

Developing Financial Intelligent Investment Methods Using an Improved-PPO Algorithm

Cheng Yang¹, Haoyang Zeng², Shilong Han², and Yifan Yu⁴

¹ Associate Professor, Director of Professional Degree Office, School of Accounting, College of Intelligent Accounting Management, Guangdong University of Finance and Economics, Guangzhou, 510320, China

² Master's Candidate in Accounting, School of Accounting, Guangdong University of Finance and Economics, Guangzhou 510320, China

³ Master's Candidate in Accounting, School of Accounting, Guangdong University of Finance and Economics, Guangzhou 510320, China

⁴ Doctoral Candidate, School of Management, Jinan University, Guangzhou, 510632, China, E-mail: YC20250227@126.com (corresponding author).

Project Management

Received December 5, 2025; revised February 5, 2026; July 2, 2026; accepted October 13, 2026
Available online October 26, 2026

Abstract: In the context of dynamic financial markets and the limitations of traditional and existing intelligent methods in balancing risk and return, this study targets the poor generalization and weak risk resistance of investment strategies in dynamic stock market environments. It proposes a Proximal Policy Optimization (PPO)-based investment model that integrates a deep sequence-encoding module and a dynamic, risk-adjusted advantage mechanism. The model employs a hybrid encoder that combines Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures to process multi-granularity time-series data (minute, daily, weekly), effectively compensating for deficiencies in state observations. Meanwhile, a dynamic risk adjustment advantage function is introduced, in which a Multi-Layer Perceptron (MLP) generates dynamic penalty coefficients to enhance risk control, working synergistically with the actor-critic structure of PPO to optimize decisions. Experimental results show that incorporating the deep sequence encoding module yields the fastest convergence in ablation studies, achieving an average round reward of 1.7. In cross-industry Exchange Traded Fund (ETF) portfolio management, the model achieves a cumulative return of 35.85% and a 99% risk tolerance of -2.50%, outperforming comparable models. During simulated high-volatility periods, peak volatility is capped at 5.5%, and the dynamic risk penalty coefficient closely tracks market fluctuations. This research provides a scientific and practical tool for A-share stock timing and portfolio management, enriches methodologies for financial time-series processing, and promotes the practical application of deep reinforcement learning in intelligent finance.

Keywords: Intelligent financial investment; improved proximal strategy optimization algorithm; deep sequence encoding module; advantages of dynamic risk adjustment; multi-granularity time-series financial data.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2025-304

1. Introduction

Against the backdrop of the deep integration of the open financial ecosystem and Artificial Intelligence (AI) technology, financial investment decision-making is accelerating its transformation from experience-driven to data-driven (Weber et al., 2023). The explosive growth of multidimensional market data and the dynamic volatility of the market environment have placed greater demands on the real-time adaptability and risk-control capabilities of investment strategies. However, the nonlinear characteristics of the financial market and sudden disturbances make it difficult for traditional decision-making models to accurately capture dynamic return opportunities, while the dynamic balance between risk and return has become a problem that urgently needs to be solved (Sahoh and Choksuriwong, 2023). Traditional financial investment decision-making has long relied on human experience, judgment and static quantitative model deduction. The former is limited by subjective cognitive biases and information-processing efficiency, making it difficult to cope with high-frequency, multidimensional data shocks and resulting in a significant lag in response to market changes (Szukits and Móricz, 2023). The latter is largely based on fixed market assumptions, ignores the time-evolution characteristics of asset allocation, and is prone to strategy failure in dynamic markets (Mazzucato, 2023). Although existing intelligent investment research has introduced machine learning or early reinforcement learning methods, it generally suffers from limitations

such as unstable training processes, low sample utilization efficiency, and insufficient generalization across heterogeneous markets, making it difficult to balance the scientific rigor and practical decision-making (Buathong et al., 2023). Therefore, how to construct an intelligent investment method that can adapt to dynamic market environments, balance training stability with decision-making efficiency, and strike a precise risk-return balance remains a key issue in intelligent financial investment.

Existing research focuses on the innovation and optimization of intelligent financial investment methods and systematically explores them from multiple dimensions. These paths include technical path exploration, strategy system construction and risk control adaptation, yielding diverse theoretical achievements and practical experience. Odonkor et al. (2024) conducted a systematic evaluation of the innovative role of AI integration in the accounting field through a systematic literature review and bibliometric analysis. The study denotes that AI can significantly enhance the accuracy and efficiency of financial reporting and help assist strategic decision-making. However, it also faces important challenges, such as a talent shortage, data privacy concerns, and high application costs, and needs to adopt a balanced strategy to address the resulting resistance to change. Li et al. (2023) developed a new set of evaluation indicators, including lag relative error, to address the lag problem in recurrent neural network models when predicting low volatility financial time series. To this end, they used state analysis and a linear regression model to examine the causes of the problem in depth. The study finds that all mainstream recurrent neural network models exhibit this lag phenomenon, whose root cause is the failure of nonlinear functions in the model, causing the prediction model to degenerate into a linear model. This conclusion offers a new approach to optimizing intelligent financial investment methods. Hambly et al. (2023) focused on the key issue of improving the quality of financial decision-making and systematically reviewed the latest progress and applications of reinforcement learning methods. First, they introduced the theoretical framework of Markov decision processes, then elaborated on value-based and policy-based algorithms, as well as deep reinforcement learning models. Results reveal that reinforcement learning technology allows new applications and improvements to various financial decisions, such as optimal execution and portfolio optimization.

Proximal Policy Optimization (PPO), as a representative deep reinforcement learning algorithm within the Actor-Critic architecture, ensures training stability by constraining the ratio between new and old policies and offers high sample efficiency and fast convergence. It has been adapted to complex continuous decision-making scenarios such as robot control and resource scheduling, and provides technical support for intelligent decision-making needs in fields such as finance (An and Wang, 2023). In response to the control problem caused by the high nonlinearity of chaotic systems, Yau et al. (2024) used the PPO algorithm to train control agents. Experiments show that this method can quickly achieve vibration control without prior knowledge of the system model and exhibits excellent robustness to random noise interference. In response to the challenge of unpredictable dynamic scheduling in production systems, Zhao et al. (2023) used the PPO algorithm and combined it with an attention mechanism policy network for real-time decision-making. This method demonstrated excellent performance and robustness in dynamic, flexible workshop scheduling tasks, offering a new approach to applying PPO to other complex resource scheduling scenarios, such as finance. To address the problem of low efficiency in continuous action processing and model interaction in autonomous driving, Wu et al. (2022) proposed a Dyna-PPO method based on a Gaussian process. This method introduced a Gaussian process model to establish a bidirectional interaction mechanism between the world and policy models. Experimental findings demonstrate that this method significantly improves the driving efficiency and robustness of vehicle control, and its successful practice also provides a reference for intelligent decision-making in the financial field. Guan et al. (2022) proposed a collaborative trajectory design method based on multi-agent PPO that effectively addresses the problem of controlling multiple Unmanned Aerial Vehicles (UAVs) for emergency communication in disaster areas. Thanks to its efficient learning mechanism, this method offers fast convergence and efficient resource allocation for UAV deployment and communication optimization. This achievement also serves as a reference for the collaborative optimization of complex distributed systems in the financial sector.

In summary, existing research has demonstrated preliminary applications of AI and reinforcement learning in multiple fields, such as finance and control, and has proven the stable and efficient advantages of the PPO algorithm in continuous decision-making. However, it still has limitations in financial scenarios, such as incomplete state observation and weak risk-return balancing mechanisms. To solve these deficiencies, this study designs an enhanced PPO algorithm that integrates a Deep Sequence Encoding Module (DSEM) and a Risk-Adjusted Advantage (RAA). Through multi-granular temporal feature fusion and real-time risk penalty, it specifically optimizes the state perception and risk management capabilities of financial decision-making. This research addresses the issues of poor generalization and limited robustness of investment strategies in dynamic stock market environments, providing feasible solutions for stock timing and portfolio management in the A-share market, while also promoting the practical application of Deep Reinforcement Learning (DRL) in intelligent finance. The novelty of this research lies in the construction of a PPO decision framework that integrates DSEM. It uses a hybrid encoder that combines Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) to encode multi-frequency financial data, thereby addressing the observability challenge for the state part. Simultaneously, it introduces the RAA function, which generates dynamic penalty coefficients via a Multi-Layer Perceptron (MLP) to adjust risk penalties and adapt to high-volatility financial scenarios. Ultimately, this provides a new paradigm for intelligent investment that combines profitability and robustness, thereby upgrading financial decision-making from data-driven to intelligent risk control.

2. Methods and Materials

2.1. Construction of PPO Decision Model Based on DSEM

The core of intelligent financial investment is to build a decision-making model adapted to dynamic markets. This scenario is essentially a partially observable Markov decision-making process, in which agents often suffer from decision bias due

to their inability to fully capture the multi-granular nature of market states. Given the PPO algorithm's advantage in training stability for continuous action decision-making, this study introduces DSEM and integrates it into the PPO Actor-Critic architecture to optimize the perception layer. The relevant interaction logic is shown in Fig. 1. This integration effectively compensates for the shortcomings of traditional PPO in observing financial states.

