

Article

Multi-Agent Deep Reinforcement Learning for V2G Scheduling and Incentive Mechanism to Enhance Distribution Network Loadability

Xiangchao Li ¹, Xiaobo Mao ¹, Ni Tang ¹, Yanming Zhu ^{2,*} and Jiaqing Lin ²

¹ State Grid Wuxi Power Supply Company, Wuxi 214000, China; li_xc_1@163.com (X.L.); mao_xb1@163.com (X.M.); tang_ni1@163.com (N.T.)

² Key Laboratory of Control of Power Transmission and Conversion of the Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China; twinkle@sjtu.edu.cn

* Correspondence: zhuyanming@sjtu.edu.cn

Abstract

The rapid integration of distributed renewable energy has substantially increased the operational uncertainty of distribution networks, making the preservation of sufficient loadability increasingly difficult. To address this problem, a bi-level coordinated scheduling framework is developed in which the interaction between the distribution network operator (DNO) and electric vehicle aggregators (EVAs) is modeled as a Stackelberg game. Unlike existing studies that treat supply security only as a constraint, allocate revenue solely by economic contribution, and prioritize experience replay by TD error alone, the proposed framework embeds the loadability directly into the upper-level objective, allocates the aggregated revenue by an improved Shapley value method that combines an S-shaped utility function reflecting user participation with a loadability contribution factor quantifying each user's contribution to supply security, and solves the bi-level model by a dual-agent Soft Actor–Critic (SAC) and Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MATD3) framework equipped with a Stackelberg-Aware Dynamic Prioritized Experience Replay (SD-PER) mechanism that augments the TD error with opponent strategy variation to make the replay aware of the leader–follower hierarchy. Case studies on the IEEE 33-bus system and a real 72-bus distribution network in Shanghai show that the proposed method raises the average peak-period loadability from 1.05 to 1.24 on the 33-bus system, lifts the minimum nodal voltage at the evening peak from 0.937 to 0.953 p.u. while reducing the average voltage-deviation index by about 20.5%, lowers the net user cost on the EVA side, and attains a scheduling quality within 1% of a KKT/MPEC bi-level benchmark, while the SD-PER mechanism accelerates convergence to about 1440 episodes and improves the final reward by about 40% over the base SAC. The same advantages are reproduced on the real distribution network in Shanghai, confirming the practical applicability and scalability of the method.



Academic Editor: Il-Yop Chung

Received: 25 May 2026

Revised: 6 July 2026

Accepted: 8 July 2026

Published: 10 July 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.

Keywords: deep reinforcement learning; improved Shapley value; Stackelberg game; vehicle-to-grid interaction; loadability; distribution network

1. Introduction

With the development of the “dual-carbon” strategy and the continuous growth of distributed renewable generation, the operational uncertainty and volatility of distribution networks have increased significantly [1]. When periods of low renewable output

coincide with peak load, the available supply capacity becomes strained, which raises the requirements for power supply security [2]. From the perspective of adequacy evaluation, the loadability reflects the system's ability to serve the actual demand while satisfying operational constraints, and is a key indicator of secure distribution network operation [3]. Electric vehicles (EVs) have the dual attributes of load and energy storage. They can absorb energy during margin-sufficient periods and provide support during margin-tight periods through coordinated charging and discharging, and are therefore regarded as an important flexibility resource for enhancing the loadability of distribution networks [4,5]. However, large-scale EV integration is characterized by strong randomness and spatiotemporal uncertainty. Without effective coordination, EVs may worsen the peak–valley difference and cause local voltage violations or line overloading [6]. There is therefore a need to design coordinated scheduling mechanisms that take into account both supply security and the economic interests of EV users.

In recent years, game-theoretic approaches have been widely used for EV-integrated distribution network scheduling, because they can effectively describe the cooperative and competitive interactions among multiple stakeholders. Among them, the Stackelberg game establishes a bi-level structure in which the leader sets the price and the follower responds optimally, and provides a suitable framework for coordinating grid-side and user-side interests. Ref. [7] builds a Stackelberg game between the DNO and load aggregators, using upper-level signals to guide flexible load response and to improve voltage performance. Ref. [8] considers competition among multiple aggregators and uses equilibrium constraint methods to optimize charging and discharging in multi-agent scenarios, with proof of equilibrium existence. Ref. [9] extends the leader–follower framework to a three-layer structure that includes an active distribution network, a microgrid cluster, and shared energy storage. Ref. [10] formulates the utility–consumer interaction as a Stackelberg–Nash dynamic game solved by asynchronous multi-agent reinforcement learning, and Ref. [11] takes this a step further with a fully distributed Stackelberg multi-agent reinforcement learning scheme that preserves user privacy while keeping the leader–follower price interaction. Ref. [12] designs a game-theoretic mechanism for V2G multilateral contracts and analyzes the benefit equilibrium among the grid, aggregators and EV users, coordinating their charging and discharging interactions through the contract and pricing design. Ref. [13] proposes a two-level Stackelberg–Nash game framework for an alternative-fuel-vehicle-integrated microgrid in a coupled power–transportation network, in which a pricing and incentive mechanism is devised to promote the participation of electric and hydrogen vehicles in V2G services. However, the upper-level objective is formulated primarily around economic cost, and supply security typically enters as a constraint rather than as an explicit objective to be maximized. When large-scale EV integration causes large changes in load characteristics, such a passive treatment of security is not sufficient to fully exploit the potential of EVs for loadability enhancement.

Within the game-based framework, the total revenue obtained by the EVA through aggregating internal user resources needs to be allocated among users in a reasonable way. The fairness and incentive effect of the allocation mechanism directly affect the sustainability of coordinated scheduling. The Shapley value, as a classical allocation method based on the marginal contribution principle in cooperative game theory, has been applied to revenue allocation in energy systems. Ref. [14] uses the Shapley value for energy management in prosumer coalitions to achieve fair compensation among coalition members. Building on this basic form, Ref. [15] introduces multi-dimensional correction factors to obtain an improved Shapley value for regional integrated energy systems, and Ref. [16] develops a comparable improved Shapley value-based profit allocation method for combined heat and power virtual power plants to balance the interests of distributed resource participants.

These methods predominantly capture the marginal economic contribution of participants, whereas each user's contribution to the loadability of the distribution network has received comparatively little attention in the allocation. As a result, it is difficult to encourage users to provide proactive discharge support during margin-tight periods.

Finding an efficient solution for the bi-level Stackelberg model is another challenge. Due to the nested optimization structure and the non-convex constraints caused by power flow equations, traditional mathematical programming methods suffer from high computational complexity and may converge to local optima. Deep reinforcement learning (DRL) can learn complex policies online without requiring an accurate environment model, and provides a feasible approach for solving such problems. Ref. [17] reviews the progress of reinforcement learning in power system decision-making and control. Ref. [18] applies the SAC algorithm to integrated energy system scheduling, and uses the maximum entropy framework to balance exploration and exploitation under uncertainty. Ref. [19] introduces prioritized experience replay (PER) into multi-agent energy management, and improves training convergence by assigning higher sampling probability to samples with larger TD errors. Ref. [20] proposes a clustering-based deep reinforcement learning method for EV charging and discharging scheduling that first builds a probabilistic behavioral model from charging transaction data to characterize the stochasticity of user charging behavior. To better respect operational security and the hierarchical decision structure, Ref. [21] further develops a data-efficient bi-level safe DRL method that coordinates the distribution network operator and a charging-station aggregator while keeping voltage and peak load within secure limits, and Ref. [22] casts the secure dispatch of multiple virtual power plants with EV clusters as a Markov game solved by a voltage-security-aware multi-agent DRL method. In addition, Ref. [23] develops a federated deep reinforcement learning framework for privacy-sensitive multi-station charging networks, in which each station trains a local Soft Actor–Critic policy and exchanges only model parameters with a global aggregator. In these PER-based solvers, however, the sampling priority is still driven mainly by TD error, and the leader–follower hierarchy of the Stackelberg game is not yet reflected in the replay mechanism. When opponent strategies change dynamically, TD error alone cannot accurately reflect the actual learning value of samples, which limits the adaptability of the agents to game dynamics.

The literature review is summarized in Table 1. Despite the progress made, several gaps can be identified:

- (1) The loadability of distribution networks is only treated as a constraint rather than an explicit objective in the upper-layer optimization problem.
- (2) The revenue allocation mechanism only accounts for users' economic contributions, while ignoring their individual contributions to distribution network loadability.
- (3) The prioritized experience replay only takes the TD error as the sole priority metric and fails to incorporate the hierarchical characteristics of the Stackelberg leader–follower game.

To address the above limitations, this paper proposes a Stackelberg game-based bi-level coordinated scheduling and revenue allocation method for the DNO and EVAs, together with a corresponding DRL solution algorithm. The main contributions are summarized as follows:

- (1) A bi-level Stackelberg game model is established in which the loadability is explicitly embedded into the upper-level objective together with DistFlow equations and physical constraints, while the lower-level model minimizes the operating cost of each EVA. By treating loadability as an objective to be maximized rather than a constraint to be merely satisfied, the model enables proactive supply security management.
- (2) An improved Shapley value-based revenue allocation mechanism is proposed, which introduces an S-shaped utility function to characterize user participation and a load-

ability contribution factor to measure each user's actual contribution to supply security. By bringing the loadability dimension into the allocation alongside economic contribution, the two factors refine the classical Shapley value and align user incentives with supply security, achieving both fairness and incentive compatibility.

- (3) A dual-agent alternating training framework based on SAC and MATD3 is designed, together with an SD-PER mechanism that constructs a dual-priority criterion by combining TD error with game strategy variation and uses adaptive weighting coefficients so that the replay priority becomes aware of the leader–follower game, enhancing sample efficiency and convergence stability in non-stationary game environments.

Table 1. Literature review.

Ref.	Coordination/Game Model	Loadability as Explicit Objective	Allocation Considers Loadability Contribution	Game-Aware Sample Prioritization
[7]	DNO–load-aggregator Stackelberg game	No	N/A	N/A
[8]	Multi-aggregator Stackelberg with equilibrium constraints	No	N/A	N/A
[9]	Three-layer Stackelberg	No	N/A	N/A
[10]	Stackelberg–Nash dynamic DR game	No	No	No
[11]	Fully distributed Stackelberg	No	No	No
[12]	V2G multilateral-contract game	No	No	N/A
[13]	Two-level Stackelberg–Nash game	No	No	N/A
[14]	Prosumer-coalition cooperative game	No	No	N/A
[15]	Regional IES cooperative game	No	No	N/A
[16]	CHP-VPP cooperative game	No	No	N/A
[20]	Clustering-based DRL scheduling	No	N/A	No
[21]	DSO–aggregator bi-level	No	N/A	No
[22]	Multi-VPP Markov game	No	N/A	No
[23]	Federated DRL	No	N/A	No
This paper	DNO–EVA bi-level Stackelberg game	SAC–MATD3 with SD-PER	Yes	Yes

2. Mathematical Formulation of the Stackelberg Game Model

In the coordinated operation scenario considered in this paper, the DNO serves as the upper-level leader and shapes the charging and discharging behavior of the EVAs through the design of electricity pricing strategies. As the lower-level follower, each EVA determines the charging and discharging schedule of its internal EV users in response to the price signals released by the DNO. This hierarchical decision-making process can be naturally described by a Stackelberg game, which provides an effective mechanism for reconciling the interests of the two levels [24]. The model framework is shown in Figure 1.

