Computer Science and Application 计算机科学与应用, 2025, 15(11), 1-9
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511277

Hans汉斯

基于深度强化学习的智能体股票投资组合自动交易模型研究

魏晓光，丁雅芳

河北金融学院河北省金融科技应用重点实验室，河北 保定

收稿日期：2025年9月29日；录用日期：2025年10月26日；发布日期：2025年11月3日

摘 要

随着金融科技的深度渗透，量化投资已成为平衡股票市场风险与收益的核心手段。传统交易模型常面临“预测与决策脱节”的痛点，而深度强化学习(DRL)凭借“智能体与环境实时交互、动态优化策略”的核心特性，为股票投资组合自动化管理提供了新路径。本文以沪深300成分股为研究对象，整合技术指标、文本情感与宏观经济三类数据，构建了A2C (Advantage Actor-Critic)、DDPG (Deep Deterministic Policy Gradient)、PPO (Proximal Policy Optimization)与TD3 (Twin Delayed DDPG)四种DRL算法智能体，并对其性能进行系统对比。通过引入动态风险厌恶系数优化奖励函数，且模拟印花税、佣金等真实交易成本，最终实现了投资组合智能化与自动化管理的高效落地。实证结果表明，PPO算法在测试集上表现最优，其风险收益平衡能力显著优于其他对比DRL算法及传统基线模型。

关键词

深度强化学习，智能体，股票投资

Research on the Automatic Trading Model of Stock Investment Portfolios Based on Deep Reinforcement Learning Agents

Xiaoguang Wei, Yafang Ding

Hebei Provincial Key Laboratory of Financial Technology Application, Hebei Finance University, Baoding Hebei

Received: September 29, 2025; accepted: October 26, 2025; published: November 3, 2025

Abstract

With the deep integration of financial technology, quantitative investment has become a core

文章引用: 魏晓光, 丁雅芳. 基于深度强化学习的智能体股票投资组合自动交易模型研究[J]. 计算机科学与应用, 2025, 15(11): 1-9. DOI: 10.12677/csa.2025.1511277

approach to balancing risks and returns in the stock market. Traditional trading models often suffer from the issue of "disconnection between prediction and decision-making." In contrast, Deep Reinforcement Learning (DRL), with its core capability of "enabling agents to interact with the environment in real-time and dynamically optimize strategies", offers a new pathway for the automated management of stock portfolios. This study focuses on the constituent stocks of the CSI 300 Index, integrating three types of data—technical indicators, textual sentiment, and macroeconomic data—to construct four DRL-based agent models: A2C (Advantage Actor-Critic), DDPG (Deep Deterministic Policy Gradient), PPO (Proximal Policy Optimization), and TD3 (Twin Delayed DDPG), and systematically compares their performance. By incorporating a dynamic risk aversion coefficient to optimize the reward function and simulating real-world transaction costs such as stamp duty and commissions, the study achieves an intelligent and automated portfolio management system. Empirical results demonstrate that the PPO algorithm performs best on the test set, with its risk-return balancing capability significantly outperforming other compared DRL algorithms and traditional baseline models.

Keywords

Deep Reinforcement Learning, Agent, Stock Investment

1. 引言

中国 A 股市场自成立以来，已发展为全球第二大资本市场。截至 2024 年底，我国 A 股总市值已突破 100 万亿元，上市公司数量在全球股票市场中位居榜首，投资者数量超 2 亿[1]。然而，市场中个人投资者群体普遍存在“追涨杀跌”的非理性行为，且受限于信息获取能力，难以及时捕捉市场中的利好或利空信号，在与机构投资者的博弈中处于弱势地位[2]。与之相对，机构投资者凭借量化交易工具实现了“去人为干扰、高执行效率”的优势，AI 驱动的自动交易占比逐年提升，国内头部券商的量化交易规模年快速增速，量化投资已成为资本市场的核心发展趋势[3]。

