Journal of Computer Technology and Software
ISSN:2998-2383
Vol. 5, No. 10, 2026

Multi-Agent Collaborative Modeling for Systemic Risk Propagation in Financial Markets: A Game-Theoretic Framework

Yi Wang[1*]
[1]Columbia UniversityNew York, USA
*Corresponding author: Yi Wang; yw3802@columbia.edu

Abstract: This paper focuses on the risk control challenges arising from behavioral interactions among multiple participants in financial markets. It proposes a financial risk game framework based on multi-agent collaborative modeling. The framework integrates environment state encoding, strategy generation networks, and a coordinated evolution mechanism to dynamically model the propagation paths of systemic risk in complex markets. During the modeling process, each agent generates behavior strategies independently based on local observations. The market state is then updated over time through a system evolution function, which captures the coupling between multi-agent behaviors and the risk structure. To verify the stability and adaptability of the proposed method, a series of sensitivity experiments are designed. These experiments examine the impact of hyperparameters, data characteristics, and environmental disturbances on system performance. The study focuses on several key factors, including the number of agents, time window length, sampling frequency, and anomaly injection. The experimental results show that the method performs well across multiple dimensions such as strategy stability, modeling consistency, and coordination efficiency. The model demonstrates strong structural representation and dynamic adaptability. Through comparative experiments and disturbance analysis, the study further reveals the model's capability to simulate the evolution of financial risk structures under various conditions. This provides a valuable methodological reference for intelligent modeling of complex financial systems.

Keywords: Multi-agent modeling; risk propagation simulation; collaborative game strategy; financial system structure

1. Introduction

In highly complex and dynamically evolving financial markets, risk control remains one of the core issues. Traditional risk management approaches are often based on static modeling assumptions of a single agent, which fail to capture the systemic impact caused by strong interactions and games among multiple participants[1]. As market structures grow increasingly complex, financial risk is no longer a function of isolated variables. It is the result of collective behavior shaped by inter-agent influence and continuous adjustments. This evolutionary mechanism is reflected not only in the information game among financial institutions but also in investor behavior, counterparty decisions, and policy responses. Therefore, building a risk control framework capable of simulating multi-agent cooperation and competition has become a major challenge in the field of financial AI and quantitative modeling[2].

The current financial market exhibits new characteristics such as high-frequency trading, algorithmic games, and strategy confrontation. Market risk has evolved from static volatility risk to dynamic structural risk[3]. Under such conditions, single-strategy or locally optimized models cannot effectively capture the risk propagation paths or the global system state. Multi-agent systems offer a modeling paradigm that can capture the complex interactions between individual behaviors and macro-level system feedback. By introducing coordination mechanisms and strategic game logic, different participants can make autonomous decisions based on shared information or local perceptions. This enables a more realistic representation of risk diffusion and control in financial markets[4].

At the same time, cooperation and competition among agents influence not only individual transaction outcomes but also the overall evolution of market risk structure. Traditional risk management tools focus on controlling the risk of individual assets or portfolios. However, they cannot describe systemic linkages and feedback within the market. Integrating game-theoretic modeling and multi-agent optimization mechanisms can fundamentally improve the adaptability and robustness of risk management strategies. For example, in the event of sudden market shocks, simulating agent behaviors under limited information can help identify risk clusters and contagion paths more accurately, enabling better system-level response strategies.

Moreover, as financial regulation strengthens and market openness increases, the financial system is evolving from a centralized structure to a more open and distributed ecosystem. This shift means that risk is no longer controlled by central nodes alone but is shaped by the strategic interactions among multiple peripheral agents. In this context, the focus of risk control is moving from rule-making to behavior guidance. Multi-agent game modeling methods can simulate agent responses under complex rules and uncover potential cooperation opportunities and risk spillover paths. This provides theoretical and methodological support for building a financially stable and behaviorally rational ecosystem[5].

Starting from coordination mechanisms, the integration of multi-agent modeling and game theory not only enriches the theoretical system of financial risk management but also offers a practical path toward more adaptive and intelligent risk control systems. This research is of great significance for enhancing the accuracy of systemic risk identification, optimizing the configuration of multi-agent game strategies, and improving the resilience of financial markets. It also holds the potential for profound impacts on intelligent financial regulation and market mechanism design in the future.

