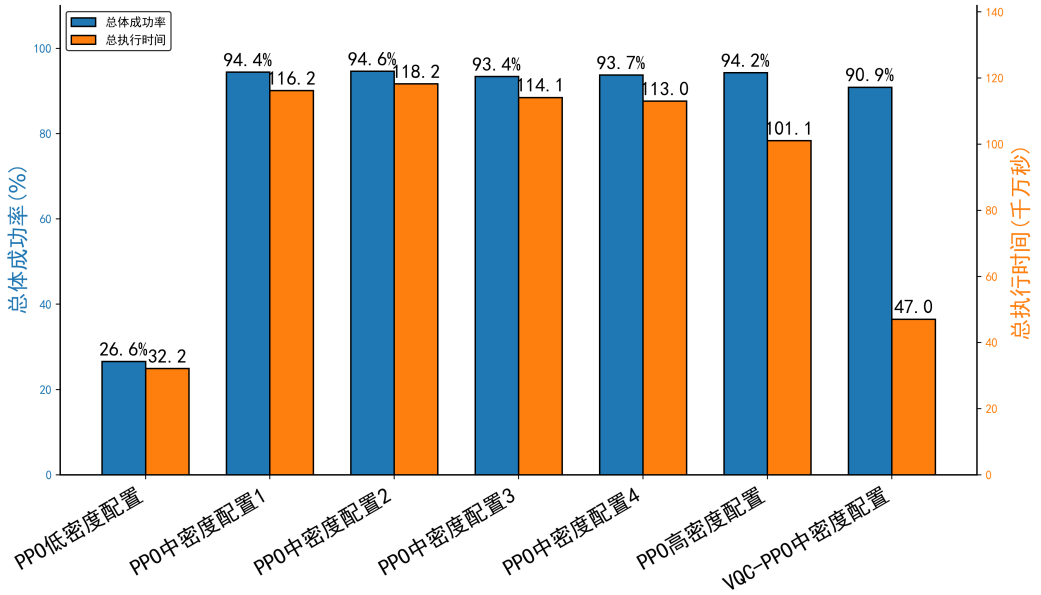

图 6.24 PPO 与 VQC-PPO 智能体配置方案性能对比

Figure 6.24 Comparative Performance Evaluation of PPO and VQC-PPO Agent Configurations

显的成功率瓶颈。本章的实验数据与对比研究不仅验证了量子-经典混合强化学习在复杂资源调度场景中的可行性，也为针对不同硬件约束环境选择最优智能体架构提供了详实的实证依据。

第 7 章 总结与展望

本文围绕含噪声中等规模量子计算环境下资源受限与硬件异构性带来的调度挑战，针对量子任务调度与量子比特映射问题开展了系统性研究，重点完成了两个方面的工作。首先，本文基于经典强化学习算法构建了多种智能调度代理，并设计实现了统一的量子任务调度仿真平台，用于对不同强化学习策略在多种复杂约束环境下的性能进行系统性评估。该调度仿真平台将量子比特的硬件属性、节点拓扑结构以及任务多维约束纳入构建调度仿真的考虑范畴，将调度问题形式化为马尔可夫决策过程，并设计了融合执行时间、截止时间约束、资源冲突以及拓扑映射开销等指标的多目标奖励函数。在此基础上，分别构建并实现了基于 DQN、A2C/A3C 以及 PPO 等多种强化学习智能体，并通过所构建的仿真实验环境对其在任务完成率及运行时间等方面的表现进行了对比分析，从而验证了强化学习方法在复杂量子任务调度场景中的有效性。

