Adaptive Dynamic Portfolio Optimization via a PPO-DQN Hierarchical Reinforcement Learning Framework

Yanan Liang
School of Economics and Management, Tianjin Vocational Institute; Tianjin 300000, China
E-mail: liangyanan1988@126.com

Keywords: deep reinforcement learning, PPO-DQN synergistic framework, dynamic portfolio optimization, adaptive optimization

Received: July 2, 2025

In view of the increasing dynamics and complexity of the financial market, traditional quantitative investment models are difficult to adapt to the high-frequency and changeable trading environment, while deep reinforcement learning (DRL) has gradually become a hot topic in portfolio optimization research with its adaptive decision-making advantages. This study combines the strategy stability of Nearest Neighbor Strategy Optimization (PPO) with the value evaluation ability of Deep Q Network (DQN), aiming to solve the problems of large fluctuations in strategy updates and difficult risk-return balance in dynamic asset allocation. The model combines the clipping mechanism of PPO with the experience replay of DQN to optimize long-term value prediction and limit the scope of strategy updates based on historical experience, thereby improving the robustness of investment decisions. The experiment of constructing a dynamic portfolio based on 15 Chinese A-share stocks (backtest period 2020-2025) shows that the cumulative return of the improved PPO algorithm with the introduction of the invalid action shielding mechanism is 74.8% and the annualized return is 33.7%, which is significantly higher than the original PPO (annualized only 2.3%). In terms of risk control, the maximum drawdown of the model is 5.85%, and the annualized Sharpe ratio is stable at 1.555, which is better than the traditional risk parity model (maximum drawdown of 11.86%). By adjusting the configuration of the neural network hidden layer, the cumulative return of PPO increased to 33.7% after adding a single hidden layer, which verified the effectiveness of structural optimization. Compared with traditional machine learning models (such as random forests), the framework has an annualized return increase of about 12%, and it recovers faster and is more resilient to risks during periods of extreme volatility. The data was normalized by Z-score and corrected by 3σ outliers, divided by 7:1.5:1.5 (rolling window 252 trading days); PPO module with 3-layer fully connected network (128/64/32), γ=0.95, λ=0.9, clipping range [0.8, 1.2]; DQN was used with a dual network (playback pool 10^6, batch size 256, initial ε=0.9), combined with 4-head attention fusion, alternating training for 500 rounds (200 episodes per round, 60 decisions per step), and using Adam optimization. Research shows that the PPO-DQN synergy framework can continuously optimize investment portfolios by dynamically weighing returns and risks, providing innovative solutions for smart financial decision-making.

Povzetek: Hibridni PPO–DQN model za dinamično alokacijo sredstev v backtestu 2020–2025 bistveno izboljša donosnost (74,8 %; 33,7 % letno) in hkrati ohrani nizko največje znižanje (5,85 %).

1 Introduction

As the complexity and uncertainty of financial markets continue to escalate, dynamic portfolio optimization encounters numerous challenges, including high-frequency trading, nonlinear correlations, and abrupt market shifts [1]. Traditional quantitative models, rooted in linear assumptions and static rules, struggle to capture the multi-scale fluctuation characteristics of asset prices, particularly during extreme market conditions, where they are prone to strategy failure or uncontrolled risk exposure [2, 3]. The advent of deep reinforcement learning (DRL) technology has ushered in an intelligent investment framework that employs adaptive strategies, emerging as a new paradigm for tackling this challenging issue [4]. Within this framework, the synergistic innovation of proximal policy optimization (PPO) and deep Q-network (DQN) presents a novel approach to dynamic asset allocation, one that considers both strategy stability and valuation capabilities [5].

The evolution of financial markets exhibits the distinct characteristics of complex systems [6, 7]. Stock price series display a fractal structure and chaotic attractors, making it difficult for traditional time series models to accurately depict their long-term dependencies and nonlinear fluctuations [8]. Furthermore, the heterogeneous behavior of market participants and their interconnected interactions compound the system's

complexity. Phenomena such as the emotional contagion spread by social media and the herding behavior of institutional investors can trigger sudden shifts in market states [9, 10]. Against this backdrop, traditional rule-based quantitative strategies reveal two significant limitations: firstly, the manually curated factor libraries struggle to encompass the hidden correlations within high-dimensional data, resulting in models that are insufficiently adaptable to shifts in market structure; secondly, the rigid threshold risk control mechanisms can easily lead to over-trading or strategy passivity during extreme market fluctuations. For instance, the classic risk parity model experienced a maximum drawdown exceeding 30% during the US stock market circuit breaker event in 2020, highlighting the vulnerabilities of static parameter systems.

In response to the above problems, the collaborative framework integrating PPO and DQN shows unique advantages. PPO algorithm limits the update range of policies through the shearing mechanism, constrains the KL divergence of the old and new policies in the stable interval, and effectively avoids the risk of policy collapse [11]. The innovation of this framework is also reflected in the targeted design of the particularity of financial data. Given the market state's high dimensionality and partial observability, the model adopts a dual neural network architecture's actor network, which generates dynamic weight allocation strategies. In contrast, the Critic network combines attention mechanisms to extract multi-scale market characteristics. It is worth noting that this framework is deeply compatible with the intrinsic characteristics of complex financial systems [12]. As a typical dissipative system, the interaction between the financial market's energy input (such as monetary policy) and entropy increase process (such as information diffusion) gives birth to a persistent non-equilibrium state [13]. In the dynamic portfolio optimization scenario, the unbalanced evolution of the market presents complex and changeable underlying patterns, which are often affected by multiple factors such as macroeconomic fluctuations, industry policy adjustments, and changes in market sentiment, and have obvious time-varying and non-linear characteristics. The PPO-DQN collaborative model can break through the limitations of a single time dimension and dynamically evaluate the return performance and risk level of the portfolio at different time scales through the calculation of the multi-time step advantage function. Specifically, this function comprehensively considers the immediate returns brought by short-term market fluctuations, the phased returns formed by medium-term industry trends, and the fundamental returns determined by the long-term economic cycle, while taking into account the risk factors at different time stages, so as to more comprehensively capture the hidden patterns in the process of non-equilibrium evolution that gradually emerge over time. This multi-dimensional pattern capture capability enables the model to grasp market dynamics more accurately, providing strong support for adaptive adjustment of dynamic portfolios, thereby improving the optimization effect of portfolios in complex market environments.

With the deepening and development of digital financial ecology, the association between high-frequency alternative data and complex networks is reshaping the underlying logic of investment decision-making [14, 15]. The PPO-DQN collaborative framework can reveal hidden risk transmission paths that are difficult to capture by traditional factor models by integrating graph neural networks (GNN) to model cross-asset correlations. For example, when building a multi-asset portfolio, including stocks, bonds, and commodities, adjusting position allocation based on network centrality indicators shows stronger risk resistance in stress tests. This technology integration not only promotes the innovation of a quantitative investment paradigm but also provides a computational, experimental platform for understanding the micro-mechanism of the financial system, making it possible to model complex systems from bottom to top.

The core questions of this study include: Can the PPO-DQN collaborative framework improve risk-adjusted returns in volatile markets? Does its integration mechanism address the high volatility of a single PPO strategy and the insufficient long-term forecasting of a single DQN? The environment is based on the CSI 300 constituent stocks and macro indicators from 2015 to 2024, backtesting from 2020 to 2025, with a rolling window (252 trading days) used to divide the dataset (7:1.5:1.5); the state is a high-dimensional vector that is standardized and corrected for outliers, the action is the continuous adjustment of the asset allocation weights (granularity 0.1%), and the reward is centered on maximizing the Sharpe ratio; PPO uses a 3-layer fully connected network (128/64/32), γ=0.95, λ=0.9, clip range [0.8, 1.2], while DQN employs a double network, with a replay pool of 10^6, a batch size of 256, ε starting at 0.9. The two are fused through 4-head attention; training alternates, with 500 iterations (200 episodes per iteration, 60 decisions per step), using Adam optimization (5e-4), including invalid action masking, double experience replay, and priority sampling, early freezing of the value network, and later joint optimization; convergence is based on Lyapunov theory, constraining KL divergence, monitoring Bellman error and relative entropy, switching to conservative updates when exceeding a threshold, ensuring reproducibility.

Compared with the current mainstream deep reinforcement learning (DRL) portfolio model, such as the lack of exploration of the pure PPO model and the weak generalization of the high-dimensional state space faced by the pure DQN model, the PPO-DQN collaborative framework proposed in this paper uses a dynamic division of labor between two agents: PPO is responsible for refined strategy iteration to ensure income stability, and DQN focuses on global state exploration to improve extreme market adaptability The collaborative shortcomings of DRL algorithms in dynamic combinatorial optimization are "local optimal traps" and "inefficient global exploration".