深度强化学习(DRL)区别于传统机器学习“分阶段预测股价–再制定交易策略”的模式，其核心逻辑是让智能体在与市场环境的实时交互中，以“长期收益最大化为目标”动态调整策略，这一特性恰好契合股票市场“非平稳性、高波动性”的本质。然而，现有 DRL 交易模型在实际应用中仍面临三大核心问题：第一，数据维度单一。多数研究仅依赖开盘价、收盘价等技术指标构建模型，忽略了文本情感与宏观经济指标对股价的影响[4]；第二，算法适应性不足。现有研究多集中于 A2C、DDPG 等经典 DRL 算法，这类算法在平稳市场中表现尚可，但在复杂市场场景中，易出现策略震荡或收益骤降的问题；第三，场景简化脱离实际。多数回测未充分模拟真实交易成本，也未验证模型在极端市场中的抗风险能力，导致回测结果偏乐观[5]，实际应用时收益大幅缩水。

基于此，本文的核心研究目标是构建“多特征输入、多算法对比、多场景验证”的 DRL 智能体模型，解决上述三大问题，为股票投资组合的自动化管理提供兼具收益性与稳健性的可行方案。

2. 研究内容与技术路线

本研究围绕深度强化学习展开智能体股票投资组合自动交易模型研究，形成“问题–方案”的对应逻辑：

首先是数据与环境优化，针对性解决“数据维度单一”与“场景简化”问题。我们选取沪深 300 成分股中的 66 只股票作为实验数据样本(涵盖能源、原材料、工业、可选消费、主要消费、医药卫生、金融、信息技术、通信服务、公用事业、房地产，共 11 个一级行业，每个行业 6 只代表性股票)，确保样本的行业代表性；同时整合三类数据——技术指标、文本情感、宏观经济，构建多维度输入特征；此外，在回测中加入印花税、佣金等交易成本，完全模拟真实交易流程。

其次是多算法 DRL 智能体开发，解决“算法适应性不足”问题。在 A2C、DDPG 两种经典算法基础上，我们引入 PPO 与 TD3 两种改进算法：PPO 通过“截断损失函数”限制策略更新幅度，缓解 A2C 的策略震荡问题；TD3 通过“双 Critic 网络 + 延迟策略更新”，解决 DDPG 的 Q 值过估计偏差。通过对四类算法的对比分析，明确不同算法在股票投资组合场景中的适用范围与优势。

最后是实证验证与效果分析。我们以 2024 年的最新市场数据为测试集，构建“收益－风险－综合”三维回测指标体系，验证 DRL 智能体的表现；同时选取传统均值方差模型与沪深 300 指数作为基线，验证 DRL 模型的优势是否具有显著性。

本文的技术路线可概括为四个连贯步骤：先通过 TusharePro、东方财富网等渠道获取多源数据，经清洗、特征工程与标准化处理后，划分训练集与测试集；再设计强化学习交易环境，明确状态、动作、奖励的定义；随后基于多层感知机(MLP)构建四类深度强化学习 DRL 智能体，用训练集完成模型迭代；最后在测试集上回测，对比不同智能体的表现，并与基线模型验证，最终输出研究结论与改进方向。

3. 国内外研究综述与评述

股票投资组合的核心目标是在控制风险的同时追求收益最大化，传统方法多依赖静态优化模型，需假设市场环境稳定且数据符合特定分布，难以应对股市的动态性与不确定性。随着人工智能技术发展，深度强化学习(DRL)凭借“智能体与环境交互试错、逐步学习最优策略”的特性，成为突破传统局限的关键技术[6]。DRL 的核心逻辑是让智能体在连续的交易决策中，通过对市场状态的感知，调整投资组合的资产配置比例，最终形成适配市场变化的自动交易策略，这一逻辑与股票投资组合“动态调整、风险收益平衡”的需求高度契合。

