Journal of Power and Energy Engineering, 2026, 14(4), 1-24
https://www.scirp.org/journal/jpee
ISSN Online: 2327-5901
ISSN Print: 2327-588X

Integrating Multi-Agent Reinforcement Learning and Evolutionary Game Theory for Adaptive Virtual Bidding Strategies in Electricity Markets

Wei Cao

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China
Email: 18803635628@163.com

How to cite this paper: Cao, W. (2026) Integrating Multi-Agent Reinforcement Learning and Evolutionary Game Theory for Adaptive Virtual Bidding Strategies in Electricity Markets. *Journal of Power and Energy Engineering*, **14**, 1-24.
https://doi.org/10.4236/jpee.2026.144001

Received: March 5, 2026
Accepted: March 27, 2026
Published: March 30, 2026

Abstract

As the electricity market is progressively liberalized, virtual bidding has emerged as a novel participation mechanism attracting increasing attention. This paper integrates evolutionary game theory with multi-agent reinforcement learning to propose an adaptive virtual bidding strategy for electricity market participants, enabling them to continuously adjust their bidding behavior in dynamic market environments to maximize individual profits. First, we apply multi-agent reinforcement learning to the fundamental virtual bidding model, incorporating risk constraints and corresponding limitations to construct a comprehensive bidding framework encompassing fixed costs, trigger conditions, and Conditional Value at Risk (CVaR) constraints. Second, we introduce heterogeneous virtual bidders with varying budget endowments and initialization configurations, rendering the model more representative of real-world electricity markets and thereby optimizing participants' strategy selection. Finally, considering the collective impact of virtual bidders' behavior on market prices and the influence of participant numbers on bidding strategies, we incorporate an evolutionary game model to analyze the evolutionary dynamics and equilibrium stability of virtual bidding strategies. Experimental results demonstrate that the MADDPG model achieves an average per-episode reward 13.7 times higher than that of independent DDPG, while exhibiting a five-fold faster convergence speed. The introduction of heterogeneous participants yields a 30.6% improvement in average returns compared with homogeneous settings. Evolutionary game analysis identifies multiple Nash equilibria across different scenarios, revealing the regulatory effects of participant numbers and price impact coefficients on equilibrium structures.

Keywords

Evolutionary Game Theory, Multi-Agent Reinforcement Learning, Electricity Market, Virtual Bidding, Strategy Optimization

1. Introduction

As the "dual carbon" targets continue to advance, the integration of high-proportion renewable energy has emerged as a core trend in global power system development. The intermittency and volatility of wind and solar generation, while exacerbating system operational uncertainty, also significantly amplify the deviation between day-ahead locational marginal price and real-time locational marginal price (LMP) in electricity markets. To promote price convergence and enhance market efficiency, mature electricity markets represented by PJM and CAISO in the United States have introduced the virtual bidding mechanism. This mechanism allows traders to engage in day-ahead and real-time price spread arbitrage without physical generation or load obligations, injecting price correction signals into the market through financial transactions and thereby optimizing resource allocation efficiency.

A substantial body of research has accumulated regarding virtual bidding. Early studies primarily focused on empirical assessments of this mechanism's impact on market efficiency, with emphasis on analyzing its effects on operational risk, transaction costs, and market performance. Regarding the market effects of virtual bidding, Reference [1] provides empirical evidence indicating that this mechanism significantly improves electricity market operational efficiency by suppressing price volatility. Reference [2] investigates the impact of virtual bidding on price volatility in the New York wholesale electricity market, finding that virtual bidding is associated with reduced volatility in both day-ahead and real-time markets. Employing game-theoretic approaches, References [3] [4] reveal the intrinsic operational dynamics of virtual bidding markets. Their findings demonstrate that as the proportion of rational virtual bidders increases, the price spread between day-ahead and real-time markets narrows significantly, with market equilibrium progressively advancing toward day-ahead price levels that achieve social welfare optimization. Most studies affirm the positive role of virtual bidding in converging locational marginal prices.

As market complexity increases, some research has begun examining traders' strategic behavior. Reference [5] formulates a chance-constrained portfolio selection problem based on day-ahead and real-time price spread distributions to determine optimal bidding strategies. Extending this optimization framework, Reference [6] proposes a machine-learning-based portfolio optimization framework for electricity market virtual bidding that incorporates risk constraints and price sensitivity, utilizing the proposed algorithmic virtual bidding trading strategy to evaluate portfolio profitability and efficiency in U.S. wholesale electricity markets.

Reference [7] introduces an online learning algorithm that allocates traders' budgets across K option bids in each trading session to maximize cumulative returns over a finite trading horizon, establishing the algorithm's convergence properties. Reference [8] presents a price-based general stochastic optimization framework for deriving optimal convergence bidding curves, with the model simultaneously generating both bid prices and quantities. Reference [9] employs a bilevel optimization approach to propose a profit-maximization model for physical bid participants, employing Conditional Value at Risk to measure the risk associated with different strategies. Reference [10] develops a bilevel stochastic optimization model for strategic retailers' optimal joint demand and virtual bidding strategies in short-term electricity markets, demonstrating how virtual bidding can enhance retailers' market power in day-ahead markets and validating the proposed model's effectiveness through case studies. Reference [11] establishes a short-term decision model for electricity retailers incorporating battery energy storage systems and virtual bidding through a two-stage stochastic optimization framework. Reference [12] employs deep reinforcement learning, utilizing deep Q-networks to interact with the environment, obtain information feedback, and optimize neural network parameters to effectively solve for optimal bidding strategies. These studies reveal that participants with information advantages or computational capabilities can design strategies that outperform simple random bidding through historical data mining or statistical arbitrage models, thereby obtaining excess returns.

Currently, most research on bidding strategies adopts the core assumption of single-agent optimization, treating the market environment as static and neglecting or simplifying the feedback effects arising from other participants' strategic adjustments. The virtual bidding market fundamentally constitutes a complex system involving multiple rational traders engaged in strategic interaction: an individual trader's payoff depends critically on the collective behavior of other participants. When multiple participants simultaneously adopt similar or divergent strategies, the market microstructure undergoes dynamic evolution, with original arbitrage opportunities potentially disappearing rapidly and new gaming equilibria subsequently forming.

In summary, this paper develops a research framework for virtual bidding strategies that integrates evolutionary game theory with multi-agent reinforcement learning. Building upon classical virtual bidding models, we incorporate a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework with additional constraints to construct a comprehensive bidding model that more accurately reflects market realities and incorporates risk-aversion mechanisms. This approach captures the dynamic characteristics of market prices while achieving optimal trade-offs between returns and risks. At the participant heterogeneity level, we design heterogeneous virtual bidders with varying budget scales and initial strategy endowments to simulate the coexistence patterns of diverse trading entities in real-world markets. Furthermore, we introduce evolutionary game theory (EGT) to extend the virtual bidding problem to a population dynamic evolutionary per-

spective, systematically examining the feedback effects of virtual bidders' collective behavior on market prices and the influence of participant number variations on strategy selection and evolutionary stability mechanisms. Empirical analysis utilizing data from the NYISO electricity market (January–December 2025) validates the effectiveness and advantages of the proposed model and algorithms.

2. A Virtual Bidding Model Based on Multi-Agent Reinforcement Learning

2.1. Multi-Agent Markov Game

The multi-agent virtual bidding problem is formulated as a partially observable Markov game $\mathcal{G}=\left\langle \mathcal{N},\mathcal{S},\{\mathcal{A}_i\}_{i\in\mathcal{N}},\mathcal{T},\{\mathcal{R}_i\}_{i\in\mathcal{N}},\gamma\right\rangle$, where $\mathcal{N}=\{1,\cdots,N\}$ is the set of competing agents, $\mathcal{A}_i\subset\mathbb{R}^{d_a}$ is the continuous action space, and $\gamma\in(0,1)$ is the discount factor. Each agent i follows a deterministic policy π_{θ_i} aiming to maximize $J(\theta_i)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_i^t\right]$.

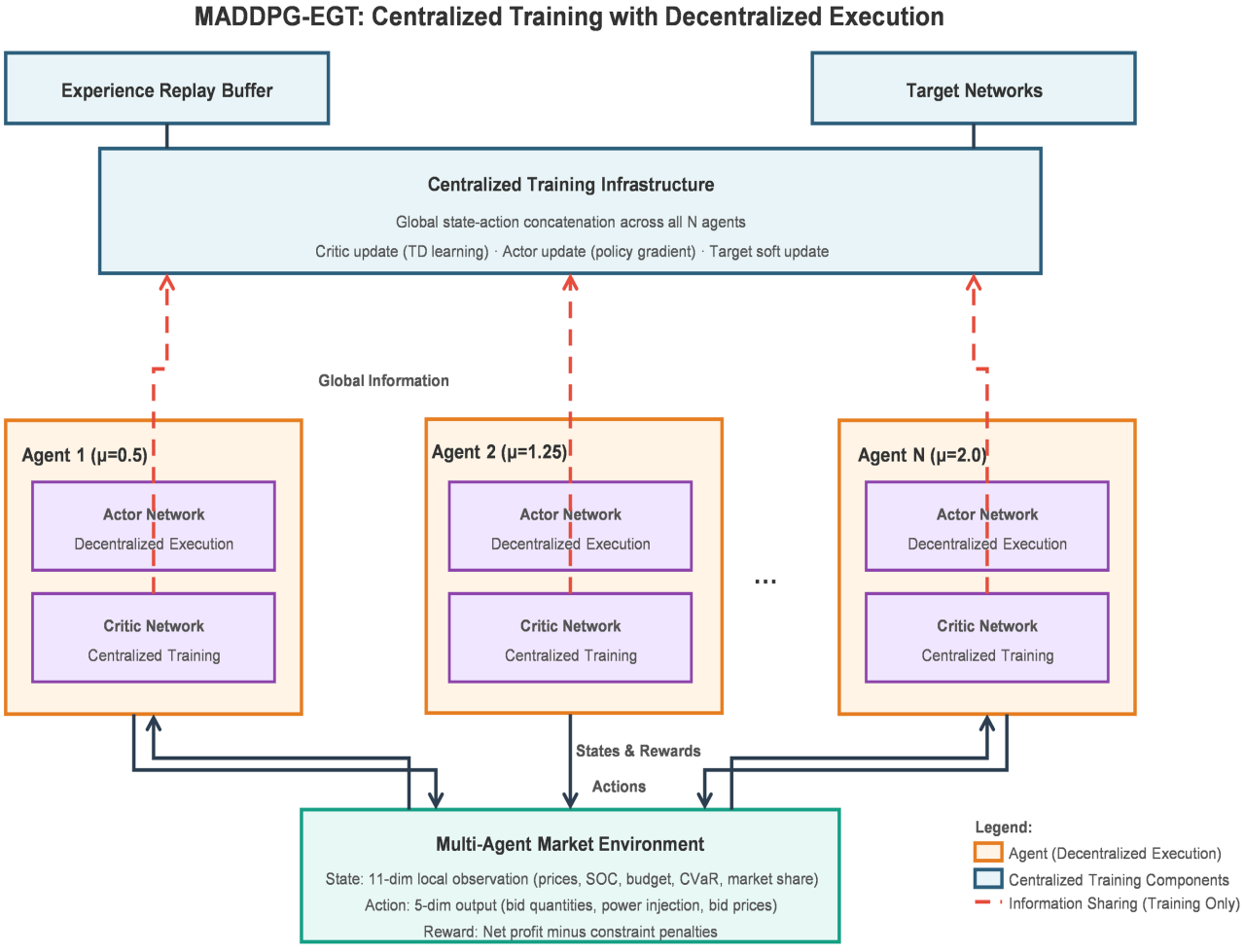

Figure 1. Overall system architecture of the MADDPG-EGT framework.