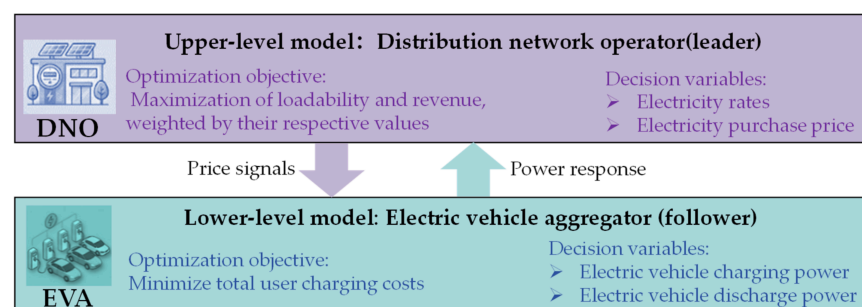


Figure 1. Stackelberg game framework of the DNO and EVAs.

2.1. Upper-Level Model: Maximization of Distribution Network Loadability

2.1.1. Objective Function

As the upper-level leader, the DNO aims to maximize the loadability of the distribution network while accounting for its own economic benefit. The loadability is described by a

single system-level variable SM_t , and the objective also includes the economic benefit of the DNO. The upper-level objective function is therefore formulated as

$$\max F_{DNO} = \lambda_1 \sum_{t \in T} \omega_t (SM_t - 1) + \lambda_2 (I_{DNO} - C_{DNO}) \quad (1)$$

$$I_{DNO} = \sum_{t \in T} \sum_{m \in N_{EVA}} \lambda_{s,m}^t \cdot P_{ch,m}^t \cdot \Delta t \quad (2)$$

$$C_{DNO} = \sum_{t \in T} \sum_{m \in N_{EVA}} \lambda_{b,m}^t \cdot P_{dc,m}^t \cdot \Delta t \quad (3)$$

where SM_t represents the loadability of the distribution network at time t , representing the maximum load multiple that the network can support relative to the actual demand while satisfying all operational security constraints; it is taken as a single value for the whole feeder. When $SM_t > 1$, it indicates that the network has sufficient loadability, when $SM_t = 1$, it indicates operation exactly at the actual load level. ω_t is the weighting coefficient assigned to the loadability at time t . I_{DNO} and C_{DNO} represent the electricity sales revenue and purchasing cost of the DNO, respectively. λ_1 and λ_2 are weighting factors used to balance loadability performance and economic benefit. λ_1 is expressed in monetary value per unit of margin, so that the loadability term and the economic benefit are measured in the same monetary unit. $P_{ch,m}^t$ and $P_{dc,m}^t$ represent the charging and discharging power of EVA m at time t , respectively. $\lambda_{s,m}^t$ and $\lambda_{b,m}^t$ represent the selling and purchasing electricity prices between the DNO and EVA m at time t , respectively. Δt represents the time interval.

2.1.2. Constraints

To ensure the secure and stable operation of the distribution network, the upper-level model is subject to the following constraints:

$$SM_t \geq SM_{min}, \forall t \in T \quad (4)$$

$$P_{jk}^t = P_{ij}^t - r_{ij} \frac{(P_{ij}^t)^2 + (Q_{ij}^t)^2}{(U_i^t)^2} - P_{L,j}^t + P_{ch,j}^t - P_{dc,j}^t \quad (5)$$

$$Q_{jk}^t = Q_{ij}^t - x_{ij} \frac{(P_{ij}^t)^2 + (Q_{ij}^t)^2}{(U_i^t)^2} - Q_{L,j}^t \quad (6)$$

$$(U_j^t)^2 = (U_i^t)^2 - 2(r_{ij}P_{ij}^t + x_{ij}Q_{ij}^t) + (r_{ij}^2 + x_{ij}^2) \frac{(P_{ij}^t)^2 + (Q_{ij}^t)^2}{(U_i^t)^2} \quad (7)$$

$$\underline{U}_i \leq U_i^t \leq \bar{U}_i, \forall i, t \quad (8)$$

$$-\bar{P}_{ij} \leq P_{ij}^t \leq \bar{P}_{ij}, \forall ij, t \quad (9)$$

$$\rho_{ij}^t = \frac{|P_{ij}^t|}{P_{ij}} \leq \rho_{max} \quad (10)$$

$$\underline{\lambda}_{s,m}^t \leq \lambda_{s,m}^t \leq \bar{\lambda}_{s,m}^t \quad (11)$$

$$\underline{\lambda}_{b,m}^t \leq \lambda_{b,m}^t \leq \bar{\lambda}_{b,m}^t \quad (12)$$

$$\frac{1}{|T|} \sum_{t \in T} \lambda_{s,m}^t = \frac{1}{|T|} \sum_{t \in T} \lambda_t^{TOU} \quad (13)$$

$$0 \leq P_{ch,m}^t \leq P_{ch,m}^{t,max} \quad (14)$$

$$0 \leq P_{dc,m}^t \leq P_{dc,m}^{t,max} \quad (15)$$

$$SM_t = \max \left\{ \rho \geq 1 \mid U_i^{\min} \leq U_i^t(\rho) \leq U_i^{\max}, |P_{ij}^t(\rho)| \leq P_{ij}^{\max}, \forall i, ij \right\} \quad (16)$$

where SM_{min} represents the predefined safety threshold of the loadability. P_{ij}^t and Q_{ij}^t represent the active and reactive power flows from node i to node j at time t , respectively. r_{ij} and x_{ij} represent the resistance and reactance of branch ij , respectively. U_i^t represents the voltage of node i at time t . $P_{L,j}^t$ and $Q_{L,j}^t$ represent the active and reactive loads at node j at time t , respectively. $\underline{U}_i$ and $\overline{U}_i$ represent the upper and lower limits of the voltage at node i , respectively. $\overline{P}_{ij}$ represents the maximum allowable transmission power of branch ij . ρ_{max} represents the maximum allowable line loading ratio. $\underline{\lambda}_{s,m}^t$ and $\overline{\lambda}_{s,m}^t$ represent the lower and upper limits of the electricity selling price, respectively. $\underline{\lambda}_{b,m}^t$ and $\overline{\lambda}_{b,m}^t$ represent the lower and upper limits of the electricity purchasing price, respectively. λ_t^{TOU} represents the original time-of-use price. $P_{ch,m}^{t,max}$ and $P_{dc,m}^{t,max}$ are the maximum charging and discharging power of EVA m at time t , jointly determined by the number of schedulable EVs and their state-of-charge conditions. $U_i^t(\rho)$ and $P_{ij}^t(\rho)$ denote the nodal voltage and branch power obtained when all nodal loads are scaled simultaneously by the factor ρ . Specifically, Equation (4) requires the loadability to remain no lower than the safety threshold so as to guarantee supply adequacy. Equations (5)–(7) describe the DistFlow power flow equations of the distribution network; at each time step, these nonlinear DistFlow equations are solved directly within the learning environment to obtain the nodal voltages and branch flows, which are then used to evaluate the loadability and the security constraints. In the active-power balance (5) the EVs at each node enter only through their charging power $P_{ch,j}^t$ and discharging power $P_{dc,j}^t$, and as the EV chargers are operated at unity power factor they exchange active power only, so the reactive balance (6) contains the nodal reactive load $Q_{L,j}^t$ alone. Equation (8) restricts node voltages within their acceptable range to maintain power quality. Equation (9) limits the branch transmission power below the line capacity to prevent line overloading. Equation (10) bounds the line loading ratio for safe operation. Equations (11)–(13) impose upper and lower bounds on the transaction prices and require the energy-weighted average selling price to equal the average original time-of-use price. This constraint fixes the average price level but not the total revenue, since the revenue also depends on the amount of energy traded, which differs across scenarios as different incentives shift different amounts of charging into the valley periods and elicit different amounts of peak-period discharging. Equations (14) and (15) constrain the charging and discharging power of each EVA according to its schedulable capability. Equation (16) defines the loadability SM_t as the largest factor by which all nodal loads can be scaled up, from the operating point produced by the charging and discharging schedule, while the voltage and branch-current limits (8) and (9) are still satisfied; proactive discharging during margin-tight periods relieves the network and raises SM_t until one of these limits becomes binding.

2.2. Lower-Level Model: Optimal Operation of EVAs

2.2.1. Objective Function

As the lower-level follower, each EVA participates in the game on behalf of its EV users, with the goal of minimizing the total electricity-related cost. The objective function is formulated as

$$\min F_{EVA} = C_{EVA} - I_{EVA} + C_{BL} \quad (17)$$

$$C_{EVA} = \sum_{t \in T} \sum_{m \in N_{EVA}} \lambda_{s,m}^t \cdot P_{ch,m}^t \cdot \Delta t \quad (18)$$

$$I_{EVA} = \sum_{t \in T} \sum_{m \in N_{EVA}} \lambda_{b,m}^t \cdot P_{dc,m}^t \cdot \Delta t \quad (19)$$

$$C_{BL} = \sum_{t \in T} \sum_{m \in N_{EVA}} (c_{ch} \cdot P_{ch,m}^t + c_{dc} \cdot P_{dc,m}^t) \cdot \Delta t \quad (20)$$

where I_{EVA} and C_{EVA} represent the discharging revenue and charging cost, respectively. C_{BL} represents battery degradation cost. c_{ch} and c_{dc} are the per-kWh battery degradation coefficients associated with charging and discharging, respectively.