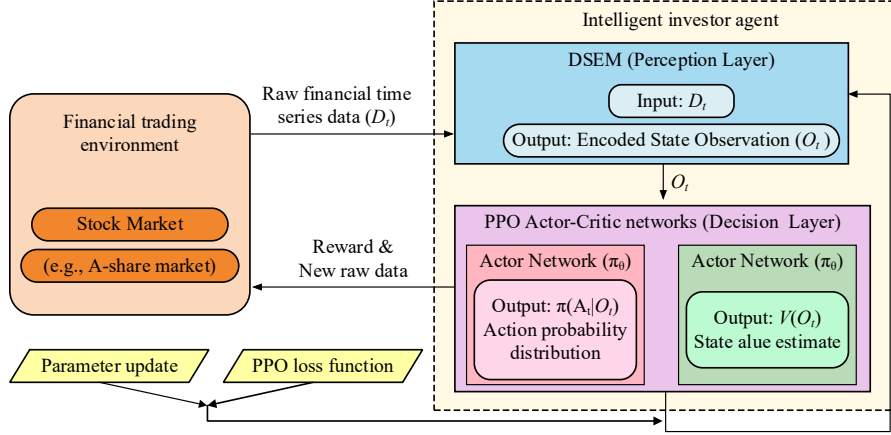


Fig. 1. Interaction framework between DRL agents and the financial environment

As shown in Fig. 1, the proposed improved method constructs a closed-loop interaction framework between the DRL agent and the financial environment. The agent receives raw time-series data from the financial trading environment. This data first enters the DSEM module, which serves as the perception layer, for deep sequence encoding to effectively capture hidden market trends and sentiment, thereby addressing the state-missing problem in certain observable Markov decision-making processes (Zhang et al., 2023). The raw financial time-series data is defined and calculated as shown in Eq. (1).

$$F_t = \{f_{t,1}, f_{t,2}, \dots, f_{t,m}\}, t \in [1, T] \quad (1)$$

In Eq. (1), F_t is the original financial time series data sample at time t . T is the total transaction time step within the study period, $f_{t,i}$ is the i th financial feature value at time t , m is the total number of features, containing key data reflecting market dynamics, and t is the time index, covering all transaction nodes within the study period. The DSEM modules encoding output formula is shown in Eq. (2).

$$O_t = \text{DSEM}(F_t; \omega) \quad (2)$$

In Eq. (2), O_t represents the final encoded state observation of the DSEM module at time t . $\text{DSEM}(\cdot)$ represents the core operation function of the DSEM, which includes sub-processes such as feature cleaning and temporal correlation capture. ω represents the set of trainable parameters for the DSEM module, which optimizes the encoding accuracy through training. The encoded state observation produced by DSEM is then used as input to the PPO Actor-Critic network (decision layer), which outputs trading actions, evaluates state value and finally updates parameters through the loss function (Li and Zhu, 2024; Li et al., 2023). This optimized design of the perception layer effectively enhances the agent's ability to understand complex financial information and provides reliable state input for stable PPO decision-making. The specific process by which DSEM fuses multi-granularity temporal features into a unified observation vector is shown in Fig. 2.

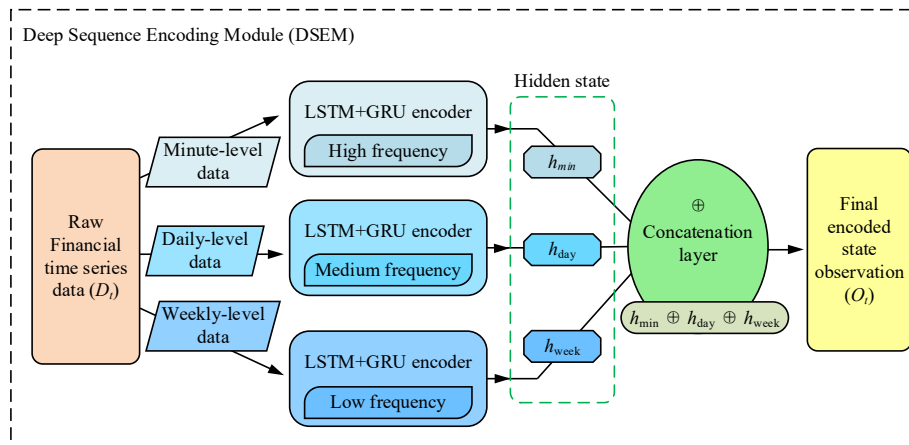


Fig. 2. Flowchart of the deep sequence state coding module

As shown in Fig. 2, DSEM uses a parallel processing mechanism to split the original financial time series data into three observation dimensions: minute, day, and week, to capture market dynamics at different time scales (Yan et al., 2023). The formula for extracting multi-granularity financial time-series subsequences is given by Eq. (3).

$$F^g = \{F_{t_k} \mid t_k = k \cdot \Delta t_g, k = 1, 2, \dots, K_g\}, g \in \{m, d, w\} \quad (3)$$

In Eq. (3), F^g is a financial time series subsequence at a certain time granularity g , used to capture the market dynamics features at that granularity. F_{t_k} is the sample with time step t_k in the original financial time series data. t_k is the time step corresponding to the k th sample at granularity g , determined by the product of k (sample index) and the time step size Δt_g at that granularity. Δt_g is the time step size corresponding to granularity g . K_g represents the total number of samples in the subsequence at granularity g . m , d , and w correspond to the observation dimensions at the minute, daily, and weekly levels, respectively. The sequence data for each dimension are input to the corresponding hybrid encoder, combining LSTM and GRU, to extract features and generate their respective hidden states. The calculation formula for the hidden state of a single granularity LSTM is shown in Eq. (4).

$$h_L^g = \text{LSTM}(F^g; \theta_L^g) \quad (4)$$

In Eq. (4), h_L^g represents the hidden state output by the LSTM encoder at granularity g , used to capture correlation features spanning a long time in the temporal data at this granularity. $\text{LSTM}(\cdot)$ represents the core operational functions of the LSTM network, including the forget gate, input gate, cell state update and output gate mechanisms. θ_L^g represents the set of trainable parameters of the LSTM encoder corresponding to granularity g , including the weight matrix from the input layer to the major gates, bias terms, etc. The LSTM parameters for different granularities are optimized independently to adapt to the data's time-scale differences at each granularity. The formula for computing the hidden state of a single-granularity GRU is given by Eq. (5).

$$h_G^g = \text{GRU}(F^g; \theta_G^g) \quad (5)$$

In Eq. (5), h_G^g represents the hidden state output by the GRU encoder at granularity g , used to efficiently capture short-term fluctuation features in time-series data at this granularity, thereby compensating for the computational redundancy of LSTM in short-term feature capture. $\text{GRU}(\cdot)$ represents the core operational function of the GRU, which simplifies the recurrent neural network structure through the reset and update gates. θ_G^g is the set of trainable parameters of the GRU encoder corresponding to granularity g . This includes the weight matrix and bias terms of the reset gate, update gate, and hidden candidate states, which are independent of θ_L^g , ensuring that GRU can autonomously learn and adapt to the data's short-term feature-extraction rules at each granularity. The formula for fusing single-granularity mixed hidden states is given in Eq. (6).

$$h^g = \alpha^g \cdot h_L^g + (1 - \alpha^g) \cdot h_G^g, \alpha^g \in [0, 1] \quad (6)$$

In Eq. (6), h^g represents the final hidden state output by the hybrid encoder at the granularity level g , which integrates the long-term dependency features of LSTM and the short-term dynamic features of GRU, and is the core feature vector at this granularity level. α^g represents the weight coefficient of the LSTM hidden state at this granularity level, with a value in $[0, 1]$ determined during model training. Specifically, the fusion weight α^g is parameterized by a learnable scalar and constrained using a sigmoid activation function to strictly ensure $\alpha^g \in [0, 1]$, thereby maintaining the stability of feature integration. This weighted fusion method can avoid overreliance on a single model (such as LSTM or GRU) for a particular type of feature. These hidden states are then concatenated and aggregated in the connection layer to form a unified final encoded state observation (Zhang and Zhou, 2022). The formula for concatenating and aggregating multi-granularity hidden states is shown in Eq. (7).

$$O_t = W \cdot [h^m; h^d; h^w] + b \quad (7)$$

In Eq. (7), O_t represents the final encoded state observation produced by the DSEM module at time t . W is the trainable weight matrix of the connection layer, which maps the concatenated high-dimensional multi-granular features to a unified dimension suitable for the decision layer. $[h^m; h^d; h^w]$ is the high-dimensional feature vector formed by concatenating the minute-level hidden state h^m , the daily-level hidden state h^d , and the weekly-level hidden state h^w column by column. $\cdot$ is the vector concatenation symbol to ensure that multi-granular features are not lost. b is the

trainable bias term of the connection layer, which is used to adjust the overall offset of the feature vector and improve the model's adaptability to the offset of financial data. This linear transformation process can effectively integrate minute-level short-term fluctuations, daily-level medium-term trends, and weekly-level long-term patterns to form a complete observation of the market state. This multi-dimensional sequence fusion design greatly enriches the agent's perception of short-term market fluctuations and long-term trends, thereby effectively making up for the shortcomings of incomplete state observation in traditional PPO when processing high-dimensional time series data, providing accurate and reliable input for the subsequent Actor-Critic network (decision layer), enabling the agent to make decisions by simultaneously utilizing short-term fluctuations and long-term trends.

2.2. Dynamic Advantage Function Evaluation Strategy With Integrated Risk Constraints

The research has developed a PPO decision model using DSEM, which effectively alleviates incomplete state observation in the stock market by fusing financial time series data across multiple time granularities. However, the advantage estimation used by traditional PPO is based solely on maximizing returns and is insufficiently sensitive to risk under high market volatility or sudden risk events, which can easily lead to a lack of strategy resilience (Jin et al., 2024). To this end, the research proposes a dynamic advantage function evaluation strategy with integrated risk constraints, which embeds the real-time risk adjustment mechanism into the advantage calculation and links it to the observed state output from the DSEM to achieve risk-return synergistic optimization. This improved logic is shown in Fig. 3.