The overall architecture of the proposed MADDPG-EGT framework is illustrated in Figure 1. Competing virtual bidding agents interact with the NYISO day-ahead and real-time electricity markets under centralized training and decentralized execution. Post-training EGT analysis verifies strategy stability.

To facilitate a better understanding of this model and method, this section lists some key symbols and explains their general meanings, as shown in Table 1.

Table 1. Key notation.

Symbol	Dimension	Description
s_i^t	$\mathbb{R}^{11}$	Local observation of agent i at time t
a_i^t	$\mathbb{R}^5$	Action of agent i at time t
π_{θ_i}	$\mathcal{S}_i \to \mathcal{A}_i$	Deterministic policy of agent i
$Q_{\phi_i}^{\pi}$	$\mathcal{S} \times \mathcal{A} \to \mathbb{R}$	Centralized Q-function for agent i
$\mathcal{D}$	—	Experience replay buffer
p_{DA}^A, p_{DA}^B	$\mathbb{R}^+$	Day-ahead prices at nodes A, B
p_{RT}^A, P_{RT}^B	$\mathbb{R}^+$	Real-time prices at nodes A, B
z_I^i, z_D^i	$[0, z_{\max}]$	Increment/decrement bid quantities
γ_I, γ_D	$\mathbb{R}^+$	Fixed transaction costs per MWh
β	$(0,1)$	CVaR confidence level
μ_i	$\mathbb{R}^+$	Budget scaling multiplier for agent i
Π	$\mathbb{R}^{N \times N}$	Empirical payoff matrix
x	Δ^{N-1}	Population strategy distribution

At each time step t, agent i observes a local state vector $s_i^t \in \mathbb{R}^{11}$:

$$s_i^t = \left[p_{DA}^A,\ p_{DA}^B,\ \Delta p,\ h,\ SOC_i,\ B_i,\ \bar{\pi}_i,\ \sigma_i,\ CVaR_i,\ \bar{q}_{-i},\ m_i \right] \tag{1}$$

where $\Delta p = p_{DA}^A - p_{DA}^B$ is the inter-nodal price spread; $h \in [0,1]$ is the normalized hour of day; $SOC_i \in [0,1]$ is the state of charge; $B_i \in [0,1]$ is the normalized cumulative budget expenditure; $\bar{\pi}_i$ and σ_i are the exponential moving average and rolling standard deviation of historical profit; $CVaR_i$ is the agent's current tail-risk estimate; $\bar{q}_{-i}$ is the average bid volume of opponents; and $m_i \in [0,1]$ is the market share of agent i.

The state incorporates three information layers: i) exogenous market signals (p_{DA}^A, p_{DA}^B, Δp, h), ii) self-monitoring variables (SOC_i, B_i, $\bar{\pi}_i$, σ_i, $CVaR_i$), and iii) opponent behavior indicators ($\bar{q}_{-i}$, m_i).

Each agent outputs a continuous action $a_i \in \mathbb{R}^5$:

$$a_i = \left[z_I^i,\ z_D^i,\ P^i,\ b_I^i,\ b_D^i \right] \tag{2}$$

where $z_I^i, z_D^i \in [0,100]$ MWh are bid quantities, $P^i \in [-50,50]$ MW is physical injection, and $b_I^i, b_D^i \in [0,1000]$ \$/MWh are bid prices. Raw network outputs are

projected onto the feasible set $\mathcal{C}$ via Euclidean projection to enforce physical constraints:

$$a_i^{feasible} = *\arg\min_{a\in\mathcal{C}} \left\| a - a_i^{raw} \right\|_2 \tag{3}$$

While virtual bidding is fundamentally a financial instrument without physical delivery obligations, the action variable p^i represents an optional physical injection capacity that certain market participants (e.g., demand response aggregators or storage operators) may possess alongside their virtual bidding activities.

In this study, agents with $p^i \neq 0$ are modeled as hybrid participants who can simultaneously engage in virtual arbitrage and physical balancing. Pure virtual bidders are represented by constraining $p^i = 0$ during training. This formulation allows the framework to accommodate both pure financial traders and physical-virtual hybrid entities commonly observed in real-world markets.

For the baseline experiments reported in Section 4, all agents are configured as pure virtual bidders ($p^i = 0$) unless otherwise stated.

2.2. Risk-Aware Virtual Bidding Model

An INC bid is accepted only when $b_I^i \le p_{DA}$; a DEC bid when $b_D^i \ge p_{DA}$. The respective revenues are:

$$R_{INC}^i = z_I^i \left(s - \gamma_I\right) 1_{\left\{b_I^i \le p_{DA}\right\}}, \quad R_{DEC}^i = z_D^i \left(-s - \gamma_D\right) 1_{\left\{b_D^i \ge p_{DA}\right\}} \tag{4}$$

where $s = p_{RT} - p_{DA}$ is the price spread and $\gamma_I = \gamma_D = 0.1$ \$/MWh are fixed transaction costs calibrated to NYISO market data. Proportional transaction costs scale with bid volume: $C_{trans}^i = \lambda\left(z_I^i + z_D^i\right)\overline{p}_{clear}$, where $\lambda = 0.001$ and $\overline{p}_{clear}$ is the average clearing price. The instantaneous net profit is thus:

$$\pi_i^t = R_{INC}^i + R_{DEC}^i - C_{trans}^i \tag{5}$$

Large bidding volumes compress price spreads. We model this via a hyperbolic impact factor ($\alpha = 0.01$, calibrated from 8,760 hourly NYISO observations):

$$\phi\left(Q_{total}\right) = \frac{1}{1 + \alpha Q_{total}}, s_{eff} = s \cdot \phi\left(Q_{total}\right) \tag{6}$$

where $Q_{total} = \sum_{i=1}^{N}\left(z_I^i + z_D^i\right)$ is aggregate bidding volume.

To control tail risk, we define the portfolio loss $L = -\sum_{i=1}^{N}\left(R_{INC}^i + R_{DEC}^i\right)$ and its Conditional Value-at-Risk (VaR):

$$CVaR_\beta\left(L\right) = \mathbb{E}\left[L \mid L \ge VaR_\beta\left(L\right)\right], \quad VaR_\beta\left(L\right) = \inf\left\{\ell \in \mathbb{R} : \mathbb{P}\left(L \le \ell\right) \ge \beta\right\} \tag{7}$$

which quantifies the expected loss in the worst $\left(1-\beta\right)\times 100\%$ of scenarios. The full reward combines profit with budget, market share, and tail-risk penalties:

$$\begin{aligned} r_i^t = \pi_i^t &- \underbrace{\kappa_B \max\left(0, U_i - B_{\max}\right)}_{\text{budget-penalty}} - \underbrace{\omega \max\left(0, m_i - m_{th}\right)\cdot 1000}_{\text{market-share-penalty}} \\ &- \underbrace{\xi \max\left(0, CVaR_\beta\left(L\right) - C\right)}_{\text{tail-risk-penalty}} \end{aligned} \tag{8}$$

where $U_i = \sum_t \left(z_I^{i,t} + z_D^{i,t}\right)\left|p_{DA}^t\right|$ is cumulative expenditure; $m_i = \left(z_I^i + z_D^i\right)/Q_{total}$ is market share; $m_{th} = 0.6$ is the concentration threshold; C is the CVaR upper bound; and penalty coefficients $\kappa_B = 0.1$, $\xi = 0.5$ are set per Section 4.

2.3. Multi-Agent Deep Deterministic Policy Gradient

MADDPG resolves multi-agent nonstationarity by training critics with global state-action information while executing actors with local observations only. The centralized critic for agent i is:

$$Q_{\phi_i}^{\pi}(s,a) = Q_{\phi_i}^{\pi}\left(\left[s_1,\cdots,s_N\right],\left[a_1,\cdots,a_N\right]\right) \tag{9}$$

The actor network maps local state to action through two ReLU hidden layers (256 units) with tanh output:

$$\begin{aligned} h_1 &= \mathrm{ReLU}\left(W_1 s_i + b_1\right) \\ h_2 &= \mathrm{ReLU}\left(W_2 h_1 + b_2\right) \\ a_i &= \mathrm{scale}\left(\tanh\left(W_3 h_2 + b_3\right)\right) \end{aligned} \tag{10}$$

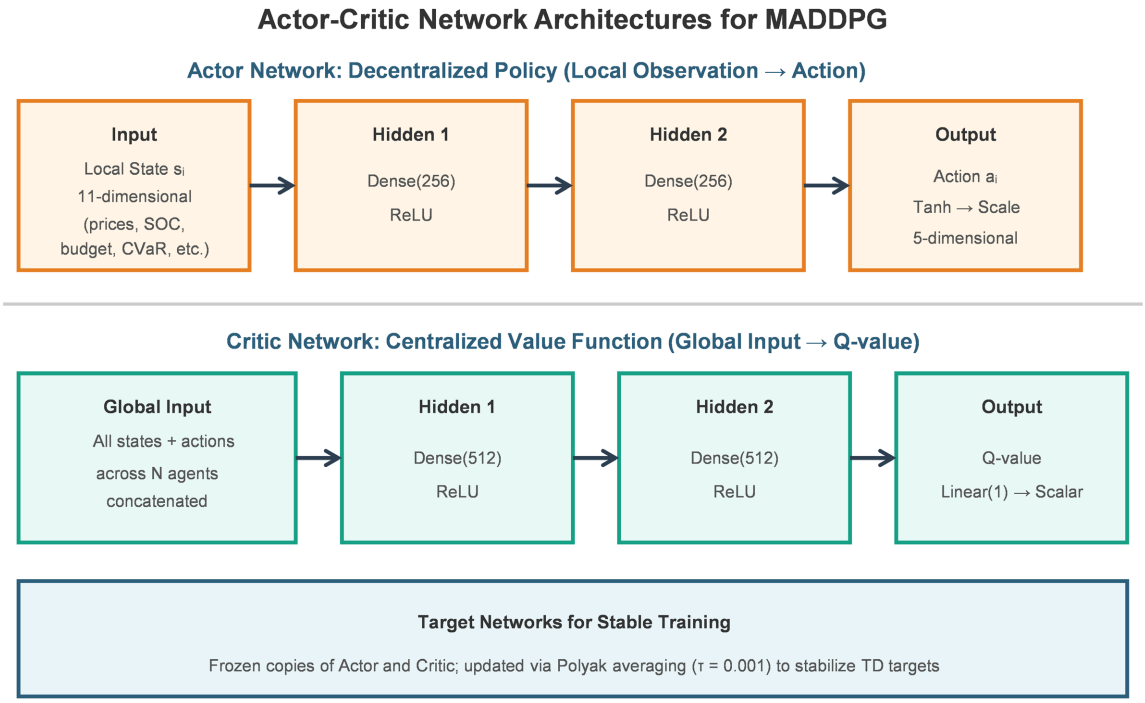

Figure 2. Decentralized actor and centralized critic network architectures.