2.2.2. Constraints

The lower-level model is subject to the following constraints:

$$\sum_{t \in T} (P_{ch,m}^t - P_{dc,m}^t) \cdot \Delta t = \sum_{t \in T} P_{ch,m}^{t,pre} \cdot \Delta t \quad (21)$$

$$P_{ch,m}^t = \sum_{n \in M} (P_{EV,ch,n}^t - P_{EV,dc,n}^t) \quad (22)$$

$$P_{dc,m}^t = \sum_{n \in M} (P_{EV,dc,n}^t - P_{EV,ch,n}^t) \quad (23)$$

$$0 \leq P_{EV,ch,n}^t \leq P_{EV,ch}^{max} \quad (24)$$

$$0 \leq P_{EV,dc,n}^t \leq P_{EV,dc}^{max} \quad (25)$$

$$P_{EV,ch,n}^t \cdot P_{EV,dc,n}^t = 0 \quad (26)$$

$$S_{EV,n}^{t+1} = S_{EV,n}^t + \frac{\eta_{EV,c} \cdot P_{EV,ch,n}^t - P_{EV,dc,n}^t / \eta_{EV,d}}{E_{EV,n}} \cdot \Delta t \quad (27)$$

$$S_{EV}^{min} \leq S_{EV,n}^t \leq S_{EV}^{max} \quad (28)$$

$$S_{EV,n}^{stop} \geq S_{EV,n}^{exp} \quad (29)$$

where $P_{ch,m}^{t,pre}$ represents the uncontrolled charging load of the EV fleet within EVA m . $P_{EV,ch,n}^t$ and $P_{EV,dc,n}^t$ represent the charging and discharging power of EV n at time t , respectively. $\eta_{EV,c}$ and $\eta_{EV,d}$ represent the charging and discharging efficiency, respectively. $E_{EV,n}$ is the battery capacity of EV n . $S_{EV,n}^t$ is the state of charge (SOC) of EV n at time t . S_{EV}^{max} and S_{EV}^{min} represent the upper and lower SOC limits, respectively. $S_{EV,n}^{stop}$ is the SOC at departure. $S_{EV,n}^{exp}$ is the expected SOC at departure. Specifically, Equation (21) enforces energy balance by requiring that the total controlled charging energy equals the original uncontrolled charging demand. Equations (22) and (23) aggregate the charging and discharging behaviors of individual EVs into the overall response power of the EVA. Equations (24) and (25) restrict the charging and discharging power of each EV to its rated values. Equation (26) prevents an EV from charging and discharging simultaneously. Equation (27) describes the SOC dynamics. Equation (28) keeps the SOC within the safe operating range. Equation (29) ensures that the SOC at the moment of departure meets the user's expected value.

3. User Profit Allocation Model Based on the Improved Shapley Value

Under the Stackelberg game framework, the aggregated profit obtained by an EVA arises from coordinating its internal EV resources to participate in the interaction with the distribution network. A fair and rational profit allocation mechanism is therefore indispensable for sustaining users' long-term willingness to cooperate. To this end, this section develops an improved Shapley value-based profit allocation model that jointly accounts for user participation and the contribution of each user to the loadability of the distribution network, so as to better incentivize active participation in coordinated dispatch.

3.1. Basic Shapley Value

The Shapley value is a well-established fair allocation method in cooperative game theory. It distributes the coalition payoff among participants in proportion to their marginal contributions. Suppose that N EV users participate in the coalition formed within an EVA. The basic Shapley value of user x is then given by

$$V_x = \sum_{Y \subseteq N \setminus \{x\}} \frac{(|Y|-1)! \cdot (N-|Y|)!}{N!} \cdot [v(Y) - v(Y \setminus x)] \quad (30)$$

where $v(Y)$ represents the total profit of coalition Y , $v(Y \setminus x)$ represents the profit of coalition Y excluding user x , $|Y|$ is the number of members in coalition Y , and N is the total number of users in the coalition.

3.2. S-Shaped Utility Function

The conventional Shapley value reflects only marginal economic contribution and cannot adequately capture each user's actual level of participation or the quality of execution. To overcome this limitation, an S-shaped utility function is adopted to describe user participation as well as the nonlinear way in which users perceive gains and losses [25]. The S-shaped utility function is shown in Figure 2.

$$\varphi(f_x - b) = \begin{cases} \xi \cdot (f_x - b)^\alpha, & f_x - b \geq 0 \\ -\zeta \cdot [-(f_x - b)]^\beta, & f_x - b < 0 \end{cases} \quad (31)$$

where f_x is the utility value of user x , and b is the reference point, which is generally taken as the expected profit level. α and β are risk-sensitivity coefficients associated with gains and losses, respectively. ξ and ζ are shape parameters controlling the steepness of the utility curve. When the profit exceeds the reference point, the curve is concave, indicating risk aversion; when the profit falls below the reference point, the curve becomes convex, indicating risk-seeking behavior. This is consistent with prospect theory and thus better captures users' actual decision-making characteristics. To quantify the deviation between actual execution and scheduled dispatch, the execution deviation function is defined as

$$f_n = 1 - \frac{1}{|T|} \sum_{t \in T} \frac{|P_{EV,n}^{t,actual} - P_{EV,n}^{t,plan}|}{P_{EV,n}^{max}} \quad (32)$$

where the closer f_n is to one, the better the execution consistency, and hence the more reliable the user's scheduling response. Based on an S-shaped utility function, the user participation factor is further defined as

$$R_{1,x} = \frac{\varphi(f_x - b)}{\sum_{n \in N} \varphi(f_n - b)} \quad (33)$$

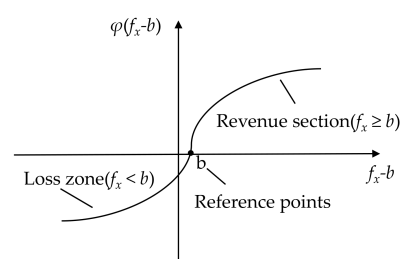


Figure 2. S-shaped utility function.

This factor represents the relative participation contribution of the user and is used as a correction term in the improved Shapley value allocation.

3.3. Definition of the Loadability Contribution Factor

Different users provide different levels of contribution to the loadability of the system, which should be reflected in their respective payoffs. To encourage users to make more regulation resources available during margin-tight periods, this paper introduces a loadability contribution factor.

1. Single-user loadability contribution.

User n 's contribution to the loadability during time period t is defined as

$$\Delta SM_n^t = \frac{P_{EV,dc,n}^t - P_{EV,ch,n}^t}{P_{cap,t}} \quad (34)$$

This formula indicates that users who provide discharge services during periods of tight grid supply will receive a positive contribution, whilst charging behavior will result in a negative impact. The user's total contribution to the grid's loadability throughout the day is the sum of the contributions for each time period:

$$\Delta SM_n = \sum_{t \in T} \omega_t \cdot \Delta SM_n^t \quad (35)$$

2. Definition of the loadability contribution factor.

The normalization of positive contributions to the loadability contribution factor takes into account the following:

$$R_{2,x} = \frac{\Delta SM_x^+}{\sum_{n \in N} \Delta SM_n^+} \quad (36)$$

A positive contribution is defined as

$$\Delta SM_x^+ = \max\{\Delta SM_x, 0\} + \varepsilon \quad (37)$$

where ε is a small positive number that ensures all users receive a basic allocation and prevents the denominator from becoming zero.

3.4. Improvements to the Shapley Value Allocation Model

The classical Shapley value implicitly assumes that all participants share an equal risk weight of $1/N$. In practice, however, users assume different levels of risk. To better reflect each user's actual risk contribution, this paper refines the classical Shapley value by introducing both the participation factor and the loadability contribution factor into the allocation.

The participation factor and the loadability contribution factor are combined into a comprehensive factor:

$$Z_x = \psi_1 \cdot \frac{R_{1,x}}{\sum_{n \in N} R_{1,n}} + \psi_2 \cdot \frac{R_{2,x}}{\sum_{n \in N} R_{2,n}} \quad (38)$$

where ψ_1 and ψ_2 are weighting coefficients that satisfy the condition $\psi_1 + \psi_2 = 1$; they can be flexibly adjusted according to actual operational requirements to balance the relative importance of user engagement and contribution to loadability in the revenue distribution. Based on the comprehensive factor, a preliminary corrected allocation is first computed as

$$\tilde{V}_x = \left(Z_x - \frac{1}{N} \right) \cdot V_{total} + V_x \quad (39)$$

where V_{total} represents the total EVA revenue, and V_x represents user x 's basic Shapley value. Since the correction term $(Z_x - 1/N)$ can be negative for users with low contribution, the preliminary value $\tilde{V}_x$ may become negative, which would conflict with the individual-

rationality requirement of a cooperative game. To keep every payoff non-negative while preserving the total revenue, the final allocation applies a non-negative projection followed by renormalization:

$$V_x^* = \frac{\max\{\tilde{V}_x, 0\}}{\sum_{n \in N} \max\{\tilde{V}_n, 0\}} V_{total} \quad (40)$$

This formula adjusts the basic Shapley value so that users with high participation levels and significant contributions to the loadability receive a larger share of the revenue allocation. The revised allocation scheme satisfies the following constraints:

$$\sum_{x \in N} V_x^* = V_{total} \quad (41)$$

$$\sum_{x \in N} Z_x = 1 \quad (42)$$

$$V_x^* \geq 0, \forall x \in N \quad (43)$$

These constraints ensure that the allocation scheme satisfies the validity and normalization conditions, and no user receives a negative payoff, so that individual rationality is preserved. By refining the Shapley value method, it is possible to allocate payoffs to participants more efficiently and equitably, thereby enhancing the stability and sustainability of cooperation.

4. Algorithm Solution

4.1. Basic Principles of the Algorithm

Given the nonconvexity and nested bi-level structure of the DNO–EVA Stackelberg game, this paper adopts a DRL approach and constructs a dual-agent alternating training framework. At the upper level, the DNO employs the SAC algorithm to learn the optimal pricing policy. Owing to its maximum entropy framework, SAC can maintain an effective balance between exploration and utilization. At the lower level, the EVA adopts the MATD3 algorithm to learn the optimal charging/discharging response policy. MATD3 inherits the advantages of TD3 in handling continuous action spaces and further addresses multi-agent coordination through a centralized training and decentralized execution architecture. Consistent with the leader–follower structure of the game, the two levels are updated on separated time scales: within each update of the upper-level DNO pricing policy, the lower-level EVA policies are trained for several iterations so that their response to the current price signal is close to optimal, and the DNO policy is then updated on the slower time scale in response to this behavior.

4.1.1. Training Objective

The objective of reinforcement learning is to identify an optimal policy π^* that maximizes the expected cumulative return of the agent in a finite Markov decision process.

$$\pi^* = \operatorname{argmax}_{\pi} \mathbb{E}_{\tau \sim \pi} [R(\tau)] \quad (44)$$

$$R(\tau) = \sum_{t=0}^T r_t \quad (45)$$

where τ represents the state–action trajectory generated under policy π , i.e., $\tau = (s_0, a_0, s_1, a_1, \dots)$, $R(\tau)$ represents the total return over an episode, and r_t is the immediate reward at time t . To enhance exploration capability and avoid premature convergence, SAC introduces entropy regularization, with the objective given by

$$J(\pi) = \sum_{t=0}^T \mathbb{E}_{\tau \sim \pi} [r(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t))] \quad (46)$$

where α is the temperature parameter that controls the relative importance of the entropy term, and H represents the entropy of the action taken under policy π at state s_t .

