# 多智能体强化学习驱动的智股协同决策研究

吴泽龙　廖思晴　杨晓雯　王枝宁*

**韩山师范学院数学与统计学院 广东 潮州 521041**

**【摘　要】：**针对传统股票投资决策方法在高波动性、多因素耦合的金融市场中面临的挑战，提出了一种基于多智能体强化学习（MARL）的智股协同决策引擎。系统采用分层多智能体架构，集成七大功能智能体，涵盖数据采集、多模态分析、协同决策与策略执行全流程。通过融合新闻舆情分析、K 线图像识别、动态风险量化多源信息，构建了包含状态空间、动作空间与分层奖励函数的 MARL 模型，并引入跨模态注意力机制实现异构数据深度融合。实验基于 A 股市场数据，采用集中训练与分布式执行范式进行模型训练。研究表明，多智能体协同机制有效提升了投资决策的收益能力和风险控制水平，为投资决策从经验驱动向数据智能驱动的转型提供了一种技术路径。

**【关键词】：**多智能体；强化学习；智股协同决策；跨模态注意力机制；年化夏普比率

**DOI:10.12417/2982-3382.25.04.014**

## 1 引言

近年来，随着全球金融市场的深度融合与信息技术的飞速发展，股票投资已成为个人与机构资产配置的重要渠道。然而，金融市场固有的高波动性、信息不对称性以及多因素耦合的复杂性，使得传统投资决策方法面临严峻挑战。在此背景下，人工智能技术，特别是强化学习（Reinforcement Learning, RL）与多智能体系统（Multi-Agent System, MAS）的交叉融合，为构建新一代智能投资决策范式提供了革命性的解决方案。多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）通过模拟分布式智能体间的协作、竞争与通信机制，能够有效刻画金融市场中多个参与者交互的复杂动态，为开发具备环境感知、协同分析与自主决策能力的投资引擎奠定了理论基础[1]。本研究旨在系统验证智股协同决策引擎的性能，通过精心构建模型，运用恰当的数据分析方法，并通过实证研究来检验模型的实际效果。

## 2 系统框架与智能体设计

### 2.1 系统整体框架

本研究开发的“智股协同决策引擎”采用了分层式多智能体系统架构，将股票投资决策流程划分为数据采集、分析、决策和执行四个层次。该架构通过明确的职能分工和高效的信息交互机制，实现了股票投资决策的智能化与自动化。系统包含七大核心智能体，各智能体在不同层次中紧密协作，共同完成投资决策任务[2]。具体框架见图 1。

### 2.2 数据采集层

数据采集层负责从多个数据源实时获取股票市场信息，并进行初步处理以确保数据质量。

该层包含两个关键智能体：（1）AI 新闻摘抄员：实时抓取新浪财经、东方财富等财经网站的新闻资讯，利用针对中文金融领域微调过的 RoBERTa-large 模型和规则引擎进行自然语言处理，提取关键事件信息并进行情感分类，生成舆情影响因子[3]。（2）AI 股票查询员：通过对接 Wind、同花顺等金融数据源，采集股价历史数据、公司财报及行业数据。

### 2.3 分析层

分析层对采集到的数据进行深度挖掘，形成多维度投资逻辑模型，为决策层提供科学依据。该层包含两个核心智能体：（1）K 线图分析员（GPT-4o 增强）：负责提取和贡献技术面特征到状态空间，实现股票走势的概率预测和技术面建议输出。（2）AI 谏官：负责监控和贡献风险评估信息到状态空间，实现对市场系统性风险的实时预警和决策风险校验。

### 2.4 决策层

决策层整合分析层输出的多维度信息，生成最终的股票买卖决策指令。该层包含两个核心智能体：（1）AI 领导：作为

作者简介：吴泽龙、廖思晴、杨晓雯均为韩山师范学院数学与统计学院 2023 级数据科学与大数据技术专业在读本科生。

*通讯作者：王枝宁，男，汉族，讲师，主要从事模糊数学优化及统计学研究。

本文得到广东省 2025 年国家级、省级大学生创新创业训练计划项目（粤教高函〔2026〕1 号、项目编号：S202510578022）、2024 年度广东省本科高校教学质量与教学改革工程项目—校企联合实验室：数据科学创新创业实验室（粤教高函〔2024〕30 号）、2025 年度广东省本科高校教学质量与教学改革工程建设项目-面向新工科的 AIGC 赋能《深度学习》智慧金课建设实践研究（粤教高函〔2026〕4 号）的资助。