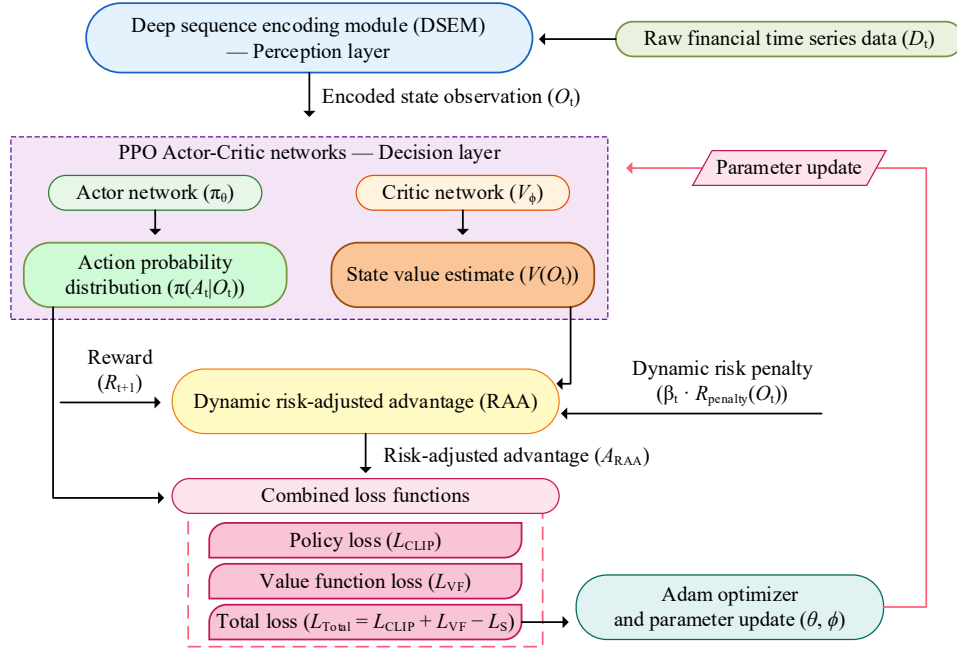


Fig. 3. Overall architecture diagram of the improved PPO algorithm

As shown in Fig. 3, the proposed PPO-RAA model clearly demonstrates the closed-loop fusion of the DSEM perception layer and the PPO decision layer. The original financial data is encoded into state observations by the DSEM and used as inputs to the Actor and Critic networks, which output action probability distributions and state values (Zhang et al., 2023). The core RAA calculation module integrates immediate rewards, value estimation, and adaptive risk-penalty terms derived from the state, yielding an adjusted advantage function that balances risk and return (Yeganeh et al., 2023). The policy network and advantage function jointly calculate the total loss function, driving parameter updates (Uysal et al., 2023). To ensure training stability and sample efficiency, policy loss optimization must be constrained by pruning the ratio of new to old policies. The specific mechanism of its pruning objective function is shown in Fig. 4.

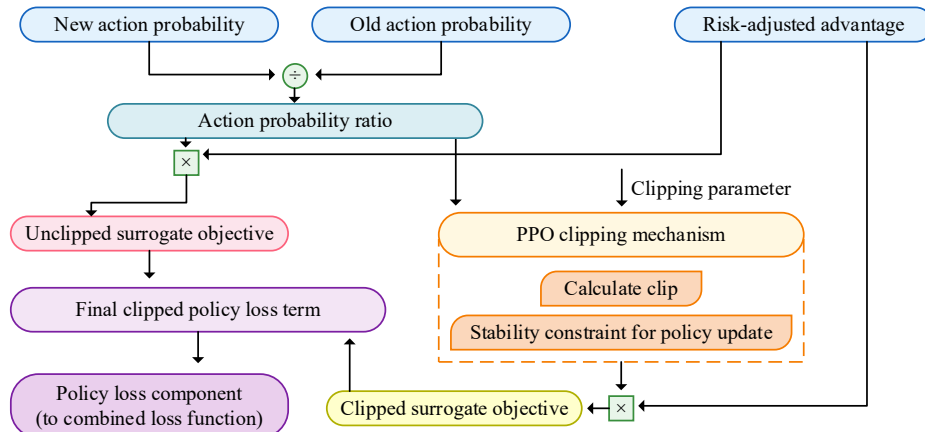
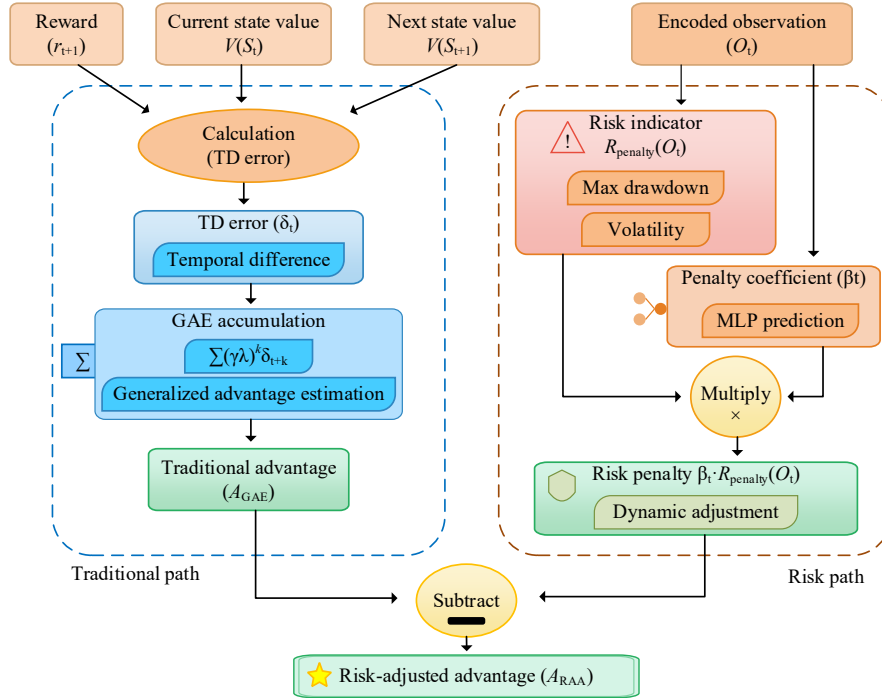


Fig. 4. Pruning objective function mechanism diagram of PPO algorithm

As shown in Fig. 4, to ensure the stability and effectiveness of policy updates, the core of policy loss calculation is to control the difference between the new and old policies. First, the action probabilities of the new and old policies are divided to obtain the action-probability ratio, which is then used as input along with the RAA proposed in the study (Rafiei et al., 2021). After that, the ratio is divided into two paths. One is the unpruned target that is directly multiplied by the dominance function. The other is first processed by the PPO pruning mechanism, which is constrained to a safe range by the pruning parameters, and then multiplied by the dominance function to obtain the pruned target (Cheng et al., 2024). The formula for the clipped surrogate objective function is given by Eq. (8).

$$L^{CLIP}(\theta) = \hat{E}_t \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon + \epsilon, 1 + \epsilon) \hat{A}_t) \right] \quad (8)$$

In Eq. (8), θ denotes the policy parameters, and $r_t(\theta) = \pi_\theta(a_t | s_t) / \pi_{\theta_{old}}(a_t | s_t)$ represents the probability ratio between the new and old policies. The term ϵ is a hyperparameter that defines the clipping range (set to 0.2 in this study) to prevent excessive policy updates. $\hat{A}_t$ is the advantage estimate, which integrates the risk constraints proposed in this paper. The final policy loss term is determined by taking the minimum of the two target values, ensuring that the policy does not fluctuate drastically when exploring new action spaces and thereby stabilizing the training process. The dominance function used in the study integrates real-time risk constraints into the evaluation. Its specific calculation logic and risk integration principle are shown in Fig. 5.


Fig. 5. Schematic diagram of risk-constrained advantage function

As shown in Fig. 5, the computational process of the proposed RAA realizes the deep coupling between the traditional Generalized Advantage Estimation (GAE) and real-time risk avoidance. The specific mechanism processes two information flows in parallel: on the one hand, the traditional path calculates the standard advantage estimate by recursively accumulating time-series difference errors. On the other hand, the risk path calculates the risk penalty index in real time from the DSEM's encoded state observations and generates a risk penalty term by dynamically weighting the adaptive penalty coefficients predicted by the (Pacal, 2024). Finally, the risk penalty term is subtracted from the standard advantage estimate to obtain the final RAA function (Shu et al., 2024). The calculation formula for the RAA function is shown in Eq. (9).

$$A_t^{RAA} = A_t^{GAE} - \beta_t \cdot R_{penalty}(O_t) \quad (9)$$

In Eq. (9), A_t^{GAE} represents the standard generalized advantage estimate calculated via the Temporal Difference (TD) error. $R_{penalty}(O_t)$ is the risk indicator derived from the current encoded state O_t , and β_t is the dynamic penalty coefficient output by the MLP network, which adaptively scales the penalty intensity in response to market volatility. This design can guide the agent to make more conservative decisions under high-risk or high-volatility conditions, effectively achieving a precise balance between risk and return. The specific policy update iteration process will guide the optimization

of the loss function based on this RAA function, as shown in Fig. 6.

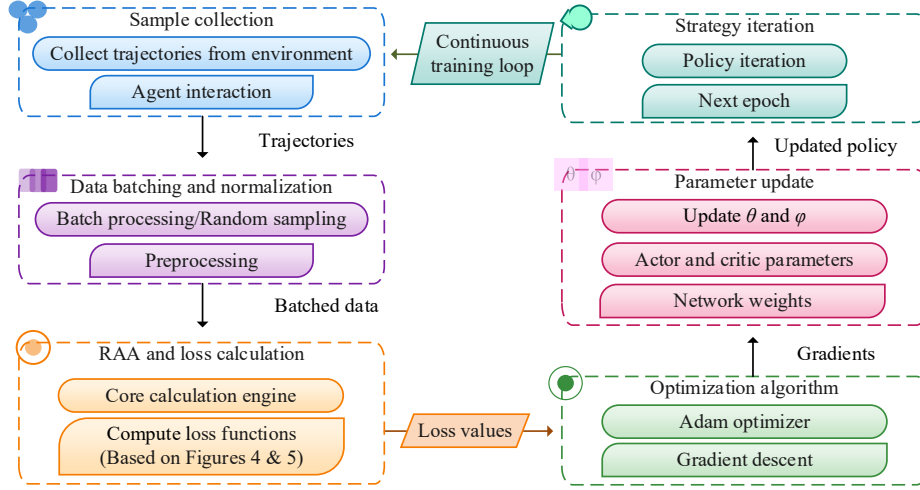


Fig. 6. Flowchart of intelligent investment strategy training

In Fig. 6, the training process of the intelligent investment strategy follows a continuous closed loop. The process begins with the interaction between the agent and the environment, which collects samples and yields transaction trajectories. The collected trajectory data is then processed with batch processing and normalization to transform it into batch data. Then, the core computing engine calculates the RAA and total loss function from the batch data using the pruning objective function mechanism and the risk-constraint advantage function logic of the PPO algorithm, thereby obtaining the loss value (Liu et al., 2023). These loss values are sent to the optimizer to compute the gradient, which is then used to update the parameters of the strategy and value networks, completing the strategy update. Finally, the updated strategy continues to interact with the environment to initiate the next training iteration, thereby forming a continuous optimization process. How the strategy network outputs continuous actions based on the current state and how the value network evaluates its value constitute the decision details of the next level, as shown in Fig. 7.