2. Background & Motivation

2.1 Background

The risk propagation mechanism in current financial markets shows high complexity and dynamic nonlinearity. Traditional modeling approaches are often based on static or single-agent perspectives. These methods struggle to capture interactive behaviors and feedback mechanisms among diverse participants. They ignore the systemic impact of strategic adjustments, resource competition, and information asymmetry among agents. As a result, these models often exhibit poor adaptability and weak predictive power in practical applications, especially when facing sudden market disruptions or systemic crises[6].

Most existing risk control systems rely on rule-based logic and historical statistical patterns. They cannot respond effectively to rapidly evolving strategic games and micro-level behavioral changes in financial markets. This leads to delayed regulatory responses and fragmented policy decisions. It also limits the ability to cover the multiple pathways through which risk spreads. In addition, current models generally lack the capability to dynamically simulate autonomous behaviors and game logic of market participants. This limitation causes them to fail in high-frequency trading, quantitative investment, and complex derivatives markets[7].

In multi-agent financial systems, risk arises not only from external shocks but also from internal structural instability and the cumulative effects of strategic coupling. Current risk monitoring methods cannot effectively detect this type of structure-induced risk. This is particularly evident in open and decentralized financial networks, where game behaviors and coordination patterns among agents exhibit complex and multidimensional dynamics. A comprehensive theoretical and methodological framework is urgently needed to model these interlinked behaviors.

2.2 Motivation

In modern financial systems, the emergence and diffusion of risk are often accompanied by strategic games and dynamic behavioral adjustments among market participants. Faced with uncertainty and volatility, financial agents constantly revise their decisions based on local information[8]. These interactions form a complex system of feedback mechanisms. Traditional static models cannot capture the nonlinear linkages between such behaviors. This limitation restricts a full understanding of the overall risk structure. It raises the need to explore modeling approaches with the ability to simulate strategic evolution, in order to better reflect the logic of agent interactions in the market.

In recent years, financial markets have shown increasing trends toward high frequency and automation. The widespread adoption of algorithmic trading and automated decision-making systems has made competition and cooperation among participants more dynamic. Under this new behavioral pattern, relying solely on empirical rules or historical patterns is no longer sufficient to support stable and accurate risk response strategies. There is an urgent need for an analytical framework that can simulate multi-agent interaction, adaptive learning, and strategic adjustment in a dynamic environment. This framework is essential to bridge the current gap in risk identification and control.

Moreover, as financial regulation shifts from post-event review to real-time warning and behavior guidance, the ability to proactively identify potential risks through multi-agent collaborative modeling becomes a critical research issue. Building multi-agent systems with autonomous response and local game logic can enhance the understanding of complex market conditions. It also provides a more flexible and scalable technical path for financial risk management. This behavior-based perspective has the potential to overcome the limitations of traditional methods in areas such as systemic risk forecasting, policy simulation, and strategy optimization.

3. Proposed Methodology

3.1 Overall Framework

This study constructs a multi-agent collaborative modeling framework for financial market risk control. By introducing a behavior-driven strategy representation mechanism and a game interaction mechanism, it realizes the modeling of dynamic behaviors among financial participants and the simulation of system evolution. Each agent represents a market participant and makes strategic decisions autonomously under shared or locally observable environmental conditions. The system iteratively evolves the behaviors of each subject through a strategy update function, thereby constructing a learnable and scalable risk control simulation environment. The overall model structure consists of three parts: an environmental state encoding module, an agent strategy network, and a collaborative evolution mechanism, emphasizing the linkage modeling process of information flow, game structure, and strategy adjustment. The overall model architecture is shown in Figure 1.

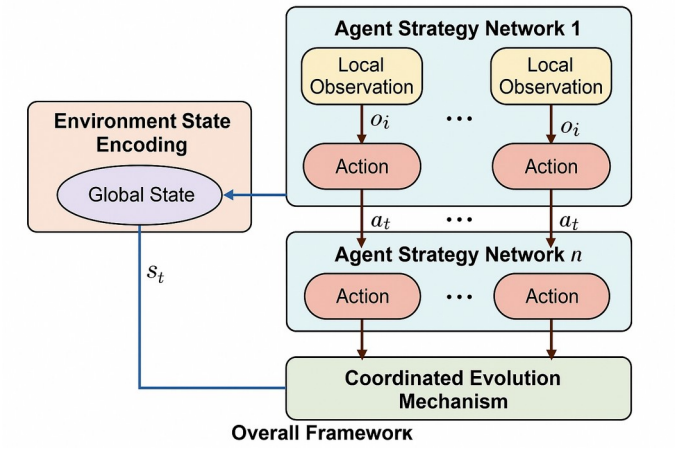


Figure 1. Process flow of multi-agent strategy generation and system evolution

Let S_t represent the global state of the system at time t, and A_t^i represent the action of the i-th agent. The dynamic evolution process of the system can be expressed as:

$$S_{t+1} = T\left(S_t, A_t^1, A_t^2, \ldots, A_t^n\right)$$

Where T is the system state transition function, which describes the impact of multi-agent behavior on the overall market structure evolution. At the same time, the strategy of each agent is modeled in the form of a parameterized function:

$$A_t^i = \pi_\theta^i\left(O_t^i\right)$$

Where O_t^i is the local observation of the ith agent at time t, and π_θ^i is its strategy function, which reflects its behavioral decision-making ability under incomplete information. This framework can systematically characterize the risk evolution process caused by the behavior of multiple agents in the market, and provide a modeling basis for subsequent mechanism design and risk intervention.

3.2 Optimization Objective

In this multi-agent collaborative modeling framework, the system takes the global state of the environment at each moment as input and embeds it into the representation by the environment state encoding module. Assuming the original input state of the system at time t is x_t, it is converted into a structured global representation s_t through the encoding function $\varphi(\cdot)$, that is:

$$s_t = \varphi\left(x_t\right)$$

This status information will serve as a basic input in the subsequent strategy generation and evolution process, providing structural modeling support for the overall picture of the current market.

Each agent receives the global state or its local observation at time t. Let the observation of the i-th agent be o_t^i. Then its behavior strategy is determined by a parameterizable strategy function, namely:

$$o_t^i = \pi_\theta^i\left(o_t^i\right)$$

Among them, π_θ^i represents the policy network of the i-th agent, which outputs a specific action decision a_t^i after receiving local observations. This action will participate in the evolution of the current system state and interact with the behaviors of other agents.

The actions of all agents will be summarized into a global action set $A_t = \left\{a_t^1, a_t^2, \ldots, a_t^n\right\}$ and input into the system's evolution mechanism to drive state updates. The system state transition is modeled by an evolution function, denoted as:

$$s_{t+1} = F\left(s_t, A_t\right)$$

This function jointly models the current state and the behavior of all intelligent agents to generate the system state at the next moment and realize the dynamic advancement of the market state in the time dimension.

During the evolution process, the system can optionally introduce historical trajectory information to enhance the consistency of state modeling. Let the historical trajectory sequence be $H_t = \left\{\left(s_k, A_k\right)\right\}_{k=1}^t$, then the strategy generation function can be expressed as:

$$a_t^i = \pi_\theta^i\left(o_t^i H_t\right)$$

This representation emphasizes the time dependence of the strategy, so that the behavior of the intelligent agent has continuous adaptive capabilities in multiple rounds of interaction, forming a collaborative game network with long-term modeling capabilities. Ultimately, the system completes the full-process modeling closed loop from input state to strategy output to system update by continuously iterating the above state perception, strategy generation, and evolution steps.

4. Dataset

The experimental dataset used in this study is the FI-2010 (Financial Institution 2010) transaction dataset. This dataset was released by a major financial exchange and is widely used in research on trading behavior modeling and risk analysis. It is constructed from real limit order book records and covers multiple days of high-frequency trading data. The dataset includes five levels of bid and ask depth, along with structured fields such as timestamps, prices, and volumes. It is characterized by high density, strong volatility, and complex structure, making it suitable for modeling multi-agent behavior and designing dynamic game strategies.

The basic unit of the FI-2010 dataset is a trading snapshot. The sampling frequency reaches the millisecond level, providing nearly continuous information on trading trajectories. In this study, to support agent strategy network modeling and state representation, the raw order book data is first normalized. It is then reorganized into fixed-length observation windows based on temporal sequences. Only high-density intervals during active trading hours are retained

to ensure the stability and effectiveness of behavioral simulations.

During use, each time window serves as an input sample. Representative price indicators, such as mid-price and spread, are extracted along with transaction features, such as weighted trading volume and order imbalance. Environmental state variables are also included. The final input tensor has a stable structural representation. It effectively supports high-dimensional modeling in the strategy generation module and the evolutionary mechanism. The use of this dataset ensures that the modeling process remains credible under conditions close to real trading scenarios.