The detailed structures of the decentralized actor and centralized critic networks are shown in Figure 2. Each Actor receives an 11-dimensional local observation and outputs a 5-dimensional action through two Rectified Linear Unit (ReLU) hidden layers (256 units) and a tanh (Hyperbolic Tangent) output layer. Each Critic receives the concatenated global state-action pair and produces a scalar Q-value through two ReLU hidden layers (512 units).

The critic minimizes the Bellman error:

$$\mathcal{L}(\phi_i) = \mathbb{E}\left[\left(Q_{\phi_i}^{\pi}(s,a) - y_i\right)^2\right], y_i = r_i + \gamma Q_{\phi_i^-}^{\pi}(s',a')\Big|_{a'_j=\pi_{\theta_j^-}(s'_j)} \tag{11}$$

The actor updates via deterministic policy gradient:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}\left[\nabla_{\theta_i}\pi_{\theta_i}(s_i)\nabla_{a_i} Q_{\phi_i}^{\pi}(s,a)\Big|_{a_i=\pi_{\theta_i}(s_i)}\right] \tag{12}$$

Target networks are updated via Polyak averaging with rate $\tau = 0.001$:

$$\phi_i^- \leftarrow \tau\phi_i + (1-\tau)\phi_i^-, \theta_i^- \leftarrow \tau\theta_i + (1-\tau)\theta_i^- \tag{13}$$

Exploration uses Ornstein-Uhlenbeck noise ($\theta_{OU} = 0.15$, $\sigma_{OU} = 0.2$) to generate temporally correlated action perturbations.

To prevent policy collapse to homogeneous equilibria, agents are initialized with differentiated budgets. The notation μ_i (distinct from CVaR confidence β) denotes the budget scaling multiplier:

$$B_i = B_{base} \cdot \mu_i, \mu_i \in \{0.5, 0.875, 1.25, 1.625, 2.0\} \tag{14}$$

In addition, the relevant hyperparameter settings are shown in Table 2.

Table 2. Hyperparameter settings.

Parameter	Value	Description
N	5	Number of agents
γ	0.99	Discount factor
τ	0.001	Polyak averaging rate
α	0.01	Price impact coefficient
γ_I, γ_D	0.1 $/MWh	Fixed transaction costs
λ	0.001	Proportional transaction cost rate
β	0.95	CVaR confidence level
κ_B	0.1	Budget penalty coefficient
ξ	0.5	CVaR penalty coefficient
m_{th}	0.6	Market share threshold
θ_{OU}	0.15	OU noise mean-reversion rate
σ_{OU}	0.2	OU noise diffusion coefficient
Actor hidden units	256	Per hidden layer
Critic hidden units	512	Per hidden layer

2.4. Evolutionary Game Analysis of Virtual Bidding Strategy Optimization

After training, all pairwise matchups are evaluated to build the empirical payoff matrix $\Pi \in \mathbb{R}^{N\times N}$:

$$\Pi_{ij} = \frac{1}{M \cdot T}\sum_{m=1}^{M}\sum_{t=1}^{T}\pi_i^{m,t}(i \text{ vs } j) \tag{15}$$

Strategy frequencies evolve according to:

$$\dot{x}_i = x_i\left[\left(\Pi x\right)_i - x^{\mathrm{T}}\Pi x\right] \tag{16}$$

A strategy x^* is an evolutionarily stable strategy (ESS) if $\forall x \neq x^*$: $x^{\mathrm{T}}\Pi x^* < \left(x^*\right)^{\mathrm{T}}\Pi x^*$, and a Nash equilibrium if $x_i^* > 0$ $\left(\Pi x^*\right)_i = \max_j\left(\Pi x^*\right)_j$. Asymptotic stability is confirmed by checking that all eigenvalues of the Jacobian $J_{ij} = \partial\dot{x}_i/\partial x_j\big|_{x=x^*}$ have negative real parts.

After MADDPG training converges, each trained policy π_i is frozen and evaluated in all $N\left(N-1\right)/2$ pairwise matchups. For each pair $\left(i,j\right)$, agents i and j compete over $K = 500$ evaluation episodes, each spanning 24 hours (one trading day).

To ensure statistical robustness, this evaluation is repeated across $R = 10$ independent random seeds (controlling initial state sampling and tie-breaking in bid acceptance), yielding 5000 total episodes per matchup. The payoff matrix entry Π_{ij} is computed as the mean cumulative reward of strategy i when facing strategy j, averaged over all $K \times R$ episodes.

In the post-training EGT analysis, a "strategy" corresponds to a distinct trained policy π_i characterized by its unique budget multiplier μ_i and network initialization. Stability analysis (eigenvalue computation of the Jacobian at fixed points) is performed numerically using the replicator dynamics solver with integration tolerance 10^{-6}.

3. Training Pipeline

The overall training procedure consists of two sequential phases, as illustrated in Figure 3.

Phase 1—MADDPG Training. Each episode begins by resetting the NYISO market environment and drawing an initial state. At every time step t, each agent independently selects an action via its local actor network augmented with Ornstein-Uhlenbeck exploration noise, then clips the output to the feasible constraint set $\mathcal{C}$. All agents execute their joint action, receive the resulting state transition and individual rewards, and store the tuple $\left(s,a,r,s'\right)$ in a shared replay buffer $\mathcal{D}$. When $\mathcal{D}$ contains sufficient samples and the update interval is reached, a mini-batch is sampled and centralized critic and actor updates are performed for every agent according to Equations (11)–(13); target networks are then soft-updated. This inner loop repeats until the episode terminates, after which a new episode begins. Training runs for M episodes in total.

Phase 2—EGT Analysis. After training converges, the learned policies are frozen and evaluated exhaustively in all $N\left(N-1\right)/2$ pairwise matchups over M episodes to construct the empirical payoff matrix Π using Equation (15). Replicator dynamics, given by Equation (16), are then integrated numerically to trace strategy-frequency trajectories and identify fixed points. Each fixed point is classified as a Nash equilibrium or ESS by verifying the stability conditions on the Jacobian eigenvalues described in Section 2.4.

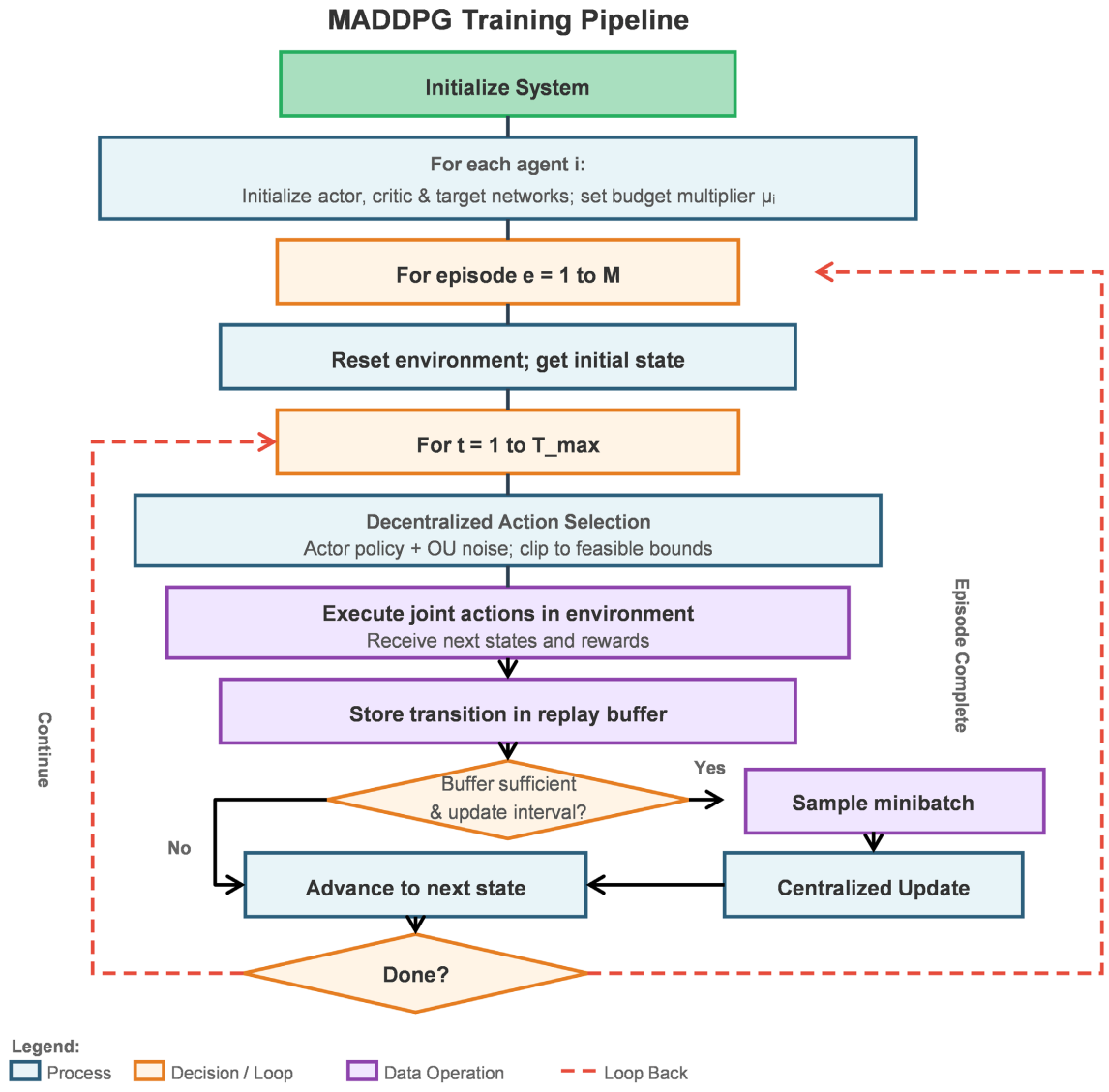

Figure 3. End-to-end training and analysis pipeline.

Phase 1 (MADDPG Training): agents interact with the NYISO market environment, storing transitions in a shared replay buffer for critic and actor updates. Phase 2 (EGT Analysis): trained policies are evaluated in all pairwise matchups to construct the payoff matrix; replicator dynamics are solved to identify Nash equilibria and ESS.

4. Case Study Analysis

4.1. Basic Settings

This study utilizes hourly locational marginal price (LMP) data from two NYISO nodes: Node A (Zone J - New York City) and Node B (Zone A - Western New York), spanning January 1 to December 31, 2025, yielding 8,760 hourly observations per node. The dataset comprises four primary variables: day-ahead LMP at

nodes A and B (p_{DA}^{A}, p_{DA}^{B}), and real-time LMP at nodes A and B (p_{RT}^{A}, p_{RT}^{B}), all denominated in \$/MWh.

Preprocessing steps include: i) removal of outliers exceeding three standard deviations from the rolling 168-hour mean, ii) forward-filling of missing values (<0.2% of observations), and iii) min-max normalization to [0,1] for neural network input stability.

