Proceedings of Business and Economic Studies, 2026, Volume 9, Issue 8
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Multi-Agent Reinforcement Learning for Dynamic Portfolio Allocation

Wanqing Feng

Rice Management Company, Houston, TX, USA

Abstract: The old portfolio model cannot be used to cope with the non-stationarity of the market, changes in asset correlation, and the pursuit of the risk-return trade-off. The problems above will be solved by applying dynamic portfolio allocation based on multi-agent reinforcement learning in this paper. The three special agents in the hierarchical heterogeneous multi-agent architecture are market perception, return prediction, and risk control; a centralized training with decentralized execution mode is employed, a graph attention network is used to obtain real-time inter-asset relationships, and thus dynamic iterative optimization of allocation weights and rebalancing strategies is achieved. Based on backtesting with CSI 300 constituents, 10-year treasury bond futures and the Nanhua Commodity Index, it was found that the proposed method can reduce the maximum drawdown and improve both the Sharpe ratio and excess return considerably compared to the traditional mean-variance model and the single-agent model; it has good market adaptability and risk resistance, and is thus a good and practical technical solution for the dynamic optimization of quantitative portfolios.

Keywords: Multi-agent reinforcement learning; Dynamic portfolio; Asset allocation; Risk control; Quantitative investment

Online publication: August 31, 2026

1. Introduction

The financial market has high volatility, is non-linear, and changes over time. Dynamic portfolio allocation is a problem that quantitative investment and asset management aim to solve in order to maximize the return at a given level of risk and achieve continuous growth of assets over time. The typical allocation methods for portfolios are the mean-variance model and the capital asset pricing model; both find the best assets at a particular time. The above models generally assume market efficiency and normal distribution of asset returns; they are far from the actual situation in the market [1]. At the same time, single-agent reinforcement learning models for investment can only optimize one objective and thus cannot achieve all the aims, such as analyzing market trends, forecasting returns, and controlling risks. They are not very flexible, perform poorly in an economic downturn, are not generalizable, and do not model multi-asset links properly. Although some studies have begun to investigate the application of multi-agent reinforcement learning in portfolio optimization, most have concentrated on single-stock markets and have not fully incorporated the cross-asset class dynamics of bonds and commodities; moreover, few have specifically applied such methods to the volatile and diverse Chinese A-share market with multi-asset allocation. To address the problems mentioned above, this paper puts forward a multi-agent reinforcement learning method to distribute the decision of complex investment among various cooperative agents and build a dynamic allocation scheme that can respond to changes in the market [2]. Thus, the deficiency of the former method is rectified, and repeated modification of the portfolio weights can be carried out to achieve intelligent asset allocation in the current volatile financial market.

2 Architecture design of the multi-agent reinforcement learning dynamic allocation model

2.1. Overall hierarchical framework

According to the operating characteristics of financial markets and the thinking logic of reinforcement learning, this paper proposes a hierarchical heterogeneous multi-agent reinforcement learning (MARL) portfolio allocation model. It is no longer a single-agent decision-making system; many agents have been added, and now cooperation, gaming, and global optimization are employed to distribute assets dynamically. The three main parts of the whole are the environment perception module, the agent decision-making unit, and the strategy execution and optimization component. The above framework is more detailed, more flexible, and can operate independently of the old way.

2.2. Environment perception layer: Market data and time-varying correlation modeling

The environment perception layer is the data-input part of the whole model, and it collects many kinds of market data to build a regularized market-state space [3]. At this level, the available data include asset prices, trading volumes, volatilities, macroeconomic indicators, and market sentiment, etc. The following are the preprocessing steps: Normalization, Outlier Removal, and Temporal Smoothing. At the same time, this paper also presents a graph attention network to model the connections among many assets more precisely and track changes in the correlations and transmission paths of these assets over time. It will solve the problem of the lack of dynamic correlation in the traditional model and provide all-around reliable data on the market for the next stage of precise agent decision-making.