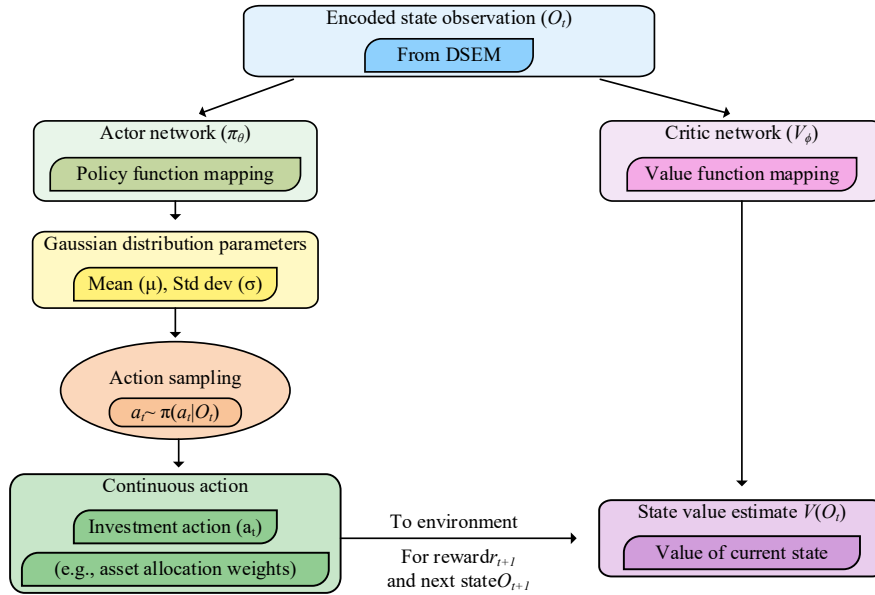


Fig. 7. Flowchart of continuous action space decision-making based on Actor-Critic

As shown in Fig. 7, this study, based on an Actor-Critic structure, details the decision-making process in a continuous action space. This process uses the encoded state observations produced by the DSEM as the sole input, which are then fed in parallel into the policy and value networks. The policy network is responsible for generating actions. It first outputs the Gaussian distribution parameters (mean and standard deviation) that represent the continuous action space, and then samples the final investment actions, such as asset allocation weights, from this distribution via an action-sampling mechanism. The value network is responsible for evaluating the current state and directly outputting the value estimate (Świechowski et al., 2022). The generated actions are executed in the financial environment, generating rewards and the next state, preparing for the next decision cycle. This mechanism is closely integrated with the training process of the intelligent investment strategy. The training process updates the policy and value network parameters by optimizing the total loss, computed from the value estimate, reward, and RAA derived from the encoded state observations, thereby enabling the strategy to make high-yield, low-risk, continuous investment decisions in complex financial markets.

The proposed method is named PPO Algorithm Fused with DSEM and RAA for Financial Intelligent Investment Method (PPO-DRF).

3. Results

3.1. Experimental Settings

3.1.1. Dataset and preprocessing

To ensure data consistency and reproducibility, the experimental dataset consolidated 20 highly liquid constituent stocks from the US S&P 500 index and 10 representative cross-industry Exchange-Traded Funds (ETFs) from the A-share market, covering the period from January 1, 2018, to December 31, 2023. Minute-level historical data, essential for the multi-granularity DSEM encoder, was acquired via the Yahoo Finance Premium API to guarantee intraday continuity. The data were preprocessed into three synchronized frequency levels (minute, daily, and weekly) and normalized using a rolling-window approach to prevent look-ahead bias.

3.1.2. Baseline models and hyperparameters

The proposed PPO-DRF model is benchmarked against a hierarchical set of baselines to comprehensively validate its efficacy. For the primary performance comparison shown in the main results, the study compared the investment strategy performance of the PPO-DRF model with three representative models: PPO, a classic DRL baseline serving as the foundational ablation target. Deep Deterministic Policy Gradient (DDPG) is a standard off-policy DRL algorithm commonly used in continuous action spaces. The Buy-and-Hold (B&H) strategy serves as a non-algorithmic benchmark for market performance. Additionally, to benchmark against broader financial standards and supervised learning methods (tabulated in the supplementary analysis). This study introduced supplementary baselines, including traditional financial strategies. These strategies include Risk Parity (RP), Mean-Variance Optimization (MVO), and a momentum strategy, as well as a supervised learning model, Gradient-Boosted Decision Tree (XGBoost), which was trained on the same multi-frequency features to distinguish RL-specific decision gains from prediction gains. Finally, to validate the architectural superiority of the proposed DSEM module (as tabulated in the ablation analysis), the study also implemented architectural baselines, including transformer-based encoders, Temporal Convolutional Networks (TCNs), and Standard LSTMs as comparative feature-extraction backbones.

For the PPO-DRF configuration, the clipping parameter δ is set to 0.2, and the GAE parameters γ (discount factor) and λ (smoothing parameter) are set to 0.99 and 0.95, respectively. The fusion weights in DSEM are constrained to $[0, 1]$.

3.1.3. Computational environment

All experiments were implemented using PyTorch (v1.10) and the Stable-Baselines3 (v1.4) library. Training was conducted on a high-performance workstation equipped with an Intel Core i9-10900K CPU @ 3.70 GHz, 64GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. The average training duration for the PPO-DRF model was approximately 14 hours for 1 million time steps.

3.2. Performance Comparison and Verification of PPO-DRF Model

To comprehensively validate the efficacy of the PPO-DRF model, this study evaluated its investment performance against the primary baselines defined in Section 3.1.2. The analysis begins with an overall performance comparison to demonstrate the proposed model's comprehensive advantages in balancing high returns and low risk. The specific results are shown in Fig. 8.

As shown in Fig. 8(a), the PPO-DRF model achieved a cumulative net asset value of approximately 3.95 at the end of 500 trading days, significantly higher than PPO (approximately 3.65), DDPG (approximately 3.35), and B&H (approximately 2.95), demonstrating optimal profitability. In Fig. 8(b), comparing key investment metrics, PPO-DRF ranked first with an annualized return of 27.5%, while maintaining the lowest Max Drawdown (only -18.0%), far lower than DDPG's -22.0% and B&H's -45.0%, ultimately resulting in a Sharpe ratio of 2.65, significantly better than PPO's 1.95, proving its superior risk-adjusted return performance. Fig. 8(c) showed that the monthly return distribution of PPO-DRF had the narrowest box and the highest median (approximately 2.5%), indicating more stable returns and the best control over the worst monthly loss (-5.0%). The rolling volatility curve in Fig. 8(d) clearly shows that the volatility of PPO-DRF consistently remained at the bottom, with a peak of only about 5.5% during periods of high volatility (around trading day 250), demonstrating that the RAA mechanism can effectively guide strategies to reduce risk. The study further validated the robustness and crisis-response capabilities of each model to assess their risk-control effectiveness during sudden market events and periods of high volatility. Results are shown in Fig. 9.

As shown in Fig. 9(a), during the simulated market crisis period (0 to 100 days), the net asset value of all models fell sharply, but the PPO-DRF model showed the strongest resilience, with its net asset value remaining at a low of approximately 0.87, far higher than B&H's approximately 0.68, and rebounding to 1.05 first after the crisis ended. Fig. 9(b) shows the distribution of maximum drawdown during the high-volatility period, further quantifying the risk-control advantage. The median maximum drawdown of PPO-DRF was approximately -10%, significantly higher than those of comparative models such as PPO (-30%) and DDPG (-25%), indicating that PPO-DRF can effectively minimize the risk of extreme losses. Fig. 9(c) clearly demonstrates the adaptability of the RAA mechanism through its dynamic risk penalty coefficient β curve. The β curve is highly synchronized with the market volatility curve on the right axis. When market volatility over the 250 trading days reached its peak (approximately 75%), the β coefficient also rose rapidly to

approximately 0.95, indicating that the model dynamically adjusts the risk penalty intensity in real time to ensure the strategy's stability during the crisis. The study continued to validate the model's adaptability and generalization capabilities, assessing its stability and effectiveness on untrained portfolios and across different market cycles. The specific results are shown in Fig. 10.

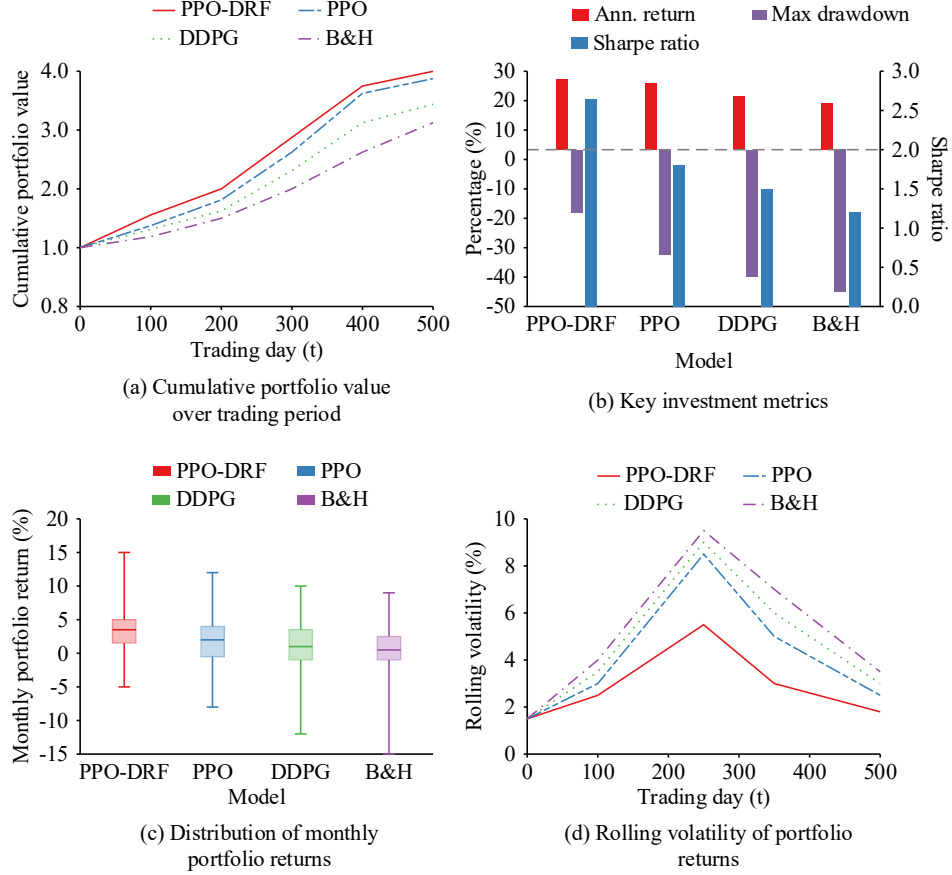


Fig. 8. Overall performance comparison on S&P 500 constituent stock dataset (training/In-sample performance)