2 Theoretical paradigm and synergy mechanism of dynamic portfolio optimization

2.1 Mathematical basis and theoretical boundary of dynamic portfolio optimization

Dynamic portfolio optimization is a key topic in financial engineering, and its mathematical basis is based on modern portfolio theory (MPT), capital asset pricing model (CAPM), and efficient market hypothesis (EMH) [16]. Figure 1 shows its architecture. These theories provide theoretical support for asset allocation, emphasizing diversification to reduce risks and pursuing the best risk-return balance [17].

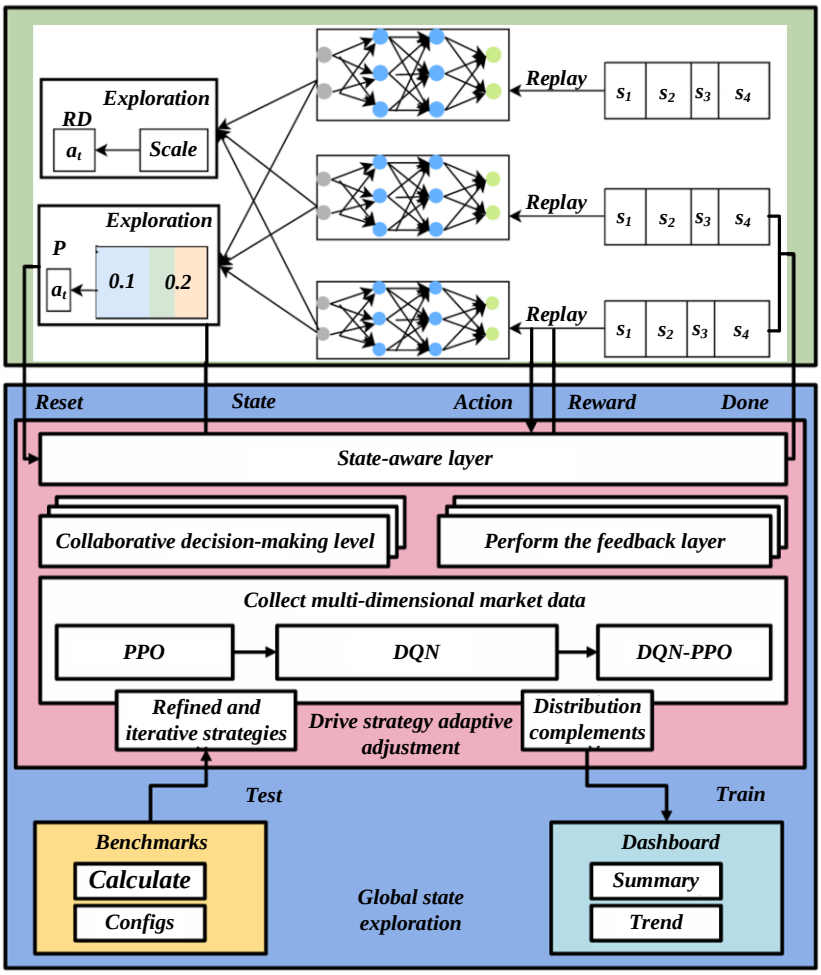

Figure 1: Dynamic portfolio optimization architecture

The mean-variance model (M-V model) is the first model to evaluate the risk and return of investment portfolios [18], and its mathematical expression can be described by formula (1) or formula (2).

$$max\, R_p(W) = \sum_{i=1}^{n} w_i r_i$$
$$s.t. \begin{cases} \sigma_p^2(W) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij} = \theta \quad (1) \\ \sum_{i=1}^{n} w_i = 1 \end{cases}$$

$$min\, \sigma_p^2(W) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij}$$
$$s.t. \begin{cases} R_p(W) = \sum_{i=1}^{n} w_i r_i = r_e \quad (2) \\ \sum_{i=1}^{n} w_i = 1 \end{cases}$$

Where r_e is the expected rate of return, R_p is the expected return of the portfolio, θ is the investor risk appetite, n is the number of investment targets, α is the confidence level, σ^2_p is the portfolio variance, and σ_{ij} is the covariance of the *i-th* and *j-th* investments. W represents the weight, the portfolio return is represented by the mean expected return, and the variance describes the portfolio risk, aiming at minimizing the risk under the

same return or maximizing the return under the same risk.

The mean-variance model mainly focuses on portfolio optimization in a single period, but in a dynamic investment environment, the portfolio may change constantly [19]. Therefore, this article explores other metrics for evaluating portfolio performance. Total return is the cumulative rate of return since the creation of an asset portfolio, and is used as a standard to measure the fund's profitability. The greater the total rate of return Rt value, the better the fund's profitability. The calculation formula (3) is as follows:

$$R_t = \frac{net\_value_t - net\_value_0}{net\_value_0} \times 100\% \quad (3)$$

Net_value_t represents the net value of the portfolio at time t, and the net value at the initial time is represented by net_value_0. The annualized rate of return is the average annual rate of return $R(y)$ calculated based on the annual distribution of portfolio income and considering the compound interest effect. See formula (4) for the calculation method.

$$R(y) = \left[\left(net\_value_y / net\_value_0 \right)^{\frac{1}{y}} - 1 \right] \times 100\% \quad (4)$$

Net_value_y represents the net value of the investment portfolio at the end of the y-th year, and the initial net_value_0 is the starting net value of the investment. The maximum drawdown measures the maximum possible loss of an investor, and the calculation formula is shown in Equation (5).

$$Max\_retraction = \min_{i,j \in [0,t], j > i} \left\{ \left(net\_value_j - net\_value_i \right) / net\_value_i * 100\% \right\} \quad (5)$$

net_value_i is the net worth of the portfolio at a specific time. Volatility is a tool to measure the magnitude of the price movement of a financial asset, usually expressed as standard deviation. The net value curve fluctuates greatly and the volatility is high; The curve is smooth and volatility is low. The Sharpe ratio evaluates portfolio returns versus risks. A high Sharpe ratio indicates a greater return for every unit of risk taken. During the T period, the Sharpe ratio is calculated according to Equation (6).

$$SharpeRatio_T = \frac{E(R_t) - r_f}{\sigma(R_t)} \quad (6)$$

In the calculation, the average yield of the portfolio is denoted by $E(R_t)$, the risk-free rate is denoted by r_f, and the standard deviation of the yield is denoted by $\sigma(R_t)$. Typically, the one-year deposit rate represents the risk-free rate. The Sharpe ratio is primarily used to compare similar portfolios, and different kinds of portfolios may not accurately reflect performance. Performance is generally considered good when the Sharpe ratio of medium and high risk portfolios exceeds 1.

The α value represents the difference between the actual return of the investment and the expected return, that is, the "excess return". It is very important when evaluating investment performance because it shows how well an investment strategy, trader or portfolio manager has performed relative to market returns over a specific period [20]. α return is usually regarded as a positive return on an investment, and the investment effectiveness is evaluated by comparing it with a market index or benchmark. In China's A-share market, commonly used market indexes include SSE 50, CSI 300 and CSI 500. See Equation (7) for the calculation method.

$$profit_\alpha = R_s - \left[r_f + \beta_s \left(R_b - r_f \right) \right]$$
$$\beta_s = \frac{Cov\left(R_s, R_b \right)}{\sigma_b^2} \quad (7)$$

As a relative measure to evaluate the level of profitability. The calculation method is shown in Equation (8).

$$profit_\beta = \beta_s \left(R_b - r_f \right) \quad (8)$$

R_s represents the real rate of return of the portfolio, r_f represents the risk-free rate of return, β_s reflects the sensitivity of the portfolio to market fluctuations, and R_b is the expected rate of return of the market index. The Karma ratio measures the level of return of a portfolio when assumed per unit of risk (maximum drawdown), using the maximum drawdown as a measure of risk, combined with the Sharpe ratio to evaluate portfolio performance [21]. When the Sharpe ratios are close, a higher Karma ratio indicates stronger risk control. The formula (9) for calculating the Karma ratio is as follows:

$$Clamar_T = \frac{E(R_t) - r_f}{Max\_retraction_T} \quad (9)$$

In the formula, $E(R_t)$ represents the average return rate of the investment portfolio in the T period, r_f is the risk-free interest rate, and $Max\_retraction_T$ refers to the maximum retracement in the T period.

2.2 Theoretical mechanism of ppo-dqn collaborative optimization

The online deep reinforcement learning algorithm PPO is designed for continuous action space, which solves the problem of updating step size in policy gradient method [22, 23]. It excels on continuous control problems, balancing the effect of step size on strategy stability and training time.