现有研究中，DRL 与股票投资组合的结合主要围绕智能体的策略学习展开。早期模型多基于简单 DRL 算法，重点解决单一资产或少量资产的交易决策问题，通过将“买入、卖出、持有”等离散动作与资产收益关联，让智能体学习基础的仓位调整规则。随着研究深入，模型逐渐向多资产组合扩展，引入一些更适用于连续动作空间的算法，可实现对多只股票持仓比例的精细化调整，同时整合多源信息以提升状态感知的全面性[7]。此外，研究还关注实际交易中的约束条件融入。例如，在模型训练中加入交易成本的影响，避免智能体因频繁交易导致收益损耗；引入风险控制机制，通过设定最大回撤阈值、资产分散度要求等，让策略更贴合实际投资场景。部分研究还尝试结合价值函数与策略梯度的双重优化，既让智能体明确“当前策略的预期收益”，又能持续优化动作选择，进一步提升策略的稳定性。

当前模型仍存在两方面核心问题：第一，算法对比与适配性分析不充分。现有研究多聚焦单一 DRL 算法的改进，缺乏对不同算法在股票投资组合场景中的系统性对比。这种单一算法研究的局限在于，无法明确不同算法对市场波动的适应性差异；适用于极端波动的算法，有可能在平稳期因保守错失收益。投资者难以据此选择适配自身需求的算法。为解决这一问题，本文同时构建 A2C、DDPG、PPO 与 TD3 四类智能体，通过量化对比它们的核心机制、超参数设置及测试集表现，明确不同算法的适配场景，为投资者提供清晰的算法选择依据。第二，场景模拟与真实交易脱节。多数研究在回测中简化了交易场景，导致部分在平稳期表现好的模型，在极端波动时可能大幅回撤。对此，本文在回测中完整模拟交易成本；同时设计动态风险厌恶系数，市场波动大时增大系数以强化风险惩罚，市场平稳时减小系数以鼓励收益

探索，让模型在不同场景下均能平衡风险与收益。

4. 多算法 DRL 智能体模型构建

4.1. 实验数据与预处理

4.1.1. 数据来源与样本选择

为确保数据的代表性、时效性与可靠性，本文选取三类公开数据，具体如下：

第一类是股票基础数据，来源于 Tushare Pro 平台，覆盖 2016 年 1 月 1 日至 2024 年 12 月 31 日的 66 只沪深 300 成分股。样本选择遵循“行业分散、流动性高”的原则，涵盖 11 个一级行业，避免行业集中导致的风险偏高。样本剔除了 2016~2024 年期间退市、调出沪深 300 指数的股票，最终保留 66 只连续交易的成分股，保证数据连续性。数据字段涵盖开盘价、收盘价、最高价、最低价、成交量、换手率，累计获得 142,565 条有效记录。

第二类是文本情感数据，来源于东方财富网“公司新闻”栏目，共收集 2016~2024 年的 108,963 条新闻文本。我们通过 BERT 预训练模型提取文本情感得分，得分范围为[−1, 1]：−1 代表“极度负面”(如公司业绩暴雷、监管处罚)，1 代表“极度正面”(如业绩超预期、重大合同签订)，0 代表“中性”。情感得分经 Z-Score 标准化(转化为均值 0、标准差 1 的分布)后，作为模型的输入特征之一。

第三类是宏观经济数据，来源于国家统计局与中国人民银行官网，包括月度 CPI (居民消费价格指数)、季度 GDP 增速、日度 10 年期国债收益率。由于 CPI 与 GDP 增速为低频数据(月度/季度)，需通过线性插值转化为日频数据，与股票数据的时间维度对齐，避免因频率不一致导致的输入偏差。

4.1.2. 多维度特征工程与标准化

特征工程是提升 DRL 智能体市场感知能力的核心环节，本文共构建 28 维特征，按类型分为三类，每类特征的选择均以“精准捕捉市场趋势、情绪或风险信号”为目标，分别是技术指标(9 维)、文本情感特征(1 维)、宏观与风险特征(18 维)，详情如表 1 所示。

