

Article

An Adaptive Bidding Strategy for Virtual Power Plants in Day-Ahead Markets Under Multiple Uncertainties

Wei Yang and Wenjun Wang * 

College of Intelligent Computing, Tianjin University, Tianjin 300072, China; philippey@tju.edu.cn

* Correspondence: wjwang@tju.edu.cn

Abstract

To address the challenges posed by multiple uncertainties in modern power systems to the market bidding of Virtual Power Plants (VPPs), this paper proposes an adaptive bidding strategy based on Deep Reinforcement Learning (DRL). First, a heterogeneous VPP aggregation model integrating dedicated energy storage, Vehicle-to-Grid (V2G), and flexible loads is constructed, incorporating complex physical and operational constraints. Second, to overcome the “myopic” local optimality problem of traditional DRL in temporal arbitrage tasks, a potential-based reward shaping mechanism linked to future price trends is designed to guide the agent toward long-term optimal strategies. Finally, multi-dimensional comparative experiments and mechanism analyses are conducted in a simulated day-ahead electricity market. Simulation results demonstrate the following: (1) The proposed algorithm exhibits robust convergence stability and effectively handles stochastic noise in market prices and renewable generation. (2) Economically, the strategy significantly outperforms the rule-based strategy and remains highly competitive with the deterministic-optimization benchmark under perfect-information assumptions. (3) Mechanism analysis further reveals that the DRL agent breaks through the rigid logic of fixed thresholds, learning a non-linear dynamic game mechanism based on “Price-SOC” states, thereby achieving full-depth utilization of energy storage resources. This work provides an interpretable data-driven paradigm for intelligent VPP decision-making in uncertain environments.

Keywords: virtual power plant; deep reinforcement learning; market bidding; uncertainty; reward shaping; mechanism analysis

1. Introduction

In response to global climate change and the pursuit of the “dual carbon” goal, the construction of a new power system, primarily based on renewable energy, is rapidly advancing [1]. However, the inherent intermittency and variability of renewable energy sources, such as wind and solar energy, pose unprecedented challenges to the real-time balancing and economic operation of power systems [2]. As an advanced energy aggregation technology, the Virtual Power Plant (VPP) has become a critical solution for enhancing system flexibility and integrating renewable energy by coordinating various heterogeneous resources, such as distributed energy storage, electric vehicles (V2G), and flexible loads [3–5].

The sustainable commercial operation of a VPP is highly dependent on its profitability in the electricity market [6]. In the day-ahead energy market, the VPP is required to formulate a bidding strategy one day in advance for the next 24 h to maximize its operating revenue. However, this decision-making process is highly complex, as VPP operators



Academic Editor: Gajendra Singh Chawda

Received: 4 March 2026

Revised: 5 April 2026

Accepted: 9 April 2026

Published: 12 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

are exposed to dual uncertainties arising from market prices and renewable energy generation [7]. Sharp fluctuations in market prices and forecasting errors make traditional arbitrage models difficult to implement accurately. Therefore, designing a VPP bidding strategy that enables adaptive decision-making and robust profit maximization under high uncertainty has become a core issue of common concern in both academia and industry [8].

Existing studies on the VPP market bidding problem can generally be classified into two categories. The first category consists of traditional optimization methods based on mathematical programming, such as stochastic programming and robust optimization. Gulotta et al. [9] developed an energy management system based on stochastic programming to optimize market bidding and real-time operation of virtual power plants under uncertainty, thereby increasing profits and reducing energy imbalances. Wang et al. [10] proposed an optimal self-scheduling strategy for a multi-energy virtual power plant providing energy and reserve services under a holistic market framework, enhancing resilience against price volatility. To address uncertainties, Wang et al. [11] proposed a distributed robust optimization strategy combined with a dual-norm uncertainty set to coordinate multi-energy VPP clusters. Meanwhile, Kong et al. [12] proposed a decentralized optimization model based on an enhanced Benders decomposition framework to determine the optimal bidding strategy for VPPs in the day-ahead market, effectively balancing privacy protection and bidding performance. Lu et al. [13] developed a distributed optimization framework integrating peer-to-peer (P2P) power sharing, and addressed uncertainty and privacy protection issues in the cooperative operation of multiple VPPs through stochastic robust modeling and distributed algorithms, thereby improving economic performance and achieving fair benefit allocation. Although the theoretical framework of such methods is relatively mature, they typically rely on accurate probability distribution assumptions for uncertain factors or conservative worst-case estimations, and require highly accurate forecasting information [14]. In real markets characterized by significant unpredictable noise, these rigid strategies based on perfect information assumptions are often fragile and struggle to adapt to dynamic environmental changes [15].

The second category comprises rule-based or heuristic strategies derived from domain expert knowledge. For real-time operations, Yu et al. [16] utilized Lyapunov optimization theory to implement time decoupling within VPPs, effectively transforming long-term scheduling into single-period online optimization to improve computational efficiency. Yu et al. [17] proposed an improved cooperative particle swarm optimization (ICPSO) algorithm for the energy management of VPPs, which significantly improves computational efficiency and scheduling profitability. In terms of multi-agent interactions, Cao et al. [18] established a cooperative alliance model for multiple VPPs based on Nash bargaining theory to optimize benefit distribution. Through P2P energy trading, the total cost of the alliance is minimized, and costs are fairly allocated according to the contributions of all participants, thereby increasing renewable energy revenues. Regarding conventional control schemes, Aboelhassan et al. [19] evaluated rule-based energy management systems (REMS), noting that while they ensure operational stability through fixed logical thresholds, they often lack the adaptability required to fully exploit complex market price fluctuations [14,19].

To address the above challenges, deep reinforcement learning (DRL), as a data-driven decision-making approach, has shown great potential [20]. DRL agents learn through trial-and-error interactions with the environment without requiring an explicit and accurate mathematical model, and are particularly effective in handling complex sequential decision-making problems under uncertainty, which aligns well with the VPP bidding scenario [21,22].

Recently, various advanced DRL algorithms have been actively deployed to optimize VPP market bidding and internal energy management. For instance, Jiang et al. utilized

a multi-agent twin delayed deep deterministic policy gradient (MATD3) algorithm to derive the optimal bidding strategy for a price-maker VPP in the day-ahead market [6]. To address the uncertainties of renewable energy, recent studies have combined condition generation adversarial networks with DRL to build robust multi-scenario scheduling strategies [23]. Furthermore, to improve robust decision-making in complex environments, recent research has proposed double-layer game models based on Soft Actor-Critic (SAC) and double-deep Q networks (DDQN) for VPP energy management [24], as well as end-to-end DRL methods focusing on feature extraction for bidding with uncertain renewable generation [25]. However, when DRL is directly applied to the VPP arbitrage problem, a key challenge is that agents are prone to converging to short-sighted local optima driven by instantaneous rewards (e.g., negative profits during charging), making it difficult to learn far-sighted, long-term arbitrage strategies [26].

Therefore, this paper proposes an adaptive day-ahead market bidding strategy for virtual power plants under multiple sources of uncertainty. First, a multi-dimensional heterogeneous VPP aggregation model is developed by integrating dedicated energy storage, V2G, and flexible loads, while key practical constraints such as the travel demand of V2G users are explicitly considered. Second, to overcome the limitations of traditional optimization methods, a DRL-based decision-making model is established for VPP bidding. To address the short-sighted behavior of DRL in arbitrage tasks, this paper proposes a potential-based reward shaping mechanism linked to the maximum forecasted electricity price over a future horizon, providing dense long-term guidance signals to encourage agents to learn the long-term optimal strategy of valley charging and peak discharging. Finally, extensive experiments are conducted on a day-ahead electricity market simulation platform, where the proposed method is benchmarked against deterministic optimization and rule-based strategies.

The main contributions of this paper are summarized as follows:

1. An adaptive VPP bidding framework based on DRL is proposed, which effectively addresses the uncertainties of market prices and renewable energy generation.
2. A novel potential-based reward shaping mechanism guided by future price signals is developed to mitigate the short-sighted behavior of DRL in arbitrage tasks, thereby significantly improving the long-term profitability of the proposed strategy.
3. A practical VPP aggregation model is established by incorporating key realistic constraints, such as V2G travel demand and scheduling costs, which enhances the applicability and engineering relevance of this study.

Simulation results demonstrate that the proposed strategy not only significantly outperforms traditional approaches, but also remains highly competitive with the deterministic-optimization benchmark under noisy and realistic market conditions, highlighting the advantage of adaptive bidding in exploiting uncertainty.