4.1.2. Action Selection

The action of each agent is determined by the output of its Actor network, and different algorithms adopt different action-selection mechanisms. For SAC, a stochastic policy is used, and the action is sampled from a probability distribution parameterized by the Actor network.

$$a_t \sim \pi_{\theta}(\cdot|s_t) = \tanh(\mu_{\theta}(s_t) + \sigma_{\theta}(s_t) \odot \xi), \xi \sim \mathcal{N}(0, I) \quad (47)$$

where $\mu_{\theta}(s_t)$ and $\sigma_{\theta}(s_t)$ are the mean and standard deviation output by the Actor network, $\odot$ denotes element-wise multiplication, and ξ is the reparameterization noise sampled from a standard normal distribution. This squashed Gaussian policy provides exploration without additional exploration noise. In contrast, MATD3 adopts a deterministic policy. To ensure sufficient exploration, Gaussian noise is added to the action output by the Actor network.

$$a_t^n = \text{clip}(\mu_{\theta_n}(o_t^n) + \epsilon, a_L, a_H), \epsilon \sim \mathcal{N}(0, \sigma) \quad (48)$$

where $\text{clip}(\cdot)$ limits the action to the admissible range $[a_L, a_H]$, ϵ is Gaussian noise with zero mean and standard deviation σ . This mechanism allows the agent to maintain adequate exploration while following a deterministic policy, thereby reducing the risk of becoming trapped in poor local optima.

4.2. Stackelberg-Aware Dynamic Prioritized Experience Replay

Experience replay improves the sample efficiency of DRL by storing interaction data in a replay buffer and randomly sampling it during training, thereby reducing temporal correlation among samples. However, conventional replay assigns equal sampling probability to all samples, regardless of their actual learning value. PER improves this by using the temporal-difference error as a priority indicator, such that samples with larger TD errors are more likely to be replayed [26]. Although effective, this criterion remains limited in Stackelberg game environments, where the learning value of a sample depends not only on TD error but also on the variation in the opponent's strategy. When the opponent's policy changes significantly, the relevance of historical experiences may change accordingly.

To address this issue, SD-PER is proposed. This mechanism jointly considers TD error and game strategy variation, and adaptively adjusts the weighting coefficients during training, thereby enabling more accurate evaluation of the importance of experience samples.

4.2.1. Game-Aware Dual-Priority Criterion

As the leader, the DNO must anticipate the strategy changes in the lower-level EVA, its experience priority is defined as a weighted combination of the TD error and the magnitude of the EVA's strategy variation:

$$p_i^{DNO} = \left| \delta_i^{DNO} \right| + \eta(k) \cdot \left\| \Delta \pi_i^{EVA} \right\| + \epsilon \quad (49)$$

$$\left\| \Delta \pi_i^{EVA} \right\| = \frac{1}{|B|} \sum_{s \in B} \left\| \pi_{(k)}^{EVA}(s) - \pi_{(k-1)}^{EVA}(s) \right\|_2 \quad (50)$$

where $\eta(k)$ is the game-awareness coefficient in training cycle k . The opponent strategy variation term $\left\| \Delta \pi_i^{EVA} \right\|$ is measured as the mean Euclidean distance between the op-

ponent's Actor outputs on a sampled mini-batch B over two consecutive update cycles, and the price-variation term is the change in the issued price between two consecutive update cycles, $|\Delta\lambda_j| = |\lambda_j(k) - \lambda_j(k-1)|$, so that both game-related terms are evaluated over the same one-cycle window. Correspondingly, the follower EVA must adapt to the upper-level price variation, so its priority combines the TD error with the magnitude of the price change:

$$p_j^{EVA} = \frac{1}{N} \sum_{n=1}^N |\delta_{j,n}^{EVA}| + \mu(k) \cdot |\Delta\lambda_j| + \varepsilon \quad (51)$$

where $|\delta|$ represents the absolute value of the TD error, reflecting the discrepancy between the predicted value from the experience sample and the actual value, and a larger TD error indicates that the sample contains more unknown information and thus has higher learning value. $|\Delta\lambda_j|$ represents the magnitude of the electricity price change. $\mu(k)$ is the price-sensitivity coefficient. ε is a small positive constant introduced to ensure a nonzero sampling probability for every sample.

4.2.2. Dynamic Weighting with Adaptive Adjustment

The relative importance of the TD error and the game-related terms changes during training. TD errors are typically large and volatile in the early stage, whereas strategy variation becomes more informative during the intermediate and later stages. To reflect this characteristic, the game-awareness coefficient $\eta(k)$ and the price-sensitivity coefficient $\mu(k)$ are increased from a small initial value toward their respective maxima as training proceeds:

$$\eta(k) = \eta_0 \cdot (1 - e^{-k/\tau_\eta}) \quad (52)$$

$$\mu(k) = \mu_0 \cdot (1 - e^{-k/\tau_\mu}) \quad (53)$$

where η_0 and μ_0 represent the maximum values of the two coefficients, τ_η and τ_μ are time constants controlling how fast they grow, and k is the current training cycle. In the early stage, two coefficients are close to zero, such that the agents mainly learn from TD-error information. As training proceeds, the opponent strategy variation and the price variation gradually gain weight, encouraging each agent to pay more attention to changes in the counterpart's strategy and thus to better adapt to the dynamic game process.

4.2.3. Sampling Probability and Importance Sampling Weights

Based on the aforementioned dual-priority criteria, sampling according to priority may lead to high-priority samples being oversampled while low-priority samples are neglected. To ensure that all experiences have the opportunity to be learned, this paper adopts a stochastic prioritized sampling strategy. The sampling probability of experience sample i is defined as

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \quad (54)$$

where p_i^α represents the priority of experience sample i , and $\alpha \in [0, 1]$ is the priority used to adjust the degree of priority influence. The priority-based sampling method alters the probability distribution of the experiences, leading to bias in the estimation of the action-value function $Q(s,a)$. To correct this bias, importance sampling weights are employed.

$$w_i = \frac{(N \cdot P(i))^{-\beta}}{\max_j(w_j)} \quad (55)$$

4.3. Network Parameter Update

The Actor–Critic framework comprises two core components: the Actor network and the Critic network. The Actor network is responsible for policy output, generating action

decisions based on the current state; the Critic network is tasked with value evaluation, assessing the quality of the actions selected by the Actor network. The two networks are updated alternately to achieve joint optimization.

4.3.1. SAC Network Update

The training objective of the Critic network is to minimize the mean squared error between the estimated value and the target value. To avoid the overestimation bias of values, SAC adopts a twin-Critic network architecture, utilizing the minimum of the outputs from the two Critic networks as the target Q -value. The update gradient for the Critic network parameters ϕ is

$$\nabla_{\phi_j} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} w_i \cdot (Q_{\phi_j}(s,a) - y(r,s',d))^2, j = 1, 2 \quad (56)$$

where the objective value $y(r,s',d)$ is calculated in an entropy-regularized form to maintain consistency with the SAC objective function.

$$y(r,s',d) = r + \gamma(1-d)(\min_{j=1,2} Q_{\phi_{\text{targ},j}}(s',\tilde{a}') - \alpha \log \pi_{\theta}(\tilde{a}'|s')) \quad (57)$$

The training objective of the Actor network is to maximize the weighted sum of the state value function and the policy entropy. The gradient for updating its parameter θ is

$$\nabla_{\theta} \frac{1}{|B|} \sum_{s \in B} (\min_{j=1,2} Q_{\phi_j}(s,\tilde{a}_{\theta}(s)) - \alpha \log \pi_{\theta}(\tilde{a}_{\theta}(s)|s)) \quad (58)$$

4.3.2. MATD3 Network Update

The EVA employs the MATD3 algorithm to address the multi-agent coordination problem, utilizing a centralized training with decentralized execution (CTDE) architecture. During the training phase, the Critic network can access global information to learn a more accurate value function; during the execution phase, each Actor network makes independent decisions based solely on local observations, which satisfies the communication constraints in practical applications. MATD3 adopts a shared centralized Critic network whose input encompasses the global state and the joint actions of all agents, thereby enabling a more accurate evaluation of the coordination effectiveness within the multi-agent system. The update gradient for the Critic network parameters is given by

$$\nabla_{\phi_j} \frac{1}{|B|} \sum_{(s,\mathbf{a},r,s',d) \in B} w_i \cdot (Q_{\phi_j}(s,\mathbf{a}) - y(r,s',d))^2, j = 1, 2 \quad (59)$$

where $\mathbf{a} = [a_1, a_2, \dots, a_N]$ is the joint action of all N user agents. The introduction of the joint action space enables the Critic network to capture collaborative or competitive relationships among the agents. The target value calculation of MATD3 inherits the design philosophy of the TD3 algorithm, utilizing a twin-Critic network architecture and adding truncated noise to the target actions to smooth the target Q -value surface.

$$y^{\text{EVA}}(r,s',d) = r + \gamma(1-d) \min_{j=1,2} Q_{\phi_{\text{targ},j}}(s',\tilde{\mathbf{a}}') \quad (60)$$

The target joint action $\mathbf{a}'$ is generated by the target Actor network of the respective agents, with truncated Gaussian noise added to achieve target policy smoothing.

$$\tilde{\mathbf{a}}^n = \text{clip}(\mu_{\theta_{\text{targ},n}}(o^n) + \tilde{\epsilon}, a_L, a_H) \quad (61)$$

$$\tilde{\epsilon} \sim \text{clip}(\mathcal{N}(0,\tilde{\sigma}), -c, c) \quad (62)$$

Under the MATD3 framework, the Actor network of each user agent n is updated independently, but the gradient calculation must be performed through the centralized

Critic network, embodying the core concept of CTDE. The update gradient for the parameter θ^n of the n th agent's Actor network is

$$\nabla_{\theta_n} \frac{1}{|B|} \sum_{s \in B} Q_{\phi_1}(s, a^1, \dots, \mu_{\theta_n}(o^n), \dots, a^N) \quad (63)$$

4.3.3. Soft Update of the Target Network

In DRL, directly using the current network to calculate target values leads to training instability because the target values change continuously as the network parameters are updated. To improve training stability, this paper adopts a target network soft update mechanism, wherein the target network parameters slowly approach the current network parameters with a small step size. The soft update rule for the DNO target network is

$$\theta_{targ} \leftarrow \tau \theta + (1 - \tau) \theta_{targ} \quad (64)$$

$$\phi_{targ} \leftarrow \tau \phi + (1 - \tau) \phi_{targ} \quad (65)$$

The soft update rule for the EVA target network is

$$\theta_{targ,n} \leftarrow \tau \theta_n + (1 - \tau) \theta_{targ,n}, n \in \{1, 2, \dots, N\} \quad (66)$$

$$\phi_{targ}^{EVA} \leftarrow \tau \phi^{EVA} + (1 - \tau) \phi_{targ}^{EVA} \quad (67)$$

where $\tau \in (0, 1)$ is the soft update coefficient. Soft updates prevent drastic changes in the target networks, thereby enhancing training stability.

As illustrated in Figure 3, the proposed solution framework is centered on the leader–follower game. The upper-level DNO applies SAC to generate transaction pricing strategies based on system state, while the lower-level EVA applies MATD3 to form charge/discharge response decisions guided by price signals. After executing the joint actions, the environment performs power flow computation, reward feedback, and state transition, storing experiences in the respective replay buffers. The SD-PER mechanism then evaluates and samples experiences by priority, followed by network parameter updates and target network soft updates. Through this iterative alternating process, the algorithm converges and outputs optimal transaction prices and scheduling results.