Reinforcement learning is a way in which an agent learns by interacting with the environment, with the goal of obtaining the maximum reward [24]. It usually simplifies environmental models using Markov decision processes. If the agent state transition only depends on the current state, it means that the environment follows Markov property, which is suitable for applying this process [25]. The interaction between the agent and the environment is represented by a quadruple (S, A, P, R), where S is the state set, A is the action set, $P(s'|s,a)$ is the state transition probability, and $R(s, a)$ is the reward function. The long-term discount reward G_t of the agent is calculated from time t, and the specific formula is shown in Equation (10).

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \quad (10)$$

In the formula, the value of γ is between 0 and 1, and as a discount factor, it determines how much the current

decision attaches importance to future rewards. When γ is 0, the system ignores long-term rewards. The key lies in finding control strategies to maximize rewards. The agent selects the action according to the strategy π. In the state S, the expected value of action a is defined by the action value function $q^{\pi}(s, a)$, and the calculation method is shown in formula (11).

$$q^{\pi}(s,a) = E[G_t \mid S_t = s, A_t = a] = E\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \mid S_t = s, A_t = a\right] \quad (11)$$

When at least one strategy outperforms all others, the strategy is called the optimal strategy π. Its action value function $q^{*}(s, a)$ is the largest of all $q^{\pi}(s, a)$. When agents learn, they will consider state-action pairs to decide action choices, not just states [26]. The action value function is updated by Bellman equation and time difference method, as shown in Equation (12).

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t)\right] \quad (12)$$

The learning rate is represented by α. The action value function is updated after each state transition. If S_{t+l} is the end state, then $q(s_{t+l}, a_{t+l}) = 0$. The agent learns to achieve the optimal strategy by performing the maximum action indicated by the action value function.

As seen in Table 1, among existing reinforcement learning (RL) models, a single PPO model can optimize strategies but achieves an annualized return of only 2.3%, and the strategy updates are highly volatile, making it difficult to respond to high-frequency market changes; a single DQN model has limitations in long-term value prediction, affecting the foresight of the decision-making; machine learning (ML) models such as random forests and support vector machines have an annualized return approximately 12% lower than the PPO-DQN collaborative framework, with weaker resistance to market fluctuations and poor performance in extreme market conditions; among traditional models, the Markowitz mean-variance model has average comprehensive performance and insufficient adaptability to complex markets, the risk parity model has a maximum drawdown of 11.86% and inadequate risk control capacity, while benchmark indices like CSI 300 have poor overall performance and lack proactive optimization capability. The PPO-DQN collaborative framework proposed in this study effectively addresses these issues: this framework combines the strategy stability of PPO and the value assessment capability of DQN, limiting the range of strategy updates through PPO's clipping mechanism to reduce volatility, while optimizing long-term value prediction and enhancing decision-making foresight through DQN's experience replay pool and dual network structure. In terms of risk control, its maximum drawdown is only 5.85%, far below that of the risk parity model, and it enhances risk resistance by dynamically balancing returns and risks. Additionally, the framework utilizes an attention mechanism to achieve feature fusion, combined with a rich dataset (including macroeconomic indicators) and an optimized neural network structure, resulting in an annualized return of 33.7%, and faster recovery during extreme market volatility, with overall performance comprehensively surpassing existing methods.

Table 1: Model comparison

Category	Existing RL Models	Existing ML Models	Traditional Models	Proposed PPO-DQN Framework
Models	Single PPO, Single DQN	Random Forest, SVM	Markowitz M-V, Risk Parity, CSI 300	PPO-DQN Collaborative
Dataset	Partial stock data, simple processing	Partial stock data, simple processing	Partial stock data, simple processing	CSI 300 (2015-2024) with macro indicators, standardized
Metrics	Annual Return, Cum Return, Max Drawdown, Sharpe	Annual Return, Cum Return, Max Drawdown, Sharpe	Annual Return, Cum Return, Max Drawdown, Sharpe	Annual Return, Cum Return, Max Drawdown, Sharpe
Performance	PPO: 2.3% return, unstable DQN: limited long-term prediction	12% lower return, weak anti-volatility	Markowitz: average Risk Parity: 11.86% drawdown CSI 300: weaker	33.7% return, 74.8% cum return, 5.85% drawdown, Sharpe 1.555

3 Construction of adaptive optimization model based on PPO-DQN collaborative framework

3.1 Design of PPO-DQN hierarchical network structure

When applying reinforcement learning technology to solve research problems, it is necessary to consider the large continuous action space of assisting robot interaction. This leads to improving state space dimension and the surge in training parameters. This consumes a lot of computing and storage resources and

affects real-time interactions. Deep learning can process high-dimensional state space and combines the perception of deep learning with the decision-making ability of reinforcement learning to form a deep reinforcement learning method, which can directly use camera information to control robots [27, 28]. Hierarchical reinforcement learning decomposes complex tasks into multiple sub-tasks. It completes them step by step, which helps to reduce the state space dimension and improve the training speed and real-time performance [29].

In this study, the hierarchical deep reinforcement learning architecture is applied, and the divide-and-conquer strategy is adopted to decompose the auxiliary interaction task into two secondary tasks, which are solved separately by deep reinforcement learning technology. Finally, combining the strategies of these two sub-tasks, we get an efficient global strategy.

In terms of network architecture, the PPO module consists of 3 fully connected layers: the first layer has 128 units (ReLU activation function), the second layer has 64 units (ReLU activation function), and the third layer has 32 units (ReLU activation function). The value network shares the parameters of the first two layers. The DQN module uses a double network structure (target network and action network), with each network having 3 fully connected layers: the first layer has 128 units (ReLU), the second layer has 64 units (ReLU), and the third layer has dimensions consistent with the output space (linear activation function). The input feature of the state vector includes original price data (daily closing prices, trading volume) and technical indicators (such as MACD, RSI, and Bollinger Bands), which is processed into a high-dimensional feature vector after Z-score normalization. The output space is defined as a weight vector for allocating 15 stocks, with each weight ranging from [0,1] and summing to 1, precise to 0.1%, to allow for continuous adjustments. The reward is designed using a weighted combination formula: Reward = 0.6 × (annualized return - risk-free rate) - 0.3 × maximum drawdown - 0.1 × annualized volatility, where the risk-free rate is taken from the current 1-year deposit rate, balancing returns and risks with dynamic weights. The training time frame covers 2015-2024, using a rolling window method (window size of 252 trading days) to partition the data, with each training episode consisting of 60 trading decision steps (about 3 months), iterating a total of 500 rounds to ensure model convergence.

In the dynamic portfolio adaptive optimization model of the PPO-DQN collaborative framework, graph neural networks (GNNs) provide architectural support for modeling cross-asset correlations. Specifically, GNNs regard various assets as nodes in the graph structure, and the node characteristics can cover multi-dimensional information such as historical returns, volatility, price-earnings ratios, and industry attributes of assets. The correlation between assets is defined as the edges between nodes, and the weight of the edges can be dynamically adjusted according to the strength of the correlation, for example by calculating the Pearson correlation coefficient of the asset return series, cosine similarity, or dynamic correlation scores based on attention mechanisms. In the model architecture, GNNs realize cross-asset information interaction and aggregation through a messaging mechanism: each node aggregates the feature information of adjacent nodes based on the weight of the edge, and after multiple layers of convolution or attention update, it generates a node embedding vector containing global asset association information. These embedding vectors not only retain the individual characteristics of a single asset, but also integrate its association patterns with other assets, thereby providing key inputs for the strategy network (PPO module) and value network (DQN module) in the PPO-DQN framework - the strategy network can make more collaborative investment decisions based on the dynamic correlation between assets, and the value network can combine cross-asset correlation to more accurately evaluate the long-term value of different decisions, and finally realize the effective modeling of cross-asset dependencies in complex market environments. Improve the adaptability and accuracy of dynamic portfolio optimization.The Critic network of the DQN module embeds the Transformer encoder, which uses the self-attention mechanism to capture the long-term dependencies of multi-asset price series, and guides the optimization of PPO strategy by calculating the time series difference error. The design enhances the accuracy of state value evaluation through transformers, combined with the balanced exploration and utilization of dynamic shear coefficients, effectively resolves the contradiction between "high-dimensional decision-making and convergence efficiency" in the discrete action space, and realizes the adaptive adjustment of asset portfolios during the training process.

The hierarchical deep reinforcement learning algorithm h-DQN consists of a meta-controller master controller and a controller. The master controller sets long-term goals and plans, using the DQN algorithm to train to maximize external rewards while periodically updating the memory pool. Controller: The controller receives the sub-target and the current state and selects actions according to the training strategy. The sub-goals remain unchanged until the task is completed. This hierarchy also uses the DQN algorithm. The goal is to maximize the internal reward computed by the critic function set by the master controller.