2. Virtual Power Plant and Market Environment Modeling

The overall framework of the DRL-based VPP market bidding system studied in this paper is illustrated in Figure 1. The system mainly consists of two components: the power market environment and the VPP intelligent decision-making agent. The electricity market environment comprises conventional generation units, renewable energy resources, and system loads, which jointly determine the market clearing price. The VPP intelligent decision-making agent perceives the market and its internal states through the reinforcement learning algorithm, and dynamically adjusts the charging and discharging behaviors of aggregated resources as well as the bidding strategy to maximize operating profit. This section presents detailed models of the key modules in the proposed system framework.

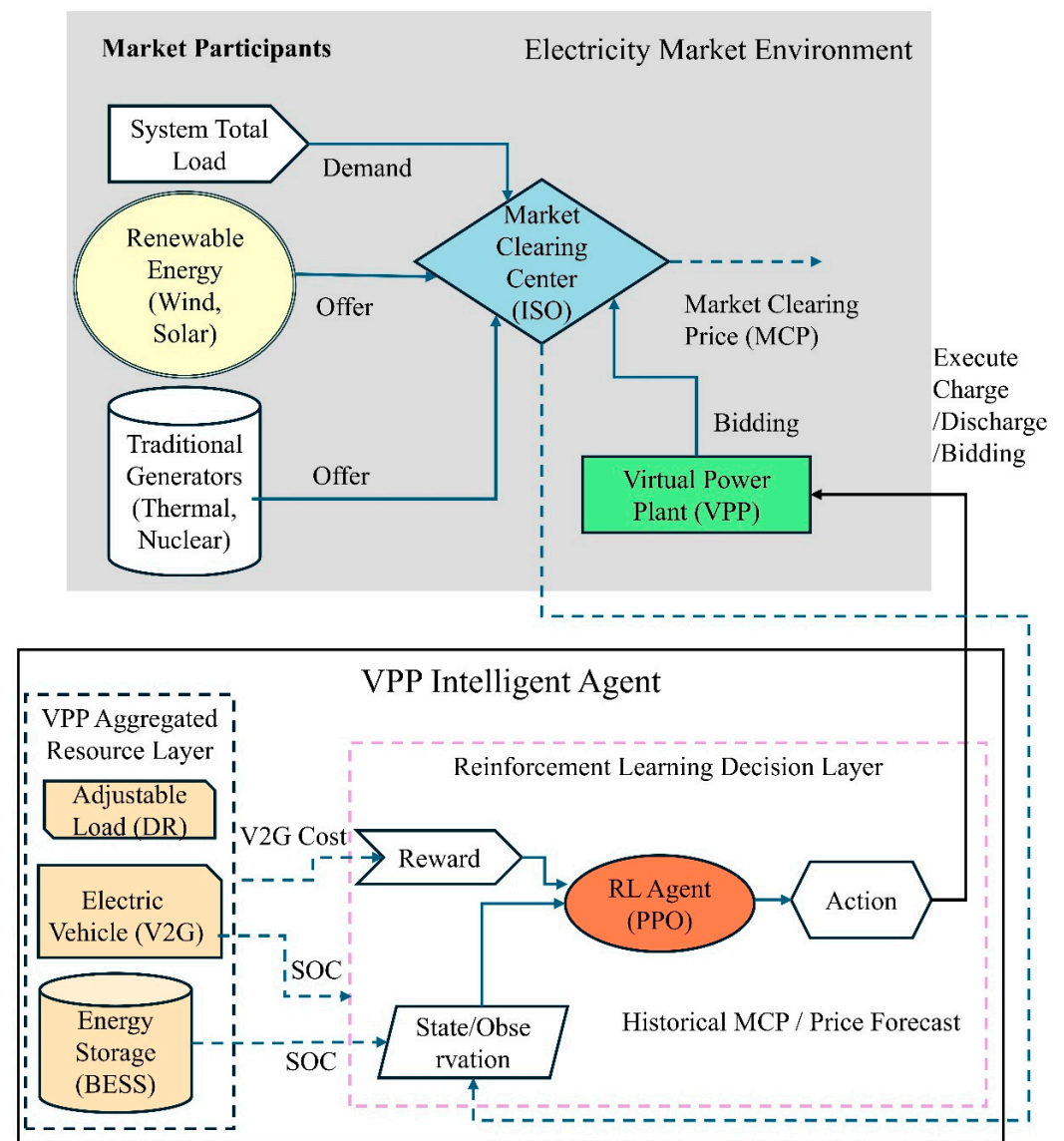


Figure 1. Framework of the DRL-based VPP market bidding system.

2.1. Day-Ahead Electricity Market Clearing Mechanism

This study is conducted within the framework of the day-ahead energy market. In this market, each generation entity (including conventional generation units, renewable energy plants, and VPPs) submits its 24 h generation schedule and bidding curve for the next day in advance. The market operator aims to maximize social welfare (or equivalently minimize electricity procurement costs) and clears all bids in a unified manner.

The process can be simplified into the following steps:

1. **Bidding collection:** Collect the quotations submitted by all market participants at each hour t ($P_i(t)$, $C_i(t)$), where $P_i(t)$ is the available power supply of unit i , and $C_i(t)$ is its marginal bid price.
2. **Order of quotations:** All quotations are sorted by the price $C_i(t)$ from low to high to form the supply curve.
3. **Market clearing:** Based on the total system load forecast $D_i(t)$ for the hour, units are selected sequentially from the lowest price on the supply curve until the cumulative supply matches the total load. The marginal price of the last selected unit is the market clearing price (MCP) for the hour, denoted as $\lambda_{mcp}(t)$. All winning units are settled at this price.

2.2. VPP Aggregated Resource Model

The VPP constructed in this study aggregates three typical distributed energy resources (DERs), which participate in market bidding as a unified entity, and the aggregated resource model forms the foundation for the observability and controllability of the VPP.

Dedicated Battery Energy Storage System (BESS) Model

The dedicated battery energy storage system is a core component of the VPP, providing fast and bidirectional power regulation. Its operating state is mainly characterized by the state of charge (SOC), and the corresponding dynamic model is given as follows:

- (1) SOC transfer equation: The discrete-time model is as follows:

$$S_{\text{bess}}(t) = S_{\text{bess}}(t-1) + \frac{P_{\text{bess,ch}}(t) \cdot \eta_{\text{ch}} - P_{\text{bess,dis}}(t) / \eta_{\text{dis}}}{E_{\text{bess}}^{\text{cap}}} \cdot \Delta t \quad (1)$$

In (1): $S_{\text{bess}}(t-1)$ is the SOC at time $t-1$; $P_{\text{bess,ch}}(t)$ and $P_{\text{bess,dis}}(t)$ are the charging and discharging power at time t , respectively. η_{ch} and η_{dis} represent charge and discharge efficiency respectively, and $E_{\text{bess}}^{\text{cap}}$ is the rated energy capacity of the BESS.

- (2) Operational constraints:

The operation of the BESS must satisfy its physical constraints, including the upper and lower limits of SOC and the charging and discharging power limits:

$$\begin{cases} S_{\text{bess}}^{\min} \leq S_{\text{bess}}(t) \leq S_{\text{bess}}^{\max} \\ 0 \leq P_{\text{bess,ch}}(t) \leq P_{\text{bess,ch}}^{\max} \\ 0 \leq P_{\text{bess,dis}}(t) \leq P_{\text{bess,dis}}^{\max} \end{cases} \quad (2)$$

In the formula, $S_{\text{bess}}^{\min}$ and $S_{\text{bess}}^{\max}$ denote the allowable minimum and maximum SOC, respectively, which are typically set to 0 and 1. $P_{\text{bess,ch}}^{\max}$ and $P_{\text{bess,dis}}^{\max}$ represent the maximum charging and discharging power respectively.

2.3. V2G Model

V2G technology forms a large-scale mobile energy storage system by aggregating a large number of electric vehicles connected to the grid. In this study, the aggregated V2G system is equivalently modeled as a centralized energy storage unit, while additional considerations are incorporated to account for its special constraints as a social resource.

- (1) Basic energy storage model: The basic V2G model is similar to that of the BESS. The aggregated energy capacity $E_{\text{V2G}}^{\text{cap}}$ and power limits $P_{\text{V2G}}^{\max}$ are determined by the fleet size, the average battery capacity per vehicle, and the rated power of charging/discharging facilities. The SOC transition equation follows the same form as (1).
- (2) Travel demand constraints: Unlike dedicated stationary storage, the primary objective of V2G is to satisfy the travel demand of EV owners. To capture this practical constraint, the following simplified assumption is adopted: During the daily morning peak travel preparation period (e.g., 06:00–08:00), the V2G fleet is prohibited from providing discharging services to ensure sufficient energy for users' subsequent travel. This hard constraint is formulated as:

$$P_{\text{V2G,dis}}(t) = 0, \forall t \in [T_{\text{start}}, T_{\text{end}}] \quad (3)$$

In (3), $[T_{\text{start}}, T_{\text{end}}]$ denotes the preset travel preparation period, which is set to [6,8] in this study.