The dataset is partitioned chronologically: January-August (5832 hours) for training, September (720 hours) for hyperparameter tuning, and October–December (2208 hours) for final evaluation. This temporal split ensures no data leakage, as future price information is strictly withheld during training and validation phases.

The price impact coefficient $\alpha = 0.01$ is estimated via ordinary least squares regression of observed price spreads ($p_{RT} - p_{DA}$) against aggregate virtual bidding volumes reported in NYISO public data (January-August 2025), yielding $R^2 = 0.68$.

Fixed transaction costs $\gamma_I = \gamma_D = 0.1$ \$/MWh are set to match NYISO's published virtual bidding transaction fees. The proportional cost rate $\lambda = 0.0001$ reflects typical bid-ask spreads observed in the dataset. The CVaR confidence level $\beta = 0.95$ follows industry-standard risk management practices [10] [12]. Penalty coefficients $\kappa_B = 0.1$ and $\xi = 0.5$ are determined through grid search over $\{0.05, 0.1, 0.2\} \times \{0.3, 0.5, 0.7\}$ to maximize training stability (measured by reward variance) while maintaining CVaR compliance rates above 90%. The market share threshold $m_{th} = 0.6$ is adopted from U.S. Federal Energy Regulatory Commission guidelines on market concentration.

4.2. Algorithm Performance Analysis and Comparison

To validate the effectiveness of the multi-agent reinforcement learning framework in virtual bidding strategy learning, this paper constructs a simulation environment based on historical data from the NYISO electricity market. A systematic comparison is conducted across four algorithms—MADDPG, Independent DDPG, Random, and Fixed—from three dimensions: learning dynamics, training performance, convergence behavior, and strategy stability.

As shown in **Figure 4**, MADDPG exhibits rapid ascent in the early training phase and converges to high-level returns after a certain number of episodes. In contrast, Independent DDPG demonstrates slower ascent and lower returns, reflecting the learning efficiency loss caused by neglecting strategic interactions. The Random strategy fluctuates at low return levels, while the Fixed strategy, although outperforming Random, is eventually surpassed by learning-based algorithms in later stages. The MADDPG curve displays minor fluctuations during the mid-training phase, attributable to strategy oscillations during the multi-agent co-adaptation process. As training progresses, the agent population gradually converges toward an evolutionarily stable equilibrium, and the curve becomes smoother, validating the adaptability of MADDPG in multi-agent interactive environments.

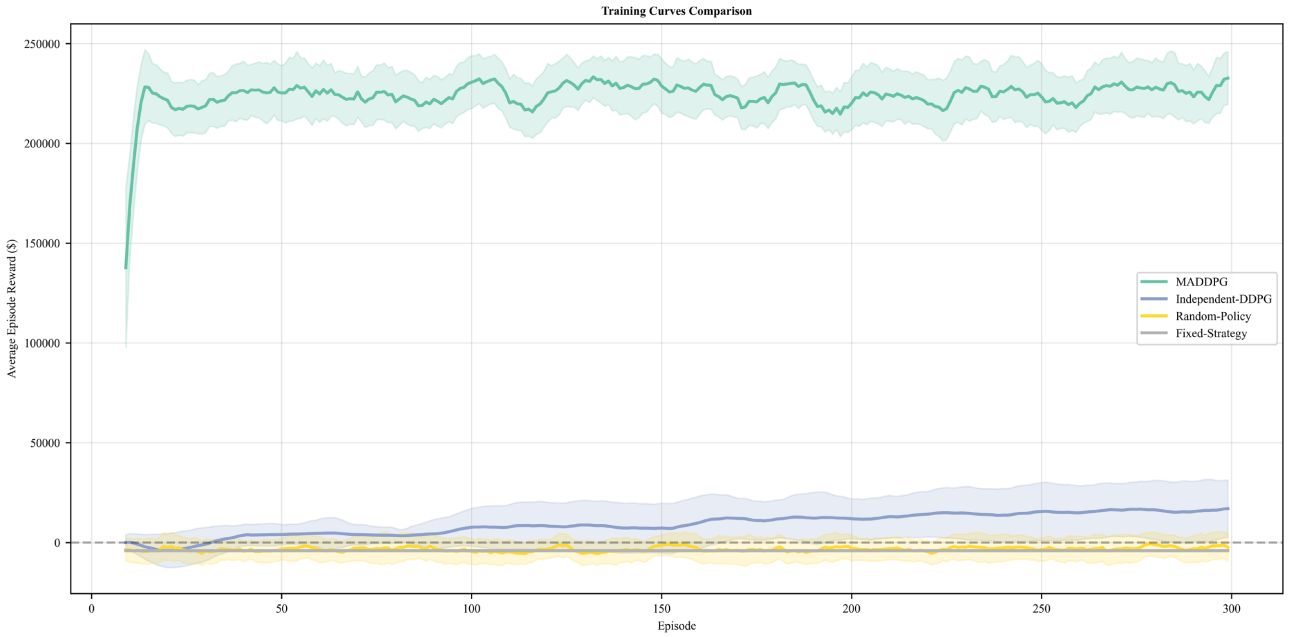

Figure 4. Dynamic comparison of training processes across different algorithms.

Figure 5 quantifies the convergence results of each algorithm at the end of training. The final average return of MADDPG significantly outperforms the other three algorithms, which is highly consistent with the training dynamics shown in Figure 4, further validating the convergence advantage of MADDPG in multi-agent virtual bidding environments. Its centralized training with decentralized execution architecture effectively models the strategic interactions among multiple agents, enabling agents to approximate Nash equilibria during the gaming process and thereby achieve higher returns in highly uncertain markets. The substantial return gap between MADDPG and Independent DDPG quantifies the performance loss resulting from neglecting strategic interactions. Furthermore, MADDPG exhibits the lowest error bars, indicating that it combines high returns with strong strategy stability—a characteristic particularly critical in real-world markets, where participants pursue not only return maximization but also tail risk control, aligning with the CVaR constraints introduced in this paper.

The simulation results in Figure 6 demonstrate that MADDPG achieves the fastest convergence speed, followed by Independent DDPG. The Random strategy fails to converge due to its stochastic nature, while the Fixed strategy is stable from initialization; both serve as baseline references, highlighting the convergence costs of learning-based algorithms. The differences in convergence speed stem from each algorithm's capacity to handle non-stationarity in multi-agent environments. MADDPG leverages global information through centralized training to guide policy updates, effectively reducing the variance of policy gradient estimates and thereby accelerating cooperative convergence. In contrast, Independent DDPG treats other agents' policy changes as environmental noise, subjecting each agent to a non-stationary Markov decision process that significantly delays con-

vergence. These results further substantiate the superiority of MADDPG from the perspective of computational efficiency.

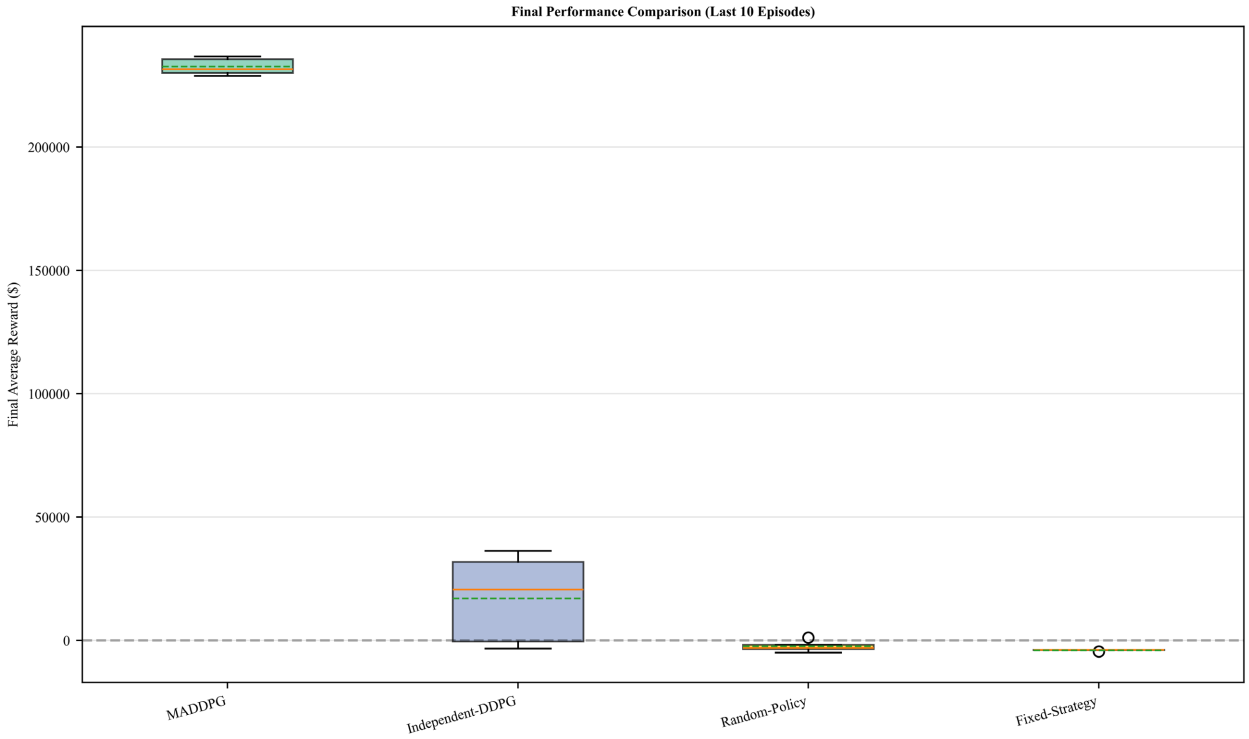

Figure 5. Final performance comparison of different algorithms.