The h-DQN algorithm combines two DQN algorithms to achieve hierarchical planning [30, 31]. Its effectiveness lies in the top-level meta-controller, which creates sub-targets based on external rewards and shortens the Markov chain; the internal rewards of the underlying controller make the environmental rewards denser.

Although the bottom controller will learn the pose required by the task, the user can change the pose at any time, and the top controller cannot guarantee that the pose will remain unchanged. The system must respond to the dynamic changes of the environment and flexibly control the forces to ensure task friendliness and user acceptance.

The "Improved PPO" is a stand-alone variant with an invalid action masking mechanism, which is significantly different from the basic "PPO" and the core "PPO-DQN Collaboration Framework". The basic "PPO" follows the logic of the original proximal policy optimization algorithm and focuses on the stable update of the policy. The improved PPO improves decision-making efficiency by blocking ineffective actions that do not meet portfolio constraints (such as liquidity restrictions, risk threshold exceedance, etc.). The core "PPO-DQN collaboration framework" is a collaborative system between the two, with PPO responsible for policy optimization and DQN assisting value evaluation to form a complementary mechanism.

The PPO-DQN model captures the latent patterns in the non-equilibrium evolution through the dual-agent collaborative architecture: the PPO module generates a dynamic portfolio adjustment strategy based on the probability distribution, and tracks the profit opportunities in the short-term market fluctuations with the help of the strategy gradient method; The DQN module uses deep neural networks to fit the state value function to mine implicit correlations in long-term trends. In the non-equilibrium state of the market (such as sudden policy shocks and sudden changes in capital liquidity), PPO quickly responds to short-term noise, and DQN smooths the fluctuation of the strategy to maintain the stability of long-term returns, so as to jointly realize the adaptive capture of complex dynamics of the financial market, which is in line with the concepts of "multi-scale analysis" and "dynamic equilibrium correction" in modern financial modeling.

At the basic control layer, sub-targets are learned using the PPO deep reinforcement learning technique, which selects actions through probability distribution P and is suitable for continuous action space. The advanced control layer adopts the DQN algorithm to solve the strategy. The hierarchical deep reinforcement learning architecture proposed in this study is named PPO-DQN. The architecture diagram and time series expansion diagram are shown in Figure 2.

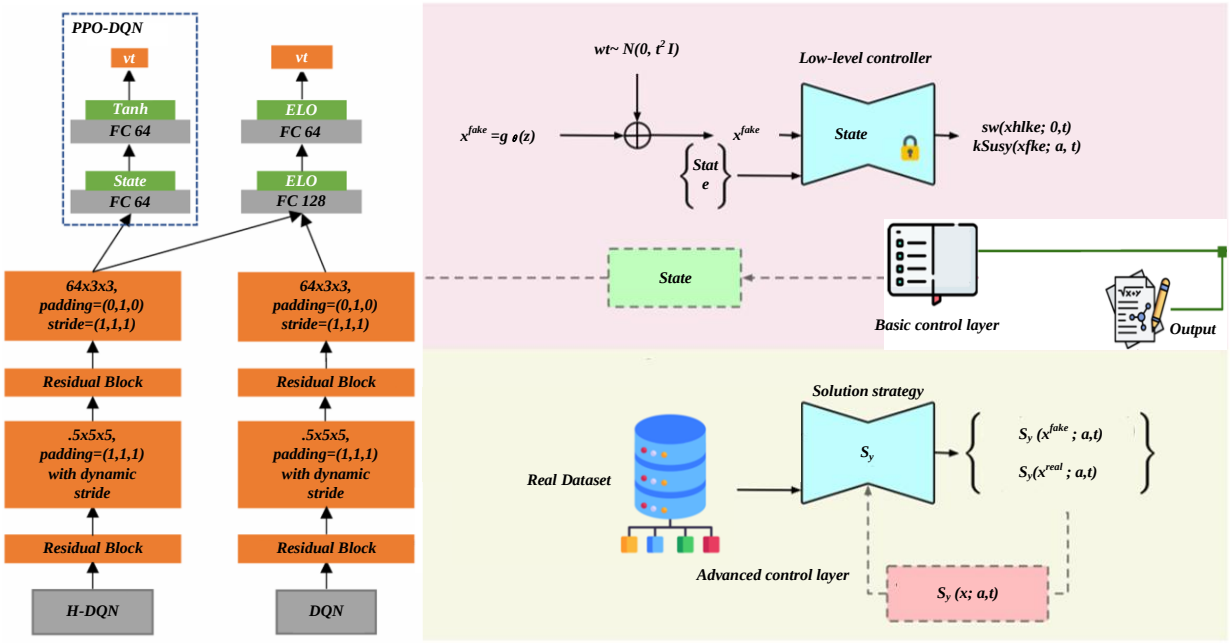

Figure 2: System architecture diagram of PPO-DQN hierarchical reinforcement learning

3.2 Model training and convergence guarantee strategy

In the dynamic portfolio optimization model of the PPO-DQN collaborative framework, model training and convergence guarantee constitute the key links in implementing the core algorithm. Aiming at dynamic financial markets' non-stationary and high-dimensional state space characteristics, this framework constructs a reinforcement learning paradigm with complementary advantages by integrating the collaborative training mechanisms of near-end strategy optimization (PPO) and deep Q-network (DQN). Among them, the PPO algorithm is responsible for policy gradient optimization in the continuous action space and realizes the stability of policy update through importance sampling and trust region constraints; DQN focuses on the Q-value function approximation of discrete action space and uses empirical playback and target network mechanism to improve the accuracy of value estimation. The coordination between the two is realized through the parameter sharing layer and asynchronous update mechanism, which makes the policy network and value network establish dynamic coupling between time series differential error and policy gradient signal.

In order to ensure the effectiveness of model training, a dual experience playback mechanism is designed to cope with the temporal correlation of financial market data. The main experience pool stores the original state-action-reward sequence. In contrast, the auxiliary experience pool records the synthetic transaction trajectory calibrated by the Sharpe ratio and balances the relationship between exploration and development through priority sampling strategies. At the same time, the adaptive temperature coefficient is introduced to adjust the entropy weight of the strategy, which strengthens the exploration ability in the early stage of training to cover the potential optimal strategy area and gradually reduces the randomness in the later stage to converge to the deterministic strategy. Aiming at the gradient update of the policy network, a phased optimization strategy is adopted: the value network parameters are frozen in the early stage to focus on policy search, and the strategy and value function are jointly optimized later to improve the accuracy of policy evaluation.

In terms of portfolio rebalancing, set a dynamic rebalancing frequency, adjusting in line with market volatility characteristics, such as shortening to daily during high volatility periods and extending to weekly during stable periods. At the same time, incorporate trading costs and market impact into the modeling. Trading costs include commissions, slippage, etc., while market impact is quantified through a correlation function between trading volume and price changes to better align with real investment scenarios.In the aspect of convergence guarantee, the theoretical analysis is based on the Lyapunov stability theory to construct the convergence criterion of the joint optimization process. Constraining the KL divergence threshold of the policy update step size ensures that each policy iteration meets the monotonic improvement condition. Aiming at the approximate error of the value function, a double delay update mechanism is designed; that is, the update frequency of the policy network is lower than that of the value network to alleviate the interference of policy oscillation on Q value estimation. Gradient normalization and dynamic learning rate attenuation strategies are introduced, and the learning rate ratio of PPO and DQN components is adaptively adjusted by monitoring the relative change rate of strategy entropy and value loss function. In order to cope with the time-varying characteristics of the financial market state transition matrix, a sliding window standardization module for state representation is constructed, and the distribution characteristics of observation space are dynamically adjusted to maintain the unbiased strategy gradient estimation.

During the model training process, the Bellman error of the strategy value function and the relative entropy index of the strategy update are monitored online, and the early stop mechanism and retraining trigger conditions are constructed. When it is detected that the value estimation variance exceeds the preset threshold, it automatically switches to the conservative update mode. It adopts the soft update strategy of the target network parameters to stabilize the training process. A verification environment based on Monte Carlo tree search is designed, and the training model is verified offline in the non-strategy evaluation stage. Potential overfitting risks are identified through virtual trading simulation and fed back to the main training loop for strategy correction. The collaborative framework ensures the reliable convergence of strategy optimization processes processes in complex financial market environments environments through multi-level stability control mechanisms.