- (3) Scheduling cost model: The dispatch of V2G resources by the VPP may accelerate battery degradation and incur additional management costs; therefore, service payments to EV owners are required. This cost is modeled as a linear function of the total scheduled energy:

$$C_{V2G}(t) = (P_{V2G,ch}(t) + P_{V2G,dis}(t)) \cdot \Delta t \cdot c_{V2G} \quad (4)$$

where $C_{V2G}(t)$ is the total scheduling cost of V2G at time t (CNY), and c_{V2G} denotes the service price per unit scheduled energy (CNY/MWh). This cost is incorporated into the subsequent reward function as an internal operating expense of the VPP.

Flexible Load Model

Adjustable loads (also referred to as demand response, DR) provide an equivalent power supply service to the system by actively reducing non-critical electricity consumption (e.g., air conditioning systems and industrial production lines) during peak periods. The model is mainly characterized by the following operational constraints:

$$\begin{cases} 0 \leq P_{dr}(t) \leq P_{dr}^{max} \\ \sum_{t=0}^{23} P_{dr}(t) \cdot \Delta t \leq E_{dr}^{daily} \end{cases} \quad (5)$$

In (5), $P_{dr}(t)$ denotes the load power that can be curtailed at time t (MW), whose effect is equivalent to a discharging operation; P_{dr}^{max} represents the maximum response capacity of the adjustable load (MW); and E_{dr}^{daily} denotes the maximum total reducible energy within one day (MWh), which is imposed to prevent excessive impact on users' energy experience.

3. Adaptive Bidding Decision Model Based on DRL

To achieve adaptive bidding and optimal scheduling of the VPP under an uncertain market environment, the sequential decision-making problem is formulated as a Markov Decision Process (MDP) and solved using a deep reinforcement learning approach. This Section elaborates on the MDP formulation, including the design of the observation space, action space, and reward function.

3.1. Problem Description and MDP Formulation

The day-ahead optimization objective of the VPP is to maximize its total operating profit over the next 24 h by determining charging/discharging schedules and bidding strategies, subject to physical constraints and market rules. The problem is formulated as a MDP defined by the tuple (S, A, P, R, γ) :

S: the state space, which contains all information required for VPP decision-making.

A: The action space, which specifies the set of actions that the VPP can take at each time step.

P: The state transition probability $P(s_{t+1}|s_t, a_t)$, which is governed by complex market dynamics and is unknown to the agent.

R: The reward function $R(s_t, a_t)$, which quantifies the immediate reward obtained after taking action a in state s .

γ : The discount factor, which balances the importance of immediate and future rewards.

Since P is unknown, a model-free deep reinforcement learning approach is adopted to learn the optimal policy through extensive interactions $\pi^*(a_t|s_t)$ with the environment.

3.1.1. Observation Space Design

The observation is a concrete representation of the system state S and serves as the direct basis for the decision-making of the DRL agent. A well-designed observation space should include all information relevant to decision-making while avoiding unnecessary redundancy. The observation vectors $o_t \in \mathbb{R}^n$ adopted in this study are summarized in Table 1.

Table 1. Design of the Observation Vector.

Components	Dimensions	Description and Design Intent
SOC of each energy storage unit	N_s	Represents the current schedulable energy reserve of the VPP and serves as a core basis for operational decision-making. N_s is the number of energy storage units.
The current hour	1	Normalized to the interval $[0, 1]$ to enable the agent to perceive daily time progression and learn time-dependent periodic strategies.
Historical market price	$H \times 1$	A vector of length H , containing the market clearing prices over the past H hours
Price signal strength	1	Normalized forecast electricity price at the current hour, helping the agent quickly identify the relative price level within a day.

Among them, the price signal strength $s_{\text{price}}(t)$ is calculated as follows:

$$s_{\text{price}}(t) = \frac{\lambda_{\text{forecast}}(t) - \mu_{\text{forecast}}}{\sigma_{\text{forecast}}} \quad (6)$$

In the formula, $\lambda_{\text{forecast}}(t)$ denotes the forecast electricity price at time t , and μ_{forecast} and σ_{forecast} represent the mean value and standard deviation of the 24 h forecast price series, respectively.

3.1.2. Action Space Design

The action space in this study is represented as a two-dimensional continuous vector $a_t = [a_{\text{power}}(t), a_{\text{price}}(t)]$, with each element normalized to the $[-1, 1]$ interval to align with the output of the reinforcement learning algorithm.

(1) Power regulation action:

This action determines the overall charging or discharging direction and magnitude of the VPP. If $a_{\text{power}}(t) > 0$, the VPP operates in discharging mode with a corresponding discharge power $P_{\text{dis}}(t) = a_{\text{power}}(t) \cdot P_{\text{dis,max}}(t)$; if $a_{\text{power}}(t) < 0$, the VPP operates in charging mode with a corresponding charging power $P_{\text{ch}}(t) = a_{\text{power}}(t) \cdot P_{\text{ch,max}}(t)$. $P_{\text{dis,max}}(t)$ and $P_{\text{ch,max}}(t)$ represent the maximum total discharge and charging power available from the VPP aggregation resources in the current hour, respectively.

(2) Quotation adjustment action $a_{\text{price}}(t)$:

This action determines the bidding strategy of the VPP in the electricity market by fine-tuning the forecast electricity price.

$$\lambda_{\text{bid}}(t) = \lambda_{\text{forecast}}(t) \cdot (1 - k \cdot a_{\text{price}}(t)) \quad (7)$$

In the formula, $\lambda_{\text{bid}}(t)$ denotes the final quotation submitted to the market by the VPP, and k is the quotation adjustment coefficient controlling the allowable fluctuation range of the quotation.

3.2. Reward Function Design with Future Potential Guidance

The reward function serves as a crucial signal for guiding the learning process of the agent. The main challenge of the VPP arbitrage problem lies in the fact that immediate rewards (e.g., negative profit during low-price charging) may mislead the agent, discouraging it from accepting short-term costs in exchange for higher long-term returns, which leads to myopic decision-making. To address this issue, a composite reward function is developed based on the Potential-Based Reward Shaping (PBRs) framework, incorporating immediate profit, future potential, and terminal penalty components. The PBRs approach provides dense guidance signals for the agent while preserving the optimal policy of the original problem, thereby significantly accelerating learning convergence.

The total reward function $R(t)$ proposed in this study consists of three components:

$$R(t) = R_{\text{profit}}(t) + R_{\text{potential}}(t) + R_{\text{penalty}}(t) \quad (8)$$

(1) Instant profit reward $R_{\text{profit}}(t)$:

This component represents the net market transaction profit of the VPP at time t , serving as the primary basis of the reward design.

$$R_{\text{profit}}(t) = \left(P_{\text{dispatched}}(t) - P_{\text{ch}}(t) \right) \cdot \lambda_{\text{mcp}}(t) - C_{\text{v2g}}(t) \quad (9)$$

In the formula, $P_{\text{dispatched}}(t)$ denotes the actual discharged energy of the VPP in the market at time t , $\lambda_{\text{mcp}}(t)$ is the market clearing price, and $C_{\text{v2g}}(t)$ represents the V2G scheduling cost defined in (4).

(2) Potential-based shaping reward:

To mitigate the myopic behavior induced by instant profit, a potential function $\Phi(s_t)$ is introduced for reward shaping.

$$\Phi(s_t) = \left(\sum_{i \in N_s} S_i(t) \cdot E_i^{\text{cap}} \right) \cdot \max_{j > t} (\lambda_{\text{forecast}}(j)) \quad (10)$$

where E_i^{cap} represents the rated energy capacity of the i -th energy storage unit.

According to the PBRs framework, the shaping reward is defined as:

$$R_{\text{potential}}(t) = w_{\text{potential}} \cdot (\gamma \cdot \Phi(s_{t+1}) - \Phi(s_t)) \quad (11)$$

In the formula, $w_{\text{potential}}$ is the weight coefficient of the shaping reward, and γ denotes the discount factor. Intuitively, if an action increases the system's future potential (e.g., charging during low-price periods), the agent receives a positive shaping reward; otherwise, a negative shaping reward is obtained. This mechanism propagates future revenue signals to the current time step, thereby encouraging long-term and far-sighted decision-making.

(3) Terminal penalty $R_{\text{penalty}}(t)$:

$$R_{\text{penalty}}(T-1) = -w_{\text{soc}} \cdot (\bar{S}(T-1) - S_{\text{target}})^2 \quad (12)$$

In the formula, $\bar{S}(T - 1)$ is the average SOC of all energy storage units at the end of the day, S_{target} denotes the target SOC, and w_{soc} is the penalty weight. For non-terminal time steps, $R_{\text{penalty}}(t) = 0$.