2.3. Agent decision-making layer: Heterogeneous agent division and collaboration mechanism

The structure of the division in the decision-making layer of the agent is heterogeneous, and the three specialized agents work together to avoid the problem of single-agent decision-making in a single dimension [4]. The Market Perception Agent will gather the latest information from the market, determine the various types of market conditions, and provide related environmental factors, such as range-bound or bull and bear markets. Temporal deep learning algorithms are employed by the return prediction agent to forecast short-term and medium-term fluctuations in the returns of various assets and determine the distribution of asset returns. First, the risk-control agent will take some risk indicators, such as portfolio volatility, maximum drawdown, and systemic risk, and then change the upper limits of these risks dynamically. The three agents can make independent, localized, reasonable decisions at the same time, and by sharing information, they will achieve the best overall outcome for the market.

2.4. Strategy execution layer: Training and dynamic rebalancing optimization logic

Centralized Training and Decentralized Execution mode is employed in the Strategy Execution and Optimization Layer. Proximal Policy Optimization (PPO) is employed in the training stage to iteratively update the parameters of all agents, and at the same time, a combined reward function that considers return, risk, and transaction costs is used to reduce the high risk exposed by a single return-based reward method. At the execution stage, each agent makes decisions according to current changes in the market; it will change the proportion of different types of assets dynamically and adjust the frequency of rebalancing to reduce trading costs caused by high-frequency trading. Thus, there will be fluctuations in the composition of the investment portfolio.

3. Core optimization mechanisms and innovative advantages of the model

3.1. Dynamic risk self-adaptation mechanism

In comparison with the traditional portfolio model and the single-agent reinforcement learning model, the multi-agent dynamic configuration model constructed in this paper has several main optimization mechanisms that can specifically address all the problems of the previous two methods [5]. The main advantage is that it can adjust the risk self-regulation mechanism dynamically, unlike the fixed risk constraint parameters in the old model; instead, the risk control agent will change the tolerance for risk according to how volatile the market is at any given time. When the market is unstable, the model will calculate various risks at the same time, such as volatility, skewness, and kurtosis, to reduce the proportion of high-risk assets and increase the allocation to bonds and defensive stocks in an effort to limit the drop in the total value of the investment [6]. When the market is sluggish, the lower bound of the risk constraint is set to a certain extent, and the return prediction agent has more freedom to choose high-yield investments. Unlike the traditional mean-variance model with a fixed covariance matrix, this approach changes the risk preference according to changes in the market and dynamically adjusts the trade-off between risk and return; thus, there will be no extended deviations in a single risk parameter during the observation period. By adjusting the threshold flexibly, the portfolio can perform well in different market environments; it will not be too conservative and miss out on rising market profits, nor will it be too risky and suffer heavy losses.

3.2. Multiagent collaborative gaming and cost constraint mechanism

A good multi-agent cooperative game and cost constraint mechanism have been built for the above model. Most traditional models have only one goal for asset allocation and do not consider all the trade-offs among trend detection, return forecasting, and risk control [7]. Based on the above framework, the three agents have different responsibilities and work together to make decisions in response to market conditions: the market perception agent issues regime labels for the current market environment, the return prediction agent generates return expectations for various assets according to this information, and the risk control agent changes the limits of risk constraints dynamically based on volatility. The three agents form a system of mutual checks and balances and cooperation; any one of their choices will be met with feedback from the others to correct the biases and defects caused by a single decision, and extreme strategies that blindly pursue high returns or are overly conservative will be avoided. Therefore, the allocation of assets will be in line with actual investment reasons. At the same time, a model of transaction costs is also introduced to quantify the trading costs that agents bear in weight updates, such as commissions and slippage. The rebalancing decision is not to maximize the return alone; rather, it is to determine how much additional return can be obtained per unit of trading cost, so agents only conduct trades that will increase the net benefit of weight updates and thus reduce inefficient high-frequency rebalancing. This mode is suitable for institutions with a large amount of money; that is to say, the market impact cost of a single large transaction is relatively high. Therefore, cost-sensitive decisions can improve the practical feasibility and capital-carrying capacity of quantitative strategies in actual trading.