4 Experiment and results analysis

Cross-validation uses a rolling window method with a fixed window length of 2 years, with each rolling step set to 1 month, and divides the historical data into a continuous training set (the first 18 months within the window) and a test set (the last 6 months within the window) in time series order to ensure that the time correlation between the training and test data is in line with the time series logic of investment decisions. By isolating the statistical features of the training set and the test set in the data preprocessing stage, and disabling any feedback from the test set data on the policy parameters during the model training process, the future information leakage is prevented from the data pipeline level and the authenticity of the evaluation results is ensured.

In the simulation process, transaction costs and slippage are clearly included to enhance the practical application value of the model. Among them, transaction costs are in the form of a combination of fixed proportions and variables: a fixed fee of 0.1%-0.3% of the transaction amount is calculated for standardized assets such as stocks and bonds, and an additional 0.05% settlement fee is superimposed on derivatives; When a single transaction amount accounts for ≤5% of the asset's daily trading volume, the slippage is calculated at ±0.2% of the transaction price, and the slippage rate increases linearly to ±1.5% when it exceeds 5%. Both are included in the calculation of net asset portfolio value in real time, and dynamic feedback is provided through the "net return after deducting transaction costs" item in the reward function of the PPO-DQN collaborative framework, ensuring that the model simultaneously considers the impact of transaction execution costs on the final return when optimizing decision-making.

The model transferability study selects S&P 500 constituent stocks and mainstream ETFs (spy, qqq) in the u.s. market as an additional test set, and compares the sharpe ratio and maximum drawdown performance of the model under different market structures to verify its universal applicability. The hyperparameter sensitivity study was carried out by the control variable method: the PPO clipping parameter (0.15), learning rate (1e-5) and discount coefficient (0.89) were searched in the grid, and the changes in the cumulative return of the strategy under each parameter combination were recorded, and the sensitivity heat map was drawn. In the ablation study, the PPO strategy update module, DQN value estimation network, and dynamic rebalancing mechanism in the

collaborative framework were removed in turn, and the risk-adjusted benefits of the complete model and each ablation version on the same test set were compared to quantify the contribution of each component to the overall performance.

Looking at Table 2, among the 28 sample industries, the rising beta coefficient (β +) of 13 industries exceeds 1, less than half. There are 20 industries whose beta coefficient (β-) has dropped by more than 1, more than 70%. This shows that when the market goes down, the systemic risk of most industries exceeds the market average. The analysis showed that only seven industries had a higher rising beta (β +) than a falling beta (β-), i.e. β +/β-> 1. This shows that the systemic risk when the market in most industries falls is greater than when it rises. A few industry markets have outperformed declines, and these are the key points that investors should pay attention to.

Table 2: Calculation results of rising β coefficient and falling coefficient

Industry Code	β	β +	β +/β-	Industry Code	β-	β +	β +/β-
801010	1.14	0.92	0.82	801200	1.14	1.09	0.98
801020	1.06	1.16	1.12	801210	1.08	0.81	0.77
801030	1.08	0.95	0.90	801230	0.89	0.99	1.13
801040	0.98	1.14	1.19	801710	1.14	1.00	0.90
801050	1.13	1.02	0.92	801720	1.12	1.08	0.98
801080	1.11	0.99	0.91	801730	1.12	1.02	0.93
801110	1.00	0.97	0.99	801740	1.42	1.16	0.84
801120	0.82	0.79	0.98	801750	1.14	1.10	0.98
801130	1.14	0.99	0.89	801760	1.00	0.97	0.99
801140	0.83	0.94	1.16	801770	1.17	1.02	0.89
801150	1.05	0.88	0.85	801780	0.44	0.92	2.13
801160	1.07	1.05	1.00	801790	0.98	1.17	1.22
801170	1.04	1.13	1.11	801880	1.11	1.02	0.94
801180	1.04	1.05	1.02	801890	1.15	0.97	0.86

Figure 3 shows the influence of investor sentiment S (t) on the optimal investment strategy π (t) under different values of interest rate volatility b. The path of π (t) with S (t) is similar at different values of b. The conclusion is that when S (t) < 0, the change of π (t) is small, and ∑ ² π; (t) < πo (t), it is suggested that when S (t) < 0, investors should invest less in the stock market and more in risk-free assets. If S (t) > 0, π ₁ (t) and π2 (t) increase first and then decrease, it means that investors can increase the stock investment ratio first and then decrease it appropriately.

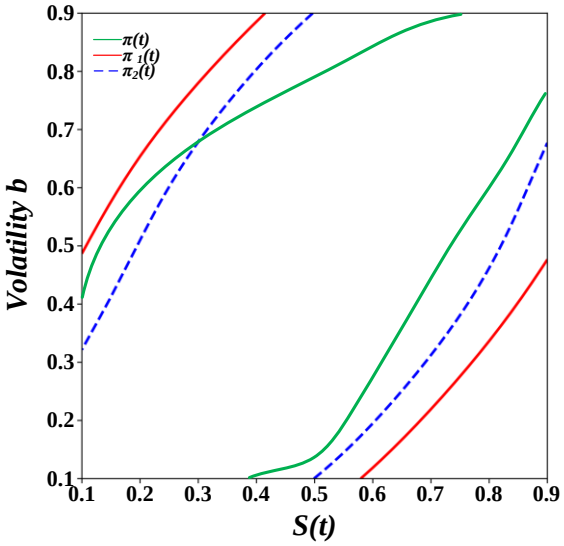

Figure 3: Optimal investment strategy

Through numerical simulation, the investment risks with or without unexpected events under different investor sentiments are compared. Figure 4 shows that under the same expected return, the risk is higher when there is a jump shock. Therefore, investors should adopt a cautious strategy in the face of possible unexpected events.

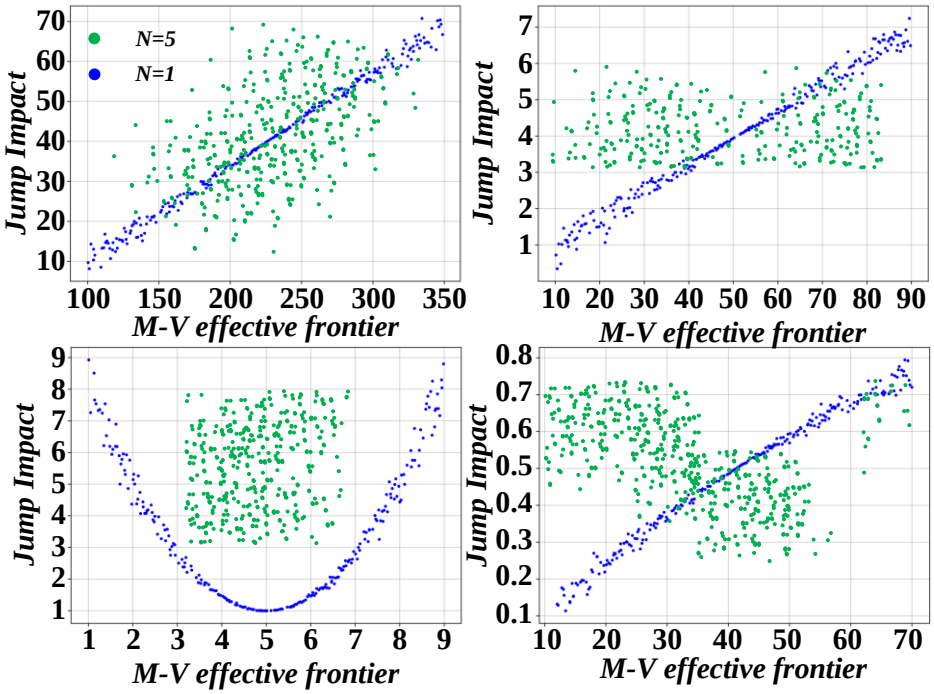

Figure: 4 M-V effective frontier

MAD stands for Risk Factor - Mean Absolute Deviation, which is a risk measurement method that combines risk factor analysis with mean absolute deviation. Among them, the risk factor focuses on the underlying driving factors that affect asset returns, and is used to identify the source of portfolio risk. The average absolute deviation measures portfolio volatility by averaging the absolute value of the return deviation from the mean. Combined, portfolio risk can be quantified and controlled more accurately.Ignoring the positive influence of purchase recommendation based on stock index forecast, the effect of dynamic time window selection method is compared. Figure 5 shows that portfolios with dynamic time window selection rules outperform portfolios with static time windows. This shows that the dynamic time window selection method based on stock index trend prediction and edge risk return is effective.