在此基础上，本文进一步探索了量子-经典融合的强化学习方法，通过基于经典强化学习的智能体结构中引入变分量子电路构建了混合量子-经典强化学习模型。该方法利用量子电路在高维希尔伯特空间中的特征表达能力，对经典神经网络的表示能力进行扩展，从而增强智能体在复杂状态空间下的决策能力。通过将经典特征嵌入映射至量子态空间，并结合参数化量子线路进行特征变换，所构建的混合模型在理论上具备更强的表达能力与潜在的优化优势。该部分工作不仅验证了量子强化学习方法在资源调度问题中的可行性，也为未来量子计算与强化学习的深度融合提供了初步探索路径。总体而言，本文从经典强化学习到量子增强模型，构建了一条完整的智能调度方法演进路径。

尽管本文在量子任务调度与强化学习方法结合方面取得了一定进展，但仍存在进一步优化与拓展的空间。首先，当前研究主要基于本文所设计的仿真平台开展，未来可考虑将所提出的调度方法部署于真实量子硬件环境中，以验证其在真实噪声、设备不稳定性以及动态变化条件下的实际表现。在实际硬件上部署过程中，需要充分考虑量子比特相干时间的严格限制、门操作误差的累积效应以及读出错误率等关键物理因素，这对调度算法的鲁棒性提出了更高要求。此外，真实量子计算平台通常采用多租户共享模式，如何在保证公平性的前提下实现高效的实时调度将是重要的研究方向。

其次，在量子-经典混合模型方面，本文仅初步引入变分量子电路，但在部分算法中的表现并没有展示出其优异性，后续可进一步研究更深层结构的量子神经网络设计及其训练稳定性问题，并探索其在更大规模任务场景中的性能表现。当前混合模型在训练过程中面临的主要挑战包括变分参数的梯度消失问题以及量子电路的局部最优陷阱，未来可以通过改进参数初始化策略、设计更具表达力的量子架构以及引入自适应学习率调度机制来缓解上述问题。

此外，也可以对仿真环境中的量子任务进行多样化扩展，探究其在更多种类的任务调度中的性能差异。例如，针对量子机器学习任务、量子化学模拟任务以及量子优化问题等不同应用场景，设计差异化的任务特征模型和评估指标，从而构建更加全面的基准测试框架。进一步地，可以考虑引入动态工作负载场景，模拟实际生产环境中任务到达率的时变特性，以评估调度策略在非平稳环境下的适应能力。

在强化学习方法上，可以引入更加高效的训练机制，如离策略算法、模型驱动方法或迁移学习策略，以提升样本效率并增强模型的泛化能力。离策略方法能够有效利用历史交互数据，显著降低对环境交互的需求，从而加速训练进程。模型驱动的强化学习方法通过学习环境动力学模型，可以实现基于想象轨迹的策略优化，进一步提升采样效率。此外，迁移学习技术可以帮助将在某一类量子硬件配置下训练好的调度策略快速适配到新的硬件环境中，降低针对不同平台的重复训练成本。同时，调度策略与量子编译过程之间的协同优化也可能是未来提升系统整体性能的重要方向，并有望进一步降低执行开销并提升系统性能。量子编译负责将高层量子算法转换为底层物理门序列，而任务调度负责确定任务的执行时机和资源分配，二者之间存在紧密的耦合关系。通过联合优化编译策略与调度决策，可以实现量子电路拓扑映射与任务调度的协同设计，从而更充分地利用硬件连通性并有效减少 SWAP 操作的开销。这种跨层次的优化思路为实现高效的量子计算资源管理提供了新的研究视角，也有望成为连接量子系统软件栈各层的关键技术突破点。

参考文献

[1] DI MEGLIO A, JANSEN K, TAVERNELLI I, et al. Quantum computing for high-energy physics: State of the art and challenges[J]. PRX Quantum, 2024, 5(3): 037001.

[2] CEREZO M, ARRASMITH A, BABBUSH R, et al. Variational quantum algorithms[J]. Nature Reviews Physics, 2021, 3(9): 625-644.

[3] BIAMONTE J, WITTEK P, PANCOTTI N, et al. Quantum machine learning[J]. Nature, 2017, 549(7671): 195-202.