Table 1. Results of multidimensional feature engineering
表 1. 多维度特征工程结果

特征类别	维度	特征名称	计算方式
技术指标	9	MACD、布林带上轨(boll_ub)、布林带下轨(boll_lb)、RSI_30、CCI_30、DX_30、close_30_sma、close_60_sma、动荡指数	TA-Lib 库计算；动荡指数以沪深 300 为基准，衡量个股相对波动
文本情感	1	新闻情感得分	BERT 提取后 Z-Score 标准化，取值[−1, 1]
宏观与风险	18	CPI、GDP 增速、10 年期国债收益率、组合波动率、11 个行业权重、3 组个股收益率均值	宏观指标线性插值至日频；组合波动率基于股票协方差矩阵；行业权重对应 11 个一级行业

第一类是技术指标(9 维)，基于 TA-Lib 库计算，包括 MACD (指数平滑异同移动平均线，用于判断买卖信号)、布林带上轨(boll_ub)与下轨(boll_lb，识别股价超买超卖)、RSI_30 (30 日相对强弱指数，衡量价格变动幅度)、CCI_30 (30 日顺势指标，判断趋势方向)、DX_30 (30 日动向指标，评估趋势强度)、30 日均线(close_30_sma)与 60 日均线(close_60_sma，判断价格偏离长期均值的程度)、动荡指数(以沪深 300 指数为基准，衡量个股相对市场的波动程度)。

第二类是文本情感特征(1 维)，即 BERT 标准化情感得分，用于捕捉市场情绪对股价的短期影响。

第三类是宏观与风险特征(18 维)，包括 3 维宏观指标(CPI、GDP 增速、10 年期国债收益率)、1 维组合波动率(基于 66 只股票的日收益率协方差矩阵计算，协方差矩阵越大，组合风险越强)、11 维行业权重(对应 11 个一级行业，反映组合的行业配置分散度)、3 维个股收益率分组均值(将 66 只股票按近 5 日收益率分为高、中、低三组，计算每组均值，反映组合内个股的收益分布)。

在特征预处理阶段，我们主要解决三类问题：

一是缺失值处理：单日缺失数据采用“前向填充”，即沿用前一交易日的特征值；连续 3 日以上的缺失数据采用“行业均值填充”，避免直接删除导致的数据损失。

二是异常值处理：通过 3σ 原则识别与修正异常值——首先计算每个特征的均值(μ)与标准差(σ)，将超出“$\mu-3\sigma$”至“$\mu+3\sigma$”范围的值定义为异常值；其中，上限异常值($>\mu+3\sigma$)用“$\mu+3\sigma$”替换，下限异常值($<\mu-3\sigma$)用“$\mu-3\sigma$”替换，降低极端值对模型训练的干扰。

三是数据标准化：由于不同特征的量级差异较大，需进行标准化处理：技术指标、文本情感、宏观指标采用 Z-Score 标准化，确保均值为 0、标准差为 1；行业权重、收益率分组均值等比例类特征采用 Min-Max 标准化，压缩至[0, 1]范围，避免量级差异影响模型参数更新。

最后，我们按时间维度划分数据集：2016 年 1 月 1 日至 2023 年 12 月 31 日为训练集，用于模型参数迭代；2024 年 1 月 1 日至 2024 年 12 月 31 日为测试集，用于验证模型在最新市场中的泛化能力。

4.2. 强化学习交易环境设计

强化学习环境的核心是定义“状态–动作–奖励”三要素，这直接决定智能体的学习方向与策略效果。本文结合股票投资组合的实际需求，设计了贴合真实交易场景的环境要素。

4.2.1. 状态空间与动作空间

状态空间的设计目标，是让智能体全面感知“市场环境 + 投资组合状态”。我们将构建的 28 维特征直接作为状态向量，涵盖技术指标、文本情感、宏观指标与组合风险特征。状态向量的更新逻辑为“每日收盘后更新”——每个交易日收盘后，根据当日的股票数据、新闻文本与宏观数据计算 28 维特征，作为智能体次日制定交易策略的输入，确保状态与实际交易节奏一致。