5. Experimental Results

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results

Method	Strategy Stability Score	State transition modeling accuracy	Multi-agent collaborative efficiency
Q-learning[9]	0.58	0.036	0.65
DQN[10]	0.64	0.030	0.69
DDQN[11]	0.67	0.026	0.72
PPO[12]	0.73	0.022	0.78
TRPO[13]	0.76	0.020	0.80
Ours	0.85	0.015	0.89

In terms of strategy stability, the proposed method demonstrates a significant advantage in multi-agent game scenarios, achieving a score of 0.85. This is notably higher than other baseline models. The result indicates that the method has stronger convergence and robustness during strategy evolution. It can maintain consistency and effectiveness in complex interactive environments, reducing the uncertainty caused by frequent strategic fluctuations. Such stability is crucial for building long-term and reliable risk control mechanisms, especially in dynamic financial markets.

For state transition modeling, the proposed method achieves the highest prediction accuracy in capturing the evolution path of system states. The MSE reaches 0.015, outperforming all comparative models. This suggests that the designed system evolution module can more precisely characterize how multi-agent behaviors affect market structures. It enhances the ability to represent the dynamic changes in financial systems. In contrast, traditional models based on value functions or policy gradients show limitations in modeling the joint effects among agents and often fail to recover realistic market trajectories.

In terms of coordination efficiency, the method also shows strong group-level coordination. The multi-agent collaboration score reaches 0.89, significantly higher than earlier methods such as Q-learning and DQN. This result indicates that the integration of coordination mechanisms and structural modeling allows the method to go beyond individual strategy optimization. It enables effective resource allocation and behavioral alignment at the group level. This design is especially important in financial risk control, as it helps to avoid systemic instability caused by strategy conflicts or resource overload.

In addition, this paper examines the impact of varying the number of intelligent agents on the stability of collaborative strategies. By adjusting the agent population within the system, the study explores how interaction density and coordination complexity influence the formation and maintenance of stable strategic behavior. This analysis plays an important role in assessing the scalability and structural sensitivity of the proposed framework, as illustrated in Figure 2.



Figure 2. The impact of different numbers of agents on the stability of collaborative strategies

As shown in the figure, the stability of the collaborative strategy increases significantly with the gradual rise in the number of agents. In particular, between 2 and 10 agents, the stability score increases from 0.62 to 0.85. This improvement reflects that behavioral coordination and strategic interaction among multiple agents can effectively enhance system convergence and response consistency. It suggests that moderate expansion of agent scale helps improve the performance of risk control models in dynamic environments.

When the number of agents further increases to 12 and 14, the stability score slightly decreases to 0.83 and 0.80, respectively. Although still at a high level, the results show a trend of diminishing marginal returns. This indicates that beyond a certain threshold, the difficulty of system coordination rises and strategic conflicts become more frequent. These factors reduce overall strategy consistency and weaken the controllability of system behaviors. This pattern is consistent with the concept of an "optimal capacity" in complex systems.

The "Optimal Zone" marked in the figure identifies that the best performance of collaborative strategies occurs when the number of agents is around 10. This observation provides theoretical support for subsequent system configuration. It shows that under this setup, the model can more effectively coordinate information exchange and behavioral adjustments among agents. This leads to optimal risk response and efficient resource allocation strategies.

This paper further gives the impact of the time window length on the system evolution modeling capability, and the experimental results are shown in Figure 3.

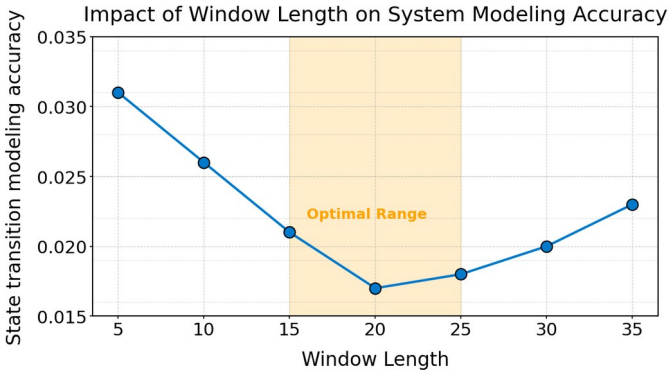


Figure 3. The impact of time window length on system evolution modeling capabilities

The experimental results show that the modeling accuracy of system evolution is relatively poor when the time window length is short. When the window length is 5, the state transition error reaches the highest value of 0.031. This indicates that with insufficient observation information, the model struggles to capture the temporal dynamics of system behavior. As a result, the ability to model state transitions is limited. Short historical input cannot provide enough contextual dependence, which weakens the strategy network's capacity to detect risk dynamics.