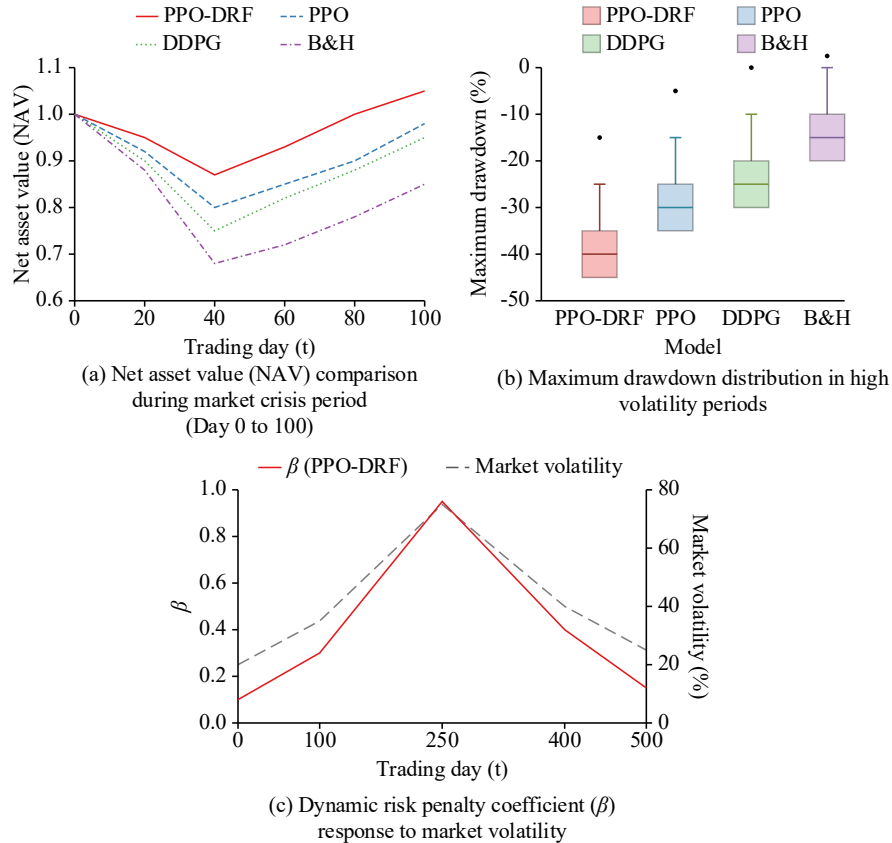


Fig. 9. Robustness analysis of various models under extreme market conditions

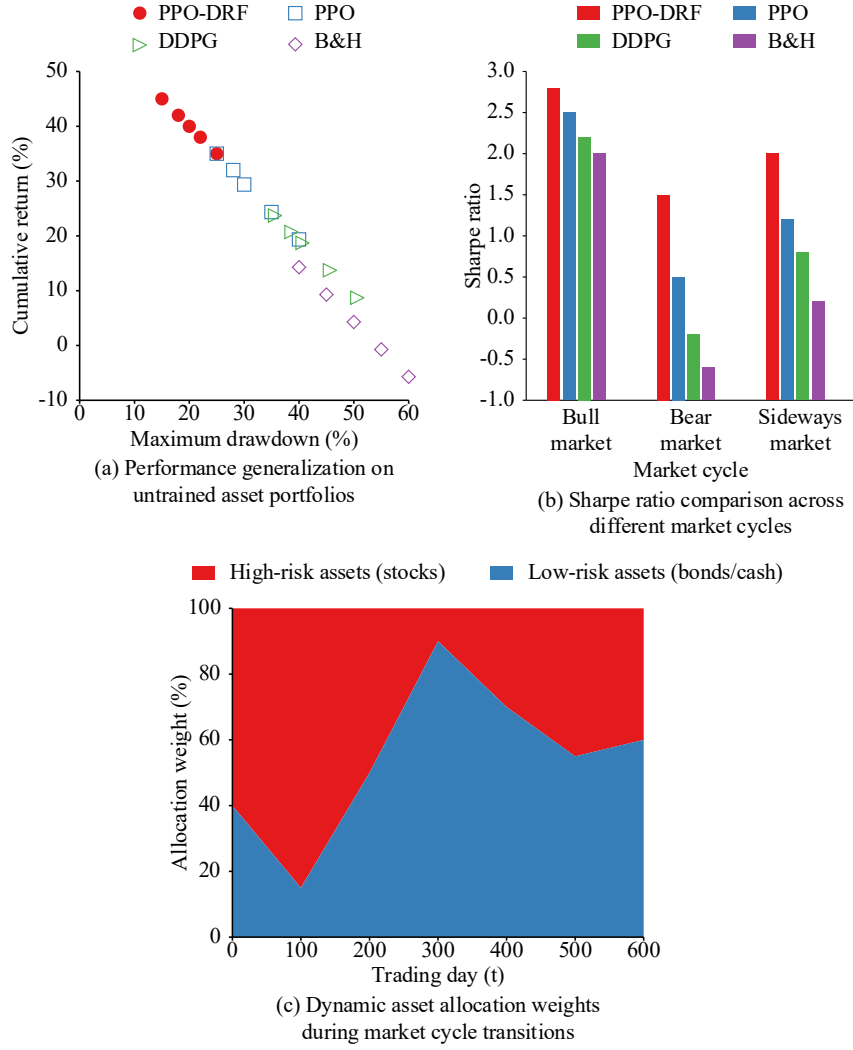


Fig. 10. Adaptability and generalization ability of each model in different markets and assets

As shown in Fig. 10(a), when generalizing to a completely new portfolio that was not used in the training, the data points of the PPO-DRF model were concentrated in the upper left corner, where the cumulative return is high (up to approximately 45%), and the maximum drawdown was low (down to approximately 15%). This clearly demonstrates that the PPO-DRF strategy can still achieve an optimal, high-return, low-risk portfolio on unseen assets, and its cross-asset generalization ability is far superior to the comparative models that are dispersed along the diagonal. The market cycle comparison in Fig. 10(b) shows that PPO-DRF led with a Sharpe ratio of approximately 2.8 in the Bull Market. In the more challenging Bear Market, PPO-DRF maintained a positive Sharpe ratio of approximately 1.5. Meanwhile, models such as B&H and DDPG fell into negative territory (approximately -0.3 and -0.2, respectively), strongly demonstrating the adaptability of the model's strategy under adverse market conditions. The dynamic allocation weight changes in Fig. 10(c) demonstrate the flexibility of its decision-making. During the Day 100 period of rising market risk, the model rapidly increased the weight of low-risk assets to about 85%, while during the Day 300 period of market optimism, it actively increased the weight of high-risk assets to about 90%, demonstrating its adaptive adjustment capability to market style.

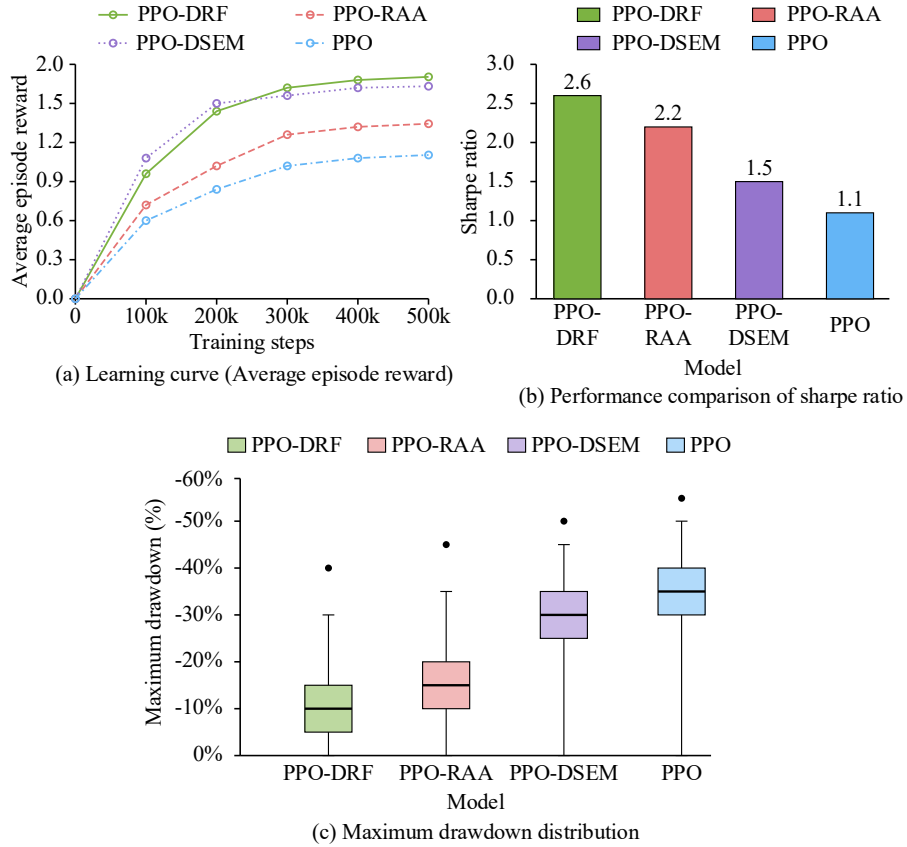
3.3. Ablation Experimental Analysis of PPO-DRF Model

To comprehensively explore the independent contributions and synergistic effects of DSEM and RAA, the two core innovations in the PPO-DRF model, a systematic ablation experiment was designed. Comparative models included: PPO-RAA (PPO with RAA, without DSEM), PPO-DSEM (PPO with DSEM, without RAA), and the Vanilla PPO (PPO) algorithm as a performance baseline. The experiments used historical data from Yahoo Finance, which underwent multi-granularity processing to ensure that the multi-frequency temporal features required by the DSEM module were fully input. All ablation models were trained in the same Python environment (PyTorch/stable-baselines3) as the full model, and performance was evaluated using metrics such as Sharpe ratio, maximum drawdown, and average round reward learning curve. The Sharpe ratio and maximum drawdown were used to assess the effect of RAA on risk-adjusted return, while the learning curve was used to evaluate the contribution of DSEM to improving training stability and convergence speed. The specific ablation experiment configuration is shown in Table 1.