动作空间的设计需满足股票投资组合的核心约束：一是“无卖空”(符合多数散户的交易习惯)，二是“资金全部分配”(不考虑现金持有，聚焦股票组合优化)。因此，我们将动作空间定义为 66 维连续空间，每个维度对应 1 只股票的资金权重，满足“单只股票权重 $\in[0,1]$、66 只股票权重和为 1”的约束。

动作的具体执行逻辑分为三步：第一步是初始权重设置：训练集与测试集的初始时刻(分别为 2016 年 1 月 1 日、2024 年 1 月 1 日)，66 只股票的权重均等，初始资金设定为 100 万元(模拟散户常见的初始投资规模)。第二步是资金调整计算：智能体每日输出新权重后，根据“当日总资金 × 新权重 − 前日总资金 × 旧权重”，计算单只股票的增资或撤资金额——若结果为正，说明需对该股票增资；若为负，则需撤资。第三步是股票数量调整与成本扣除：根据当日股票收盘价，计算需增持或减持的股票数量；同时，从当日总资金中扣除交易成本——卖出股票时征收 0.1%的印花税，买卖双向征收 0.02%的佣金，确保动作执行贴合真实交易流程。

4.2.2. 动态风险厌恶系数的奖励函数优化

奖励函数是智能体学习的“指挥棒”，需同时兼顾“收益最大化”与“风险最小化”。现有研究多采用固定风险厌恶系数，但股票市场的波动程度随时间变化，固定系数会导致风险惩罚失准：市场平稳时可能过度惩罚风险，错失收益机会；市场波动大时又惩罚不足，导致回撤扩大。为此，本文提出“动态风险厌恶系数”，根据当日市场动荡指数与近 30 日平均动荡指数的比值调整，公式定义为：

$$\lambda_t \begin{cases} 1, if \dfrac{V_t}{\overline{V}_{t-30:t-1}} < 1 \\ \dfrac{V_t}{\overline{V}_{t-30:t-1}}, if 1 \le \dfrac{V_t}{\overline{V}_{t-30:t-1}} \le 5 \\ 5, if \dfrac{V_t}{\overline{V}_{t-30:t-1}} > 5 \end{cases} \tag{1}$$

其中，V_t 为第 t 日的市场动荡指数，$\overline{V}_{t-30:t-1}$ 为第 $t-30$ 日至第 $t-1$ 日的平均动荡指数。该系数取值范围被限制在[1, 5]，核心逻辑为：市场波动越剧烈($V_t/\overline{V}_t$ 越大)，系数越大，风险惩罚越强；市场越平稳，系数越小，鼓励适度收益探索。这样设计的核心目的是：当市场出现大幅波动时，系数增大以强化风险惩罚，避免智能体过度投机；当市场平稳时，系数减小以鼓励适度探索，捕捉更多收益机会。

基于动态风险厌恶系数，我们将奖励函数定义为“对数收益率减去动态风险惩罚”。其中，对数收益率能避免收益的非线性偏差，更适合长期收益计算；动态风险惩罚则通过“系数 × 组合方差”实现，确保智能体在追求收益的同时，不会忽视风险控制。动态风险厌恶奖励函数定义公式为：

$$r_t = \ln\left(\frac{A_t}{A_{t-1}}\right) - \lambda_t \times \sigma_{t-1}^2 \tag{2}$$

其中，A_t 为第 t 日的总资产，$\ln\left(\dfrac{A_t}{A_{t-1}}\right)$ 为第 t 日的对数收益率(避免收益非线性偏差)；σ_{t-1}^2 为第 $t-1$ 日的组合方差(反映组合风险)；$\lambda_t \times \sigma_{t-1}^2$ 为动态风险惩罚项，确保智能体在追求收益的同时，不会忽视风险控制。

4.3. 多算法 DRL 智能体实现

4.3.1. 基础网络参数设置

所有智能体均采用多层感知机(MLP)作为 Actor/Critic 网络，因 MLP 对高维特征拟合能力强，且超参数少，避免与 DRL 算法干扰，参数设置如表 2：