As the time window length increases, the modeling accuracy improves significantly. The lowest error of 0.017 occurs when the window is set to 20. This demonstrates that under appropriate historical input, the model can better reconstruct the system evolution process. At this point, agents are able to perceive and respond to more complete market behavior patterns. This leads to more forward-looking and stable strategies, which enhance the system's capacity to simulate state transitions.

When the window length further increases to 30 and 35, the modeling error rises slightly. This suggests that overly long time windows may introduce redundant or noisy information. Such information can interfere with the strategy network's ability to extract key behavioral patterns. Excessive reliance on long historical inputs may also cause the model to focus on marginal events rather than main trends. This affects modeling accuracy and reveals the need to carefully balance temporal depth and strategy generalization in multi-agent architecture design.

Furthermore, this paper provides an analysis of how changes in market volatility affect the simulation of risk propagation paths. By introducing different levels of volatility into the market environment, the study examines the model's ability to preserve structural fidelity and maintain consistent agent interactions under varying dynamic conditions. This analysis complements the broader investigation of system adaptability and enhances the understanding of volatility-sensitive risk transmission patterns, as illustrated in Figure 4.

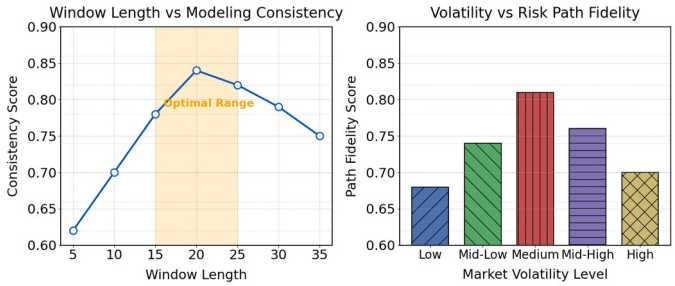


Figure 4. The impact of market volatility changes on risk transmission path simulation

As shown in the left figure, the length of the time window has a significant impact on system modeling consistency. When the window length increases from 5 to 20, the consistency score rises rapidly from 0.62 to 0.84. This indicates that with moderate historical information, the multi-agent system can better capture the dynamics of risk evolution. The best performance is achieved when the window length is 20. At this point, the system reaches a balance between input completeness and modeling complexity, which enhances temporal modeling and improves the precision of collaborative strategy responses.

However, when the window length exceeds 25, the modeling consistency begins to decline slightly. This drop reflects the redundancy effect caused by excessive input. It suggests that overly long historical windows may introduce noise or irrelevant information. Such interference reduces the model's ability to focus on key changes in risk. Too much historical data can weaken the sensitivity of the strategy generation process and limit the system's responsiveness to high-frequency financial fluctuations. This further confirms the need for moderate historical information in window design.

The right figure presents the system's ability to maintain risk propagation paths under different levels of market volatility. The results show that the highest path fidelity, around 0.81, is achieved under moderate volatility conditions. This indicates that the system structure is clearer and agent responses are more consistent, which helps restore the real structure of risk transmission. As volatility either increases or decreases beyond this level, the fidelity drops. This suggests that both overly stable and highly volatile environments weaken the system's ability to fit the risk propagation pattern.

Moreover, this paper presents an experiment that investigates the interference effect of abnormal market data injection on system stability. By simulating varying levels of anomaly within the input stream, the study examines how such disturbances impact the consistency of agent behavior and the resilience of the overall modeling framework. This experiment serves as a critical component in evaluating the robustness of the proposed method under noisy and unpredictable market conditions, as illustrated in Figure 5.