3.3. Model Training Algorithm

In this study, the Proximal Policy Optimization (PPO) algorithm is adopted for model training. As an advanced policy gradient method, PPO demonstrates strong performance in handling continuous action space problems. Its core advantage lies in the introduction of a clipped surrogate objective function, which constrains the update step size of each policy iteration, effectively preventing policy collapse caused by overly large updates and thereby ensuring training stability and high sample efficiency. These characteristics make PPO particularly suitable for solving complex engineering optimization problems, such as VPP scheduling.

4. Case Study and Experimental Analysis

4.1. Experimental Environment and Parameter Settings

To evaluate the effectiveness of the proposed method, a day-ahead electricity market simulation platform is developed using Python 3.9. This section describes the key experimental parameters and their configurations. All parameters are configured according to typical grid operation data and relevant literature to ensure the realism and validity of the simulation results.

4.1.1. Market and VPP Parameters

The physical and economic parameters of the electricity market and the internal resources of the VPP used in the experiments are summarized in Tables 2 and 3, respectively. The market generation parameters are configured with reference to typical day-ahead market settings reported in [6], while the internal resource parameters of the BESS, V2G, and demand response are determined based on representative values adopted in [4,17,22].

Table 2. Market Generation Parameters.

Generation Type	Capacity (MW)	Marginal Cost (CNY/MWh)
Nuclear	1100	80
Coal-fired	600	360
Coal-fired	550	420
Gas-fired	200	750
Gas-fired	180	900

Table 3. Parameters of VPP Internal Resources.

Resource Type	Parameter	Value
Battery Energy Storage System (BESS)	Rated energy capacity (MWh)	100
	Maximum power (MW)	25
	Charging/discharging efficiency	0.95/0.95
	Initial SOC	0.5
Electric Vehicles (V2G)	Fleet size	500
	Average battery capacity per EV (kWh)	60

Table 3. *Cont.*

Resource Type	Parameter	Value
	Average charging/discharging power (kW)	7
	Charging/discharging efficiency	0.90/0.85
	Initial SOC	0.6
	Scheduling cost (CNY/MWh)	35.0
Demand Response (DR)	Maximum curtailment power (MW)	15
	Daily energy limit (MWh)	60

4.1.2. Reinforcement Learning Model Hyperparameters

The hyperparameters of the PPO algorithm and the environment-related reward function are summarized in Table 4. These parameters are determined through preliminary experiments to balance learning efficiency and final performance.

Table 4. Hyperparameters of the Reinforcement Learning Model.

Category	Parameter	Value
Training settings	Total training steps	200,000
	Learning rate	3×10^{-4}
	Discount factor	0.99
PPO algorithm	Rollout steps per update	2048
	Mini-batch size	64
	Training epochs	10
	GAE Lambda	0.95
Reward function	Potential-based reward weight	0.8
	Terminal SOC penalty weight	200.0
	Target terminal SOC	0.5
Observation space	Historical price window size	6

All simulation experiments are implemented in a Python 3.9 environment using PyTorch 1.13 and the Stable-Baselines3 2.0 library. The hardware platform consists of an Intel Core i7-10700 CPU and 16 GB RAM.

4.1.3. Baseline Strategy Settings

To comprehensively evaluate the performance of the proposed reinforcement learning (DRL) method, two representative baseline strategies are designed for comparison:

- (1) **Deterministic Optimization:** This strategy represents a traditional optimization approach under perfect-information assumptions. It assumes that the electricity price curve for the next 24 h is perfectly known in advance. Based on this information, a linear programming model is formulated to maximize the total daily profit, yielding a fixed optimal charging and discharging schedule for the entire day. This strategy serves as a reference benchmark under perfect-information assumptions.
- (2) **Rule-based Strategy:** This strategy mimics the intuitive decision-making of domain experts and represents a typical heuristic approach. The control rules are hard-coded as follows: when the predicted electricity price is lower than the predefined charging threshold (380 CNY/MWh), the VPP is charged at full power; when the predicted electricity price exceeds the predefined discharging threshold (700 CNY/MWh), the VPP discharges at full power. No active charging or discharging actions are performed when the price falls between the two thresholds.

All strategies are evaluated under the same realistic scenario with uncertainty to ensure a fair comparison.

4.2. Model Training and Convergence Analysis

Figure 2 illustrates the learning process of the VPP agent over 8500 training episodes. The horizontal axis represents the number of training episodes, while the vertical axis denotes the normalized total reward per episode.

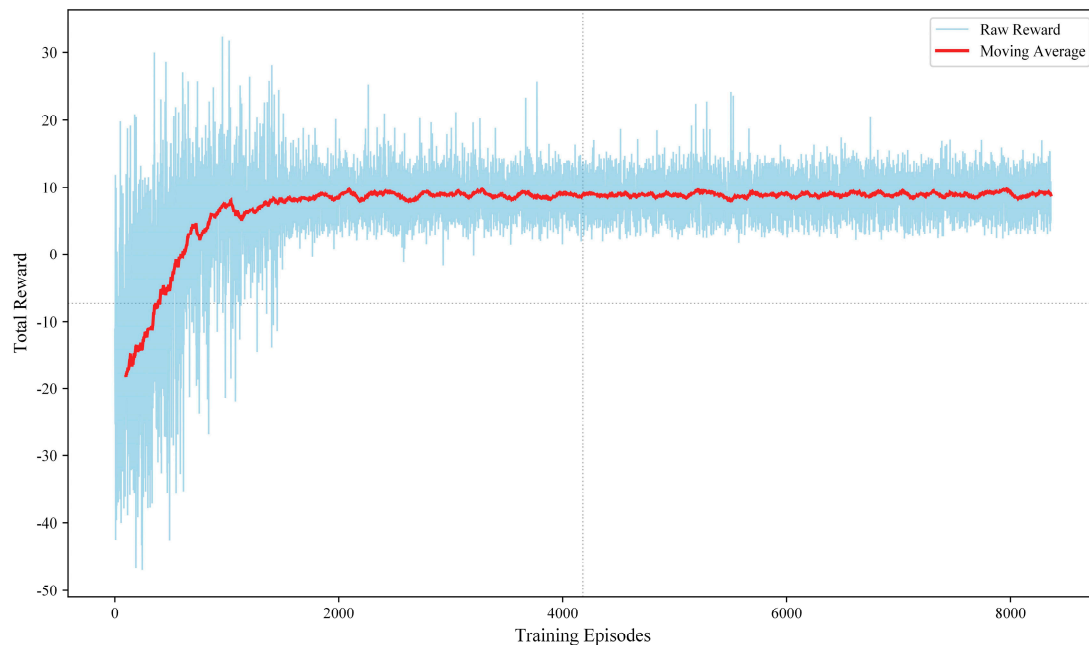


Figure 2. Convergence of the VPP Market Strategy Based on Deep Reinforcement Learning.

During the initial training stage (the first 500 episodes), the agent primarily explores the environment through random actions. Due to the absence of effective charging and discharging policies, the agent frequently violates operational constraints or discharges at unfavorable prices, resulting in low reward values. Subsequently, as the PPO algorithm effectively leverages historical experience, the reward curve exhibits a rapid increase between 500 and 1500 episodes. This indicates that the proposed potential-based reward shaping mechanism provides dense guidance signals and significantly accelerates the early learning process.

After approximately 2000 episodes, the red curve representing the moving average reward enters a clear plateau, indicating that the training process has converged. Meanwhile, the raw reward values shown in the light-blue background continue to exhibit noticeable fluctuations. This phenomenon does not indicate a lack of convergence but rather reflects the inherent stochasticity of the electricity market environment. Even under an optimal policy, fluctuations in electricity prices and variations in renewable energy output across different days inevitably lead to variability in daily profits. The ability of the agent to maintain a stable average reward under such a high-noise environment demonstrates the strong robustness of the proposed strategy.

4.3. Comparison of Economic Benefits of Different Strategies

To verify the effectiveness and robustness of the proposed strategy under uncertain environments, the proposed method is compared with deterministic optimization and a rule-based strategy in a realistic scenario with stochastic disturbances. The cumulative net profit of each strategy over a single-day scheduling horizon is summarized in Table 5.

Table 5. Comparison of Daily Net Profit in a Representative Realistic Scenario.

Scheduling Strategy	Daily Net Profit (CNY)	Profit Improvement Rate (vs. Rule-Based)
DRL-based strategy	76,206.64	94.65%
Deterministic optimization	66,426.78	69.67%
Rule-based strategy	39,150.00	--

As shown in Table 5, the proposed DRL-based strategy achieves the highest daily net profit in this representative realistic scenario, reaching CNY 76,206.64. In comparison, the deterministic optimization strategy and the rule-based strategy yield CNY 66,426.78 and CNY 39,150.00, respectively.