3.3. Advantages of timevarying correlation modeling based on graph attention networks

A Graph Attention Network (GAT) model is employed to improve the representation of time-varying asset correlations; that is to say, it can dynamically adjust the weight given to the correlation among different assets and be sensitive to changes in this weight caused by sector rotation or other external market conditions. GAT constructs a graph of all the assets and takes the initial node embeddings of these assets as the asset price return series and volatility characteristics [8]. At each time step, attention is paid to how strongly connected the nodes are at that time, and thus the adjacency matrix is changed according to the degree of correlation. Therefore, if there is a sudden change in the transmission channel of risk among assets during a period of market liquidity crunch, for instance, many assets that are usually weakly correlated may suddenly become highly correlated; in such cases, the model can detect this structural break in time and inform the decision module of the agent accordingly [9]. At the same time, we should reduce the concentration of risk in groups of highly correlated assets during allocation optimization. The allocation distortion caused by fixed correlation matrices in the traditional model has been reduced; thus, the rationality and stability of multi-asset portfolio allocation have improved, the robustness of the strategy in times of market turmoil has increased, and high-level agents now have a more realistic and dynamic view of market interdependence [10].

4. Empirical results and analysis

4.1. Data description and experimental setup

To comprehensively verify the practical effect of the proposed model, daily trading data for January 2022 to June 2026 of the main components of the CSI 300 Index, the main contract of 10-year Treasury bond futures, and the Nanhua Commodity Composite Index will be taken as the research sample. The data will be preprocessed to handle missing values and outliers, and the logarithm of the raw price series will be taken, etc. A rolling window method is used to create the training, validation, and test sets: The training period is from January 2022 to December 2024; the validation period is from January 2025 to June 2025; and the test period is from July 2025 to June 2026, etc. The traditional mean-variance model and a single-agent PPO-based reinforcement learning model are taken as the references. The four main quantitative indicators of comparison are cumulative return, Sharpe ratio, maximum drawdown, and annualized volatility. For the purpose of comparison, all models use the same asset pool and the same time window for training and backtesting, and the rebalancing frequency is uniformly set to every five trading days to avoid any influence from different trading periods.

4.2. Backtest results and comparative analysis

According to the backtesting results for the whole volatile period of the sample, the multi-agent reinforcement learning model performed better than both benchmark models in terms of cumulative return; its annualized return was about 28.6% higher than that of the mean-variance model and around 15.3% higher than that of the single-agent model. In terms of the control of risk, the maximum drawdown of the proposed model is about 12.4 percentage points lower than that of the traditional mean-variance model, the annualized volatility has dropped considerably, and the Sharpe ratio reached 1.82 in the backtest period; thus, it is a relatively low-risk, high-return option. Based on the above analysis, the old mean-variance model fixes the risk parameters and asset weight benchmarks for the entire observation period and cannot adjust the allocation structure in response to changes in the market environment; thus, it is likely to be out of touch with the direction of the market during periods of sudden market fluctuations and suffer from volatile returns and larger drawdowns. The single-agent model performs well on the training set but is significantly worse on the validation and test sets; therefore, it has been determined that the single network cannot effectively handle market assessment, return forecasting, and risk constraints in all market environments at the same time, and lacks good generalization ability. At the same time, the three agents in the proposed model are responsible for market regime identification, discovery of return opportunities, and adjustment of risk constraints, so each module can be optimized independently for its own sub-problem and show good cross-cycle adaptability as a whole.