Table 1. Experimental configuration for PPO-DRF model ablation

Configuration Item	PPO-DRF (Full Model)	PPO-RAA (w/o DSEM)	PPO-DSEM (w/o RAA)	PPO (Vanilla)
Perception Layer	DSEM (LSTM/GRU + Multi-frequency)	Standard MLP / Simple RNN	DSEM (LSTM/GRU + Multi-frequency)	Standard MLP / Simple RNN
Sequence Encoding	Applied	Removed	Applied	Removed
Input Data Type	Multi-frequency (Min, Day, Week)	Multi-frequency (Min, Day, Week)	Multi-frequency (Min, Day, Week)	Multi-frequency (Min, Day, Week)
Advantage Function	RAA	RAA	GAE	GAE
Risk Constraint	Applied	Applied	Removed	Removed
Core Optimization Target	Maximize ARAA (Risk-Adjusted)	Maximize ARAA (Risk-Adjusted)	Maximize AGAE (Return Maximization)	Maximize AGAE (Return Maximization)

As shown in Table 1, four model configurations were designed for ablation comparison. The PPO-RAA model removed the multi-frequency temporal coding mechanism to evaluate DSEM's contribution to state awareness. The PPO-DSEM model removed the dynamic risk constraint mechanism to evaluate RAA's effect on risk control, and the vanilla PPO model served as a baseline without any optimization components. By using different combinations of DSEM or traditional MLP for the perception layer and RAA or traditional GAE for the dominance function, the independent gains of DSEM in improving model training stability and RAA in balancing risk and reward were systematically verified. Ultimately, it was demonstrated that the superior performance of the complete PPO-DRF model stems from the effective synergy of the two innovative modules. Specific results are shown in Fig. 11.

**Fig. 11.** Ablation experimental results of PPO-DRF model

As shown in Fig. 11(a), the models containing the DSEM module (PPO-DRF and PPO-DSEM) exhibited the fastest convergence, with the final average round reward approaching 1.7, which was better than the model without DSEM,

thereby verifying the contribution of DSEM to improving learning efficiency. The Sharpe ratio comparison in Fig. 11(b) shows that PPO-DRF and PPO-RAA with the RAA advantage function performed best (2.6 and 2.2, respectively), significantly higher than PPO-DSEM (1.5) and PPO (1.1) without RAA, proving that RAA is key to improving risk-adjusted returns. The maximum drawdown distribution in Fig. 11(c) further quantified the role of RAA. The model with RAA controlled the median maximum drawdown between -10% and -15%, while the model without RAA experienced a larger drawdown (approximately -30% to -35%), confirming that RAA effectively suppresses the risk of extreme losses. To further anchor the gains of the proposed RL framework against broader financial standards and supervised learning methods, a supplementary comparative experiment was conducted. This analysis aimed to distinguish the specific decision-making advantages of the PPO-DRF agent from those of purely prediction-based models and traditional static allocation strategies. The supplementary performance comparison with financial standards and supervised baselines is shown in Table 2.

Table 2. Supplementary performance comparison with financial standards and supervised baselines

Model / Strategy	Cumulative Return (%)	Annualized Return (%)	Max Drawdown (%)	Sharpe Ratio
PPO-DRF (Proposed)	35.85	23.50	-15.20	2.55
XGBoost (Supervised)	28.60	19.20	-21.50	2.15
Risk Parity (Traditional)	22.40	14.80	-11.20	1.85
Mean-Variance (Traditional)	20.10	13.50	-24.60	1.45
Momentum (Traditional)	24.50	16.10	-28.40	1.55

As indicated by the data in Table 2, while the supervised XGBoost model outperforms the B&H benchmark with a Sharpe ratio of 2.15 due to its strong feature extraction capabilities, it lacks the dynamic risk adjustment mechanism inherent to RL agents, resulting in a significantly higher maximum drawdown (-21.50%) compared to the -15.20% achieved by PPO-DRF. Similarly, in terms of traditional strategies, although Risk Parity effectively controls volatility (10.50%), its conservative allocation logic limits its total cumulative return to 22.40%, which is notably lower than the 35.85% realized by PPO-DRF. These findings confirm that PPO-DRF surpasses both traditional allocation heuristics and prediction-based supervised models by effectively synthesizing deep state perception with active, volatility-linked risk management. To further justify the architectural selection of the proposed DSEM relative to other prevalent sequence modeling techniques in finance, such as Transformer-based encoders and Temporal Convolutional Networks (TCNs), a comparative analysis focused on feature-encoding efficiency was conducted. The performance comparison of different encoding architectures is shown in Table 3.

Table 3. Performance comparison of different encoding architectures

Encoding Architecture	Cumulative Return (%)	Sharpe Ratio	Max Drawdown (%)	Training Stability (Variance)
DSEM (Proposed Hybrid)	35.85	2.55	-15.2	Low
Transformer (Attention)	31.1	2.15	-19.45	High
Temporal CNN (TCN)	28.5	1.85	-21.3	Moderate
Standard LSTM (Vanilla)	30.2	2.1	-18.6	Moderate

As shown in Table 3, the DSEM architecture achieves the highest Sharpe ratio (2.55) and the lowest maximum drawdown (-15.20%), outperforming both the Transformer (Sharpe ratio 2.15) and TCN (Sharpe ratio 1.85) variants. While Transformers utilize global attention mechanisms to capture long-range dependencies, they often suffer from overfitting and high computational noise when processing high-frequency, low-signal-to-noise ratio financial data. In contrast, the LSTM/GRU hybrid structure of DSEM offers a superior inductive bias for sequential processing, effectively balancing long-term trend retention with short-term fluctuation sensitivity. This confirms that, given the specific characteristics of multi-granular financial time series, the recurrent inductive bias of DSEM yields a more robust state representation than purely attention-based or convolutional approaches.

3.4. Analysis of Application Effect on PPO-DRF Model

Following verification of the robustness and adaptability of the PPO-DRF model, the present study focused on evaluating its practical value in financial investment. The study used historical data from Yahoo Finance, meticulously selected representative A-share stocks, and obtained multi-frequency time series data from cross-industry ETFs. This approach was undertaken to ensure the authenticity of the application scenarios and the compatibility of the data input with the DSEM module. To demonstrate the model's effectiveness in solving specific investment problems, all application experiments were back tested in a unified Python/PyTorch environment. Firstly, the study used a visual analysis of individual stock timing decisions to intuitively demonstrate the PPO-DRF model's effectiveness in identifying buying and selling opportunities and mitigating risk for a single asset, as illustrated in Fig. 12.

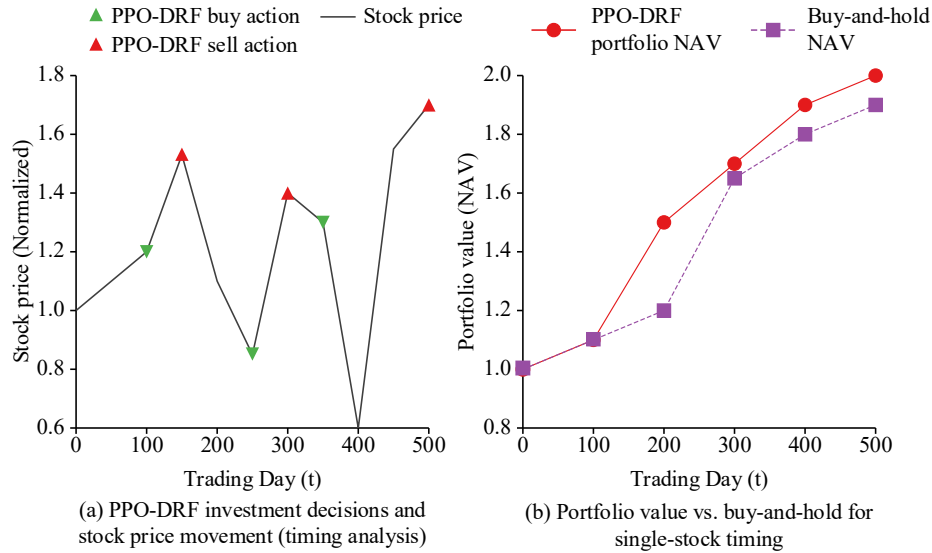


Fig. 12. Application results of PPO-DRF in timing and decision analysis of single stocks

In Fig. 12(a), the PPO-DRF model demonstrated accurate decision-making capabilities in individual stock timing. For instance, the investment was initiated when the stock price reached a nadir of 0.85 on Day 250, and was divested when it attained a local peak of 1.40 on Day 300, thereby capitalizing on market turning points. The portfolio net value comparison in Fig. 12(b) showed that the PPO-DRF timing strategy achieved a cumulative value of 2.0 over 500 days, which is significantly higher than the 1.9 achieved by the Buy-and-Hold strategy. This finding demonstrates the model's efficacy in capturing excess returns through flexible buying and selling operations, thereby substantiating its practical value. The study further demonstrated the model's ability to balance returns, risks, and industry distribution across multiple asset allocations by evaluating cross-industry portfolio management benefits. The specific outcomes are delineated in Table 4.