Compared with the rule-based strategy, the DRL-based strategy achieves a profit improvement of 94.65%. This significant improvement can be attributed to the inherent limitations of the rule-based strategy, which relies on fixed charging and discharging thresholds. As a result, it fails to exploit arbitrage opportunities when electricity prices fluctuate within intermediate ranges. In contrast, the DRL agent learns a more flexible nonlinear control policy through continuous interaction with the environment, enabling it to effectively capture small price spreads and adapt to stochastic market conditions.

In addition, the revenue achieved by the DRL-based strategy exceeds that of the deterministic optimization strategy based on perfect price forecasts (CNY 66,426.78), with an improvement of approximately 14.7%. This result highlights the advantages of the DRL model in coordinating heterogeneous resources within the VPP. When addressing the strongly coupled constraints between the BESS and the V2G fleet, traditional deterministic optimization methods often adopt decoupling techniques or sequential solution procedures to reduce computational complexity, which may lead to suboptimal solutions. In contrast, the DRL agent learns an end-to-end control policy that enables global cooperative scheduling of BESS and V2G resources. In particular, it effectively exploits the discharge potential of the V2G fleet during morning and evening peak periods, thereby achieving additional economic gains beyond the conventional optimization benchmark.

Figure 3 illustrates the dynamic evolution of power response behaviors and cumulative economic benefits of different scheduling strategies under a typical daily scenario. As shown in Figure 3a, all three strategies generally follow the basic arbitrage principle of charging during low-price periods and discharging during high-price periods. During the low electricity price period (00:00–06:00) in the morning, each strategy controls the energy storage unit to charge, so as to improve the state of charge at a lower cost. However, notable differences emerge among the strategies during periods of significant price fluctuations.

Specifically, the rule-based strategy exhibits clear rigidity in its decision-making due to its reliance on fixed price thresholds. As indicated by the green dotted line, during certain sub-peak periods (e.g., around 18:00), although the market price has reached a relatively high level, the VPP remains inactive because the predefined discharging threshold is not triggered, resulting in missed arbitrage opportunities.

In contrast, the proposed DRL-based strategy (red line) demonstrates stronger adaptability to the market environment. Rather than relying on a single predefined threshold, the DRL agent makes decisions based on the current system state and learned implicit representations of future price trends. During the evening peak period (18:00–21:00), the DRL strategy accurately identifies transient high-price signals and performs timely high-power discharging. Its operational behavior closely aligns with that of the deterministic optimization strategy (blue dotted line), which benefits from a global planning perspective.

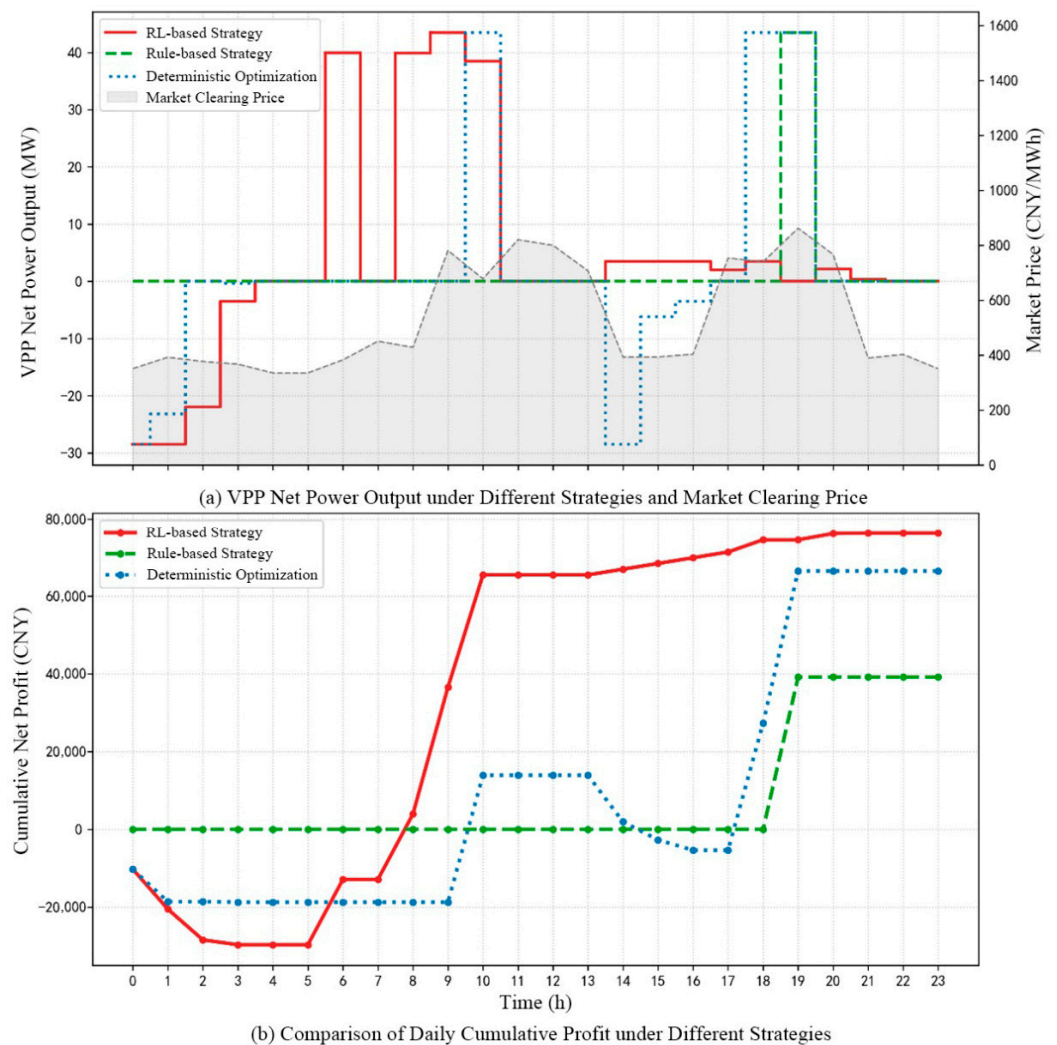


Figure 3. Comparison of 24-Hour Scheduling Behaviors and Cumulative Profits under Different Strategies.

This behavioral distinction is directly reflected in the cumulative economic performance shown in Figure 3b. Before the end of the morning peak, the profit differences among the strategies remain relatively small. With the arrival of the evening peak, the cumulative profit curve of the DRL-based strategy exhibits a steep upward trend due to its accurate timing of discharging actions, rapidly widening the gap with the rule-based strategy. Ultimately, the DRL-based strategy achieves a significantly higher daily cumulative net profit than the rule-based strategy and remains highly competitive with deterministic optimization. These results demonstrate that the proposed algorithm attains strong decision-making performance and robustness in an uncertain electricity market environment.

4.4. In-Depth Analysis of Policy Scheduling Mechanism

To further explore the physical logic and intelligent characteristics underlying the ‘black-box’ decisions of the deep reinforcement learning model, this section provides an in-depth analysis from two dimensions: energy storage state of charge (SOC) management and price-action response mechanisms.

4.4.1. In-Depth Analysis of Energy Storage Resource Utilization

Figure 4 shows the full-day variation in the energy storage unit's SOC under different control strategies in the same market environment. The comparative analysis demonstrates that the DRL strategy has a significant advantage in resource utilization.

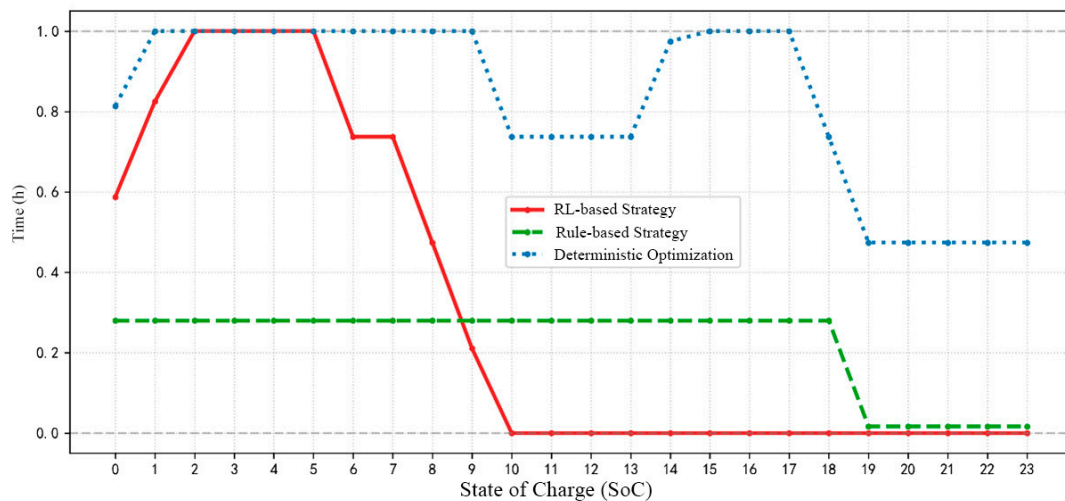


Figure 4. Full-day Trajectories of Energy Storage State of Charge (SOC) under Different Strategies.