Table 2. Basic network parameter settings
表 2. 基础网络参数设置

网络类型	输入维度	隐藏层配置	激活函数	输出维度
Actor 网络	28	2 层，每层 160 个节点	隐藏层：ReLU	66
Critic 网络	94	2 层，每层 160 个节点	隐藏层：ReLU	1

Actor 网络：输入维度为 28 (对应 28 维状态特征)，隐藏层为 2 层、每层 160 个节点，激活函数选用 ReLU；输出维度为 66 (对应 66 只股票的权重)，输出层激活函数选用 Sigmoid，并通过“输出值/总和”的归一化处理，确保 66 只股票权重和为 1，满足动作空间约束。

Critic 网络：输入维度为 94 (28 维状态 + 66 维动作)，隐藏层与 Actor 网络一致(2 层、每层 160 个节点，ReLU 激活)；输出维度为 1 (输出该状态 − 动作对应的 Q 值，即预期长期收益)，输出层无需激活函数(Q 值可为任意实数，无需范围限制)。

4.3.2. 四类 DRL 算法的核心差异与参数设置

A2C、DDPG、PPO、TD3 四类 DRL 算法的核心差异在于“策略更新方式”与“风险控制机制”，这直接决定其在股票市场中的适应性。本文聚焦每类算法的关键改进、超参数设置与适用场景，通过量

化对比突出优势，具体如表 3 所示(超参数通过网格搜索法优化确定)。

Table 3. Comparison of core parameters and characteristics of four types of DRL algorithms
表 3. 四类 DRL 算法核心参数与特性对比

核心参数	A2C	DDPG	PPO	TD3
核心改进机制	策略梯度 + 价值函数	经验回放池 + 目标网络	截断损失函数	双 Critic 网络 + 延迟策略更新
关键超参数设置	学习率：0.0002； 折扣因子：0.95； 熵系数：0.01	学习率：0.001； 经验回放池大小：5 万条； 批次大小：128； 软件更新系数：0.005	学习率：0.0003； 折扣因子：0.95； 优势估计系数：0.9； 截断范围：[−0.2, 0.2]	学习率：0.001； 经验回收池大小：10 万条； 批次大小：256； 延迟更新步数：2
优势	训练速度快，在线学习效率高	支持连续动作，风险控制优于 A2C	策略更新稳定，平衡收益与风险	解决 Q 值过度估计，极端风险控制稳健
局限	无经验回放，对市场波动适应性弱	Q 值过估计，易高估预期收益)	极端波动下收益略低于 TD3	策略更新滞后，市场转折时响应慢
适用场景	平稳市场	温和震荡市场	多阶段市场	极端波动市场

A2C 算法的核心逻辑是“策略梯度 + 价值函数”的结合，通过“优势函数”减少策略更新的方差，实现高效的在线学习。其关键超参数为：Actor 与 Critic 的学习率均为 0.0002，折扣因子 0.95，熵系数 0.01。A2C 的优势是训练速度快，但缺点是无经验回放机制，对市场波动的适应性弱，仅适合平稳市场场景。

DDPG 算法针对 A2C 在连续动作空间中的不足，引入“经验回放池”与“目标网络”：经验回放池存储历史交互数据，每次训练随机采样，缓解样本相关性问题；目标网络固定目标 Q 值的计算基准，避免参数频繁更新导致的震荡。其关键超参数为：学习率 0.001，经验回放池大小 5 万条数据，批次大小 128，软更新系数 0.005。DDPG 的风险控制能力优于 A2C，但存在 Q 值过估计问题。

PPO 算法在 A2C 基础上，引入“截断损失函数”这一核心改进，通过限制策略更新幅度，解决 A2C 策略震荡的问题。其关键超参数为：学习率 0.0003，折扣因子 0.95，优势估计系数 0.9，截断范围 0.2。PPO 的优势在于平衡了收益与稳定性。