4.3. Extreme market stress tests and mechanism discussion

In the event of a sudden drop in the market and changes in the structure of the economy, the new plan can adjust asset allocations quickly to limit the damage from economic problems and reduce overall risk. In the early 2024 market correction, the risk-control agent quickly noticed that volatility had increased significantly and promptly reduced the proportion of equity investment from 65% to around 42%, while raising the proportion of bonds to 48% to avoid a serious drop. In the structural rebound in the third quarter of 2025, the return prediction agent detected strengthened return signals from some leading sector stocks, and after the market perception agent confirmed the improvement in sentiment, it gradually increased equity weights to about 58% to seize the opportunities in the market recovery. At the same time, due to rebalancing lag, the traditional mean-variance model also fell to an even lower level; the single-agent model was set at a certain risk level and did not promptly adjust the equity allocation, so it failed to achieve all the benefits of the rebound. Based on the above comparison, both the response speed and the decision advantage of the multi-agent system near the market turning point are relatively high. Based on the above analysis, it is known that traditional static allocation models cannot keep up with the changes in the market and are slow to respond to sudden market fluctuations; at the same time, single-agent models have a single-dimensional decision system and are prone to overfitting and poor adaptation. At the same time, the multi-agent model also has a multi-dimensional cooperative decision system that covers the assessment of market trends, return generation, risk control, and reduction of transaction costs; thus, it can achieve a good balance of profitability and stability and meet the actual asset allocation needs of real financial markets more closely.

5. Conclusion

To solve the problems of a fixed form, one-sidedness, and weak market adaptability in the traditional portfolio allocation method, this paper puts forward a dynamic portfolio allocation model based on multi-agent reinforcement learning. The three general approaches are hierarchical heterogeneous agent division and cooperation, time-varying asset correlation modeling, and dynamic risk-adaptive optimization; together, they will achieve intelligent and dynamic weight allocation of multi-asset portfolios. Based on the above empirical evidence, the model can achieve a good balance between portfolio return and risk, improve the risk-adjusted return, strengthen the application of the strategy in unstable and extreme market conditions, and thus correct the shortcomings of traditional models and single-agent approaches. This paper has made some contributions to the technical research system of intelligent quantitative investment and has provided a good technical reference for institutional asset management, individual quantitative investment, and multi-asset portfolio optimization. In the future, various motivating factors such as macroeconomic policies and market sentiment will be continuously added to improve the agent interaction and game mechanism, enhance the generalization ability of the model in different market cycles and situations, and promote the practical application and wider use of intelligent investment allocation technology.

Disclosure statement

The author declares no conflict of interest.

References

[1] Zhang B, 2025, Graph Attention-Based Heterogeneous Multi-Agent Deep Reinforcement Learning for Adaptive Portfolio Optimization. Scientific Reports, 16(1): 2674.

[2] Shen Y, Li C, Scaillet O, et al., 2026, Dynamic Portfolio Allocation Under Market Incompleteness and Wealth Effects. Operations Research, 74(1): 93.

[3] Cousin A, Lelong J, Picard T, 2023, Mean-variance Dynamic Portfolio Allocation with Transaction Costs: A Wiener Chaos Expansion Approach. Applied Mathematical Finance, 30(6): 313–353.

[4] Siyang P, Shaojun G, Yonghong L, 2022, Large Dynamic Covariance Matrix Estimation with an Application to Portfolio Allocation: A Semiparametric Reproducing Kernel Hilbert Space Approach. Journal of Systems Science and Complexity, 35(4): 1429–1457.

[5] Siyang P, Shaojun G, Yonghong L, 2022, Large Dimensional Portfolio Allocation Based on a Mixed Frequency Dynamic Factor Model. Econometric Reviews, 41(5): 539–563.

[6] Haitao L, Chongfeng W, Chunyang Z, 2021, Time-Varying Risk Aversion and Dynamic Portfolio Allocation. Operations Research, 70(1).

[7] Jin X, Lehnert T, 2018, Large Portfolio Risk Management and Optimal Portfolio Allocation with Dynamic Elliptical Copulas. Dependence Modeling, 6(1): 19–46.

[8] Garcia CR, González V, Contreras J, et al., 2017, Applying Modern Portfolio Theory for a Dynamic Energy Portfolio Allocation in Electricity Markets. Electric Power Systems Research, 150: 11–23.

[9] Pinar CM, 2014, About the Dynamic Allocation of Robust Portfolio Against the Uncertainty of the Average Yields. INFOR: Information Systems and Operational Research, 52(1): 14–19.

[10] Thomaidis SN, Roumpis E, Kondakis N, 2010, Optimal Portfolio Allocation Strategies with Dynamic Factor Models. Int. J. of Financial Markets and Derivatives, 1(4): 352–370.