系统监管者，AI 领导负责统筹智能体协作流程，通过系统内部的分布式消息队列调度数据传输，并动态分配算力资源，同时记录决策日志用于事后分析。（2）AI 巴菲特：作为最终决策制定者，AI 巴菲特基于多智能体输出的基本面指标、技术面信号和舆情评分，运用层次分析法加权融合生成决策指令，并模拟巴菲特价值投资逻辑筛选长期优质标的。

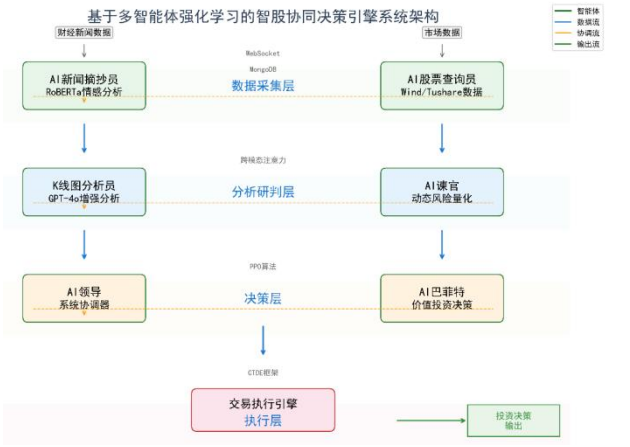


图 1 基于多智能体强化学习的智股协同决策引擎系统架构

### 2.5 执行层

本层负责将决策指令转化为自适应交易操作，通过智能体学习机制优化执行效率。作为学习型智能体，其观测空间包括市场实时流动性、订单簿深度和交易滑点历史数据。动作空间涵盖订单类型选择和仓位分配策略，奖励函数设计为最小化执行成本与跟踪误差。

### 2.6 智能体协同：信息交互决策与冲突高效解决

系统采用 WebSocket 实现智能体间实时通信，定义标准化消息格式，并构建 MongoDB 6.0 分布式数据库作为信息中枢，支持多智能体并发访问。协同决策流程包括数据汇聚、多维度分析、决策融合和执行反馈四个环节，在此过程中，分析层智能体产生的信号、预测和风险指标均通过消息队列传递至决策层，作为决策融合与奖励计算的直接输入，确保决策的科学性和一致性。当智能体出现决策分歧时，系统自动触发三级验证机制：历史案例匹配、蒙特卡洛模拟[4]和专家规则校验，最终由 AI 领导裁定最优方案。

## 3 模型构建与算法设计

为构建一个面向股票投资决策的多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）模型，将前述的系统框架中的多个功能智能体，抽象并映射为 MARL 中的智能体角色：环境感知智能体（对应数据采集层与分析层）、决策智能体（对应决策层）和执行智能体（对应执行层）。模型通过严谨的环境建模、精巧的协作机制设计以及高效的算法框架，实现了多个专业智能体在复杂金融市场环境中的协同感知、决策与优化[5]。

### 3.1 状态空间（State Space）

状态空间 S 是对金融投资环境的完整刻画，包含市场环境状态和投资组合状态两大部分，确保智能体能够基于可观测的环境信息做出决策：

$$S=\left\{S_{market},S_{portfolio}\right\} \tag{1}$$

**（1）市场环境状态 $(S_{market})$**

市场环境状态整合了多源金融市场信息，为智能体提供全面的市场感知能力：

$$S_{market}=\left\{S_{price},S_{news},S_{technical},S_{fundamental},S_{macro}\right\} \tag{2}$$

价格序列特征 $S_{price}$ 包含标的资产过去 N 个交易日的开盘价、最高价、最低价、收盘价和成交量，经过标准化处理：

$$S_{price}=\left\{\frac{P_t-\mu_{price}}{\sigma_{price}}\,|\,t=1,2,\ldots,N\right\} \tag{3}$$

新闻舆情特征 $S_{news}$ 基于实时新闻文本的情感分析和事件提取：

$$S_{news}=\left\{sentiment\_score,event\_vector,urgency\_level\right\} \tag{4}$$