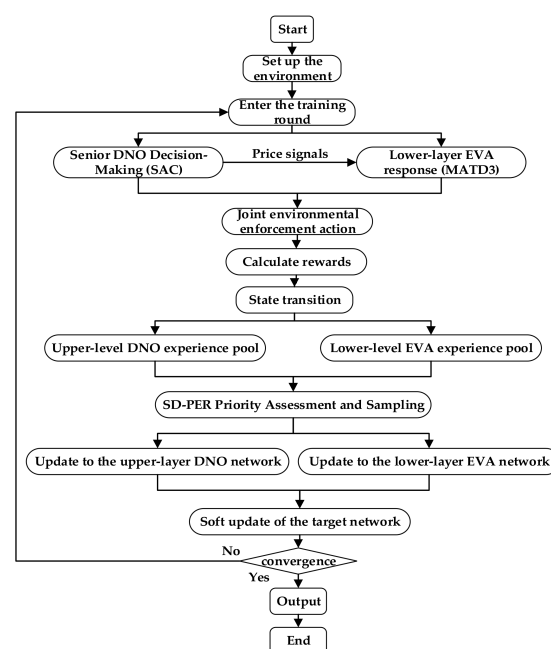


Figure 3. Flowchart of the DNO-EVA two-layer DRL algorithm incorporating the SD-PER mechanism.

5. Case Study

5.1. Case Study Setup and Assumptions

The IEEE 33-bus distribution system is adopted to verify the effectiveness of the proposed Stackelberg game-based coordinated optimization model and its solution method. The system has a rated voltage of 12.66 kV and a total active load of 3.715 MW. One DNO and two EVAs are configured in the test system, with each aggregator managing 100 electric vehicles. EVA1 is connected at bus 18 and EVA2 at bus 33. The arrival time, departure time and initial state of charge of the EVs are characterized by truncated Gaussian distributions fitted to data from the U.S. National Household Travel Survey [27], and the travel-charging profile of each vehicle is generated by Monte Carlo sampling. Acting as the leader, the DNO designs time-of-use transaction prices to guide the charging and discharging decisions of the aggregators; as followers, the EVAs adjust the response of their internal users according to the announced prices, and the profit obtained by each aggregator is then shared among its users through the proposed improved Shapley value mechanism. The scheduling horizon is one day, discretized into 24 periods with a time step of one hour. The relevant EV and aggregator parameters are listed in Table 2, and the time-of-use prices are listed in Table 3, set with reference to the time-of-use price policy for EV charging and swapping facilities in Shanghai and the actual billing ranges of public charging stations.

Table 2. Distribution network, EV and aggregator parameters.

Parameter	Value	Parameter	Value
Number of buses/branches	33/32	Nominal voltage	12.66 kV
Total active/reactive load	3.715 MW/2.30 MVar	Voltage security limits	[0.95, 1.05] p.u.
Battery capacity	50 kWh	Charging/discharging efficiency	0.95/0.95
Max charging/discharging power	7 kW/7 kW	SOC limits	[0.2, 0.9]
Charging/discharging degradation coeff.	0.05/0.05 RMB/kWh	Expected departure SOC	0.90
Arrival time (mean $\pm$ std)	18:00 $\pm$ 1.5 h	Departure time (mean $\pm$ std)	07:30 $\pm$ 1.5 h

Table 3. Time-of-use prices.

Period	Time	Price (RMB/kWh)
Valley	23:00–07:00	0.30
Flat	07:00–08:00, 12:00–17:00, 21:00–23:00	0.68
Peak	08:00–12:00, 17:00–21:00	1.18

All cases are run on a workstation equipped with an Intel Core i7-12700K CPU, 32 GB of RAM and an NVIDIA RTX 3080 GPU, and the algorithms are implemented in Python 3.9 and PyTorch 1.12. The upper-level DNO adopts the SAC algorithm and the lower-level EVA adopts the MATD3 algorithm, with hyperparameters set with reference to [28]; both the Actor and Critic networks are fully connected networks with two hidden layers of 256 neurons each, the batch size is 256, the replay buffer capacity is 1×10^6 , the discount factor γ is 0.99, the target network soft update coefficient τ is 0.005, and the number of training episodes is 3000. The SAC learning rate is 3×10^{-4} with automatically tuned entropy temperature, while the MATD3 learning rate is 1×10^{-3} with exploration noise standard deviation 0.1, target policy smoothing noise clip 0.5, and policy delay of two steps. For the SD-PER mechanism, the maximum game-awareness coefficients are 0.5, the time constants are 1000, the priority exponent is 0.6, and the importance sampling exponent is annealed linearly from 0.4 to 1.0. In the objective function, the loadability weight λ_1 is 50 RMB per margin unit, the economic weight λ_2 is 1.0, and the peak-period loadability time weight is two while the weight of the remaining periods is one; for the improved

Shapley value, the participation and loadability contribution weights are both 0.5, and the allocation constant is 1×10^{-6} .

5.2. Results and Discussion

5.2.1. Convergence Analysis

The cumulative reward trajectories of the DNO and the EVAs throughout the dual-agent alternating training process are plotted in Figure 4. A clear pattern can be observed: the reward signals of both the upper-level and lower-level agents exhibit pronounced fluctuations at the beginning of training. This is because the agents have not yet established a stable understanding of the distribution network environment or the relevant state–action space in the initial phase of the Stackelberg interaction. At the upper level, the DNO, driven by the SAC, relies on the maximum entropy framework to explore pricing alternatives and search for a satisfactory trade-off between loadability improvement and economic return. At the lower level, the EVA, governed by MATD3, repeatedly modifies its charging and discharging actions under exploration noise to respond to the varying price signals. The reciprocal adaptation between the two levels gives rise to the noticeable reward oscillations seen during the early episodes.

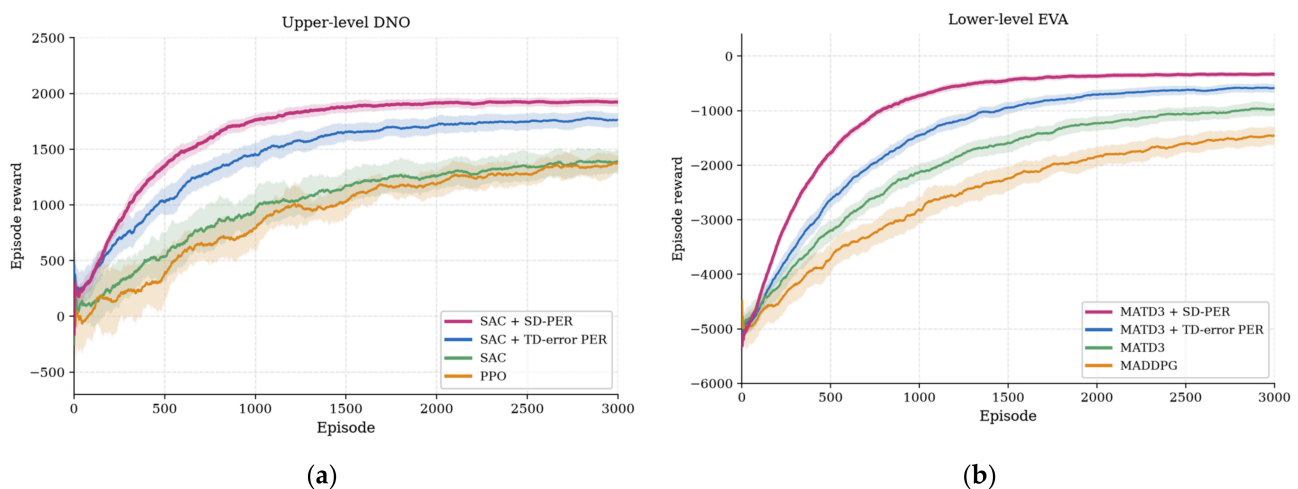


Figure 4. Training convergence curves of the compared schemes (a) upper-level DNO; (b) lower-level EVAs.

This subsection analyzes the contribution of each component of the proposed method to the convergence performance through ablation experiments. To quantitatively evaluate convergence, modules are introduced incrementally from a base algorithm at both levels. At the upper level, the base SAC, SAC with TD-error PER, and the proposed SAC with SD-PER are compared, with PPO as a reference. At the lower level, the base MATD3, MATD3 with TD-error PER, and the proposed MATD3 with SD-PER are compared, with MADDPG as a reference. The base algorithms use uniform-sampling replay to assess the gain brought by prioritized experience replay itself, while the TD-error-PER variants add TD-error-based prioritized replay. All compared schemes use identical state and action definitions, network structures and training budgets, and are trained independently under five random seeds; accordingly, in Figure 4 each solid line is the mean episode reward over the five seeds and the shaded band spans ± 1 standard deviation across seeds, and the convergence episode and final reward in Tables 4 and 5 are reported as the mean and standard deviation over the same five seeds.

As shown in Figure 4a and summarized in Table 4, at the upper level the base SAC, relying only on uniform sampling, converges most slowly and stabilizes after about

2600 episodes with relatively large late-stage fluctuations. The PPO reference converges smoothly but, owing to its on-policy updates and lower sample efficiency, attains a final reward close to that of the base SAC and clearly below the schemes with prioritized replay. After introducing TD-error PER, samples with larger errors are replayed preferentially and convergence accelerates to about 2000 episodes with an improved final reward. After further introducing SD-PER, the replay priority additionally accounts for the variation in the opponent's strategy on top of the TD error, and the scheme stabilizes after about 1440 episodes with the highest final reward and the smallest inter-seed standard deviation, improving by about 8.0% over the TD-error-PER scheme and by about 40.0% over the base SAC.

Table 4. Convergence performance comparison for upper-level DNO (total 3000 episodes).

Algorithm & Replay Strategy	Convergence Episode	Final Reward
SAC + SD-PER	1440 ± 55	1960 ± 38
SAC + TD-Error PER	2000 ± 86	1740 ± 64
SAC + Uniform Replay	2600 ± 140	1480 ± 103
PPO	>2600	1610 ± 96

Table 5. Convergence performance comparison for lower-level EVA (total 3000 episodes).

Algorithm & Replay Strategy	Convergence Episode	Final Reward
MATD3 + SD-PER	1560 ± 60	−300 ± 45
MATD3 + TD-Error PER	2210 ± 95	−560 ± 78
MATD3 + Uniform Replay	2700 ± 150	−950 ± 120
MADDPG	>2700	−1380 ± 165

As shown in Figure 4b and Table 5, the lower-level EVA exhibits a consistent pattern: the MADDPG reference, using a single Q-network and suffering from value overestimation, converges most slowly with the lowest final reward and the largest fluctuation. The base MATD3, benefiting from twin Q-networks and delayed updates, outperforms MADDPG and converges after about 2700 episodes. Introducing TD-error PER accelerates convergence to about 2210 episodes, and the proposed MATD3 with SD-PER stabilizes after about 1560 episodes with a final reward markedly superior to the other schemes and the smallest inter-seed standard deviation among all compared methods. Across both levels, the proposed SD-PER schemes attain not only the highest final rewards but also the smallest standard deviations across seeds, whereas the uniform-replay and single-network baselines show markedly larger inter-seed variability, indicating that the improvement is consistent across seeds rather than the result of a single favorable run. These results indicate that prioritized experience replay and its game-aware extension are both necessary for achieving good sample efficiency and convergence stability in the non-stationary Stackelberg environment.