[4] SCHULD M, SINAYSKIY I, PETRUCCIONE F. An introduction to quantum machine learning [J]. Contemporary Physics, 2015, 56(2): 172-185.

[5] PRESKILL J. Quantum computing in the NISQ era and beyond[J]. Quantum, 2018, 2: 79.

[6] PATIL K T, BORSE K, KULKARNI M. Quantum computing-based cybersecurity applications: Case studies[M]//Quantum Algorithms for Enhancing Cybersecurity in Computational Intelligence in Healthcare. CRC Press, 2025: 296-306.

[7] ABUGHANEM M. IBM quantum computers: evolution, performance, and future directions: M. AbuGhanem[J]. The Journal of Supercomputing, 2025, 81(5): 687.

[8] ZHU Y, CHENG J, LI B, et al. Hardware-aware calibration protocol for quantum computers[C]// Proceedings of the 52nd Annual International Symposium on Computer Architecture. 2025: 241-256.

[9] GUERRESCHI G G, HOGABOAM J, BARUFFA F, et al. Intel Quantum Simulator: A cloud-ready high-performance simulator of quantum circuits[J]. Quantum Science & Technology, 2020, 5(3): 034007.

[10] ASPURU-GUZIK A, DUTOI A D, LOVE P J, et al. Simulated quantum computation of molecular energies[J]. Science, 2005, 309(5741): 1704-1707.

[11] TANNU S S, QURESHI M K. Not all qubits are created equal: A case for variability-aware policies for NISQ-era quantum computers[C]//Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems. 2019: 987-999.

[12] SIRAICHI M Y, SANTOS V F D, COLLANGE C, et al. Qubit allocation[C]//Proceedings of the 2018 international symposium on code generation and optimization. 2018: 113-125.

[13] MATSUO Y, LECUN Y, SAHANI M, et al. Deep learning, reinforcement learning, and world models[J]. Neural Networks, 2022, 152: 267-275.

[14] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. nature, 2017, 550(7676): 354-359.

[15] NGUYEN T T, NGUYEN N D, NAHAVANDI S. Deep reinforcement learning for multiagent

systems: A review of challenges, solutions, and applications[J]. IEEE transactions on cybernetics, 2020, 50(9): 3826-3839.

[16] ARUTE F, ARYA K, BABBUSH R, et al. Quantum supremacy using a programmable superconducting processor[J]. nature, 2019, 574(7779): 505-510.

[17] PINO J M, DREILING J M, FIGGATT C, et al. Demonstration of the trapped-ion quantum CCD computer architecture[J]. Nature, 2021, 592(7853): 209-213.

[18] EBADI S, WANG T T, LEVINE H, et al. Quantum phases of matter on a 256-atom programmable quantum simulator[J]. Nature, 2021, 595(7866): 227-232.

[19] ZHONG H S, WANG H, DENG Y H, et al. Quantum computational advantage using photons[J]. Science, 2020, 370(6523): 1460-1463.

[20] 魏世杰, 王涛, 阮东, 等. 量子算法的一些进展[J]. 中国科学: 信息科学, 2017, 47(10): 1277.

[21] BENEDETTI M, LLOYD E, SACK S, et al. Parameterized quantum circuits as machine learning models[J]. Quantum Science and Technology, 2019, 4(4): 043001.

[22] JORDAN S P. Quantum Algorithm Zoo[Z]. https://quantumalgorithmzoo.org.

[23] MONTANARO A. Quantum algorithms: an overview[J]. npj Quantum Information, 2016, 2(1): 1-8.

[24] FARHI E, GOLDSTONE J, GUTMANN S. A quantum approximate optimization algorithm[J]. arXiv preprint arXiv:1411.4028, 2014.

[25] ZHOU L, WANG S T, CHOI S, et al. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices[J]. Physical Review X, 2020, 10(2): 021067.