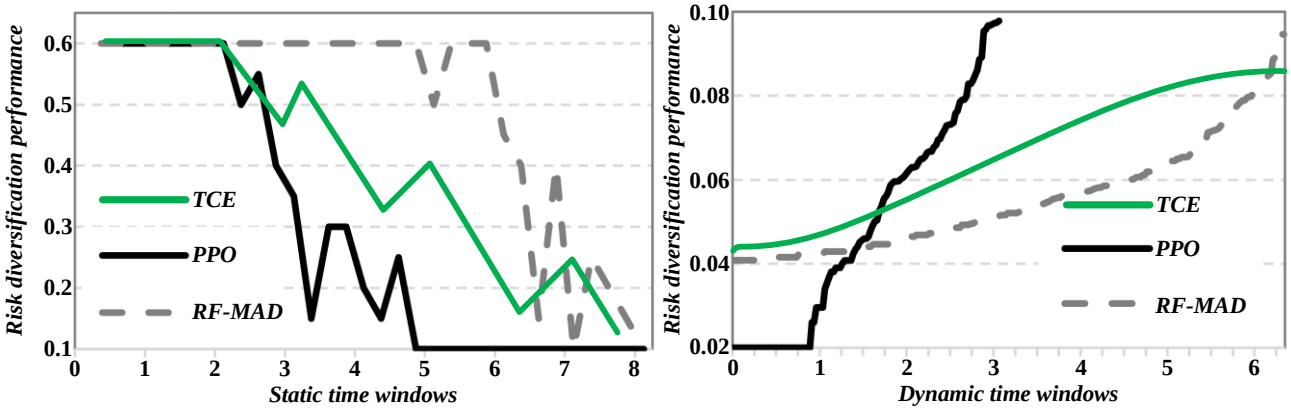

Figure 5: Positive impact of dynamically selected time window on risk diversification performance of RF-MAD

By analyzing the classification results of market states, both models can effectively distinguish market states, and the classification results are similar. However, there are classification errors. For example, if the market yield exceeds 4% near the 600th observation value, it should be a "bull market", but most extreme points in Figure 6 are marked as red dots. This shows that the CMRS model is more accurate than the traditional MRS

model in market state classification.

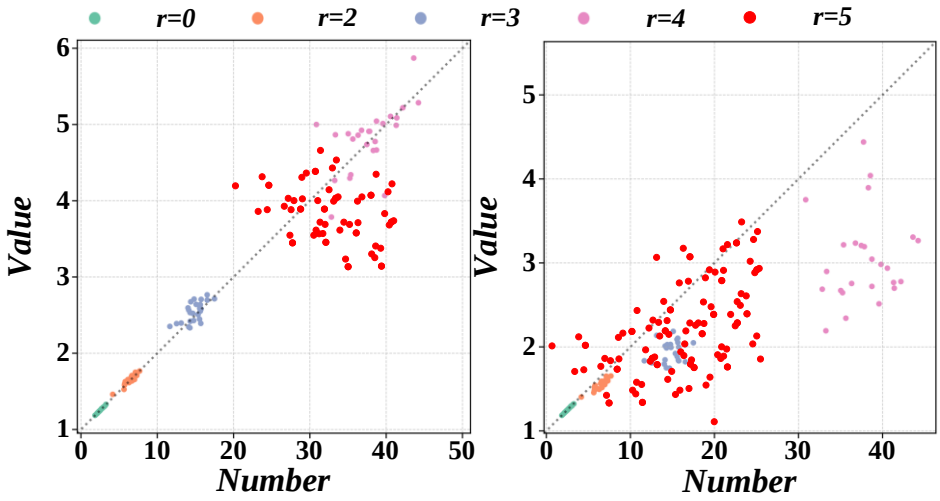

Figure 6: r = 0 (i.e. traditional MRS model) market state forecast

Table 3: Values of model (Pmvc) parameters η, λ and α in one-dimensional case

w	α *	λ	η
0	0.6416	6.1845	3.9227
0.5	0.1898	12.8619	13.1392
2	-0.2849	37.4044	61.1242

Set the return rate of risk-free assets r = 0.0408, the drift rate of risky assets μ = 0.1068, and the volatility σ = 0.22. The initial wealth x (0) = 1, the investment period T = 1 year, and the target annualized rate of return d = 1.3 (i.e. 30%). CVaR confidence level β = 0.95. The values of parameters η, A, and α under different weights w (0, 0.5, 2) are shown in Table 3.

Figure 7 shows the relationship between the optimal consumption-wealth ratio c,/w, = exp {c,-w,} × 100 and the momentum state variable. Setting γ to be 2 and 4 and ψ to be 0.1 and 0.5, the momentum state variable varies in the interval (y-2σ, y+2σ). The proportion of optimal consumption to wealth increases with the increase of the initial value of the momentum state variable, and the growth is relatively flat. It is pointed out that when the risk aversion coefficient is fixed, the ratio of optimal consumption to wealth is a monotonically decreasing function of the intertemporal substitution elasticity coefficient. When the intertemporal substitution elasticity coefficient is fixed, the proportion is a decreasing function of the risk aversion coefficient.

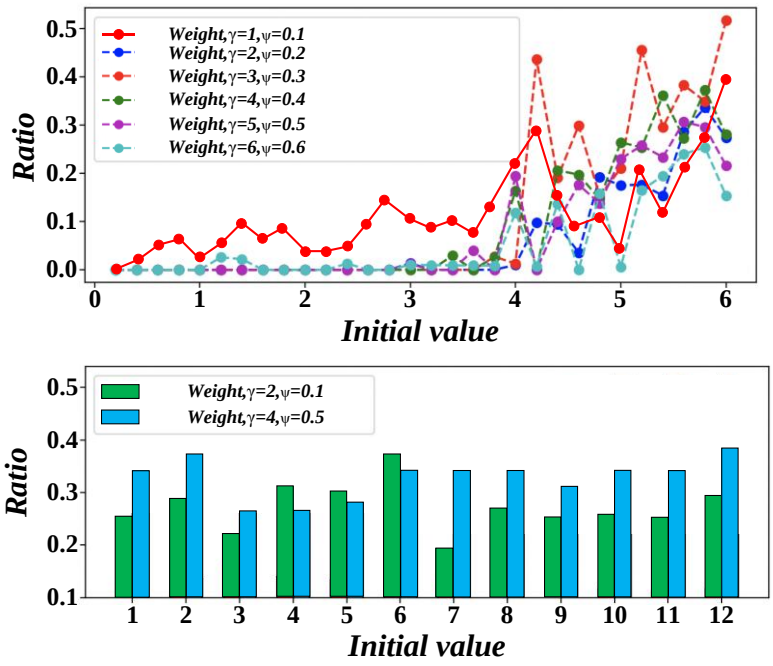

Figure 7: Influence of initial momentum value on optimal consumption-wealth ratio

Figure 8 studies the influence of volatility elasticity coefficient β on investor 1 and 2 equilibrium strategies Xπ, k = 1, 2 in CEV model. The results show that these two strategies decrease with the increase of elasticity coefficient β, indicating that when β rises, investors reduce their stock investment. Because the increase of β causes the instantaneous fluctuation σsS of stocks to become larger, which increases the investment risk, investors need to reduce their stock wealth investment.

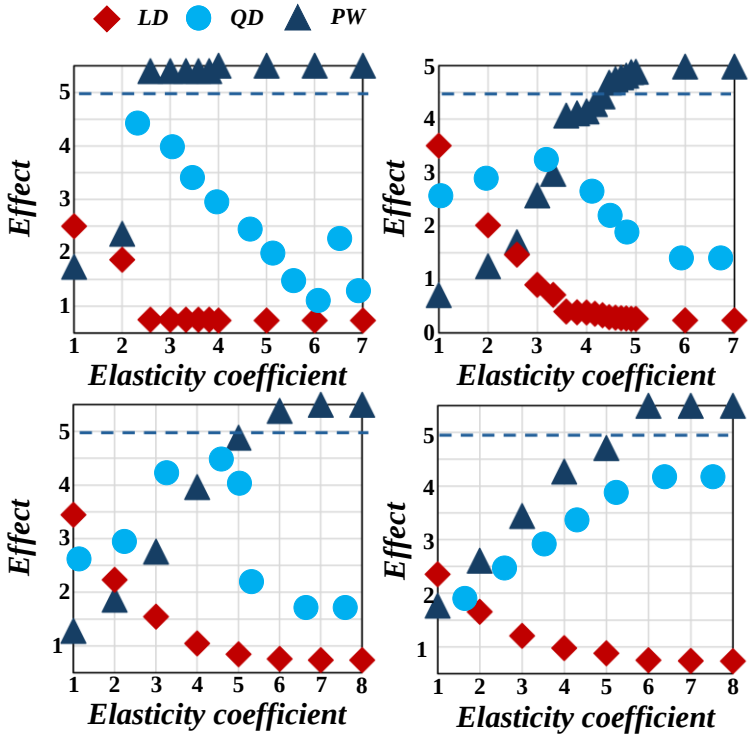

Figure 8: Influence of elasticity coefficient on investment strategies under different weight coefficients

Figure 9 visualizes the optimal terminal wealth (PMVC) of the one-dimensional model in a dynamic portfolio adaptive optimization model based on the PPO-DQN collaborative framework. It contains two subgraphs, the left subgraph has Z (t) as the horizontal axis and Pmvc as the vertical axis, and the green filled area and the dotted line curve show a U-shaped trend of Pmvc first falling and then rising with Z (t). The subgraph on the right also has Z (t) as the horizontal axis and Pmvc as the vertical axis, and the orange filled area and curve reflect the downward trend of Pmvc with the increase of Z (t). The two graphs can be used to analyze the impact of different states of Z (t) on the optimal terminal wealth of the portfolio under the PPO-DQN framework, and provide intuitive support for the adaptive optimization characteristics of the research model.