5.2.2. Comparison with a Benchmark Solution Method

To verify the quality of the scheduling scheme obtained by the proposed deep reinforcement learning method, it is compared with a bi-level optimization benchmark. This benchmark equivalently replaces the lower-level follower problem by its KKT optimality conditions, transforming the model into a mathematical program with equilibrium constraints (MPEC) solved by a commercial solver, which yields a near-optimal reference solution of the leader–follower equilibrium under an exact analytical model. The comparison results are listed in Table 6. On the two key indicators of peak-period loadability and user cost, the proposed method attains 1.240 and 3299 RMB/day, with relative deviations within 1% from the 1.246 and 3258 RMB/day of the KKT/MPEC benchmark,

indicating that although the proposed method does not rely on an exact analytical model, the scheduling policy it learns is already very close to the near-optimal solution of bi-level optimization in terms of solution quality. Moreover, unlike KKT/MPEC, whose solution scale expands rapidly as the system grows, the proposed method requires no re-solving of an optimization problem during execution and can generate scheduling schemes directly by policy-network inference, making it more suitable for the non-convex full nonlinear power flow environment and for online application.

Table 6. Comparison of solution quality between the proposed method and the KKT/MPEC benchmark.

Method	Peak-Period Loadability	User Cost (RMB/Day)
KKT/MPEC (near-optimal bi-level)	1.246	3258
Proposed method	1.240	3299

5.2.3. Time-Varying Transaction Electricity Prices

The 24 h charging and discharging prices that the DNO sets for EVA1 and EVA2 are displayed in Figure 5. As shown, charging prices are kept noticeably lower during off-peak hours and become considerably higher during peak periods. This pricing structure restrains uncoordinated charging when the system is heavily loaded and motivates EVs to recharge during light-load intervals. At the same time, the discharging prices in peak periods are placed above the corresponding charging prices, which reinforces the economic incentive for users and aggregators to feed power back to the grid when the system is under stress. It can also be observed that the peak–valley spread of EVA1 is wider than that of EVA2, implying that the DNO sends a stronger incentive signal to EVA1, while the pricing curve for EVA2 is comparatively flatter and reflects a more moderate regulation strategy. This behavior shows that, within the bi-level game framework, the DNO is able to tailor its pricing decisions to the response capability, resource characteristics, and marginal operational impact of each individual aggregator rather than imposing a uniform price across all participants. Overall, the optimized transaction prices deliver a more precise and targeted economic signal for unlocking EV flexibility.

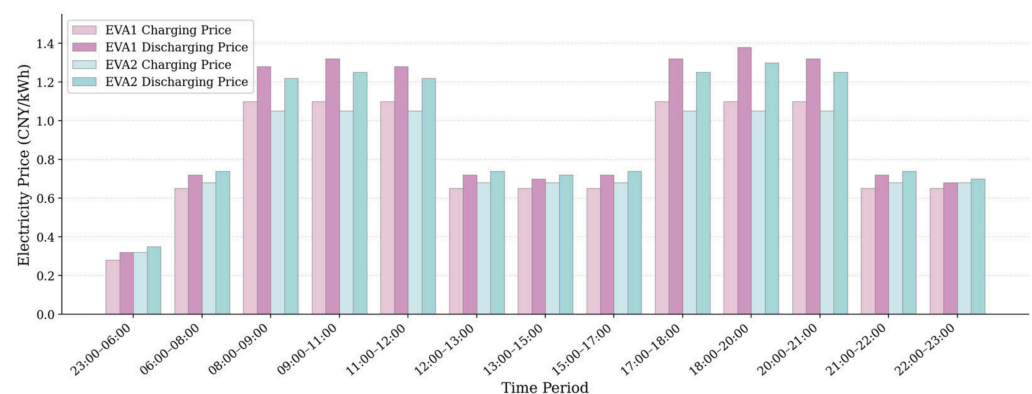


Figure 5. Time-varying transaction electricity prices for EVA1 and EVA2.

5.2.4. Load Profiles and Loadability Enhancement

The 24 h load profiles of the two EVA regions are compared in Figure 6. Guided by the time-of-use prices and the coordinated response of the aggregators, the peak load of both regions is reduced and the load curves become smoother, and part of the charging demand is shifted to lower-load periods. The evening-peak load of EVA1 decreases from 1858 kW to 1703 kW, a reduction of 8.3%, and that of EVA2 decreases from 1645 kW to 1557 kW, a reduction of 5.3%; meanwhile, the valley load rises correspondingly, and the peak–valley differences in the two regions narrow by 14.7% and 11.2%, respectively.

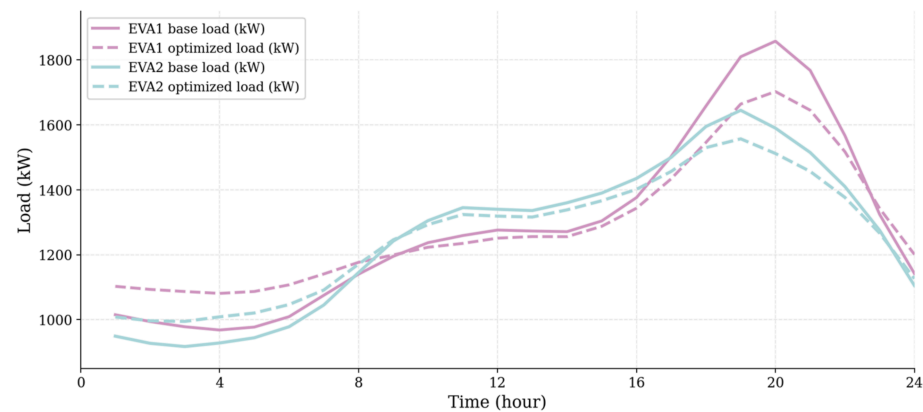


Figure 6. 24-h load profiles of the two aggregator zones before and after optimization.

A comparison of the 24 h loadability of the distribution network before and after V2G interaction is further presented in Figure 7. Before optimization, uncoordinated EV charging coincided with the morning and evening peaks, and the average peak-period loadability was held at only 1.05. After the proposed optimization mechanism is introduced, aggregators provide coordinated discharging support during peak periods, lifting the average peak-period loadability to 1.24. At the same time, a sizeable share of the charging demand is shifted to off-peak hours, which slightly increases the valley-period loading and causes a modest decline in the average valley-period margin from 1.42 to 1.37. On the whole, by directing EVs to charge in low-load intervals and discharge in peak intervals, the proposed V2G framework strengthens the operational security of the distribution network and keeps the loadability above the safety threshold across the full scheduling horizon.

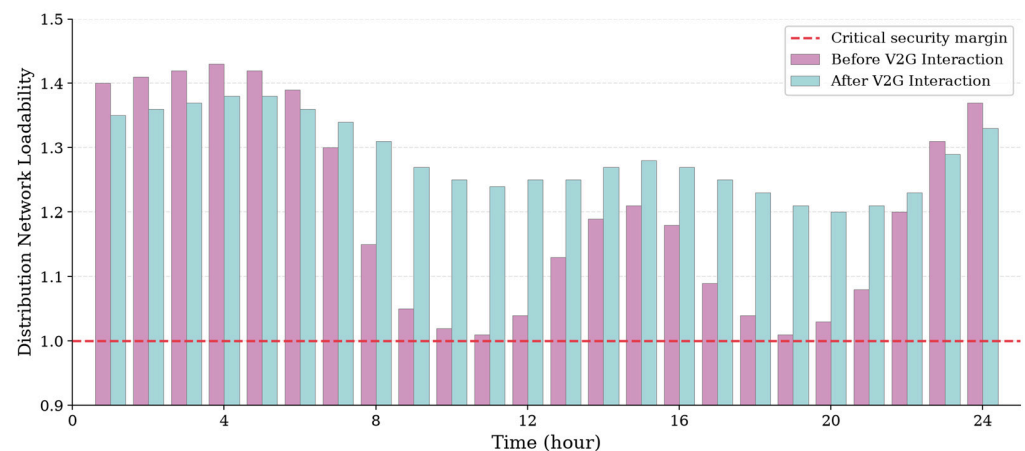


Figure 7. Comparison of 24 h distribution network loadability before and after vehicle-grid interaction.

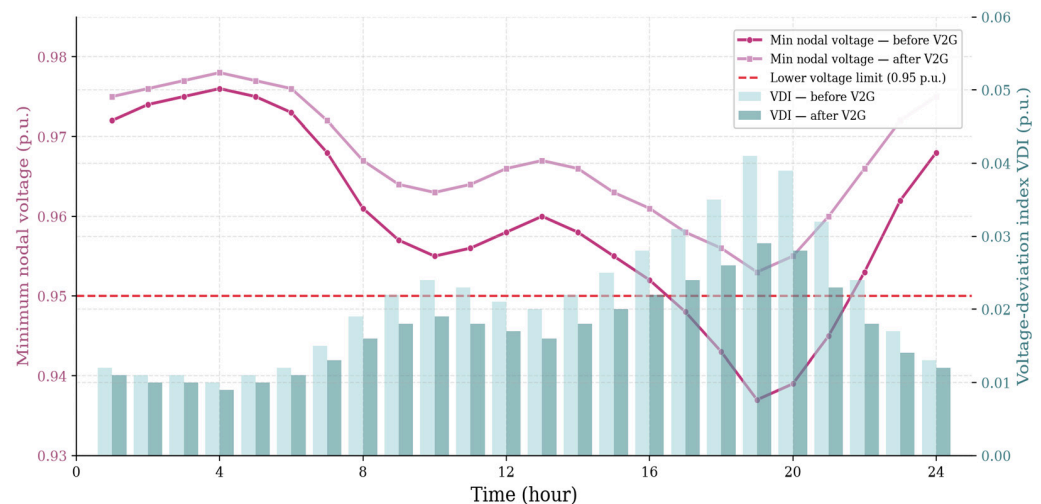
A statistical comparison of the loadability before and after coordination is given in Table 7. During the morning and evening peaks, the average loadability increases from 1.05 to 1.24, corresponding to a relative improvement of 18.1%, while the daily minimum margin rises from 1.01 to 1.20. In practical terms, at its most stressed hour the feeder can accommodate only about 1% of additional load before coordination, whereas after coordination this figure rises to about 20%, so that the usable security margin at the binding hour is enlarged roughly twentyfold. The valley-period margin decreases moderately from 1.42 to 1.37 as part of the charging demand is shifted into the low-load hours, and the difference between the valley and peak margins narrows by 65.2%, yielding a substantially flatter and more balanced loadability profile over the day.