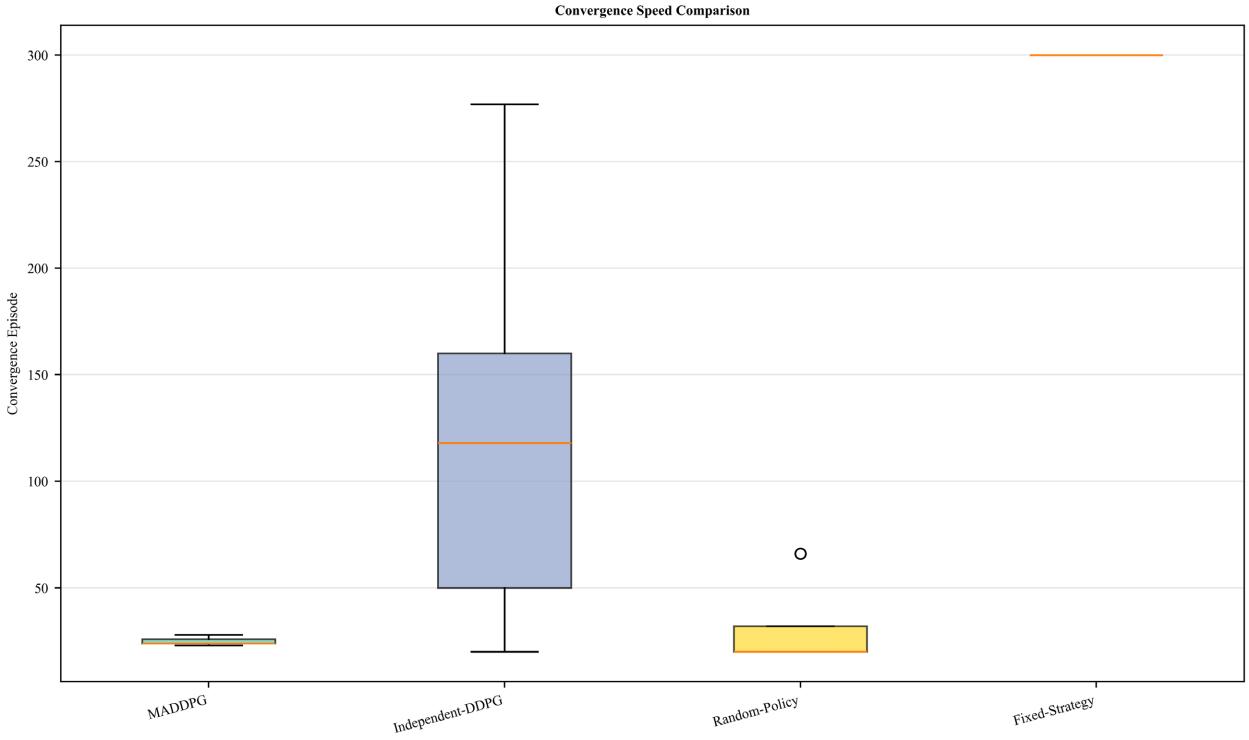

Figure 6. Comparison of convergence rates among different algorithms.

In summary, the MADDPG framework effectively models strategic interactions and gaming dynamics in contexts involving multiple coexisting traders. It guides the agent population to converge toward equilibrium states, exhibits resilience against minor individual policy perturbations, and possesses evolutionary stability, thereby providing a powerful computational tool for analyzing micro-level behavior in virtual bidding markets.

Table 3. Model and algorithm parameter settings.

Method	Final Reward ($)	Standard Deviation ($)	Convergence Rate	Total Reward ($)
MADDPG	2.33e+5 ± 3.1e+3	3.1e+3	25	6.67e+7
Independent DDPG	1.7e+4 ± 1.63e+4	1.63e+4	125	2.75e+6
Random	−2437.09 ± 2.06e+3	2.06e+3	32	−9.89e+5
Fixed	−3.99e+3 ± 307.93	307.93	300	−1.19e+6

Table 4. Statistical significance test results.

Comparison	t-Statistic	p-Value	Cohen's d	Significant
MADDPG vs Independent DDPG	26.05	$p < 0.001$	16.48	√
MADDPG vs Random	126.24	$p < 0.001$	79.84	√
MADDPG vs Fixed	151.78	$p < 0.001$	95.99	√
Independent DDPG vs Random	2.37	0.05	1.5	√
Independent DDPG vs Fixed	2.58	0.03	1.63	√
Random vs Fixed	1.49	0.17	0.94	×

Table 3 quantifies the comprehensive performance of each algorithm across four dimensions: return level, stability, convergence speed, and cumulative returns. MADDPG achieves an average per-episode return of $2.32e+5, substantially outperforming Independent DDPG by a factor of 13.7, which is consistent with the results presented in Figure 4 and Figure 5. Regarding stability, the standard deviation of MADDPG is considerably lower than that of Independent DDPG, corroborating the observation in Figure 5 that MADDPG exhibits the lowest error bars. This finding reflects the robust tail risk control capability of strategies incorporating CVaR constraints. In terms of convergence efficiency, MADDPG achieves a five-fold faster convergence speed compared with Independent DDPG, aligning with the results in Figure 6. The cumulative return metric further amplifies these performance differences.

Table 4 provides statistical validation for the aforementioned differences through pairwise tests. All comparisons between MADDPG and the baseline algorithms exhibit extreme statistical significance ($p < 0.0001$), with Cohen's d effect sizes ranging from 16 to 96, indicating that the advantages of MADDPG are both statistically significant and practically meaningful. The differences between Independent DDPG and both Random and Fixed strategies also reach statistical significance ($p < 0.05$). Collectively, Table 3 and Table 4 systematically establish the

comprehensive superiority of MADDPG in multi-agent virtual bidding scenarios from the dual perspectives of quantitative performance and statistical verification.

4.3. Evolutionary Game Analysis

4.3.1. Dynamic Analysis of Strategy Evolution

To intuitively illustrate the evolutionary paths and equilibrium convergence processes of agent strategies in the virtual bidding market, Figure 7 and Figure 8 depict the strategy evolution phase diagrams for the $N = 2$ and $N = 3$ scenarios, respectively.

Figure 7 depicts the two-dimensional phase space for the two-agent system. Both axes represent the strategy proportion of Agent 1, which, by symmetry, reflects the population strategy distribution. The red marker at (1,1) indicates a pure-strategy equilibrium, while the green trajectories show evolutionary paths from different initial states. All trajectories converge to a stable region near the diagonal, confirming the existence of evolutionarily stable equilibria: agents consistently converge toward uniform bidding behavior regardless of initial conditions. The distribution of convergence points along the diagonal suggests a continuum of equilibria rather than isolated points.

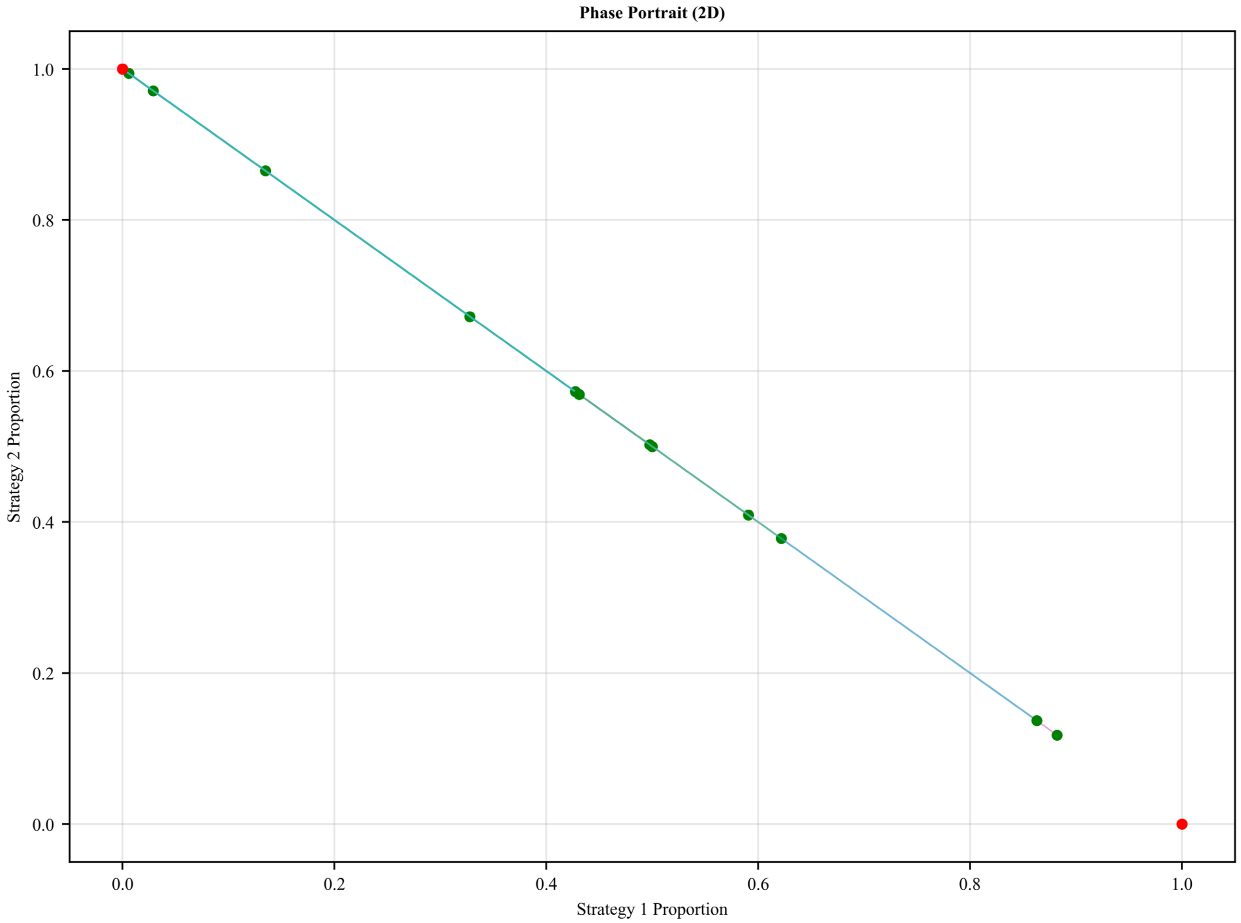

Figure 7. Policy evolution phase diagram of dual-agent systems.

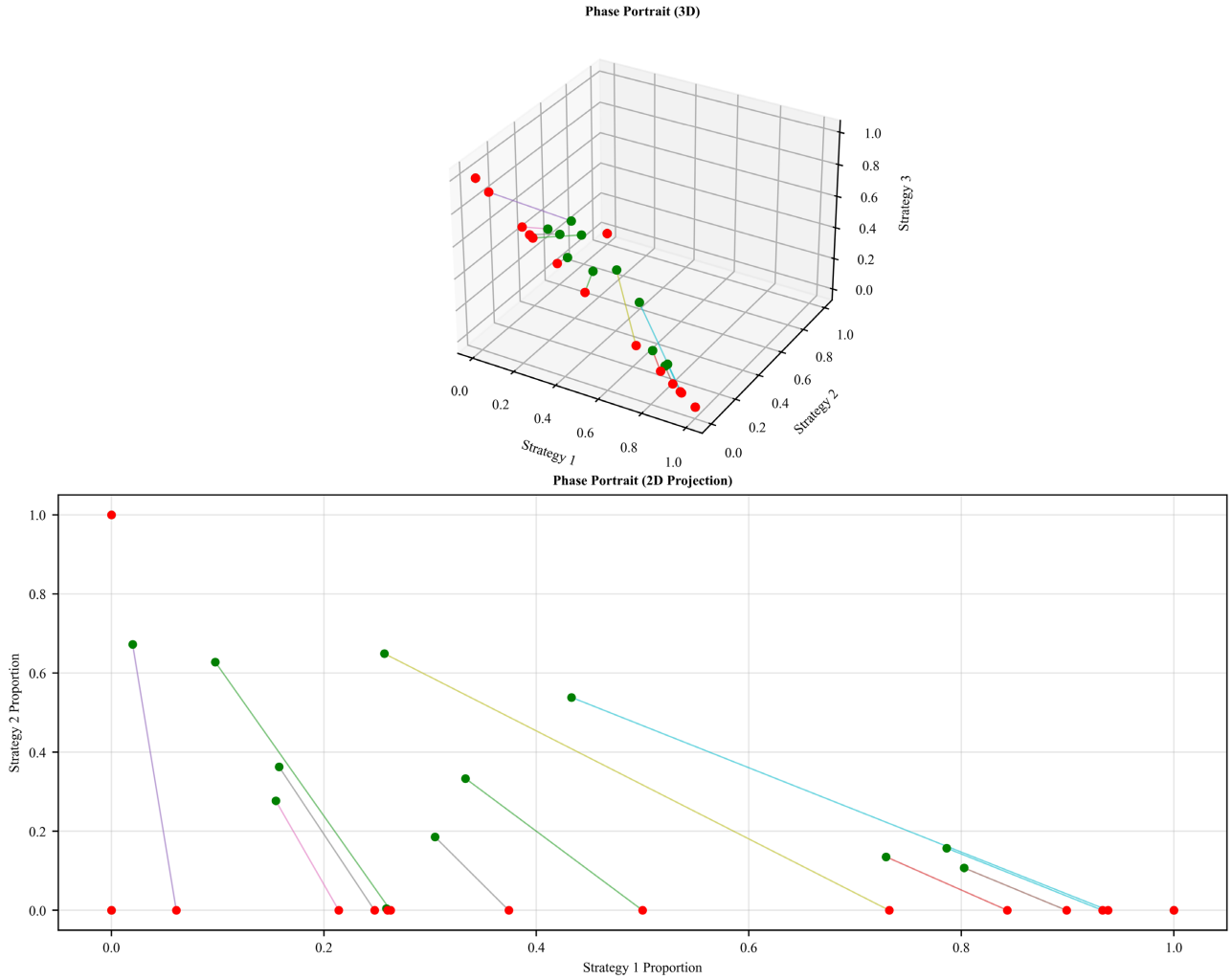

Figure 8. Strategy evolution phase diagram of a three-agent system.