其中，情感得分基于金融领域微调的 RoBERTa 模型计算，事件向量捕捉关键财经事件编码。

技术指标特征 $S_{technical}$ 包含 15 个核心技术指标：

$$S_{technical}=\left\{RSI,MACD,Bollinger\_Bandas,Volume\_MA,\ldots\right\} \tag{5}$$

基本面特征 $S_{fundamental}$，即公司财务和估值指标：

$$S_{fundamental}=\left\{PE,PB,ROE,Debt\_to\_Equity,Revenue\_Growth\right\} \tag{6}$$

宏观环境特征，即宏观经济和市场整体状况：

$$S_{macro}=\{market\_volatility, interest\_rate, economic\_ind ex, industry\_trend\} \quad (7)$$

**（2）投资组合状态 $(S_{portfolio})$**

投资组合状态反映当前持仓情况和历史表现，为风险控制和仓位管理提供依据：

$$S_{portfolio}=\{S_{position}, S_{performance}, S_{constraints}\} \quad (8)$$

持仓状态 $S_{position}$：

$$S_{position}=\{current\_holdings, cash\_balance, position\_concentration\} \quad (9)$$

绩效状态 $S_{performance}$：

$$S_{performance}=\{cumulative\_return, daily\_return, max\_drawdown, win\_rate\} \quad (10)$$

约束状态 $S_{constraints}$：

$$S_{constraints}=\{leverage\_ratio, sector\_exposure, liquidity\_constraints\} \quad (11)$$

**（3）对应关系**

状态空间的设计与前述的系统架构具有明确的映射关系。其中数据采集层：负责获取 $S_{market}$ 中的原始数据。分析层：对原始数据进行特征提取，生成技术指标、情感分析等衍生特征。决策层：基于完整的状态空间 $S$ 进行投资决策。系统内部状态：作为智能体的内部参数，不纳入环境状态空间。

**3.2 多智能体观测空间**

在分布式执行阶段，各个智能体基于局部观测进行决策。

**（1）K 线图分析员观测空间：**

$$O_{chart}=\{S_{price}, S_{technical}, S_{news}\} \quad (12)$$

专注于价格形态和技术面分析。

**（2）AI 谏官观测空间：**

$$O_{risk}=\{S_{portfolio}, S_{market\_volatility}, VaR\_metrics\} \quad (13)$$

专注于风险识别和监控。

**（3）AI 巴菲特观测空间：**

$$O_{decision}=\{S_{market}, S_{portfolio}\} \quad (14)$$

基于完整市场信息进行最终决策。

**3.3 动作空间（Action Space）**

动作空间 $A$ 是智能体所能执行的所有操作的集合。本系统采用分层动作设计，对应不同智能体的职能：

$$A=A_{分配}\bigcup A_{决策}\bigcup A_{生成} \quad (15)$$

（1）算力分配策略 $(A_{分配})$：这是一个由 AI 领导智能体执行的连续或离散动作，其维度等于可用计算资源与待分配任务的组合，用于动态优化系统计算资源的分配效率。

（2）交易决策 $(A_{决策})$：这是由 AI 巴菲特智能体执行的核心动作，定义为一个三维离散动作空间，即 A 决策={买入,卖出,持有}。智能体基于融合后的多维度分析结果，从此空间中选择最优动作。

（3）代码生成模板选择 $(A_{生成})$：由 AI 程序员智能体执行，其动作空间是一个预定义的代码模板库，每个动作对应一个特定的 Python 代码块 ID，用于自动化生成数据预处理、模型训练或交易执行脚本。

**3.4 奖励函数（Reward）**

奖励函数 $R$ 的设计旨在引导智能体学习以实现长期累积收益最大化和风险控制。我们采用分层加权奖励机制，确保智能体在追求收益的同时有效管理风险：

$$R_t=\lambda_1 R_{return,t}+\lambda_2 R_{risk,t}+\lambda_3 R_{cooperation,t} \quad (16)$$

其中，$\lambda_1,\lambda_2,\lambda_3$ 为超参数，用于平衡收益、风险与协作的权重，$t$ 表示时间步（交易日）。

（1）收益奖励 $(R_{return})$：收益奖励基于投资组合的日收益率，鼓励智能体做出能够带来正收益的决策：

$$R_{return,t}=sign(r_t)\cdot|r_t|^{\alpha}\cdot 100 \quad (17)$$

其中，$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ 表示第 $t$ 日的投资组合收益率，$P_t$ 为当日投资组合净值，$P_{t-1}$ 为前一日净值，$\alpha \in [0.5, 1.0]$ 为非线性变换参数，用于平滑极端收益的影响，$sign(\cdot)$ 为符号函数，保持收益方向性，乘以 100 是为了数值稳定性。