Table 7. Statistical comparison of distribution network loadability before and after V2G interaction.

Indicator	Before V2G	After V2G	Relative Change
Peak-period average	1.05	1.24	+18.1%
Flat-period average	1.24	1.29	+4.0%
Valley-period average	1.42	1.37	−3.5%
Daily minimum margin	1.01	1.20	+18.8%
Daily average	1.21	1.29	+6.6%
Valley-to-peak margin difference	0.37	0.13	−65.2%

5.2.5. Voltage Profile and Operational Indicators

To evaluate the improvement of the proposed method in voltage security and voltage quality, Figure 8 presents the hourly variation in the minimum nodal voltage of the feeder and the system voltage-deviation index; before V2G coordination the minimum nodal voltage of the feeder drops to 0.937 p.u. around the evening peak, below the 0.95 p.u. safety limit, and the system voltage-deviation index rises markedly during the heavily loaded evening hours, at which time a total of nine buses at the feeder end exceed the voltage limit. After the proposed coordination mechanism is introduced, the active discharging of the aggregators during the peak periods relieves the branch loading, so that the minimum nodal voltage is maintained above 0.95 p.u. in all 24 periods of the day and the evening-peak minimum recovers to 0.953 p.u., the voltage-limit violations are completely eliminated, and the daily average voltage-deviation index decreases from 0.0216 p.u. to 0.0172 p.u., a reduction of about 20.5%, with a flatter voltage profile along the feeder. Together with the loadability enhancement over the 24 h, this shows that while raising the average peak-period loadability from 1.05 to 1.24, the proposed method simultaneously strengthens the static voltage security and the voltage quality of the distribution network.

**Figure 8.** Comparison of the 24 h minimum nodal voltage and the system voltage-deviation index before and after V2G interaction.

5.2.6. Economic Results

Under the optimized transaction prices, the daily economic outcomes of the DNO and the two EVAs under the proposed improved Shapley value scheme are summarized in Tables 8 and 9. These results suggest that the proposed leader–follower game model achieves system-level optimization while preserving a reasonable degree of economic fairness across all participants.

Table 8. DNO daily economic performance metrics.

Stakeholders	Electricity Sales Revenue (RMB/Day)	Electricity Procurement Cost from Aggregators (RMB/Day)	Net Profit (RMB/Day)
DNO	3198.33	720.05	2478.29

Table 9. Daily operational economics of the aggregator.

Aggregator	Discharging Revenue (RMB/Day)	Charging Cost (RMB/Day)	Battery Depreciation Cost (RMB/Day)	Net User Economics (RMB/Day)
EVA 1	399	1570.6	107	−1278.6
EVA 2	321.05	1627.73	100.9	−1407.58

On the DNO side, revenue mainly originates from electricity sales to the EVAs, whereas costs are largely attributable to discharging compensation paid back to the aggregators. The DNO retains a positive net return, which confirms that the designed transaction mechanism does not undermine its economic viability while introducing V2G incentives; instead, it improves grid-side earnings together with system supply adequacy.

On the EVA side, the revenue chiefly comes from peak-period discharging compensation, while the costs consist of charging expenditures together with battery degradation costs. Combining these terms, the net user economics, that is the discharging revenue minus the charging and degradation costs, are −1278.60 RMB/day for EVA1 and −1407.58 RMB/day for EVA2. Both aggregators thus remain net electricity consumers over the day, but the peak-period discharging revenue offsets a considerable part of their charging and degradation costs and lowers their overall electricity expenditure relative to uncoordinated charging. These observations confirm that the proposed framework can generate differentiated operational outcomes that match the response characteristics of each aggregator, thereby reinforcing the overall collaborative optimization capability of the system.

5.2.7. Improved Shapley Value Revenue Allocation Results

The profit allocation outcomes for the 100 EV users within each aggregator are illustrated in Figure 9. The horizontal axis records the user participation level, the vertical axis records the discharging contribution, and the profit value indicates the share that each user ultimately receives. A clearly differentiated pattern emerges among individual users: only a small fraction of users obtain relatively high profits, whereas the majority of users fall within the low-to-medium profit range. This pattern indicates that the allocation results are capable of effectively distinguishing the differences in contribution among users during coordinated dispatch.

As both the participation level and the discharging contribution grow, the corresponding profit share also tends to grow. That is, users positioned in the region of high participation and high contribution end up with greater returns, while those with low participation and limited support obtain comparatively modest returns. Although the overall profit levels of EVA1 and EVA2 are not identical, the allocation patterns of the two aggregators share the same fundamental trend: users who actively participate in coordinated scheduling and provide stronger support to the distribution network ultimately receive a larger share of the total profit. These findings demonstrate that the proposed improved Shapley value method is able to faithfully reflect the joint impact of user participation and loadability contribution and to keep the profit allocation highly consistent with each user's actual contribution.

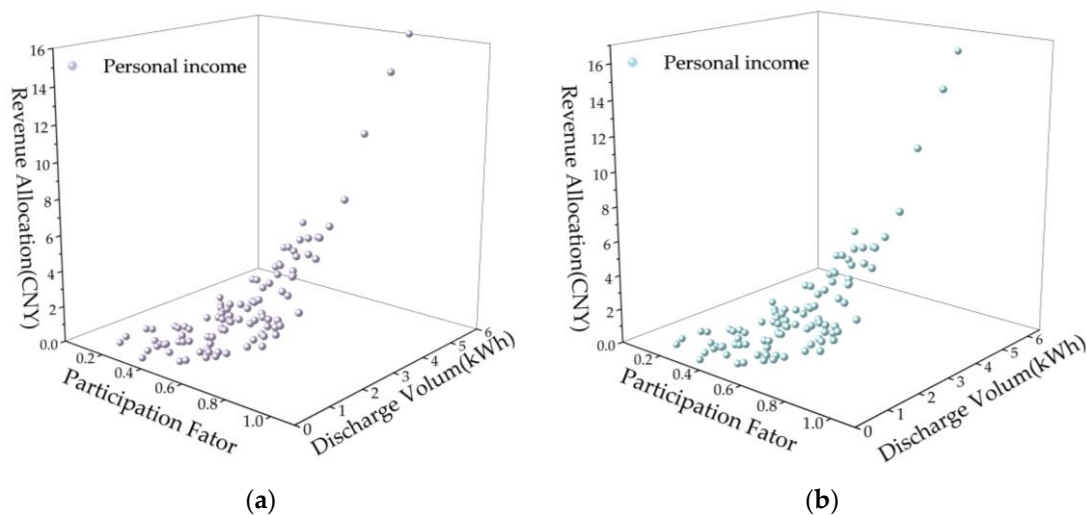


Figure 9. Improving the distribution of user benefits under the Shapley value. (a) EVA1 user revenue distribution results; (b) EVA2 user revenue distribution results.

5.2.8. Impact of Revenue Allocation Mechanism on System Operational Metrics

To further validate the effectiveness of the proposed improved Shapley value-based profit allocation mechanism, three schemes are compared with the bi-level game fixed: no revenue allocation, classical Shapley value allocation, and the proposed improved Shapley value allocation. The comparative results are reported in Tables 10 and 11.

Table 10. Comparison of DNO economic performance and supply adequacy indicators under different revenue-sharing mechanisms.

DNO Operational Metrics	Without Benefit Allocation	Classical Shapley	Improved Shapley Value-Based Benefit Allocation
Electricity sales revenue (RMB/day)	3454.2	3351.85	3198.33
Electricity procurement cost from aggregators (RMB/day)	432.03	547.24	720.05
Net profit (RMB/day)	3022.17	2804.61	2478.29
Average peak-period loadability	1.16	1.19	1.24

From the standpoint of distribution network operation, after adopting an allocation scheme with stronger incentives, more users participate in peak-period V2G discharging, so the procurement cost of the DNO rises and its sales revenue decreases somewhat, but the average peak-period loadability increases successively from 1.16 without allocation to 1.19 under classical Shapley value and 1.24 under improved Shapley value. This cost increment corresponds to an overall improvement in dispatch quality: a more robust V2G discharging response narrows the peak–valley difference and raises the loadability of the distribution network, thereby strengthening the security margin of the grid under heavy load.

From the standpoint of EVA cluster returns, the discharging revenue and discharging volume of both aggregators increase as the incentive strengthens, while the charging cost and battery degradation cost decrease as charging is shifted to the low-price valley periods. Specifically, under the improved Shapley value scheme the discharging volume of EVA1 increases from 190 to 310 kWh/day and its discharging revenue from 239.40 to 399.00 RMB/day, although its charging volume increases from 2464 to 2584 kWh/day and its average charging price decreases, so the total charging cost only rises slightly from 1696.25 to 1735.80 RMB/day, consistent with the higher charging

volume. EVA2 exhibits the same trend. This shows that the improved Shapley value allocation, by tying returns to the actual participation level and loadability contribution of the users, effectively incentivizes users to provide discharging support during the peak periods and makes the charging–discharging coordination across the aggregator cluster more orderly. The above comparison illustrates that a properly designed revenue-sharing mechanism not only secures the efficient and well-coordinated operation of the EVA cluster but also strengthens the alignment of interests between the DNO and end-users.

Table 11. Comparison of EVA economic indicators under different revenue-sharing mechanisms.

Stakeholders	Allocation Method	Revenue from Electricity Sales (RMB/Day)	Discharging Volume (kWh/Day)	Charging Cost (RMB/Day)	Charging Volume (kWh/Day)	Battery Depreciation Cost (RMB/Day)
EVA 1	Without benefit allocation	239.4	190	1696.25	2464	88.4
	Classical Shapley	303.24	238	1713.18	2512	95.84
	Improved Shapley value-based benefit allocation	399	310	1570.6	2584	107
EVA 2	Without benefit allocation	192.63	150	1757.95	2150	83.2
	Classical Shapley	244.00	190	1710.04	2198	90.28
	Improved Shapley value-based benefit allocation	321.05	250	1627.73	2269	100.9

The 24 h loadability trajectories under the three allocation schemes are further contrasted in Figure 10, where the no-allocation case already includes V2G coordination and serves as the baseline with a peak-period loadability of 1.16. Relative to the no-allocation and classical Shapley value schemes, the proposed improved Shapley value scheme produces a more pronounced loadability lift during the peak periods and attains the highest average peak-period loadability. Specifically, the average peak-period loadability rises from 1.16 to 1.24. This implies that the improved allocation mechanism not only reinforces user incentives but also enhances the system’s capability of supplying support under heavy-load conditions.