Figure 8 extends the analysis to the three-dimensional phase space for the three-agent system. The three axes represent the population proportions of Strategies 1, 2, and 3, subject to the constraint that they sum to 1. Trajectories illustrate the migration paths of strategy combinations over evolutionary iterations. Multiple attractor regions emerge, with trajectories flowing from initial points toward specific convergence domains. Unlike the continuous equilibrium in the $N = 2$ scenario, the $N = 3$ scenario exhibits discrete equilibrium clusters, indicating that increased participant numbers elevate strategy space dimensionality and complicate equilibrium structure, with some trajectories displaying spiral oscillations before convergence.

These two phase diagrams provide a dynamic perspective on the profound influence of participant scale on evolutionary paths: the $N = 2$ system features a simple structure with strategies smoothly converging to a continuous equilibrium band; the $N = 3$ system exhibits a marked increase in complexity, characterized by multi-equilibrium coexistence and strategy oscillations.

4.3.2. Analysis of the Impact of Participant Numbers on Bidding Strategies

Changes in the number of participants reshape the game-theoretic interactions among agents, thereby influencing the direction of strategy evolution and equilibrium characteristics. This section conducts evolutionary game simulations under three scenarios ($N = 2, 3, 5$) to analyze the mechanism through which participant scale affects market evolution from three dimensions: strategy diversity, equilibrium complexity, and payoff distribution.

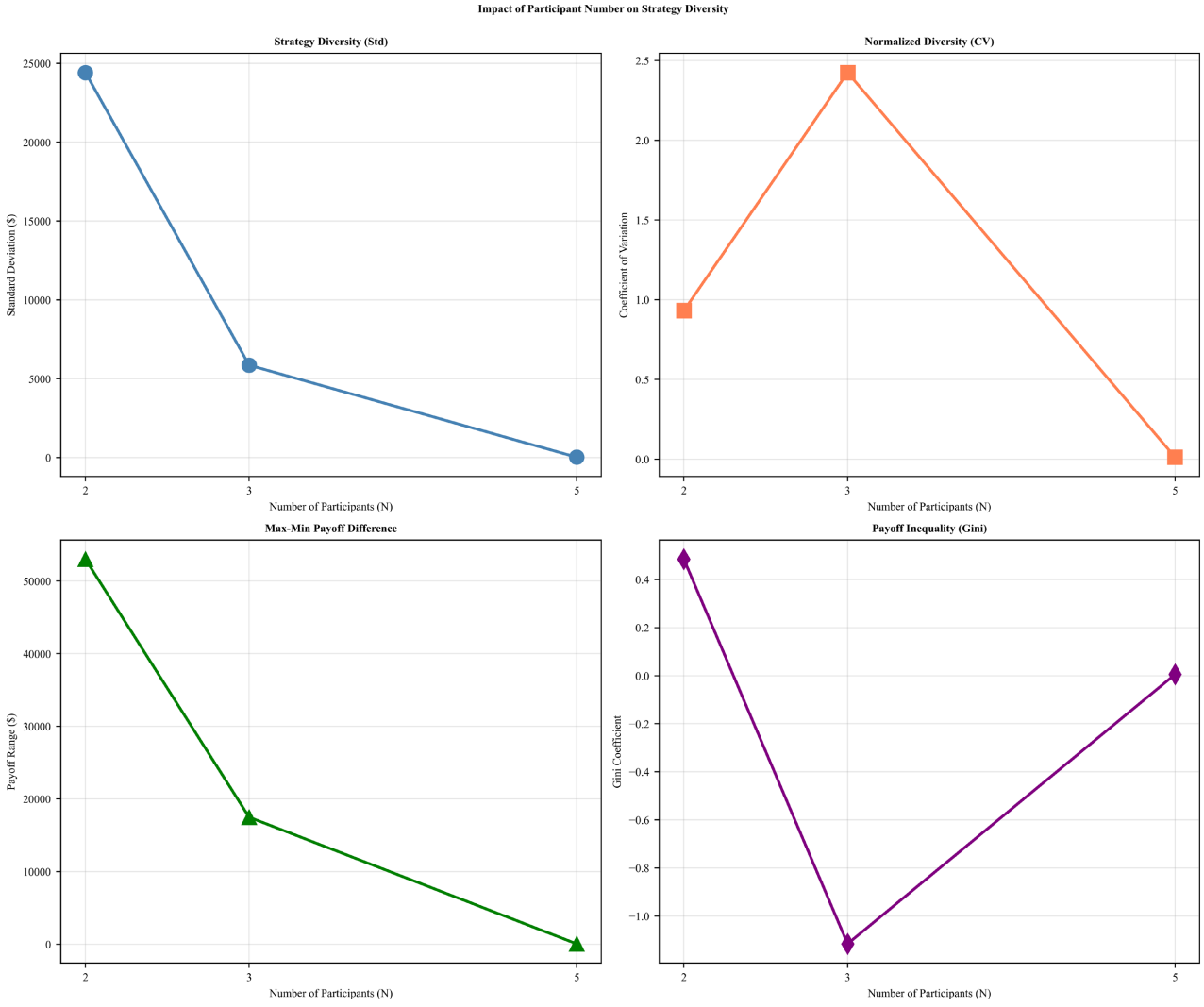

Figure 9. The correlation between strategy diversity and the number of participants.

Figure 9 reveals the moderating effect of participant number on strategy diversity and payoff inequality. As participant number increases, strategy diversity (standard deviation) decreases significantly, indicating that larger populations drive agent strategies toward convergence. The normalized diversity metric (coefficient of variation) peaks at $N = 3$, confirming intensified strategy differentiation in medium-sized markets: with a moderate number of participants, game in-

teractions become most complex, and the strategy space is fully explored. Regarding payoff distribution, both the max-min payoff difference and the Gini coefficient decline sharply as participant number increases, revealing that intensified competition dilutes individual market power, narrows excess arbitrage opportunities, and equalizes population payoffs.

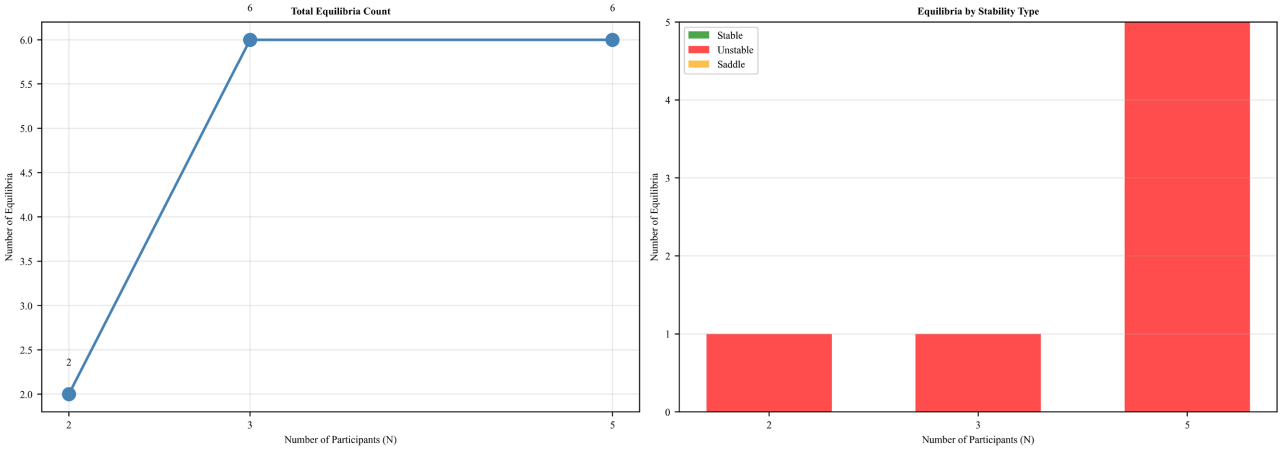

Figure 10. The correlation between the number of equilibrium points and the number of participants.

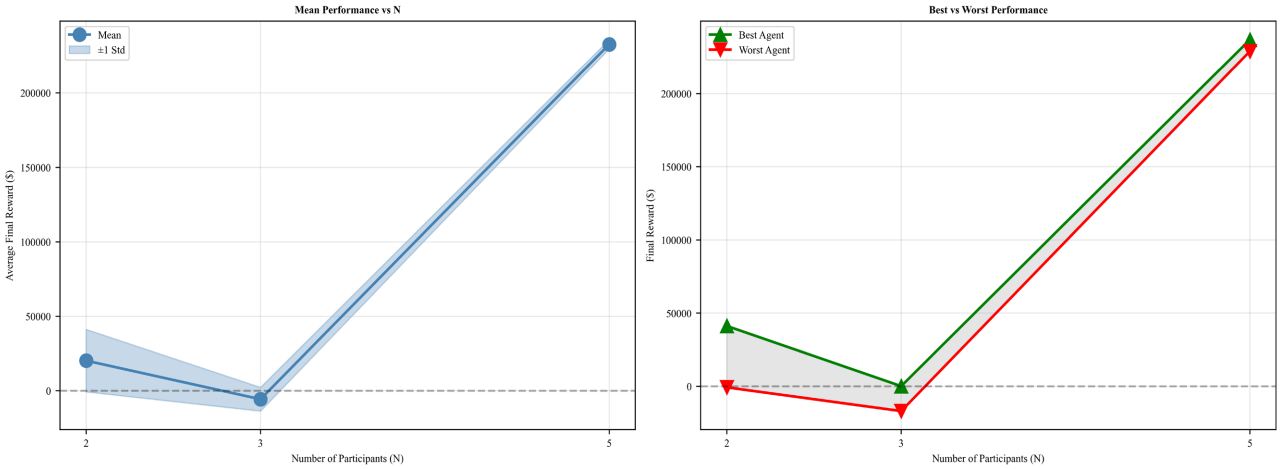

Figure 11. Correlation between performance and number of participants.