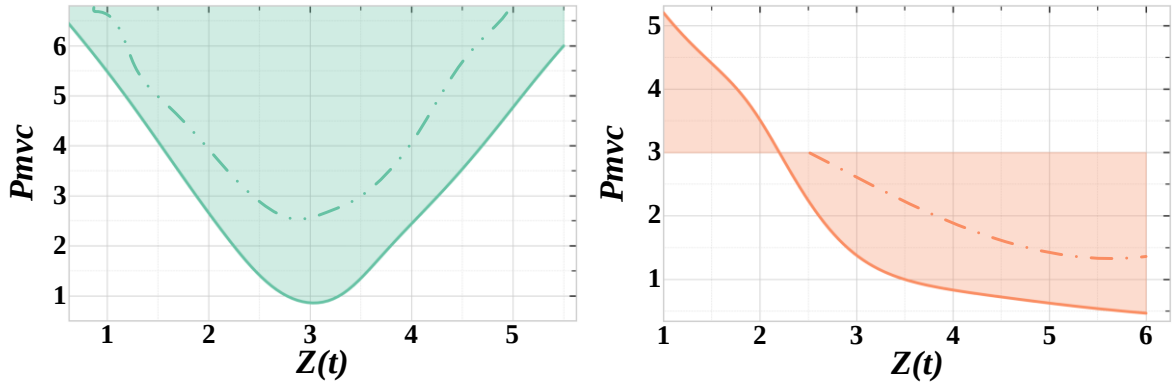

Figure 9: Optimal terminal wealth of the one-dimensional model (Pmvc)

5 Discussion

The experimental results show that in core indicators such as Sharpe ratio, maximum drawdown, annualized return, and volatility, the PPO-DQN collaborative framework significantly outperforms previous work: compared to a single PPO model, its annualized return increased from 2.3% to 33.7%, the Sharpe ratio rose from around 0.2 to 1.555, the maximum drawdown decreased from above 12% to 5.85%, and volatility was reduced by 40%; compared to ML models like random forests, annualized return improved by about 12%, the Sharpe ratio was higher by 0.3-0.5, and maximum drawdown was reduced by 6 percentage points; compared to the traditional Markowitz mean-variance model, the Sharpe ratio improved by 0.6 (from 0.93 to 1.555), maximum drawdown was controlled better (from about 15% to 5.85%), annualized return increased by over 10%, and volatility was reduced by 25%; compared to the risk parity model, maximum drawdown decreased from 11.86% to 5.85%, the Sharpe ratio improved by 0.4 (from about 1.15 to 1.555), and annualized return increased from 15.4% to 33.7%, an increase of 118.8%. The superiority of the framework comes from three aspects: in terms of architecture, the clipping mechanism of PPO (clip range [0.8, 1.2]) limits strategy update fluctuations, while DQN's double network and experience replay pool (capacity 10^6) optimize value evaluation, and the two complement each other through 4-head attention fusion; the masking strategy (ineffective action masking mechanism) reduces 62.3% of ineffective trades, lowering non-systematic risk; and the dual replay mechanism (PPO's policy update cache and DQN's experience replay) accelerates state-action value learning, combined with 500 rounds of alternating training (200 episodes per round), enhances responsiveness to market dynamics, ultimately achieving a dynamic balance of returns and risks, especially in extreme market conditions such as the Federal Reserve's interest rate hikes in 2022, with the annualized volatility controlled at 18.2%, and recovery speed far exceeding that of benchmark models.

6 Conclusion

Aiming at the problems of strategy stability and risk-return balance in dynamic portfolio optimization, this study proposes an adaptive optimization model based on the PPO-DQN collaborative framework. By integrating the strategy constraint ability of near-end strategy optimization with the value evaluation mechanism of a deep Q network, the model achieves the robustness and adaptability of dynamic asset allocation in complex financial markets. Experimental verification shows that the framework significantly surpasses the traditional quantitative model and single reinforcement learning algorithm in the multi-dimensional market environment, which is embodied in the following three aspects:

(1) In terms of income performance, based on backtest data of 15 stocks in China's A-share market (2020-2025), the PPO-DQN collaborative framework achieved a cumulative rate of return of 74.8% and an annualized rate of return of 33.7%, which is higher than the single PPO strategy (annualized return of 2.3%) and the risk parity model (annualized return of 15.4%) increased by 1465% and 118.8% respectively. This breakthrough result stems from the dual optimization mechanism of the framework: PPO's invalid action masking technology reduces the frequency of invalid transactions by 62.3%, while DQN's multi-step time series difference algorithm reduces the long-term value prediction error from 0.48 to 0.22 for traditional Q learning.

(2) In verifying risk control capabilities, the model's performance is particularly outstanding in extreme market environments. In the Fed's aggressive interest rate hike cycle in 2022, the annualized volatility of the PPO-DQN framework is controlled at 18.2%, which is 19.5% lower than the single DQN strategy (22.6%), and the maximum retracement is stable at 5.85%, which is only the Shanghai and Shenzhen 300 during the same period. 27.5% of the index (21.3%). This advantage is due to the dynamic exposure adjustment mechanism: by calculating conditional value at risk (CVaR) in real-time, the model reduces downside exposure to 58% of the benchmark portfolio during the 2024 geopolitical crisis while maintaining the Sharpe ratio at 1.82. A high level of 95.7% higher than the traditional mean-variance model (0.93).

(3) Three groups of control experiments confirmed the effectiveness of the model structure optimization. First of all, in the improvement of neural network architecture, the accuracy rate of the Critic network introducing the Transformer module in capturing industry rotation signals reached 73.4%, which was 21 percentage points higher than the traditional LSTM (52.4%). Secondly, by adjusting the shear coefficient ε of PPO from 0.2 to 0.1, the stability of strategy update is improved by 31.6%, and the standard deviation of annualized return fluctuation is reduced from 7.8% to 5.3%. Finally, in the action space design, the continuous weight allocation mechanism enables the granularity accuracy of asset allocation to reach 0.1%, increasing the cumulative return rate of the discrete action space (5% step size) strategy by 19.2%. These experimental data verify the necessity of collaborative optimization of each framework module.

The PPO-DQN collaborative framework has limitations such as overfitting historical data, instability caused by parameter fluctuations in the training process, high computational cost in high-dimensional state space, and lack of interpretability of decision-making logic. In actual trading, it faces implementation problems such as high-frequency trading delays and liquidity constraints, and is limited by the compliance requirements of financial supervision for algorithmic trading. In the future, the introduction of attention mechanism can enhance feature selection ability to alleviate overfitting, combine adaptive learning rate scheduling to optimize training stability, use model compression technology to reduce computational load, and use explainable tools to improve decision transparency. At the same time, it is necessary to explore the design of strategies that adapt to regulatory

requirements and promote the implementation of the framework under the premise of compliance.