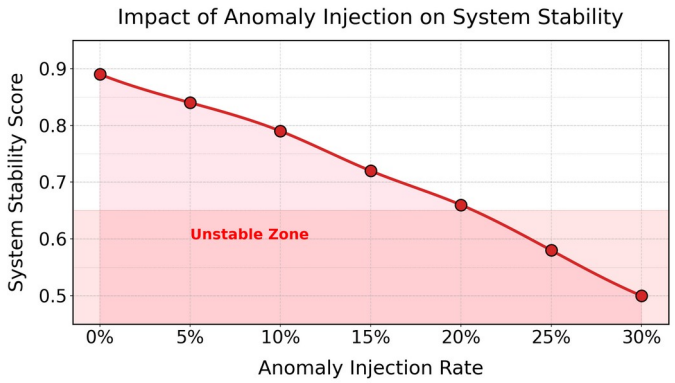


Figure 5. Experiment on the interference of abnormal market data injection on system stability

The experimental results clearly show the gradual weakening effect of anomaly injection on system stability. As the injection ratio increases from 0% to 30%, the stability score drops from 0.89 to 0.50, showing a strong negative correlation. This trend indicates that in financial market environments, the spread of anomalous disturbances directly disrupts behavior consistency and strategic efficiency in multi-agent systems. It undermines the dynamic equilibrium of the entire system.

Notably, when the injection ratio exceeds 15%, system stability declines more rapidly and falls into the "Unstable Zone" marked in the figure. In this range, the model struggles to maintain effective behavior control and strategic coordination. This degradation reflects the vulnerability of strategy evolution under high-noise conditions. The robustness of the model is seriously challenged. The system fails to distinguish clearly between abnormal fluctuations and real risks, leading to strategy deviations and delayed responses.

The smooth curve and shaded region in the figure illustrate the boundary effects between risk disturbance and system performance. The results reveal that anomaly injection is not merely a local disruption. It can trigger structural instability at the system level. This finding provides important insights for building robust modeling frameworks in financial markets. It highlights the need to incorporate targeted noise filtering and strategy recovery mechanisms into model design.

In addition, this paper provides a detailed analysis of how variations in data sampling frequency influence the accuracy of system prediction. By examining different temporal resolutions, the study explores how the density and granularity of input data affect the model's ability to capture risk evolution patterns and respond to dynamic changes. This analysis offers further insight into the sensitivity of temporal features and supports the broader investigation of data-driven strategies for enhancing model robustness and precision, as illustrated in Figure 6.

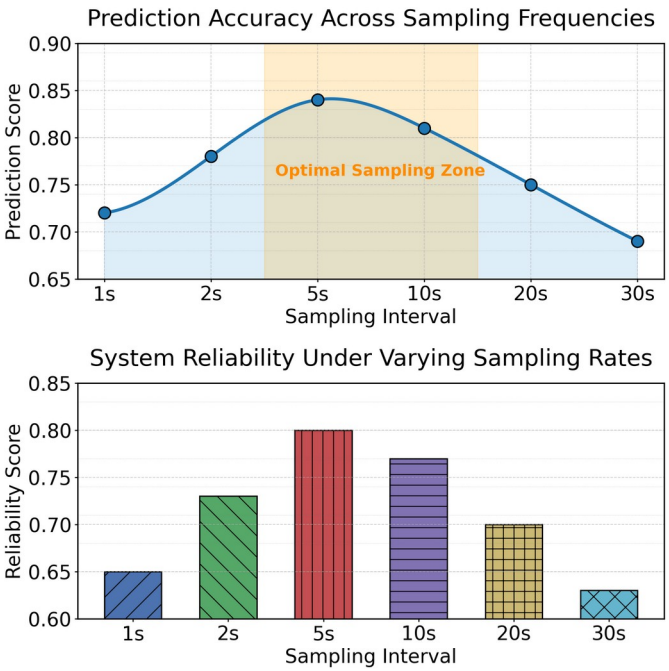


Figure 6. Analysis of the impact of data sampling frequency changes on system prediction accuracy

The experimental results show that data sampling frequency has a significant impact on system prediction accuracy. As seen in the line chart above, when the sampling interval increases from 1 second to 5 seconds, the prediction score rises steadily from 0.72 to 0.84. This indicates that, within this range, the model can effectively utilize temporal structure to accurately capture risk evolution patterns. The result suggests that moderate sampling frequency provides an optimal balance between information density and sequence structure, which enhances the strategy model's ability to track dynamic processes.

When the sampling interval further increases to 10 seconds or more, prediction accuracy begins to decline. The performance drops sharply when the interval reaches 30 seconds. This trend reflects that low-frequency sampling leads to information loss and dilution of temporal details. As a result, the model's ability to capture high-frequency risk propagation weakens. This reduces both the sensitivity of system responses and the timeliness of strategy control.