Table 4. Results of PPO-DRF validation on cross-industry ETF portfolios (testing/out-of-sample performance)

Metric/Asset Class	PPO-DRF (Research Model)	PPO (Baseline Model)	Buy-and-Hold (Market Benchmark)
Total Cumulative Return (%)	35.85	25.10	18.23
Annualized Return (%)	23.50	17.50	12.00
Maximum Drawdown (%) (MaxDD)	-15.20	-25.80	-35.50
Sharpe Ratio	2.55	1.60	1.10
Value at Risk (VaR, 99%) (%)	-2.50	-3.80	-4.50
Information Ratio	0.85	0.35	N/A
Avg. Weight: Tech ETF (%)	30.0	35.0	20.0
Avg. Weight: Financial ETF (%)	15.0	20.0	20.0
Avg. Weight: Healthcare ETF (%)	25.0	25.0	20.0
Avg. Weight: Energy ETF (%)	10.0	10.0	20.0

In Table 4, the PPO-DRF model denoted significant application benefits in cross-industry ETF portfolio management. The model achieved a cumulative return of 35.85% and a Sharpe ratio of 2.55, both of which greatly outperformed the baseline PPO model (Sharpe ratio 1.60) and the market benchmark Buy-and-Hold (Sharpe ratio 1.10). In terms of risk control, the max drawdown of PPO-DRF was only -15.20%, far lower than Buy-and-Hold's -35.50%, and its 99% Value at Risk (VaR) was also the lowest, fully demonstrating the risk aversion capability of the RAA mechanism in practical application. Regarding asset allocation, PPO-DRF dynamically allocated 20.0% of its funds to cash/low-risk assets to achieve proactive risk management, rather than blindly averaging out, proving the rationality of the model's decision structure. Finally, through multi-objective optimization analysis in practical application, the study explored how the PPO-DRF model can meet the customized needs of investors with different risk appetites by adjusting the parameters of the RAA module. Specific results are shown in Table 5.

Table 5. Results of PPO-DRF cross-industry ETF portfolio management effectiveness

Metric/Risk Preference	Aggressive($\beta=0.05$)	Balanced($\beta=0.50$)	Conservative($\beta=0.95$)
Total Cumulative Return (%)	45.20	35.85	25.10
Annualized Return (%)	28.00	23.50	18.00
Maximum Drawdown (%) (MaxDD)	-28.50	-15.20	-8.50
Sharpe Ratio	1.80	2.55	2.05
Annualized Volatility (%)	15.60	9.20	8.50
Average Holding Period (Days)	25	45	70
Trade Frequency (Trades/Month)	12	8	4
Avg. Weight: Cash (%)	5.0	20.0	45.0

As shown in Table 5, by adjusting the dynamic risk penalty coefficient (β) of the RAA module, the PPO-DRF model could achieve multi-objective optimization to meet the needs of investors with different risk appetites. Under the aggressive allocation with $\beta=0.05$, the model achieved the highest annualized return (28.00%) and the fastest trading frequency, but at the cost of a relatively high drawdown (-28.50%). Under the conservative allocation with $\beta=0.95$, the model aimed for the lowest drawdown (-8.50%) and annualized volatility (8.50%), increasing the cash allocation weight to 45.0% and extending the holding period. The balanced allocation ($\beta=0.50$) achieved the best risk-adjusted return, with a Sharpe ratio of 2.55, thereby demonstrating that the PPO-DRF can be applied as a customized investment tool in practice.

4. Discussion

To address incomplete state perception and weak risk control in intelligent financial investment, this study proposes an improved PPO algorithm, the PPO-DRF model, which integrates DSEM and RAA. The DSEM module enhances state perception through multi-granularity time-series data fusion, while the RAA mechanism implements real-time risk penalty, thereby optimizing the risk-return balance and improving the strategy's robustness and risk resistance. Empirical results show that PPO-DRF achieved a cumulative net value of 3.95 over 500 trading days, with an annualized return of 27.5%, a Sharpe ratio of 2.65, and a maximum drawdown of -18.0%, significantly outperforming PPO, DDPG, and buy-and-hold strategies. This performance advantage stems from two modules that precisely address the shortcomings of traditional reinforcement learning: DSEM processes minute, daily, and weekly data in parallel, using an LSTM/GRU encoder to extract multi-scale market dynamics and construct a complete observation vector, effectively mitigating the problem of missing states in some observable environments; RAA dynamically adjusts the risk penalty coefficient to achieve real-time linkage with market fluctuations, thereby controlling the peak volatility to 5.5% during periods of high volatility and achieving a minimum net asset value of only 0.87 during simulated crises. Furthermore, the model's excellent performance in stock timing and cross-industry ETF allocation validates its ability to flexibly explore the continuous action space and to enforce risk constraints within the Actor-Critic framework, demonstrating significant practical investment value.

The academic contributions of this research lie in supplementing and expanding existing theories and methods of intelligent financial investment, while also establishing an effective dialogue with related research. Hambly et al. (2023) found that the application of reinforcement learning in financial decision-making was systematically reviewed for its potential in portfolio optimization and highlighted its limitations, including insufficient generalization and a poor risk-reward balance. Their research, through the fusion design of DSEM and RAA, precisely addressed this research gap, providing a new technical path for the practical application of reinforcement learning in finance. Li et al. (2023) researched the lag problem in recurrent neural network time-series prediction, revealing the shortcomings of traditional time-series models in financial data processing. Their research, employing multi-granularity encoding and a combination of LSTM/GRU, effectively alleviated prediction lag and enriched the methodological system of time-series financial data processing. Odonkor (2024) reported on the team's research on the application of AI in accounting and finance, highlighting the challenges of balancing risk management and efficiency in AI applications. Their research on the RAA mechanism, using a dynamic risk penalty, achieved synergistic optimization of return efficiency and risk control, providing an

operational solution to this challenge. At the theoretical level, the research developed a closed-loop framework of "multi-granularity perception - dynamic risk adjustment - stable strategy optimization", which advanced the decision-making theory of DRL in certain observable financial environments. At the practical level, based on the application results of stock timing and cross-industry portfolio management in the A-share market, it provides financial institutions with investment tools that are both scientific and practical, and promotes the upgrade of financial decision-making from data-driven to "data + intelligent risk control" driven.

While this study has yielded significant results, it still has certain limitations that point the way for future research. First, the dataset is primarily based on historical data from Yahoo Finance. Although it covers a relatively long period, it does not adequately account for extreme "black swan" events (such as global crises or policy upheavals). The model's adaptability in such unforeseen scenarios remains to be verified, aligning with the core concern in research on the robustness of financial models. Second, the model was only tested on A-share stocks and S&P 500 constituent stocks, and its generalization ability across markets (such as Hong Kong stocks and US small-cap stocks) and asset classes (such as bonds and derivatives) has not been fully verified. Different market rules and data characteristics may affect model performance. Finally, model parameters (such as the PPO pruning factor and RAA penalty factor) rely on initial manual tuning and lack adaptive adjustment mechanisms, potentially limiting their responsiveness in complex markets. Future research can be advanced in the following three aspects: First, introduce real-time public opinion and macroeconomic indicators, optimize the DSEM module, and enhance the ability to capture sudden information; second, build a multi-agent collaboration mechanism, design dedicated agents for different assets, and improve the efficiency of complex portfolio management; third, conduct long-term real-world testing, combine blockchain technology to realize real-time traceability and verification of transaction data, further test the long-term robustness of the model in the actual market, and provide support for the industrial application of intelligent investment.

5. Conclusion

In the context of volatile financial markets and the inadequacy of conventional investment decision-making methodologies for aligning with emerging intelligent investment approaches, this study examined the challenges of insufficient generalization and diminished risk aversion in investment strategies in a dynamic stock market environment. It developed a proximal policy optimization (PPO) model (PPO-DRF) that fuses DSEM and RAA for intelligent financial investment. The study used DSEM to process multi-granularity time-series financial data, combined it with RAA to dynamically adjust risk penalties, and achieved decision optimization using the PPO Actor-Critic architecture. Performance validation denoted that within 500 trading days, the PPO-DRF achieved a cumulative net asset value of 3.95, an annualized return of 27.5%, a maximum drawdown of -18.0%, and a Sharpe ratio of 2.65. In simulation tests, the simulated crisis period reached a mini net asset value of 0.87, and the total cumulative return of the cross-industry ETF portfolio was 35.85%, significantly outperforming the comparative models. In conclusion, this study proves the superiority of PPO-DRF in terms of return and risk control, providing a feasible solution for intelligent investment. However, the research dataset does not fully cover extreme "black swan" events, and its cross-market generalization needs to be verified. In the future, public opinion and macro data will be integrated, and long-term live testing will be carried out to improve the model.

Author Contributions

Cheng Yang contributed to the study conception and design. Material preparation, data collection and analysis. Haoyang Zeng contributed to the study of conception and design. Material preparation, data collection and analysis. Yifan Yu contributed to the study of conception and design. Material preparation, data collection and analysis. The first draft of the manuscript was written by Shilong Han. Manuscript review and revision were performed by Yifan Yu. All authors commented on previous versions of the manuscript. All authors read and approve of the final manuscript.

Funding

This research is supported by Guangdong Provincial Audit Department's Key Scientific Research Project for 2024 to 2025, "Investigation of Internal Control Deficiencies and Audit Issues of Special Funds for Rural Revitalization" (No.24GDSJZZ012620). Guangdong University of Finance & Economics, School of Accounting/School of Intelligent Financial and Accounting Management, Scientific Research Cultivation Project from 2025 to 2026, "Scientific Research Cultivation Project for 2025 to 2026 Organized Scientific Research Innovation Team for Financial Governance and Coordinated Development in Guangdong-Hong Kong-Macao Region" (PYXM-2025-5). Guangdong University of Finance & Economics, School of Accounting/School of Intelligent Financial and Accounting Management, Cultivation Project of Teaching Quality and Teaching Reform Engineering for 2025 to 2026: "Exploration and Practice of Man-Machine Co-Teaching Model for Digital-Intelligent Financial Management" (PYXM-2025-41).