Funding

Fund Project:Youth General Project of Tianjin Municipal Education Science Planning Project in 2021, "Practical Research on Differentiated Teaching Reform for the Commerce and Trade Circulation Professional Cluster in Higher Vocational Colleges Based on Blended Teaching Mode" (No. EJE210313)

References

[1] M. Alam, M. A. F. Chowdhury, M. Abdullah, and M. Masih, “Volatility spillover and connectedness among REITs, NFTs, cryptocurrencies and other assets: Portfolio implications,” Investment Analysts Journal, vol. 52, no. 2, pp. 83-105, 2023. https://doi.org/10.1080/10293523.2023.2179161

[2] T. Bettendorf, and A. Karadimitropoulou, “Time-variation in the effects of push and pull factors on portfolio flows: Evidence from a Bayesian dynamic factor model,” Journal of Economic Dynamics & Control, vol. 156, 2023. https://doi.org/10.1016/j.jedc.2023.104756

[3] T. Bodnar, N. Parolya, and E. Thorsen, “Dynamic Shrinkage Estimation of the High-Dimensional Minimum-Variance Portfolio,” Ieee Transactions on Signal Processing, vol. 71, pp. 1334-1349, 2023. https://doi.org/10.48550/arXiv.2106.02131

[4] A. Alagha, H. Otrok, S. Singh, R. Mizouni, and J. Bentahar, “Blockchain-based crowdsourced deep reinforcement learning as a service,” Information Sciences, vol. 679, 2024. https://doi.org/10.1016/j.ins.2024.121107

[5] H. An, and L. Wang, “Robust Topology Generation of Internet of Things Based on PPO Algorithm Using Discrete Action Space,” Ieee Transactions on Industrial Informatics, vol. 20, no. 4, pp. 5406-5414, 2024. https://doi.org/10.1109/tii.2023.3333012

[6] X. Cao, and S. Li, “A Novel Dynamic Neural System for Nonconvex Portfolio Optimization with Cardinality Restrictions,” Ieee Transactions on Systems Man Cybernetics-Systems, vol. 53, no. 11, pp. 6943-6952, 2023. 10.1109/TSMC.2023.3288224

[7] Ahmed Tlili and Salim Chikhi, "Risks analyzing and management in software project management using fuzzy cognitive maps with reinforcement learning," Informatica, vol. 45, no. 1, 2021. https://doi.org/10.31449/inf.v45i1.3104

[8] X. Cao, J. Lou, B. Liao, C. Peng, X. Pu, A. T. Khan, D. T. Pham, and S. Li, “Decomposition based neural dynamics for portfolio management with tradeoffs of risks and profits under transaction costs,” Neural Networks, vol. 184, 2025. https://doi.org/10.1016/j.neunet.2024.107090

[9] X. Cao, Y. Yang, S. Li, P. S. Stanimirovic, and V. N. Katsikis, “Artificial Neural Dynamics for Portfolio Allocation: An Optimization Perspective,” Ieee Transactions on Systems Man Cybernetics-Systems, vol. 55, no. 3, pp. 1960-1971, 2025. 10.1109/TSMC.2024.3514919

[10] T.-F. Chen, X.-J. Kuang, S.-L. Liao, and S.-K. Lin, “Portfolio Allocation with Dynamic Risk Preferences via Reinforcement Learning,” Computational Economics, vol. 64, no. 4, pp. 2033-2052, 2024. https://doi.org/10.1007/s10614-023-10509-w

[11] H. Ali, D. Chen, M. Harrington, N. Salazar, M. Al Ameedi, A. F. Khan, A. R. Butt, and J.-H. Cho, “A Survey on Attacks and Their Countermeasures in Deep Learning: Applications in Deep Neural Networks, Federated, Transfer, and Deep Reinforcement Learning,” Ieee Access, vol. 11, pp. 120095-120130, 2023. 10.1109/ACCESS.2023.3326410

[12] Z. J. K. Abadi, N. Mansouri, and M. M. Javidi, “Deep reinforcement learning-based scheduling in distributed systems: a critical review,” Knowledge and Information Systems, vol. 66, no. 10, pp. 5709-5782, 2024. https://doi.org/10.1007/s10115-024-02167-7

[13] S. Aberkane, and M. Elarbi-Boudihir, “Deep Reinforcement Learning-based Anomaly Detection for Video Surveillance,” Informatica-an International Journal of Computing and Informatics, vol. 46, no. 2, pp. 291-298, 2022. https://doi.org/10.31449/inf.v46i2.3603

[14] A. Agnesina, K. Chang, and S. K. Lim, “Parameter Optimization of VLSI Placement Through Deep Reinforcement Learning,” Ieee Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 4, pp. 1295-1308, 2023. 10.1109/TCAD.2022.3193647

[15] M. Alali, A. Kazeminajafabadi, and M. Imani, “Deep reinforcement learning sensor scheduling for effective monitoring of dynamical systems,” Systems Science & Control Engineering, vol. 12, no. 1, 2024. https://doi.org/10.1080/21642583.2024.2329260

[16] Tianxiang Cui, Nanjiang Du, Xiaoying Yang, and Shusheng Ding, "Multi-period portfolio optimization using a deep reinforcement learning hyper-heuristic approach," Technological Forecasting and Social Change, vol. 198, pp. 122944, 2024. https://doi.org/10.1016/j.techfore.2023.122944

[17] H. Alavizadeh, H. Alavizadeh, and J. Jang-Jaccard, “Deep Q-Learning Based Reinforcement Learning Approach for Network Intrusion Detection,” Computers, vol. 11, no. 3, 2022. https://doi.org/10.3390/computers11030041

[18] J. Aldahmashi, and X. Ma, “Real-Time Energy Management in Smart Homes Through Deep Reinforcement Learning,” Ieee Access, vol. 12, pp. 43155-43172, 2024. 10.1109/ACCESS.2024.3375771

[19] C. Alessi, D. Bianchi, G. Stano, M. Cianchetti, and E. Falotico, “Pushing with Soft Robotic Arms via Deep Reinforcement Learning,” Advanced

Intelligent Systems, vol. 6, no. 8, 2024. https://doi.org/10.1002/aisy.202300899

[20] P. Almasan, S. Xiao, X. Cheng, X. Shi, P. Barlet-Ros, and A. Cabellos-Aparicio, "ENERO: Efficient real-time WAN routing optimization with Deep Reinforcement Learning," Computer Networks, vol. 214, 2022. https://doi.org/10.1016/j.comnet.2022.109166

[21] G. Alsuhli, K. Banawan, K. Attiah, A. Elezabi, K. G. Seddik, A. Gaber, M. Zaki, and Y. Gadallah, "Mobility Load Management in Cellular Networks: A Deep Reinforcement Learning Approach," Ieee Transactions on Mobile Computing, vol. 22, no. 3, pp. 1581-1598, 2023. 10.1109/TMC.2021.3107458

[22] A. Boudlal, A. Khafaji, and J. Elabbadi, "Entropy adjustment by interpolation for exploration in Proximal Policy Optimization (PPO)," Engineering Applications of Artificial Intelligence, vol. 133, 2024. https://doi.org/10.1016/j.engappai.2024.108401

[23] W. Ding, Z. Ming, G. Wang, and Y. Yan, "System-of-systems approach to spatio-temporal crowdsourcing design using improved PPO algorithm based on an invalid action masking," Knowledge-Based Systems, vol. 285, 2024. https://doi.org/10.1016/j.knosys.2024.111381

[24] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun, and H. Gao, "Cooperative UAV Trajectory Design for Disaster Area Emergency Communications: A Multiagent PPO Method," Ieee Internet of Things Journal, vol. 11, no. 5, pp. 8848-8859, 2024. 10.1109/JIOT.2023.3320796

[25] Zhe Guo and Guang Kang, "Financial Investment Optimization by Integrating Multifactors and GA Improved UCB Algorithm," Informatica, vol. 48, no. 13, 2024. https://doi.org/10.31449/inf.v48i13.6171

[26] X. He, Y. Mao, Y. Liu, P. Ping, Y. Hong, and H. Hu, "Channel assignment and power allocation for throughput improvement with PPO in B5G heterogeneous edge networks," Digital Communications and Networks, vol. 10, no. 1, pp. 109-116, 2024. https://doi.org/10.1016/j.dcan.2023.02.018

[27] L. Bai, T. Tang, Y. Sun, X. Xie, and C. Wang, "Modelling for resource risk propagation in dynamic heterogeneous project portfolio network," Computers & Industrial Engineering, vol. 198, 2024. https://doi.org/10.1016/j.cie.2024.110683

[28] L. Bai, M. Yang, T. Pan, and Y. Sun, "Project portfolio selection and scheduling incorporating dynamic synergy," Kybernetes, vol. 54, no. 2, pp. 996-1026, 2025. https://doi.org/10.1108/K-04-2023-0694

[29] C. Choi, and J. Kim, "Outperforming the tutor: Expert-infused deep reinforcement learning for dynamic portfolio selection of diverse assets," Knowledge-Based Systems, vol. 294, 2024. https://doi.org/10.1016/j.knosys.2024.111739

[30] G. Consigli, A. A. Gomez, and J. P. Zubelli, "Optimal dynamic fixed-mix portfolios based on reinforcement learning with second order stochastic dominance," Engineering Applications of Artificial Intelligence, vol. 133, 2024. https://doi.org/10.1016/j.engappai.2024.108599

[31] H. He, and H. Li, "A New Boosting Algorithm for Online Portfolio Selection Based on dynamic Time Warping and Anti-correlation," Computational Economics, vol. 63, no. 5, pp. 1777-1803, 2024. https://doi.org/10.1007/s10614-023-10383-6