Figure 10 further illustrates the impact of participant number on equilibrium structure complexity. The total equilibrium count increases with participant number, indicating that larger populations enrich the strategy combination space and create a pattern of multi-equilibrium coexistence. However, the proportion of stable equilibria declines, consistent with evolutionary game theory predictions: although larger populations increase the number of potential equilibria, the stability of most equilibria diminishes, making the system more susceptible to phase transitions under perturbations.

Figure 11 quantifies the effect of participant number on agent welfare distribution from an individual payoff perspective. Average population payoff improves

significantly as participant number increases, indicating that more traders jointly exploiting arbitrage opportunities promotes price discovery and market efficiency. Simultaneously, the payoff gap among individuals narrows considerably. Combined with the Gini coefficient approaching zero in Figure 9, this leads to the conclusion that increasing participant numbers not only enhances overall returns but also promotes payoff distribution fairness. Notably, the $N = 3$ scenario exhibits negative average payoffs. Together with the peak strategy diversity and increased equilibrium count in this scenario, this suggests that medium-sized markets operate at the chaotic edge of evolutionary games: strategies compete fully but have yet to form a stable order, with some agents experiencing temporary losses due to high strategy exploration costs.

In summary, the impact of participant number on virtual bidding markets can be characterized as follows: duopoly ($N = 2$) exhibits pronounced strategy differentiation and substantial payoff disparities; medium scale ($N = 3$) represents a critical point of complex gaming, with peak strategy diversity and the most complex equilibrium structure, accompanied by risks of individual losses; perfect competition ($N = 5$) trends toward strategy convergence, payoff equalization, and optimal market efficiency.

Table 5. Market performance and evolutionary characteristics statistics across different participant scales.

N	Mean Reward ($)	Standard Deviation ($)	CV	Gini
2	2.03e+4	2.44e+4	0.93	0.48
3	−5.66e+3	5.86e+3	2.43	−1.12
5	2.33e+5	23.32	0.012	0.005
N	N Equilibria	Best ($)	Worst ($)	Gap ($)
2	2	4.12e+4	−701.91	4.194+4
3	6	118.4	−1.69e+4	1.7e+4

Table 5 quantifies the evolutionary characteristics of the virtual bidding market under different participant scales. Regarding market efficiency, the average payoff is $20,263 for $N = 2$, drops sharply to –$5664 for $N = 3$, and then surges to $2.33e+5, for $N = 5$, indicating that medium-scale markets are in a period of strategic chaos, whereas perfect competition ($N = 5$) substantially enhances overall efficiency. In terms of strategy diversity, the coefficient of variation increases from 0.93 for $N = 2$ to 2.43 for $N=3$ before plummeting to 0.012 for $N = 5$, corroborating the conclusion in Figure 9: strategy differentiation is most pronounced in medium-scale markets, and strategies become highly convergent after perfect competition is achieved. Regarding payoff fairness, the Gini coefficient declines from 0.48 for $N = 2$ to nearly zero for $N = 5$, and the gap between the best and worst payoffs narrows from $4.19e+4 to $7.89e+3, revealing that increasing participant numbers significantly improves payoff distribution. At the equilibrium structure level, the equilibrium count increases from two for $N = 2$ to six for $N = 5$, con-

sistent with the trend observed in Figure 10. Collectively, $N = 5$ achieves an optimal trade-off among efficiency, fairness, and stability.

4.3.3. Analysis of the Modulating Effects of Price Influence Coefficients on Strategy Evolution

To investigate the feedback effect of virtual bidders' collective behavior on market prices, the price impact coefficient α is introduced to quantify the marginal influence of individual bidding strategies on market clearing prices. The moderating mechanism of α on market evolution is systematically examined from three dimensions: learning dynamics, strategy diversity, and equilibrium structure.

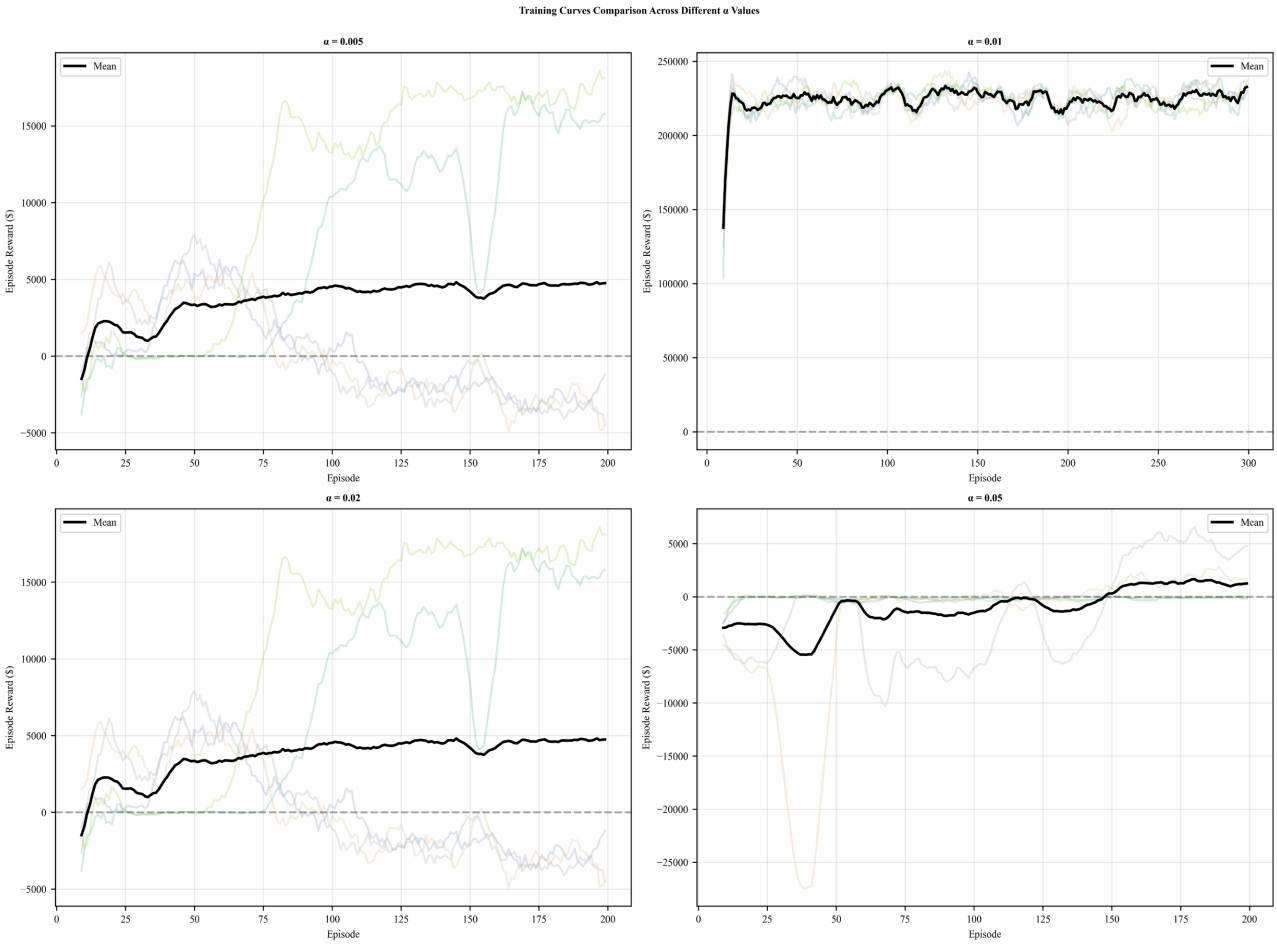

Figure 12. Dynamic comparison of training under different price impact coefficients.

Figure 12 presents training curves under different α values. Under low α, reward curves fluctuate sharply and converge slowly, indicating weak market feedback signals. Under high α, curves become smoother and converge faster, demonstrating that price feedback accelerates learning. However, excessively high α induces mid-curve fluctuations, suggesting that strong price impact may trigger strategy oscillations.

Figure 13 quantifies the effect of α on payoff matrix diversity and payoff dis-

tribution. Strategy diversity decreases monotonically with increasing α: when individual actions significantly affect prices, deviating from consensus becomes more costly. Normalized diversity exhibits an opposite trend, reflecting that average payoff declines faster than standard deviation, making the market more sensitive to strategy differences under high α. The max-min payoff difference and the self-play versus cross-play payoff gap both narrow as α increases, confirming that price feedback suppresses excess arbitrage and drives the game toward symmetric equilibrium.

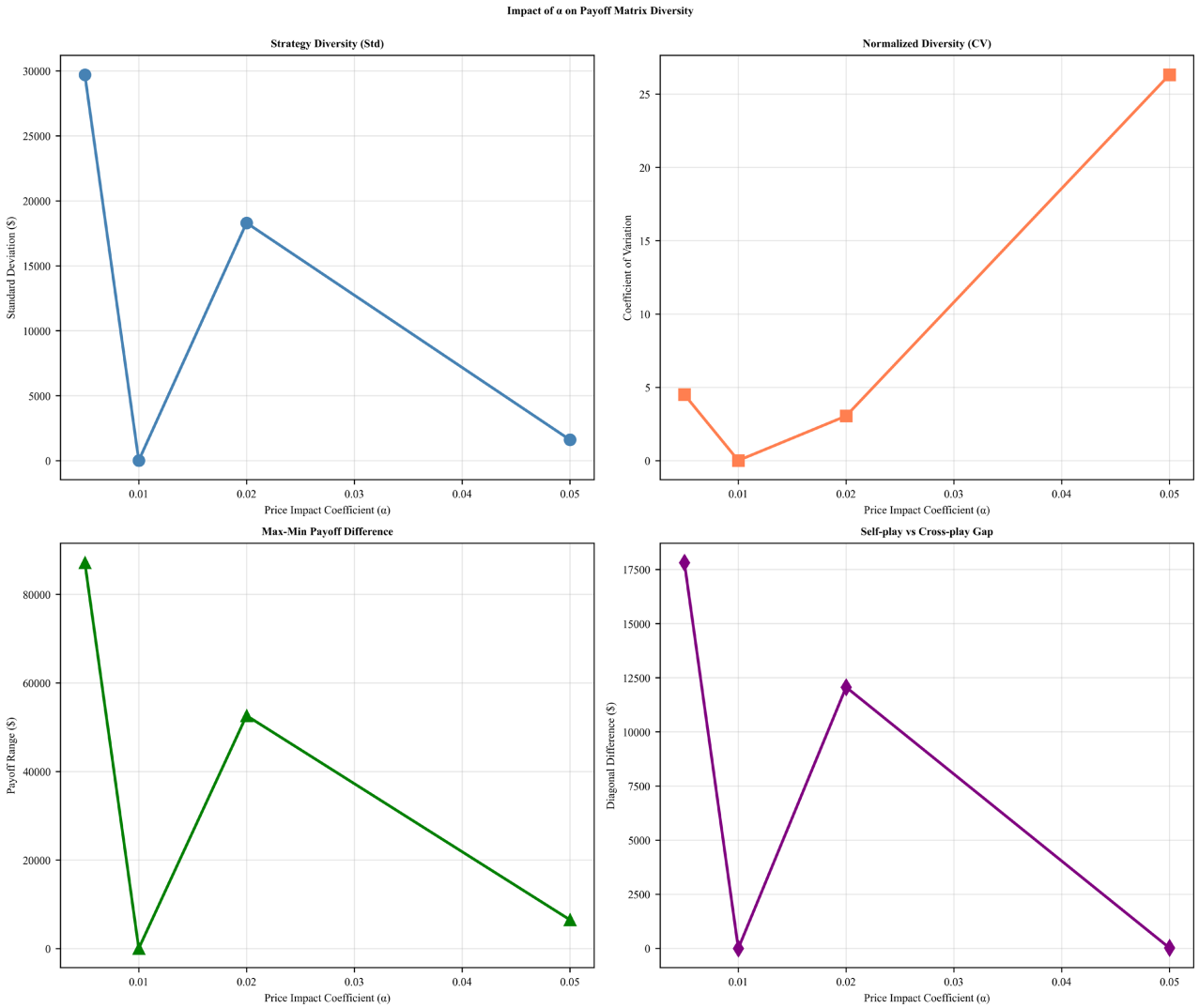

Figure 13. The moderating effects of price influence coefficients on strategy diversity and return distribution.