TD3 算法针对 DDPG 的 Q 值过估计问题，提出两项改进：一是“双 Critic 网络”，同时训练两个网络，取最小值作为 Q 值评估结果；二是“延迟策略更新”，每更新 2 次 Critic 网络，再更新 1 次 Actor 网络。其关键超参数为：学习率 0.001，经验回放池大小 10 万条数据，批次大小 256，延迟更新步数 2。TD3 的风险控制与 DDPG 持平，但因延迟更新导致策略调整滞后，更适合风险厌恶型投资者。

5. 实证分析与模型验证

5.1. 回测指标体系

为全面评估智能体的“收益能力”“风险控制能力”与“风险收益平衡能力”，本文构建“收益－风险－综合”三维回测指标体系，核心指标的定义如下：

收益指标：包括年化收益率与累计收益率。年化收益率将测试期收益按年折算，反映长期收益能力；累计收益率则是测试期总收益与初始资金的比例，反映短期收益水平。

风险指标：包括最大回撤率与年化波动率。最大回撤率是“资金峰值－资金谷值”与峰值的比例，反映极端风险承受能力；年化波动率是日收益率标准差的年化值，反映收益波动程度。

综合指标：主要是夏普比率，计算方式为“(年化收益 − 无风险利率)/年化波动率”，衡量“每承受 1 单位风险获得的超额收益”，夏普比率越高，风险收益比越优。

5.2. 单智能体回测

回测基于 2024 年测试集(初始资金 100 万元)，核心结果如表 4 所示，直接验证“算法适应性”问题的解决效果：

Table 4. Key backtesting results of four types of DRL agents
表 4. 四类 DRL 智能体回测关键结果

指标	A2C 智能体	DDPG 智能体	PPO 智能体	TD3 智能体
年化收益率	13.8%	14.9%	17.1%	15.2%
累计收益率	13.8%	14.9%	17.1%	15.2%
最大回撤率	−5.5%	−4.8%	−5.2%	−4.9%
年化波动率	17.8%	16.2%	17.5%	16.0%
夏普比率	1.45	1.73	1.84	1.80
期末资金(万元)	113.8	114.9	117.1	115.2

通过表 4 数据，我们发现：(1) PPO 在多阶段适配与收益平衡上最优：年化收益比 TD3 高 1.9 个百分点，夏普比率比 A2C 高 0.39，虽最大回撤比 DDPG 高 0.4 个百分点，但全年收益稳定性更强；(2) TD3 风险控制表现稳健：全年最大回撤−4.9%，但延迟更新导致策略对市场变化的响应滞后，收益略低于 PPO；(3) A2C 表现最弱：因无经验回放机制，难以适配多阶段波动市场，年化收益与夏普比率均为四类算法中最低。

5.3. 与基线模型对比验证

本文针对表现较好的 PPO 与 TD3 模型，进一步选取两类传统模型为基线进行对比验证：均值方差模型(静态策略)、沪深 300 指数(市场基准)，2024 年全年对比结果如表 5 所示。

Table 5. Backtest comparison between DRL agents and baseline models
表 5. DRL 智能体与基线模型回测对比

指标	PPO 智能体	TD3 智能体	均值方差模型	沪深 300 指数
年化收益率	17.1%	15.2%	14.3%	10.8%
累计收益率	17.1%	15.2%	14.3%	10.8%
最大回撤率	−5.2%	−4.9%	−8.8%	−10.2%
年化波动率	17.5%	16.0%	24.0%	23.5%
夏普比率	1.84	1.80	1.30	1.05
期末资金(万元)	117.1	115.2	114.3	110.8

上表数据显示 DRL 智能体全面超越传统模型：(1) 收益能力：PPO 年化收益是均值方差模型的 1.20 倍、沪深 300 指数的 1.58 倍；(2) 风险控制能力：PPO 最大回撤仅为沪深 300 指数的 51%，年化波动率比均值方差模型低 27.1%；(3) 风险收益比：PPO 夏普比率是均值方差模型的 1.42 倍、沪深 300 指数的 1.75 倍。尤其在多阶段市场中，DRL 的动态策略优势显著，解决了传统模型“静态策略、风险控制弱”的痛点。