（2）风险惩罚 $(R_{risk})$：风险惩罚项旨在控制投资组合的下行风险，由三部分组成：

$$R_{risk,t} = -\left[\beta_1 \cdot I_{VaR} \cdot L_t + \beta_2 \cdot DD_t + \beta_3 \cdot I_{loss} \cdot |r_t|\right] \quad (18)$$

其中，$I_{VaR}$ 为指示函数，当日内最大回撤超过 AI 谏官计算的 VaR 阈值时取 1，否则取 0。$L_t$ 为 VaR 超限的严重程度：

$$L_t = \max(0, (\text{实际损失} - \text{VaR阈值}) / \text{VaR阈值}) \quad (19)$$

$DD_t$ 为当日最大回撤率：

$$DD_t = \frac{P_{t,high} - P_{t,low}}{P_{t,high}} \quad (20)$$

$I_{loss}$ 为指示函数，当日收益率为负时取 1，否则取 0，$r_t$ 为当日收益率，$\beta_1, \beta_2, \beta_3$ 为惩罚系数。

（3）协作奖励 $(R_{cooperation})$：协作奖励促进智能体间的高效信息交换，但权重应显著低于收益和风险项。

$$R_{cooperation,t} = -\eta \cdot (D_{avg,t} - D_{target}) \quad (21)$$

其中，$D_{avg,t}$ 为第 $t$ 日智能体间通信的平均延迟（毫秒），$D_{target}$ 为目标延迟阈值，$\eta$ 为缩放参数。当实际延迟低于目标阈值时，该奖励项为正，鼓励高效协作。

**3.5 跨模态注意力机制（Cross-Modal Attention）**

为使分析层智能体能深度融合视觉、文本和数值等异构数据，我们采用了跨模态注意力机制。该机制的核心在于计算图像特征向量 $s_{cnn}$（查询, Query）与文本特征向量 $s_{bert}$（键, Key）之间的关联权重：

$$\alpha_{i,j} = \frac{\exp(Q_{image}(s_{cnn}^{(i)}) \cdot K_{text}(s_{bert}^{(j)}))}{\sum_k \exp(Q_{image}(s_{cnn}^{(i)}) \cdot K_{text}(s_{bert}^{(k)}))} \quad (22)$$

其中，$Q_{image}(\cdot)$ 和 $K_{image}(\cdot)$ 分别是别是图像查询和文本键的线性变换函数。注意力权重 $\alpha_{i,j}$ 量化了在解读某个 K 线形态视觉特征（$s_{cnn}^{(i)}$）时，一条新闻语义特征（$s_{bert}^{(j)}$）的重要性。最终，加权融合后的特征向量为下游决策提供了更全面、上下文丰富的市场状态表示。

**3.6 动态权重调整机制（Dynamic Weighting）**

市场风格瞬息万变，不同智能体的决策权重也应随之动态调整。我们设计了一个由强化学习策略网络 $f_\theta$ 驱动的动态权重分配机制：

$$\omega = f_\theta\left(s_{market\_representation}\right) \quad (23)$$

其中，市场状态表征 $s_{market\_representation}$ 是一个低维特征向量，由历史波动率、市场情绪指数、宏观经济指标多种基础数据聚合而成。策略网络 $f_\theta$ 以 $s_{market\_representation}$ 为输入，输出一个代表各智能体建议权重的概率分布 。

**3.7 算法框架**

本系统采用集中式训练与分布式执行（Centralized Training with Decentralized Execution, CTDE）的经典 MARL 范式，该框架完美契合本项目中智能体各司其职又需协同的特点[6]。

（1）集中式训练（Centralized Training）：设置一个中央协调器，它在训练阶段可以获取全局状态 $s$ 和所有智能体的动作 $(a_1, \ldots, a_n)$，并据此学习一个全局价值函数

$V(s,a_1,\ldots,a_n)$。该函数用于准确评估联合动作的长期期望回报，从而通过梯度回传指导各个智能体策略网络的更新。

（2）分布式执行（Decentralized Execution）：每个智能体在执行阶段仅依靠自身的局部观察 $s_i$ 和来自协调器的全局指导，独立地执行其策略 $\pi_i(a_i \mid s_i)$。