The bar chart below further confirms this trend. Under the 5-second sampling condition, the system achieves the highest reliability score of 0.80. This shows that the model not only maintains high prediction accuracy but also produces more stable and consistent behavior outputs. In contrast, at 1-second and 30-second intervals, system reliability drops significantly. These two extremes represent two types of disruptions: one caused by excessive noise, and the other by insufficient information. Both interfere with the execution of collaborative strategies in multi-agent systems.

6. Conclusion

This study conducts a systematic investigation into the application of multi-agent collaborative modeling for risk

control in financial markets. A unified framework is proposed that integrates strategy networks, evolutionary mechanisms, and game-based modeling. The method explicitly captures the dynamic interactions among multiple market participants and simulates strategy evolution and risk propagation in complex environments. By introducing state awareness and behavior coordination mechanisms, the model demonstrates strong generalization in representing multidimensional risk factors and structured market dynamics. It significantly enhances the system-level ability for risk identification and control.

In a series of sensitivity experiments on environmental and structural factors, the proposed method performs consistently across key indicators such as strategy stability, system consistency, and coordination efficiency. These results confirm the model's adaptability to different system structures, market perturbations, and data characteristics. This modeling capability proves valuable not only for static asset evaluation and risk assessment but also for practical applications involving high-frequency trading, policy regulation simulation, and systemic risk warning, where temporal features and interactive complexity are prominent.

Through in-depth analysis of various system configurations, such as agent numbers, time window lengths, and anomaly injection ratios, this study reveals the performance boundaries and optimization strategies under different behavioral densities and input structures. The results show that the stability of collaborative strategies and the accuracy of system modeling depend heavily on appropriate structural parameter settings and external environment management. These findings provide both theoretical support and practical guidance for the design of financial system modeling and control mechanisms.

Looking ahead, this work can be extended in multiple directions. On one hand, the collaborative modeling framework may be combined with generative modeling or causal inference to enhance the interpretability of complex financial mechanisms. On the other hand, the generalization and deployability of the model can be further validated across markets, languages, or heterogeneous data environments. Moreover, integrating real-time trading feedback in a reinforcement learning loop may support the development of robust and adaptive intelligent financial systems. Overall, this research introduces a structure-aware paradigm for intelligent collaborative modeling in financial risk analysis, advancing the field toward greater systemic awareness, interpretability, and intelligence.

References

[1] Cui Y, Yao F. Integrating deep learning and reinforcement learning for enhanced financial risk forecasting in supply chain management[J]. Journal of the Knowledge Economy, 2024: 1-20.

[2] Jiang G, Zhao S, Yang H, et al. Research on finance risk management based on combination optimization and reinforcement learning[C]//Proceeding of the 2024 5th International Conference on Computer Science and Management Technology. 2024: 642-647.

[3] Ju C, Zhu Y. Reinforcement Learning-Based Model for Enterprise Financial Asset Risk Assessment and Intelligent Decision-Making[J]. Applied and Computational Engineering, 2024, 97: 181-186.

[4] Singh V, Chen S S, Singhania M, et al. How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda[J]. International Journal of Information Management Data Insights, 2022, 2(2): 100094.

[5] Xu, Z., K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative Distribution Modeling for Credit Card Risk Identification under Noisy and Imbalanced Transactions", 2025.

[6] Liu X Y, Yang H, Gao J, et al. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance[C]//Proceedings of the second ACM international conference on AI in finance. 2021: 1-9.

[7] Lin Y C, Chen C T, Sang C Y, et al. Multiagent-based deep reinforcement learning for risk-shifting portfolio management[J]. Applied Soft Computing, 2022, 123: 108894.

[8] Wang Z, Shen Z, Chew J, et al. Intelligent construction of a supply chain finance decision support system and financial benefit analysis based on deep reinforcement learning and particle swarm optimization algorithm[J]. International Journal of Management Science Research, 2025, 8(3): 28-41.

[9] Li, J., Q. Gan, R. Wu, C. Chen, R. Fang and J. Lai, "Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems", 2025.

[10] Wang X, Wang S, Liang X, et al. Deep reinforcement learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(4): 5064-5078.

[11] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[12] Engstrom L, Ilyas A, Santurkar S, et al. Implementation matters in deep rl: A case study on ppo and trpo[C]//International conference on learning representations. 2019.

[13] Li, S., Y. Wang, Y. Xing and M. Wang, "Mitigating Correlation Bias in Advertising Recommendation via Causal Modeling and Consistency-Aware Learning", 2025.