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used Grammarly solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

References

An, H. and Wang, L. (2023). Robust topology generation of Internet of Things based on PPO algorithm using discrete action space. *IEEE Transactions on Industrial Informatics*, 20(4), 5406–5414. <https://doi.org/10.1109/TII.2023.3333012>.

- Buathong, W., Sieng-Ek, P., and Jarupunphol, P. (2023). Measuring the performance of machine learning forecasting models to support bitcoin investment decisions. *Journal of Data Science and Intelligent Systems*, 2(2), 100–112. <https://doi.org/10.47852/bonviewJDSIS3202677>.
- Cheng, Y., Guo, Q., and Wang, X. (2024). Proximal policy optimization with advantage reuse competition. *IEEE Transactions on Artificial Intelligence*, 5(8), 3915–3925. <https://doi.org/10.1109/TAI.2024.3354694>.
- Guan, Y., Zou, S., Peng, H., Ni, W., Sun, Y., and Gao, H. (2023). Cooperative UAV trajectory design for disaster area emergency communications: A multiagent PPO method. *IEEE Internet of Things Journal*, 11(5), 8848–8859. <https://doi.org/10.1109/JIOT.2023.3320796>.
- Hambly, B., Xu, R., and Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3), 437–503. <https://doi.org/10.1111/mafi.12382>.
- Ji, S., Li, X., Sun, W., Dong, H., Taalas, A., Zhang, Y., and Marttinen, P. (2024). A unified review of deep learning for automated medical coding. *ACM Computing Surveys*, 56(12), 1–41. <https://doi.org/10.1145/3664615>.
- Jin, Y., Song, X., Slabaugh, G., and Lucas, S. (2024). Partial advantage estimator for proximal policy optimization. *IEEE Transactions on Games*, 17(1), 158–166. <https://doi.org/10.1109/TG.2024.3408298>.
- Li, J., Song, L., Wu, D., Shui, J., and Wang, T. (2023). Lagging problem in financial time series forecasting. *Neural Computing and Applications*, 35(28), 20819–20839. <https://doi.org/10.1007/s00521-023-08879-1>.
- Li, L. and Zhu, Y. (2024). Boosting on-policy actor–critic with shallow updates in critic. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3), 5644–5653. <https://doi.org/10.1109/TNNLS.2024.3378913>.
- Li, L., Li, W., Wang, J., Chen, X., Peng, Q., and Huang, W. (2023). UAV trajectory optimization for spectrum cartography: A PPO approach. *IEEE Communications Letters*, 27(6), 1575–1579. <https://doi.org/10.1109/LCOMM.2023.3265214>.
- Liu, Y., Wang, L., Li, D., and Wang, K. (2023). State-of-health estimation of lithium-ion batteries based on electrochemical impedance spectroscopy: A review. *Protection and Control of Modern Power Systems*, 8(3), 1–17. <https://doi.org/10.1186/s41601-023-00314-w>.
- Mazzucato, M. (2023). Governing the economics of the common good: From correcting market failures to shaping collective goals. *Journal of Economic Policy Reform*, 27(1), 1–24. <https://doi.org/10.1080/17487870.2023.2280969>.
- Odonkor, B., Kaggwa, S., Uwaoma, P. U., Hassan, A. O., and Farayola, O. A. (2024). The impact of AI on accounting practices: A review. *World Journal of Advanced Research and Reviews*, 21(1), 172–188. <https://doi.org/10.30574/wjarr.2024.21.1.2721>.
- Oroojlooy, A., and Hajinezhad, D. (2022). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11), 13677–13722. <https://doi.org/10.1007/s10489-022-04105-y>.
- Pacal, I. (2024). A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. *International Journal of Machine Learning and Cybernetics*, 15(9), 3579–3597. <https://doi.org/10.1007/s13042-024-02110-w>.
- Rafiei, N., Asadzadeh, S., and Niaki, S. T. A. (2021). Multi-objective design of risk-adjusted control chart in healthcare systems with economic and statistical considerations. *Communications in Statistics - Simulation and Computation*, 52(7), 2967–2984. <https://doi.org/10.1080/03610918.2021.1923744>.
- Sahoh, B. and Choksuriwong, A. (2023). The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 14(6), 7827–7843. <https://doi.org/10.1007/s12652-023-04594-w>.
- Shu, Y., Yu, C., and Mulvey, J. M. (2024). Downside risk reduction using regime-switching signals: A statistical jump model approach. *Journal of Asset Management*, 25(5), 493–507. <https://doi.org/10.1057/s41260-024-00376-x>.
- Świechowski, M., Godlewski, K., Sawicki, B., and Mańdziuk, J. (2022). Monte Carlo tree search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56(3), 2497–2562. <https://doi.org/10.1007/s10462-022-10228-y>.
- Szukits, Á. and Móricz, P. (2023). Towards data-driven decision making: The role of analytical culture and centralization efforts. *Review of Management Science*, 18(10), 2849–2887. <https://doi.org/10.1007/s11846-023-00694-1>.
- Uysal, A. S., Li, X., and Mulvey, J. M. (2023). End-to-end risk budgeting portfolio optimization with neural networks. *Annals of Operations Research*, 339(1), 397–426. <https://doi.org/10.1007/s10479-023-05539-4>.
- Weber, P., Carl, K. V., and Hinz, O. (2023). Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, 74(2), 867–907. <https://doi.org/10.1007/s11301-023-00320-0>.
- Wu, G., Fang, W., Wang, J., Ge, P., Cao, J., Ping, Y., and Gou, P. (2022). Dyna-PPO reinforcement learning with Gaussian process for the continuous action decision-making in autonomous driving. *Applied Intelligence*, 53(13), 16893–16907. <https://doi.org/10.1007/s10489-022-04354-x>.
- Yan, B. Y., Yang, C., Chen, F., Takeda, K., and Wang, C. (2023). FDNet: A deep learning approach with two parallel cross encoding pathways for precipitation nowcasting. *Journal of Computer Science and Technology*, 38(5), 1002–1020. <https://doi.org/10.1007/s11390-021-1103-8>.
- Yau, H. T., Kuo, P. H., Luan, P. C., and Tseng, Y. R. (2024). Proximal policy optimization-based controller for chaotic systems. *International Journal of Robust and Nonlinear Control*, 34(1), 586–601. <https://doi.org/10.1002/rnc.6988>.
- Zhang, B., and Zhou, W. (2022). Transformer-encoder-GRU (TE-GRU) for Chinese sentiment analysis on Chinese comment text. *Neural Processing Letters*, 55(2), 1847–1867. <https://doi.org/10.1007/s11063-022-10966-8>.
- Zhang, C., Sjarif, N. N. A., and Ibrahim, R. (2023). Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(1), e1519. <https://doi.org/10.1002/widm.1519>.
- Zhang, X., Zheng, K., Wang, C., Chen, J., and Qi, H. (2023). A novel deep reinforcement learning for POMDP-based autonomous ship collision decision-making. *Neural Computing and Applications*, 37(21), 15963–15977. <https://doi.org/10.1007/s00521-023-08908-z>.

Zhao, L., Fan, J., Zhang, C., Shen, W., and Zhuang, J. (2023). A DRL-based reactive scheduling policy for flexible job shops with random job arrivals. *IEEE Transactions on Automation Science and Engineering*, 21(3), 2912–2923. <https://doi.org/10.1109/TASE.2023.3271666>.



Cheng Yang (Ph.D.) was born in Xingning City, Guangdong Province, China, in July 1990. He obtained a Bachelor of Science in National Accounting from Liaoning University, Shenyang, Liaoning, China, in 2013. He then earned a Bachelor of Science in Accounting and Finance from De Montfort University, Leicester, UK, in the same year. In 2014, he received a Master of Science in International Accounting and Financial Management from the University of Glasgow, Glasgow, UK. He earned his Ph.D. in financial management from the School of Management, Jinan University, Guangzhou, Guangdong, China, in 2019. He began his academic career in July 2019 as a lecturer at the School of Accounting, Guangdong University of Finance and Economics (GDUFE). From April 2021 to October 2023, he served as a lecturer, department head, and assistant dean at the School of Intelligent Accounting and Management, Guangdong University of Finance and Economics. He was promoted to Associate Professor in November 2023 while retaining his positions as Department Head and Assistant Dean at the School of Intelligent Accounting and Management. From July 2024 to August 2025, he worked as an Associate Professor and Director of the Professional Degree Office at the School of Accounting/School of Intelligent Accounting and Management, GDUFE. Since August 2025, he has been an Associate Professor and Director of the Professional Degree Office at the School of Accounting, Guangdong University of Finance and Economics. He has published fifteen academic papers and three textbooks and has led or participated in seven research projects. His research interests have centered on financial management, corporate governance, and auditing, and he continues to focus on these areas currently.



Shilong Han was born in Ruzhou City, Henan Province, China, in April 2000. He studied accounting as an undergraduate student at the School of Business, Zhengzhou University of Aeronautics and Administration, Zhengzhou, Henan, China, from 2018 to 2022. Since 2023, he has been pursuing a postgraduate degree in accounting at the School of Accounting, GDUFE, Guangzhou, Guangdong, China.



Haoyang Zeng was born in Guangzhou, Guangdong, China, in September 2000. He received a Bachelor of Science in international economics and trade from Zhongkai University of Agriculture and Engineering in Guangzhou, China, in 2023. He has been pursuing a graduate degree in accounting at the School of Accounting, Guangdong University of Finance and Economics, Guangzhou, China, since 2024. He is currently a graduate student in the School of Accounting at Guangdong University of Finance and Economics in Guangzhou, China. His current academic interests are in accounting-related topics.



Yifan Yu was born in Guangzhou, Guangdong Province, China, in October 1999. He obtained his Bachelor of Arts from Jinan University in Guangzhou, China, in 2021, and his Master of Arts from the same university in 2025. Since 2025, he has been pursuing his Doctor of Management at Jinan University in Guangzhou, China. He is currently a doctoral candidate at the School of Management, Jinan University, Guangzhou, China. His research interests focus on international business.