（3）经验回放与优化：我们采用带优先级的经验回放机制，将异构交互数据分桶存储，并根据其时序差分误差赋予不同采样优先级。策略优化采用近端策略优化（PPO）算法。

## 4 数据分析与实验验证

为系统评估“基于多智能体强化学习的智股协同决策引擎”的实际性能，本研究设计了完整的实验验证方案。

### 4.1 实验环境配置

本研究的实验平台基于 Google Colab Pro 与 Kaggle 云端 GPU 算力，采用 Docker 容器化技术确保环境一致性。系统核心采用 Python 3.9 为开发语言，依托 PyTorch 2.0 深度学习框架构建强化学习模型，并集成 Gymnasium 模拟股票交易环境。智能体间的通信通过 WebSocket 协议实现，数据存储于 MongoDB 6.0 分布式数据库，以满足高并发读写需求。

### 4.2 数据获取与预处理

实验数据来源于多渠道，确保了数据的全面性与真实性。（1）行情与基本面数据：通过 Wind 金融终端及 Tushare 开源库获取 2018 年 1 月至 2024 年 6 月 A 股市场（以上证 50、沪深 300 成分股为主）的日频历史数据。（2）新闻舆情数据：基于新浪财经与东方财富网的公开 API 接口，爬取同期与标的股票相关的新闻标题、正文及评论数据，每日平均处理新闻文本约 5000 条。为应对资源限制，采用 Docker 容器化技术动态分配算力，并使用 TF-IDF 算法过滤低质量文本。（3）宏观与行业数据：从国家统计局及同花顺 iFinD 获取宏观经济指标及行业研报数据。

数据预处理包括：（1）缺失值处理：采用时间序列线性插值法补全缺失数据。（2）标准化：价格序列使用 Z-score 标准化，技术指标采用 min-max 归一化，文本特征通过 BERT 模型编码为 768 维向量。（3）数据集划分：按时间顺序划分训练集（2018.01-2021.12）、验证集（2022.01-2022.12）和测试集（2023.01-2024.06）。

### 4.3 多模态特征提取

将 OHLC 价格序列及成交量转化为 512×512 三通道张量（价格形态/成交量热力/技术指标），利用预训练的 ResNet-18 卷积神经网络提取 128 维空间视觉特征。新闻文本通过微调的 RoBERTa 模型生成 768 维嵌入向量，并基于跨模态注意力机制（公式 6）与视觉特征融合。使用 GRU 网络捕捉成交量序列的长期依赖关系，输出 50 维时序特征向量。最终状态空间整合了视觉、文本和数值特征[7]，形成共计 946 维的市场状态表征。

### 4.4 基于 MARL 的决策模型训练

本研究采用集中式训练-分布式执行架构。使用 PPO 算法[8]，设置折扣因子 $\gamma = 0.99$，学习率 $3\times10^{-4}$，批量大小 1024。奖励函数设置为：

$$R_t = 0.6\times R_{\text{return}} + 0.3\times R_{risk} + 0.1\times R_{cooperation} \quad (24)$$

其中超参数通过网格搜索确定。经过 500 回合训练后，系统奖励曲线收敛（图 2），表明智能体成功学习了协同决策策略。

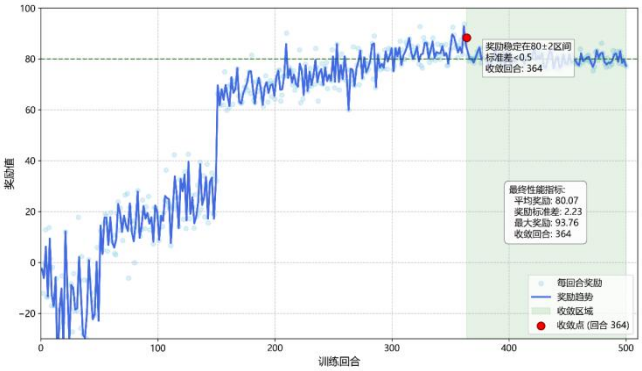


图 2 多智能体强化学习训练奖励曲线

### 4.5 模拟交易与绩效评估

使用 2023 年 1 月至 6 月的数据作为测试集，进行模拟交易回测。初始资金设为 100 万元人民币，单笔交易仓位不超过总资金的 5%，交易成本（佣金和印花税）设为 0.1%。