[26] KUROWSKI K, PECYNA T, SLYSZ M, et al. Application of quantum approximate optimization algorithm to job shop scheduling problem[J]. European Journal of Operational Research, 2023, 310(2): 518-528.

[27] PRASAD Y J D S, MASTHAN S F. Quantum Scheduling Optimization for Airline and Space Missions: A Hybrid Quantum-Classical Approach[J]. Authorea Preprints, 2025.

[28] SCHWORM P, WU X, KLAR M, et al. Multi-objective Quantum Annealing approach for solving flexible job shop scheduling in manufacturing[J]. Journal of Manufacturing Systems, 2024, 72: 142-153.

[29] GROVER L K. Quantum mechanics helps in searching for a needle in a haystack[J]. Physical review letters, 1997, 79(2): 325.

[30] BENNETT C H, BERNSTEIN E, BRASSARD G, et al. Strengths and weaknesses of quantum computing[J]. SIAM journal on Computing, 1997, 26(5): 1510-1523.

[31] CHEN Y, GILYÉN A, de WOLF R. A quantum speed-up for approximating the top eigenvectors of a matrix[C]//Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). 2025: 994-1036.

[32] HAVLÍČEK V, CÓRCOLES A D, TEMME K, et al. Supervised learning with quantum-enhanced

feature spaces[J]. Nature, 2019, 567(7747): 209-212.

[33] SCHULD M, KILLORAN N. Quantum machine learning in feature Hilbert spaces[J]. Physical review letters, 2019, 122(4): 040504.

[34] JERBI S, FIDERER L J, POULSEN NAUTRUP H, et al. Quantum machine learning beyond kernel methods[J]. Nature Communications, 2023, 14(1): 517.

[35] PAINE A E, ELFVING V E, KYRIIENKO O. Quantum kernel methods for solving regression problems and differential equations[J]. Physical Review A, 2023, 107(3): 032428.

[36] XU L, WANG J, CHEN J, et al. Quantum Support Vector Machines and Quantum Kernel Methods [J]. Software: Practice and Experience, 2026.

[37] MCCLEAN J R, BOIXO S, SMELYANSKIY V N, et al. Barren plateaus in quantum neural network training landscapes[J]. Nature communications, 2018, 9(1): 4812.

[38] WANG S, FONTANA E, CEREZO M, et al. Noise-induced barren plateaus in variational quantum algorithms[J]. Nature communications, 2021, 12(1): 6961.

[39] GRANT E, WOSSNIG L, OSTASZEWSKI M, et al. An initialization strategy for addressing barren plateaus in parametrized quantum circuits[J]. Quantum, 2019, 3: 214.

[40] MEYER J J, MULARSKI M, GIL-FUSTER E, et al. Exploiting symmetry in variational quantum machine learning[J]. PRX quantum, 2023, 4(1): 010328.

[41] PALANIVEL R, MUTHULAKSHMI P. Quantum prioritized experience replay with MaDi-based priority and quantum circuit mechanisms for optimizing reinforcement learning[J]. International Journal of Advanced Technology and Engineering Exploration, 2024, 11(121): 1664.

[42] YUN W J, PARK S, KIM J, et al. Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control[J]. IEEE Transactions on Industrial Informatics, 2022, 18(10): 7086-7096.

[43] De SOUSA U G. ARDNS-FN-Quantum: A Quantum-Enhanced Reinforcement Learning Framework with Cognitive-Inspired Adaptive Exploration for Dynamic Environments[J]. arXiv preprint arXiv:2505.06300, 2025.

[44] MAO H, ALIZADEH M, MENACHE I, et al. Resource management with deep reinforcement learning[C]//Proceedings of the 15th ACM workshop on hot topics in networks. 2016: 50-56.

[45] WANG Y, YANG X. Research on edge computing and cloud collaborative resource scheduling optimization based on deep reinforcement learning[C]//2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE). 2025: 2065-2073.