As shown in the figure, the DRL strategy (red line) demonstrates great scheduling flexibility and depth in charge–discharge operations. During the early scheduling phase (00:00–02:00), the DRL strategy quickly charges the SOC from its initial value to 1.0 (full power state) and maintains it until the 5th hour, fully utilizing the low-price period in the morning for energy storage. Subsequently, during the 6th to 10th hour, the DRL strategy performs a decisive discharge operation, reducing the SOC to 0.0, thereby achieving full utilization of the energy storage capacity. This bold strategy illustrates that the agent has successfully learned an arbitrage model that maximizes the use of physical resources while satisfying constraints.

In contrast, the SOC curve of the rule-based strategy (green dotted line) shows minimal fluctuations throughout the entire process, maintaining a low level of around 0.3 for an extended period, with only a small discharge after the 18th hour. This reflects the conservatism of the fixed-threshold strategy: because the predicted price does not reach the preset charging threshold, the energy storage unit remains ‘idle’ for a prolonged period, leading to significant resource wastage and lost opportunity costs. Furthermore, the DRL strategy and the deterministic optimization strategy (blue dotted line) converge near 0 on the final SOC value, further verifying its effective adherence to intra-day resource clearing boundary conditions.

4.4.2. Price Signal and Action Response Mechanism

To reveal the decision logic of the DRL agent, Figure 5 visualizes the net output action distribution of the DRL agent under different market clearing prices (MCPs), where the color of each scatter point represents the output magnitude, with red indicating discharging and blue indicating charging.

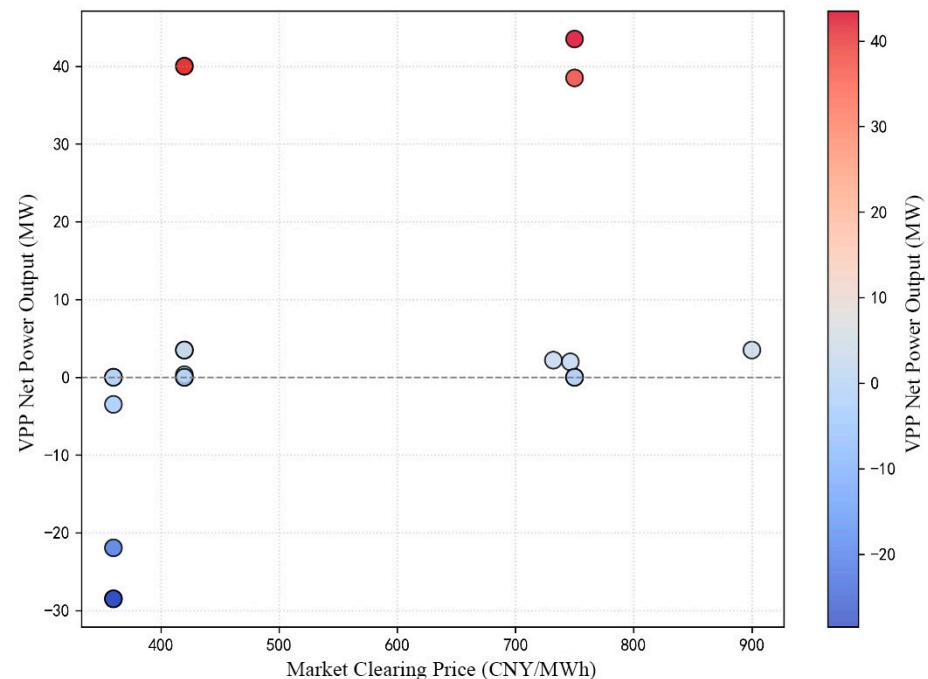


Figure 5. Action Response Distribution of the DRL Strategy under Different Market Prices.

- (1) **Nonlinear hierarchical response:** The action distribution exhibits a clear polarization pattern. When the electricity price is lower than 400 CNY/MWh (lower-left region of the figure), the scatter points are mainly concentrated in the negative output range (dark blue dots), indicating a strong tendency toward charging behavior. Conversely, when the electricity price exceeds 750 CNY/MWh (upper-right region of the figure), the scatter points are concentrated in the positive peak output range (dark red dots), indicating full discharging behavior.
- (2) **State-dependent decision-making:** Notably, at an intermediate price around 420 CNY/MWh, a high-power discharging outlier (red dot) can be observed. This indicates that the DRL strategy does not follow a simple linear price-action mapping, but instead makes decisions by jointly considering the current SOC state (e.g., when the battery is fully charged) and learned expectations of future price movements. This flexibility in handling atypical price conditions constitutes the core advantage of the DRL strategy over rigid rule-based approaches.

4.5. Statistical Robustness and Confidence-Interval Analysis

To further evaluate the reliability of the proposed method under stochastic market disturbances, repeated evaluations were conducted under the same realistic market setting using 20 different random seeds. For each strategy, the daily net profit was recorded and summarized in terms of the mean value, standard deviation, and 95% confidence interval, as reported in Table 6.

The results show that the proposed DRL-based strategy achieves the highest average daily net profit of CNY 84,374.88, with a standard deviation of 10,685.48 and a 95% confidence interval of [79,691.76, 89,058.00]. In comparison, deterministic optimization achieves an average daily net profit of CNY 59,730.45, with a much larger standard deviation of 30,412.47 and a 95% confidence interval of [46,401.60, 73,059.30], indicating that its performance is considerably less stable under stochastic disturbances. The rule-based strategy yields the lowest average daily net profit of CNY 20,024.96, with a standard deviation of 20,939.98 and a 95% confidence interval of [10,847.61, 29,202.31].

Table 6. Statistical Comparison of Daily Net Profit under Repeated Stochastic Evaluations.

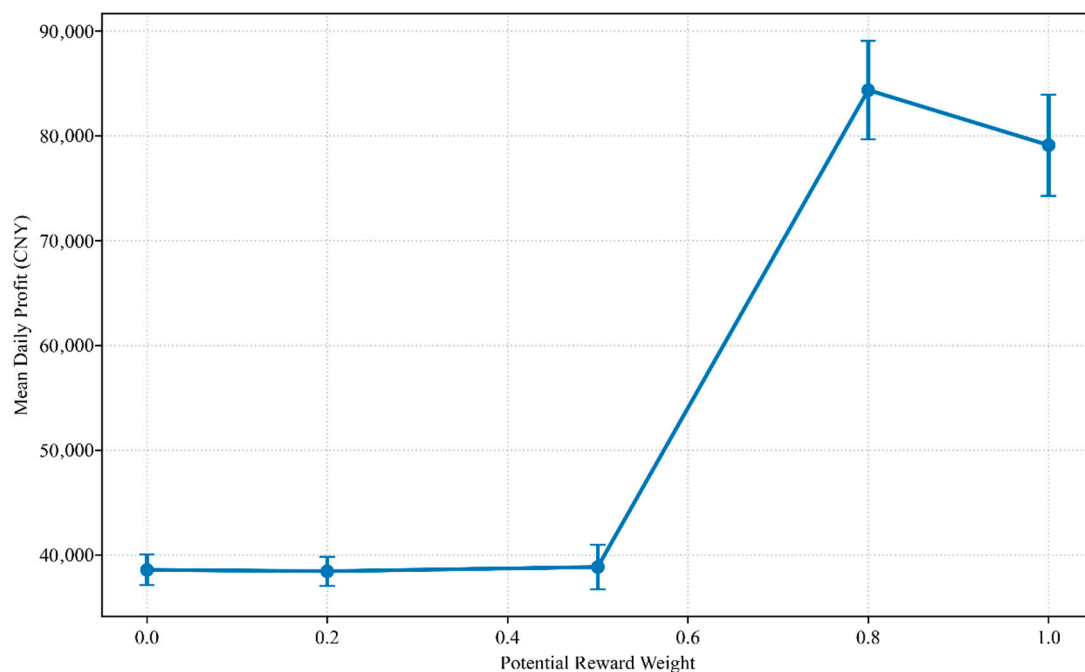
Method	Evaluation Runs	Mean Daily Net Profit (CNY)	Standard Deviation (CNY)	95% Confidence Interval (CNY)
DRL-based Strategy	20	84,374.88	10,685.48	[79,691.76, 89,058.00]
Deterministic Optimization	20	59,730.45	30,412.47	[46,401.60, 73,059.30]
Rule-based Strategy	20	20,024.96	20,939.98	[10,847.61, 29,202.31]

These repeated-evaluation results confirm that the proposed DRL-based strategy not only achieves the highest expected economic return, but also maintains a relatively stable performance under uncertainty. In particular, its confidence interval is clearly separated from those of the rule-based strategy and deterministic optimization, which provides stronger statistical support for the superiority of the proposed method in noisy realistic market scenarios.