Figure 14 reveals the moderating effect of α on equilibrium structure complexity. Total equilibrium count follows an inverted U-shaped pattern. Low α is dominated by stable equilibria; medium α yields exclusively unstable equilibria, corresponding to strategic chaos where active exploration prevents stable order formation; high α restores stable equilibria, consistent with the smooth

convergence in Figure 12, indicating that strong price feedback drives the system toward predictable equilibria.

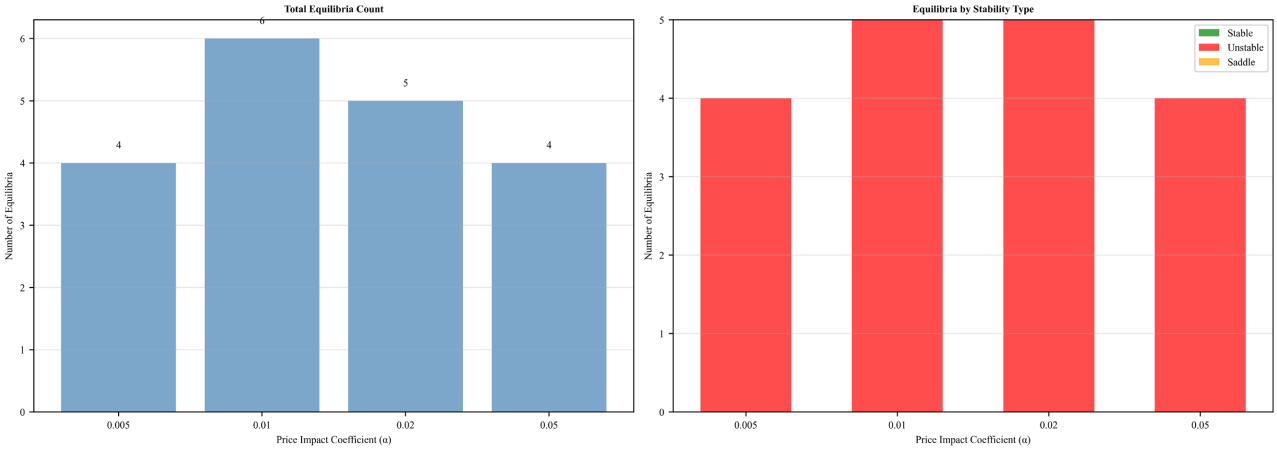

Figure 14. The impact of price influence coefficients on the complexity of equilibrium structures.

In summary, the moderating effect of α can be characterized as follows: low α exhibits weak feedback, with thorough strategy exploration but slow convergence, resulting in inefficient chaos; medium α represents a critical point of complex gaming, with peak strategy diversity and highly unstable equilibria, resembling the "edge of chaos"; high α strengthens feedback, accelerating convergence and narrowing payoff disparities, leading the market toward stable equilibrium. Moderate price feedback enhances market efficiency and stability, but the strategic chaos risk in the medium α range warrants caution.

Table 6. Market performance and evolutionary characteristics statistics across different participant scales.

α	Mean Reward ($)	Standard Deviation ($)	CV	Range ($)
0.005	6.59e+3	2.97e+4	4.508	8.72e+4
0.01	1.98e+3	23.32	0.012	60
0.02	−6e+3	1.83e+4	3.049	5.26e+4
0.05	−61.38	1.62e+3	26.332	6.52e+3
α	N Equilibria	Best Agent ($)	Worst Agent ($)	
0.005	4	1.81e+4	−4.51e+3	
0.01	6	2.37e+5	2.29e+5	
0.02	5	1.81e+4	−4.51e+3	
0.05	4	4.8e+3	−148.41	

Table 6 demonstrates that strategy differences and market outcomes are highly sensitive to α. Regarding market efficiency, average payoff initially decreases then increases with α, reaching its lowest negative value at $\alpha = 0.02$, indicating that medium α may induce overall losses due to strategic chaos. The payoff gap

between best and worst agents narrows significantly as α increases, confirming the suppressive effect of price feedback on excess returns. In terms of strategy diversity, standard deviation drops sharply at $\alpha = 0.01$, corresponding to smooth training curve convergence; the coefficient of variation peaks at $\alpha = 0.05$, reflecting amplified relative dispersion as average payoff approaches zero. Regarding equilibrium structure, the equilibrium count follows an inverted U-shaped pattern, peaking at $\alpha = 0.01$ with exclusively unstable equilibria corresponding to strategic chaos, while $\alpha = 0.005$ and $\alpha = 0.05$ yield stable equilibria, confirming that both weak and strong feedback drive the system toward stable order.

5. Conclusions

1) At the multi-agent learning level, the MADDPG algorithm demonstrates significant advantages: average per-episode reward improves by a factor of 13.7 compared with independent DDPG, convergence speed increases fivefold ($p < 0.0001$), and it exhibits the lowest reward standard deviation, combining high returns with strong robustness.

2) Regarding participant scale, market structure critically shapes strategy evolution. $N = 2$ exhibits duopoly characteristics with pronounced strategy differentiation; $N = 3$ represents a critical point of complex gaming, with peak strategy diversity accompanied by individual loss risks; $N = 5$ approaches perfect competition, characterized by strategy convergence and payoff equalization. Equilibrium count increases with participant number, while the proportion of stable equilibria declines.

3) Concerning price feedback, the price impact coefficient α exhibits nonlinear moderating effects. Low α enables thorough strategy exploration but slow convergence; medium α operates at the "edge of chaos," where peak strategy diversity coincides with equilibrium instability; high α strengthens feedback, accelerating convergence and narrowing payoff disparities, with equilibrium count following an inverted U-shaped pattern.

While this study provides valuable insights into virtual bidding dynamics, several limitations should be acknowledged. First, the price impact function employed is a stylized linear specification (coefficient α), which may not fully capture the complex, nonlinear price feedback mechanisms observed in real-world electricity markets. Second, the analysis is conducted under fixed population sizes ($N = 2$, 3, and 5), whereas actual markets feature dynamically changing participant numbers that could influence equilibrium stability. Third, the empirical validation relies solely on NYISO data; generalizing these findings to other market designs (e.g., PJM, European markets) requires further investigation.

Nevertheless, practical implications emerge. For traders, the MADDPG framework offers superior profitability with CVaR-based risk control. For regulators, the findings identify critical thresholds—moderate competition ($N \approx 3$) maximizes diversity but risks instability, and the inverted-U relationship between price impact and stability provides actionable guidance for market design.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Ledgerwood, S.D. and Pfeifenberger, J.P. (2013) Using Virtual Bids to Manipulate the Value of Financial Transmission Rights. *The Electricity Journal*, **26**, 9-25. https://doi.org/10.1016/j.tej.2013.09.014

[2] Hadsell, L. (2007) The Impact of Virtual Bidding on Price Volatility in New York's Wholesale Electricity Market. *Economics Letters*, **95**, 66-72. https://doi.org/10.1016/j.econlet.2006.09.015

[3] Mather, J., Bitar, E. and Poolla, K. (2017) Virtual Bidding: Equilibrium, Learning, and the Wisdom of Crowds. *IFAC-PapersOnLine*, **50**, 225-232. https://doi.org/10.1016/j.ifacol.2017.08.038

[4] Kazempour, J. and Hobbs, B.F. (2018) Value of Flexible Resources, Virtual Bidding, and Self-Scheduling in Two-Settlement Electricity Markets with Wind Generation—part I: Principles and Competitive Model. *IEEE Transactions on Power Systems*, **33**, 749-759. https://doi.org/10.1109/tpwrs.2017.2699687

[5] Li, R., Svoboda, A.J. and Oren, S.S. (2015) Efficiency Impact of Convergence Bidding in the California Electricity Market. *Journal of Regulatory Economics*, **48**, 245-284. https://doi.org/10.1007/s11149-015-9281-3

[6] Li, Y., Yu, N. and Wang, W. (2022) Machine Learning-Driven Virtual Bidding with Electricity Market Efficiency Analysis. *IEEE Transactions on Power Systems*, **37**, 354-364. https://doi.org/10.1109/tpwrs.2021.3096469

[7] Baltaoglu, S., Tong, L. and Zhao, Q. (2019) Algorithmic Bidding for Virtual Trading in Electricity Markets. *IEEE Transactions on Power Systems*, **34**, 535-543. https://doi.org/10.1109/tpwrs.2018.2862246

[8] Mones, L. and Lovett, S. (2023) A General Stochastic Optimization Framework for Convergence Bidding. *IEEE Transactions on Energy Markets, Policy and Regulation*, **1**, 60-72. https://doi.org/10.1109/tempr.2023.3243765

[9] Ding, W., Zhang, F., Zhelin, Y., Liu, R., Liu, Y. and Jing, Z. (2021) Supervision Mechanism of Virtual Bidding in Electricity Market: A Review. *IOP Conference Series: Earth and Environmental Science*, **675**, Article ID: 012130. https://doi.org/10.1088/1755-1315/675/1/012130

[10] Xiao, D.L., do Prado, J.C. and Qiao, W. (2021) Optimal Joint Demand and Virtual Bidding for a Strategic Retailer in the Short-Term Electricity Market. *Electric Power Systems Research*, **190**, Article ID: 106855. https://doi.org/10.1016/j.epsr.2020.106855

[11] Do Prado, J.C. and Chikezie, U. (2021) A Decision Model for an Electricity Retailer with Energy Storage and Virtual Bidding under Daily and Hourly CVaR Assessment. *IEEE Access*, **9**, 106181-106191. https://doi.org/10.1109/ACCESS.2021.3100815

[12] Han, D., Huang, W., Ren, H., Zhao, W. and Li, Y. (2022) Machine Learning Analytics for Virtual Bidding in the Electricity Market. *International Journal of Electrical Power & Energy Systems*, **143**, Article ID: 108489. https://doi.org/10.1016/j.ijepes.2022.108489