6. 结论与展望

6.1. 核心研究结论

本研究得出如下结论：

第一，PPO 算法是全年多阶段市场的最优 DRL 选择。实证结果表明，PPO 凭借“截断损失函数 + 熵正则化”，实现 17.1%年化收益、1.84 夏普比率、−5.2%最大回撤，适配性显著优于 A2C/DDPG/TD3，为投资者提供明确算法选择依据。

第二，多特征融合与动态系数提升全年实用性。整合技术、文本、宏观三类特征后，模型对政策信号的感知提前 1~2 个交易日；动态风险厌恶系数使全年最大回撤比固定系数模型显著降低；交易成本模拟让回测收益更真实。

第三，DRL 全面超越传统模型且适配全年行情：PPO 年化收益是均值方差模型的 1.20 倍，最大回撤显著低于沪深 300 指数，为散户与中小机构提供对抗全年多阶段市场非理性的工具。

6.2. 实践启示与未来改进方向

对个人投资者而言，可基于本文模型框架优先选择 PPO 智能体或 TD3 智能体，实际应用中需结合自身风险承受能力调整核心参数，在极端市场阶段需辅以手动干预，避免模型因短期极端波动出现策略偏差；对机构投资者来说，可在现有模型基础上进一步扩展样本覆盖范围，将创业板、科创板股票纳入研究以提升市场代表性，同时引入龙虎榜资金流向、机构持仓变化等另类数据丰富特征维度，并通过对接实时行情系统实现“策略生成–订单执行”的全自动化流程，从而提升全年交易效率与策略落地速度。

未来可从三方面优化模型：一是构建多智能体协作架构，设计“行业智能体 + 个股智能体 + 风控智能体”的三层体系，其中行业智能体负责跨行业权重动态配置，个股智能体聚焦行业内优质个股筛选，风控智能体实时监控组合风险敞口，以此适配跨市场、跨资产的全年资产配置需求；二是推动大语言模型(LLM)与 DRL 的深度融合，利用 GPT-4 等工具解析央行货币政策报告、行业监管文件等文本信息，提取“利率变动预期”“产业扶持信号”等前瞻性特征作为模型输入，解决当前宏观经济数据滞后的问题，进而提升后续市场的预判能力；三是加强全年极端场景的对抗训练，通过对抗训练强化智能体的抗风险能力，同时引入定期压力测试机制，持续验证模型在极端场景下的最大回撤与收益损失情况，确保模型在实际应用中的全年稳健性。

基金项目

2025 年度河北省金融科技应用重点实验室课题《基于大语言模型的用户画像与金融营销场景应用技术研究》(课题编号：2025002)。

参考文献

[1] 白彦锋, 袁贵博. 健全我国直接税体系的系统性对策研究[J]. 税务研究, 2025(9): 22-29.

[2] 董文昊, 刘春林. CEO 强制变更对盈余管理的溢出效应[J]. 经济与管理研究, 2025, 46(3): 128-144.

[3] 陆蓉, 张瑞瑞, 闵思凯. 量化交易的市场价值效应——信息优势的作用[J]. 管理世界, 2025, 41(6): 55-97, 157.

[4] 师应来, 訾轩. 基于投资者情绪的动态 Copula-小波 SVR 模型构建与应用[J]. 统计与决策, 2024, 40(16): 140-145.

[5] 王建. 量化交易与黑天鹅现象漫谈: 应对不确定性的策略[J]. 中国信用卡, 2025(7): 79-84.

[6] 林升, 綦科, 魏楷聪, 等. 机器学习在股价预测中的研究综述[J]. 经济师, 2019(3): 71-73, 78.

[7] 张倩玉, 严冬梅, 韩佳彤. 结合深度学习和分解算法的股票价格预测研究[J]. 计算机工程与应用, 2021, 57(5): 56-64.