4.6. Sensitivity and Ablation Analysis

4.6.1. Sensitivity to the Potential-Based Reward Weight

To examine whether the effectiveness of the proposed method depends on a narrowly tuned shaping coefficient, a sensitivity analysis was performed on the potential-based reward weight. Specifically, five representative values, namely 0.0, 0.2, 0.5, 0.8, and 1.0, were tested under the same training and evaluation protocol. The corresponding results are summarized in Figure 6 and Table 7.

**Figure 6.** Sensitivity of the Proposed Method to the Potential-based Reward Weight.

When the shaping weight is set to 0.0, 0.2, and 0.5, the average daily net profits are CNY 38,573.94, CNY 38,443.63, and CNY 38,843.41, respectively, indicating that weak or absent potential guidance fails to support profitable long-horizon arbitrage behavior. In contrast, when the shaping weight increases to 0.8, the average daily net profit rises

sharply to CNY 84,374.88, which is the best performance among all tested settings. When the weight is further increased to 1.0, the average daily net profit remains high at CNY 79,116.51, although it becomes slightly lower than the result at 0.8.

Table 7. Sensitivity Analysis of Daily Net Profit under Different Potential-based Reward Weights.

Potential-Based Reward Weight	Evaluation Runs	Mean Daily Net Profit (CNY)	Standard Deviation (CNY)	95% Confidence Interval (CNY)
0	20	38,573.94	3354.29	[37,103.86, 40,044.02]
0.2	20	38,443.63	3189.57	[37,045.75, 39,841.52]
0.5	20	38,843.41	4835.89	[36,723.99, 40,962.83]
0.8	20	84,374.88	10,685.48	[79,691.76, 89,058.00]
1	20	79,116.51	11,027.55	[74,283.47, 83,949.55]

These results indicate that the shaping weight has a substantial impact on the final policy performance. More importantly, the proposed method does not rely on a single fragile parameter point. Rather, it performs strongly within a moderate-to-high shaping range, and the selected value of 0.8 provides the best balance between effective long-term guidance and training stability under the current experimental setting.

4.6.2. Ablation Study on the Reward-Shaping Mechanism

To further verify the actual contribution of the proposed potential-based reward shaping (PBRS) mechanism, an ablation study was conducted by comparing three PPO variants under the same training and repeated stochastic evaluation protocol: PPO with the proposed PBRS, PPO with a simpler heuristic shaping baseline, and vanilla PPO without potential shaping. In the heuristic shaping baseline, the future maximum-price anchor used in the proposed shaping term was replaced by the current forecast price, thereby providing a simpler but more short-sighted guidance signal. The corresponding results are summarized in Table 8.

Table 8. Ablation Comparison among PPO Variants with Different Reward-shaping Designs.

Method	Evaluation Runs	Mean Daily Net Profit (CNY)	Standard Deviation (CNY)	95% Confidence Interval (CNY)
PPO with proposed PBRS	20	84,374.88	10,685.48	[79,691.76, 89,058.00]
PPO with heuristic current-price shaping	20	38,443.63	3189.57	[37,045.75, 39,841.52]
PPO without potential shaping	20	38,573.94	3354.29	[37,103.86, 40,044.02]

Note: The ablation isolates the contribution of the proposed PBRS term. The heuristic shaping baseline reuses the same training and repeated evaluation protocol, but replaces the future maximum-price anchor with the current forecast price.

As shown in Table 8, PPO with the proposed PBRS achieves the highest average daily net profit of CNY 84,374.88, with a 95% confidence interval of [79,691.76, 89,058.00]. By contrast, PPO with heuristic current-price shaping achieves only CNY 38,443.63, with a 95% confidence interval of [37,045.75, 39,841.52], while PPO without potential shaping achieves CNY 38,573.94, with a 95% confidence interval of [37,103.86, 40,044.02]. Notably, the heuristic shaping baseline provides no meaningful improvement over the no-shaping

baseline, indicating that simply adding an ad hoc shaping term is insufficient to produce strong long-horizon arbitrage behavior.

These results demonstrate that the proposed PBRS mechanism is not merely an auxiliary modification, but a key factor underlying the strong performance of the DRL-based strategy. By explicitly propagating future price potential to the current decision step, the proposed design effectively alleviates myopic behavior and yields a substantial profit improvement of more than 100% over both the heuristic shaping baseline and the no-shaping baseline.

4.7. Additional DRL Baseline and Practical Deployment Discussion

To provide an additional reference beyond PPO, a representative off-policy DRL baseline, namely Soft Actor-Critic (SAC), was also tested under the same environment and repeated stochastic evaluation protocol. Considering the higher computational cost of SAC in this environment, SAC was trained for 50,000 timesteps as a lightweight representative baseline, whereas PPO used the default 200,000-timestep setting. The corresponding comparison is summarized in Table 9. Under the tested configuration, PPO with the proposed PBRS achieves an average daily net profit of CNY 84,374.88, whereas the SAC-based strategy achieves CNY 53,275.18, with a 95% confidence interval of [45,187.40, 61,362.96]. These results indicate that PPO remains a competitive and effective choice for the current day-ahead VPP bidding problem. It should be noted that this comparison is intended as a representative algorithmic benchmark rather than an exhaustive survey of all DRL architectures.

Table 9. Comparison between PPO and an Additional DRL Baseline (SAC).

Method	Training Timesteps	Evaluation Runs	Mean Daily Net Profit (CNY)	Standard Deviation (CNY)	95% Confidence Interval (CNY)
PPO with proposed PBRS	200,000	20	84,374.88	10,685.48	[79,691.76, 89,058.00]
SAC (additional DRL baseline)	50,000	20	53,275.18	18,453.91	[45,187.40, 61,362.96]

From a practical deployment perspective, it is important to distinguish between offline training and online decision-making. The computational burden of the proposed DRL framework is mainly concentrated in the offline training phase, where the model learns from historical or simulated market interactions. In contrast, once the policy has been trained, online deployment only requires a forward pass of the neural network to generate the bidding action. The corresponding runtime statistics are summarized in Table 10.

Table 10. Runtime Characteristics of PPO Training and Online Inference.

Item	Value
Training algorithm	PPO
Training timesteps	200,000
Total training time (s)	127.24
Total training time (h)	0.0353
Average inference time per forward pass (ms)	0.339
Number of inference repeats used for measurement	2000

As shown in Table 10, the total training time of the PPO-based model is 127.24 s, whereas the average inference time of a single forward pass is only 0.339 ms. This result indicates that, although DRL model training requires offline computation, the trained policy can be deployed efficiently for practical day-ahead bidding support. Therefore, the

proposed framework is computationally feasible for real-world VPP operation in scenarios where periodic offline retraining is acceptable.

4.8. Discussion and Limitations

The experimental results presented in Sections 4.2–4.7 collectively demonstrate the practical viability of the proposed DRL-based bidding framework for VPP operations. As shown in Table 10, the PPO-based model can be trained offline in approximately two minutes and subsequently deployed with a sub-millisecond inference latency of 0.339 ms per forward pass. This computational profile makes the framework well-suited for real-world day-ahead market operations, where bidding decisions are made on an hourly or daily basis and periodic offline retraining with updated market data is entirely feasible. Moreover, the sensitivity analysis (Section 4.6.1) confirms that the proposed method performs robustly across a range of shaping weights, reducing the need for exhaustive hyperparameter tuning in practical deployment.

It should be noted that the current study adopts several deliberate modeling simplifications to maintain a clear experimental focus on the proposed reward-shaping mechanism. Specifically, the VPP is modeled as a price-taker, and the market-clearing process does not incorporate detailed network constraints such as transmission line capacities or voltage limits. These simplifications are commonly adopted in the DRL-based energy management literature to isolate algorithmic contributions from environmental complexity. While they may not fully capture all operational factors encountered in real-world markets, the core algorithmic findings—particularly the effectiveness of potential-based reward shaping in overcoming myopic behavior—are expected to remain valid under more detailed market models.

Building on the current work, future research will extend the proposed framework in two directions: (1) incorporating a multi-agent DRL formulation to capture strategic interactions among multiple price-maker participants, and (2) integrating AC network constraints to ensure physically feasible dispatch solutions. These extensions represent natural next steps toward bridging the gap between the simulation environment and full-scale real-world deployment.