将本系统（MARL-System）与以下基准策略进行对比：

（1）基准 1（Buy & Hold）：买入并持有沪深 300ETF。

（2）基准 2（Technical Strategy）：基于传统技术指标（MA+RSI）的简单规则策略。

（3）基准 3（Single-Agent RL）：采用相同状态空间但无智能体协作的单一智能体强化学习策略。

此外，为验证系统各组件的贡献，设计了消融实验：（1）

消融模型 1：无跨模态注意力机制（w/o Attention）；（2）消融模型 2：无风险惩罚机制（w/o Risk Penalty）；（3）消融模型 3：无智能体协同（w/o Cooperation）。

回测结果如下表所示：

表 1 模拟交易绩效评估结果（2023.01-2024.06）

| 模型 | 累计收益率 | 年化夏普比率 | 最大回撤 | VaR | 决策一致性 |
|---|---|---|---|---|---|
| MARL-System | 24.5% | 2.8 | -7.2% | -1.8% | 95.3% |
| Single-Agent RL | 15.3% | 1.5 | -12.1% | -2.5% | 88.7% |
| Technical Strategy | 9.8% | 0.9 | -15.4% | -3.1% | 82.4% |
| Buy & Hold | 6.2% | 0.5 | -18.9% | -3.5% | - |
| w/o Attention | 18.2% | 1.9 | -10.3% | -2.3% | 90.1% |
| w/o Risk Penalty | 26.1% | 2.1 | -14.5% | -3.0% | 91.5% |
| w/o Cooperation | 16.8% | 1.7 | -11.8% | -2.4% | 89.2% |

数据分析：

（1）收益能力：本研究提出的 MARL-System 获得 24.5% 的累计收益率和 2.8 的年化夏普比率，显著优于所有基准模型。

（2）风险控制：系统最大回撤仅-7.2%，VaR 值-1.8%，均为最低水平，证实了风险动态量化机制的有效性。消融实验显示，风险惩罚机制对下行风险控制贡献最大。

（3）系统有效性：决策一致性达 95.3%，智能体协同与冲突解决机制确保决策稳定。消融实验中，无注意力机制模型性能下降显著，证明跨模态融合对信息整合至关重要。

（4）各组件贡献排序为：风险惩罚 > 跨模态注意力 > 智能体协同。完整系统在所有指标上均优于消融模型，证实了系统设计的必要性和协同效应。

实验结果表明，所提方法在多源信息融合与协同决策方面展现出潜力，但在极端市场条件下的适应性仍需加强。后续研究可进一步拓展资产类别并优化实时性能。

## 5 结论

本文通过多智能体强化学习技术，成功构建了兼具高性能、低风险与高可解释性的智股协同决策引擎，为金融科技领域的跨学科融合提供了实证范例。系统核心创新在于通过多智能体协同机制与多模态数据分析，为投资决策从经验驱动向数据智能驱动的转型提供了一种技术路径。未来工作将聚焦于算法轻量化、市场适应性拓展及合规性提升，推动智能决策技术在更广泛金融场景中的落地应用。

## 参考文献：

[1] 杜威,丁世飞.多智能体强化学习综述[J].计算机科学,2019,46(08):1-8.

[2] 殷昌盛,杨若鹏,朱巍,等.多智能体分层强化学习综述[J].智能系统学报,2020,15(04):646-655.

[3] 齐甜方,蒋洪迅. 基于 Seq2Seq 文本摘要和情感挖掘的股票波动趋势预测[J].管理评论,2021,33(05):257-269.

[4] 李承奥.基于机器强化学习与蒙特卡洛树的基本原理及其应用[J].通讯世界,2019,26(02):212-213.

[5] 黎麟玉.基于强化学习的股票自动交易策略研究[D].哈尔滨工业大学,2024.

[6] 邹启杰,蒋亚军,高兵,等.协作多智能体深度强化学习研究综述[J].航空兵器,2022,29(06):78-88.

[7] 任泽裕,王振超,柯尊旺,等.多模态数据融合综述[J].计算机工程与应用,2021,57(18):49-64.

[8] 刘一鸣.基于奖励设计的深度强化学习算法研究与应用[D].北京邮电大学,2020.