[46] SHENG S, CHEN P, CHEN Z, et al. Deep reinforcement learning-based task scheduling in IoT edge computing[J]. Sensors, 2021, 21(5): 1666.

[47] LV L, ZHANG S, DING D, et al. Path planning via an improved DQN-based learning policy[J]. IEEE Access, 2019, 7: 67319-67330.

[48] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J].

arXiv preprint arXiv:1707.06347, 2017.

[49] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[50] TAN H. Reinforcement learning with deep deterministic policy gradient[C]//2021 International conference on artificial intelligence, big data and algorithms (CAIBDA). 2021: 82-85.

[51] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International conference on machine learning. 2018: 1861-1870.

[52] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[J]. Handbook of reinforcement learning and control, 2021: 321-384.

[53] WEI X, GAO X, YE K, et al. A quantum reinforcement learning approach for joint resource allocation and task offloading in mobile edge computing[J]. IEEE Transactions on Mobile Computing, 2024, 24(4): 2580-2593.

[54] ASTITI N M E P, LEE B M. A Survey of Quantum Deep Reinforcement Learning for Resource Allocation in Future Wireless Networks[J]. IEEE Communications Surveys & Tutorials, 2025, 28: 109-148.

[55] NGUYEN T T H, NGUYEN T T, CAO H. Variational Quantum Rainbow Deep Q-Network for Optimizing Resource Allocation Problem[J]. arXiv preprint arXiv:2512.05946, 2025.

[56] DAI S, SAURABH N, WANG Q, et al. Quantum Reinforcement Learning for QoS-Aware Real-Time Job Scheduling in Cloud Systems[J]. IEEE Systems Journal, 2025.

致 谢

时光荏苒，行文至此，我的硕士生涯也已接近尾声。回望在上海交通大学的求学时光，心中充满感激。

首先，我谨以最诚挚的敬意，感谢我的导师计算机学院马汝辉老师。马老师不仅在学术研究上给予我悉心的教导，更是在人生道路上为我指明了方向。您严谨的治学态度、深厚的专业造诣以及对学生无微不至的关怀，让我受益匪浅，这种指引将成为我未来前行中最重要的精神财富。

其次，我也要感谢实验室里所有的同门师兄弟。那些并肩作战、互相帮助、共同进步的日日夜夜，共同构成了我研究生岁月中最珍贵的记忆。特别要感谢蔡子诺学长自本科至硕士多年来的无私帮助与悉心引领，在每一个迷茫与挑战的时刻给予了我许多宝贵的建议和经验分享。

此外，我由衷感谢巴黎卓越工程师学院的培养，为我提供前往法国攻读双学位的宝贵机会，让我能够走出原有的边界，在异国的学术氛围中体会截然不同的学习与生活。这段跨越国界的求索之旅，不仅极大地开阔了我的国际视野，更让我学会了在多元文化中思考与成长。

最后，我要将最深的情感献给我的家人与挚友。求学之路并非总是坦途，是你们无条件的爱与包容，构成了我最坚实的后盾。感谢父母多年来的默默付出，是你们给了我追求梦想的勇气和底气。感谢好友们一路的陪伴与倾听，分享我的喜悦，也分担我的烦忧，在我遇到困难的时候与我站在一起。此外，也感谢那个不曾轻言放弃、始终保持好奇与探索欲的自己。

学术论文和科研成果目录

学术论文

[1] Guo S, Liu Z, Qu Z. QRAP: A Quantum Resource Allocation Platform For Adaptive Scheduling Under Topology Constraints[J]. Software: Practice and Experience, 2026, early accessed. DOI:10.1002/spe.70080.

专利

[1] 第二发明人，“一种具备拜占庭鲁棒性的智能交通量子联邦学习系统”，专利申请号202511545198.6.

[2] 第二发明人，“一种基于强化学习的异构量子云平台任务调度方法”，专利申请号202610405680.8.