5. Conclusions

Aiming at the bidding decision-making problem of VPPs in the uncertain day-ahead electricity market, this paper proposes an adaptive deep reinforcement learning strategy incorporating potential function-based reward shaping. Through multidimensional simulation experiments and mechanism analysis, the following conclusions are drawn:

- (1) **Convergence and effectiveness of the algorithm:** By introducing the potential function-based reward shaping mechanism, the sparse reward issue and the short-sighted behavior of reinforcement learning in long-horizon decision-making are effectively mitigated. Training process analysis shows that the model converges to a stable performance region after approximately 2000 episodes (roughly 50,000 interaction steps), demonstrating good learning efficiency and stability.
- (2) **Economic performance and robustness:** In realistic market scenarios with random noise, the proposed DRL strategy demonstrates significant economic advantages. Its single-day cumulative net profit not only substantially exceeds that of the traditional rule-based strategy, but also exhibits decision-making behavior highly consistent with deterministic optimization during key peak arbitrage periods, indicating strong robustness under uncertain market conditions.
- (3) **Interpretability of the decision-making mechanism:** Through in-depth analysis of scheduling behaviors, this paper reveals the internal decision logic underlying the

DRL agent's superiority over rule-based strategies. SOC trajectory analysis shows that the DRL strategy achieves full-depth utilization of energy storage capacity from 0% to 100%, avoiding resource idleness caused by conservative thresholds in rule-based strategies. Price–action distribution analysis further confirms that the agent has learned a nonlinear, state-dependent strategy, which flexibly adjusts charging and discharging behaviors according to its state of charge under atypical price conditions, reflecting advanced decision-making capability beyond rigid rule-based control.

Author Contributions: Methodology, W.Y.; Software, W.W.; Validation, W.Y.; Writing—original draft, W.Y.; Writing—review & editing, W.W.; Funding acquisition, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Research Program of Yazhou Bay Innovation Research Institute, Hainan Tropical Ocean University. Grant No: 2023CXYZD001.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Due to privacy settings, the authors do not have permission to provide the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, X.; Gao, C. Review and prospects of artificial intelligence technology in virtual power plants. *Energies* **2025**, *18*, 3325. [\[CrossRef\]](#)
2. Xiao, D. A Review on Risk-Averse Bidding Strategies for Virtual Power Plants with Uncertainties: Resources, Technologies, and Future Pathways. *Technologies* **2025**, *13*, 488. [\[CrossRef\]](#)
3. Li, Y.; Chang, W.; Yang, Q. Deep reinforcement learning based hierarchical energy management for virtual power plant with aggregated multiple heterogeneous microgrids. *Appl. Energy* **2025**, *382*, 125333. [\[CrossRef\]](#)
4. Feng, B.; Liu, Z.; Huang, G.; Guo, C. Robust federated deep reinforcement learning for optimal control in multiple virtual power plants with electric vehicles. *Appl. Energy* **2023**, *349*, 121615. [\[CrossRef\]](#)
5. Alabi, T.M.; Lu, L.; Yang, Z. Data-driven optimal scheduling of multi-energy system virtual power plant (MEVPP) incorporating carbon capture system (CCS), electric vehicle flexibility, and clean energy marketer (CEM) strategy. *Appl. Energy* **2022**, *314*, 118997. [\[CrossRef\]](#)
6. Jiang, Y.; Dong, J.; Huang, H. Optimal bidding strategy for the price-maker virtual power plant in the day-ahead market based on multi-agent twin delayed deep deterministic policy gradient algorithm. *Energy* **2024**, *306*, 132388. [\[CrossRef\]](#)
7. Sun, Q.; Wang, X.; Liu, Z.; Mirsaeidi, S.; He, J.; Pei, W. Multi-agent energy management optimization for integrated energy systems under the energy and carbon co-trading market. *Appl. Energy* **2022**, *324*, 119646. [\[CrossRef\]](#)
8. Tang, X.; Wang, J. Deep Reinforcement Learning-Based Multi-Objective Optimization for Virtual Power Plants and Smart Grids: Maximizing Renewable Energy Integration and Grid Efficiency. *Processes* **2025**, *13*, 1809. [\[CrossRef\]](#)
9. Gulotta, F.; Crespo del Granado, P.; Piscicella, P.; Siface, D.; Falabretti, D. Short-term uncertainty in the dispatch of energy resources for VPP: A novel rolling horizon model based on stochastic programming. *Int. J. Electr. Power Energy Syst.* **2023**, *153*, 109355. [\[CrossRef\]](#)
10. Wang, J.; Ilea, V.; Bovo, C.; Xie, N.; Wang, Y. Optimal self-scheduling for a multi-energy virtual power plant providing energy and reserve services under a holistic market framework. *Energy* **2023**, *278*, 127903. [\[CrossRef\]](#)
11. Wang, Z.; Guo, H.; Zhu, R.; Liu, Z. Distributed robust optimization strategy for multi-energy virtual power plant clusters. *Sci. Rep.* **2026**, *16*, 12. [\[CrossRef\]](#)
12. Kong, Y.; Chen, Y.; Du, J.; Yang, Y.; Xu, Q. A Bidding Strategy for Virtual Power Plants in the Day-Ahead Market. *Energies* **2025**, *18*, 4874. [\[CrossRef\]](#)
13. Lu, X.; Xia, S.; Sun, G.; Hu, J.; Zou, W.; Zhou, Q.; Shahidehpour, M.; Chan, K.W. Hierarchical distributed control approach for multiple on-site DERs coordinated operation in microgrid. *Int. J. Electr. Power Energy Syst.* **2021**, *129*, 106864. [\[CrossRef\]](#)
14. Hu, R.; Zhou, K.; Yin, H. Reinforcement learning model for incentive-based integrated demand response considering demand-side coupling. *Energy* **2024**, *308*, 132997. [\[CrossRef\]](#)
15. Gao, S.; Hou, X.; Li, C.; Sun, Y.; Du, M.; Wang, D. Intelligent Decision-Making for Multi-Scenario Resources in Virtual Power Plants Based on Improved Ant Colony Algorithm-Simulated Annealing Algorithm. *Sustainability* **2025**, *17*, 8600. [\[CrossRef\]](#)

16. Yu, J.; Fan, Y.; Hou, J. Research on Distributed Optimization Scheduling and Its Boundaries in Virtual Power Plants. *Electronics* **2025**, *14*, 932. [\[CrossRef\]](#)
17. Yu, D.; Zhao, X.; Wang, Y.; Jiang, L.; Liu, H. Research on energy management of a virtual power plant based on the improved cooperative particle swarm optimization algorithm. *Front. Energy Res.* **2022**, *10*, 785569. [\[CrossRef\]](#)
18. Cao, J.; Yang, D.; Dehghanian, P. Cooperative operation for multiple virtual power plants considering energy-carbon trading: A Nash bargaining model. *Energy* **2024**, *307*, 132813. [\[CrossRef\]](#)
19. Aboelhassan, I.; Almetwally, R.; Jin, Y.; Xu, S.; Chang, L. Modified energy management system for virtual power plant applications. In Proceedings of the 2025 IEEE Electrical Power and Energy Conference (EPEC), Waterloo, ON, Canada, 15–17 October 2025; pp. 242–248.
20. Li, J.; Wang, C.; Liu, Y. AI-Driven Virtual Power Plants: A Comprehensive Review. *Energies* **2026**, *19*, 1084. [\[CrossRef\]](#)
21. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement learning and its applications in modern power and energy systems: A review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. [\[CrossRef\]](#)
22. Gao, Z.; Kang, W.; Chen, X.; Gong, S.; Liu, Z.; He, D.; Shi, S.; Shangguan, X.-C. Optimal economic dispatch of a virtual power plant based on gated recurrent unit proximal policy optimization. *Front. Energy Res.* **2024**, *12*, 1357406. [\[CrossRef\]](#)
23. Ma, H.; Li, G.; Zhang, Y.; Hu, J.; Wu, Q. Research on optimal scheduling strategy of virtual power plant multi-operation scenarios based on deep reinforcement learning. In Proceedings of the 2024 Second International Conference on Cyber-Energy Systems and Intelligent Energy (ICCSIE), Shenyang, China, 17–19 May 2024; pp. 1–6.
24. Zhang, X.; Yan, G.; Zhou, M.; Tang, C.; Qu, R. Energy management method for virtual power plant based on double-layer deep reinforcement learning game. In Proceedings of the 2024 6th International Conference on Energy, Power and Grid (ICEPG), Guangzhou, China, 27–29 September 2024; pp. 650–655.
25. Li, J.; Guo, Y.; Sun, H.; Wen, X. Bidding strategic of virtual power plant based on end-to-end deep reinforcement learning. In Proceedings of the 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2), Hangzhou, China, 15–18 December 2023; pp. 3044–3049.
26. Lu, R.; Jiang, Z.; Wu, H.; Ding, Y.; Wang, D.; Zhang, H.-T. Reward shaping-based actor–critic deep reinforcement learning for residential energy management. *IEEE Trans. Ind. Inform.* **2023**, *19*, 2662–2673. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.