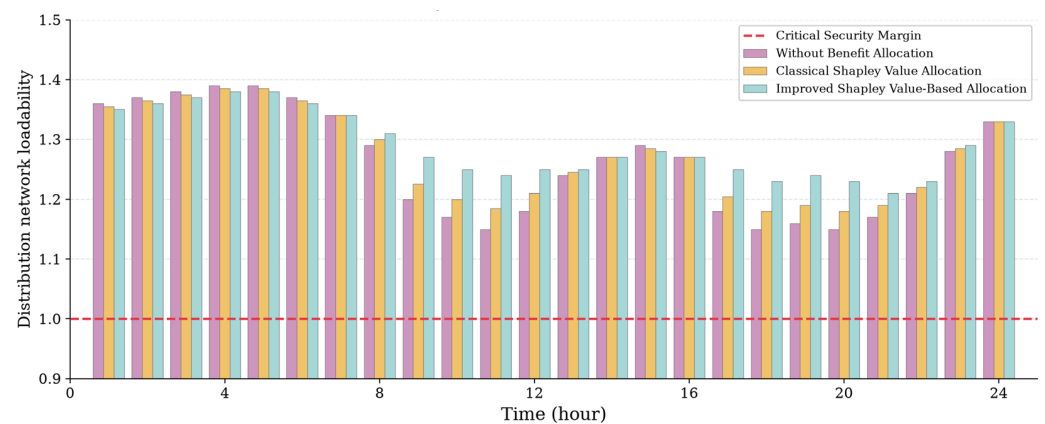


Figure 10. Comparison of 24 h distribution network loadability under the three revenue allocation schemes.

5.3. Validation on a Real Distribution Network in Shanghai and Scalability Analysis

5.3.1. Test System and Scenario

To verify the practical applicability, scalability and generalizability of the proposed framework in a real distribution network, this section tests it on a real medium-voltage distribution network in Shanghai. The network is a 10 kV radial feeder system with 72 buses and 71 branches, supplying the park loads from the substation bus through several feeders, with a total active load of about 5.6 MW. Compared with the standard test system, the real distribution network has longer feeders, more laterals and a more uneven load distribution, and therefore places higher demands on the solution capability and scalability of the coordinated scheduling method.

One DNO and three EVAs are configured in this network. Given the larger network scale, the three aggregators are connected to the electrically farthest weak end buses of three feeders: EVA1 is connected at bus 6, EVA2 at bus 39 and EVA3 at bus 59. Each aggregator still manages 100 EVs, and the EV and aggregator parameters, the time-of-use prices, and the hyperparameters of SAC, MATD3 and SD-PER are all kept identical to those of the IEEE 33-bus case above.

5.3.2. Load Profiles and Loadability Enhancement

The 24 h load profiles of the three EVA regions are compared in Figure 11. Guided by the time-of-use prices and the coordinated response of the aggregators, the peak load of all three regions is reduced and the load curves become smoother, and part of the charging demand is shifted to lower-load periods. The evening-peak load of EVA1 decreases from 1860 kW to 1731 kW, that of EVA2 from 1640 kW to 1584 kW, and that of EVA3 from 1400 kW to 1318 kW. Meanwhile, the valley load of the three regions rises correspondingly and their peak–valley differences narrow by about 23%, 17% and 22%, respectively.

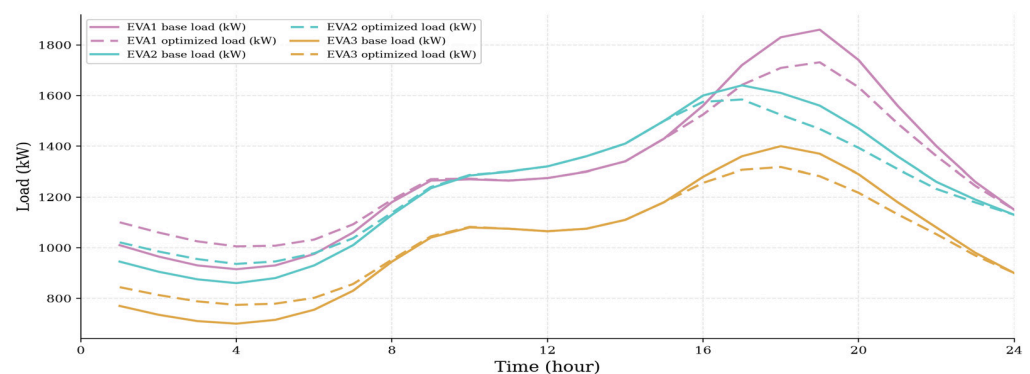


Figure 11. 24 h load profiles of the three aggregator regions of the real 72-bus network before and after optimization.

The comparison of the 24 h loadability of this network before and after V2G interaction is shown in Figure 12. Before optimization, disorderly charging coincides with the evening peak and the feeder loadability drops substantially in the evening hours, with an average peak-period loadability of only 1.05 and a minimum margin around 18:00 falling to 1.03, close to the safety threshold; after the proposed optimization mechanism is introduced, the aggregators provide coordinated discharging support during the peak periods, raising the average peak-period loadability to 1.20 and recovering the evening-peak minimum margin to 1.19, with the loadability remaining above the safety threshold throughout the day. This indicates that the proposed method can still effectively relieve the peak-period loadability stress in a larger and more complex real distribution network, verifying its practical applicability and generalizability.

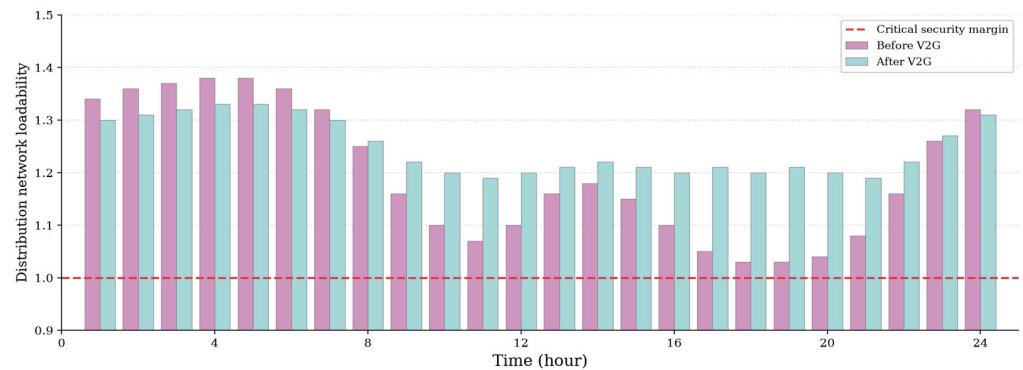


Figure 12. Comparison of 24 h distribution network loadability before and after vehicle–grid interaction.

5.3.3. Impact of Revenue Allocation Mechanism on System Operational Metrics

With the bi-level game fixed, the three schemes of no allocation, classical Shapley value allocation and the proposed improved Shapley value allocation are compared on the real distribution network, and the comparative results are reported in Tables 12 and 13, respectively. From the perspective of distribution network operation, as the allocation incentive strengthens, more users participate in peak-period V2G discharging, so the procurement cost of the DNO increases successively from 612.33 RMB/day without allocation to 790.06 RMB/day under classical Shapley value and 1060.1 RMB/day under improved Shapley value, with the sales revenue decreasing accordingly, while the average peak-period loadability increases from 1.13 through 1.16 to 1.20. This cost increment corresponds to an overall improvement in dispatch quality: a more robust V2G discharging response narrows the peak–valley difference and raises the loadability, thereby strengthening the security margin of the grid under heavy load.

Table 12. Comparison of DNO economic performance and supply adequacy indicators under different revenue-sharing mechanisms.

DNO Operational Metrics	Without Benefit Allocation	Classical Shapley	Improved Shapley Value-Based Benefit Allocation
Electricity procurement cost from aggregators (RMB/day)	612.33	790.06	1060.1
Net profit (RMB/day)	4367.84	4055.43	3552.44
Average peak-period loadability	1.13	1.16	1.2

On the EVA side, the discharging revenue and discharging volume of the three aggregators increase as the incentive strengthens, while the charging cost and battery degradation cost decrease as charging is shifted to the low-price valley periods. Under the improved Shapley value scheme, the discharging volume of EVA1 increases from 258 to 417 kWh/day and its discharging revenue from 321.27 to 525.65 RMB/day; EVA2 and EVA3 exhibit the same trend. This shows that, by tying returns to the actual participation level and loadability contribution of the users, the improved Shapley value allocation can effectively incentivize users to provide discharging support during the peak periods in a larger real network as well, and make the charging–discharging coordination across the aggregator cluster more orderly.

The 24 h loadability trajectories of this network under the three allocation schemes are compared in Figure 13. It can be seen that, relative to the no-allocation and classical Shapley value schemes, the proposed improved Shapley value scheme produces a more pronounced loadability lift during the peak periods and attains the highest average peak-

period loadability, which further verifies the effectiveness and generalizability of the proposed allocation mechanism.

Table 13. Comparison of EVA economic indicators under different revenue-sharing mechanisms.

Stakeholders	Allocation Method	Revenue from Electricity Sales (RMB/Day)	Discharging Volume (kWh/Day)	Charging Cost (RMB/Day)	Charging Volume (kWh/Day)	Battery Depreciation Cost (RMB/Day)
EVA 1	Without benefit allocation	321.27	258	1822.19	2642	112.63
	Classical Shapley	400.94	316	1818.9	2678	116.89
	Improved Shapley	525.65	417	1874.43	2772	123.05
	value-based benefit allocation					
EVA 2	Without benefit allocation	259.15	204	1830.34	2253	96.4
	Classical Shapley	324.38	260	1771.16	2308	99.54
	Improved Shapley	419.54	336	1690.95	2361	107.66
	value-based benefit allocation					
EVA 3	Without benefit allocation	221.48	176	1543.29	1957	79.86
	Classical Shapley	269.74	216	1543.67	2040	86.53
	Improved Shapley	370.14	292	1470.67	2067	93.09
	value-based benefit allocation					

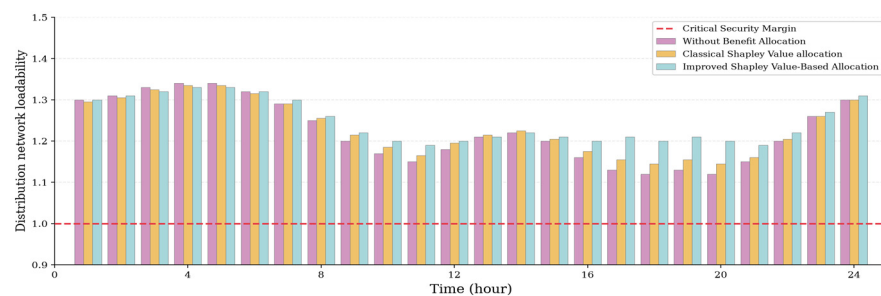


Figure 13. Comparison of the 24 h loadability of the real 72-bus network under the three revenue allocation schemes.

6. Conclusions

To address the difficulty of maintaining supply adequacy in distribution networks under deepening renewable penetration and large-scale EV integration, this paper has developed a Stackelberg game-based bi-level coordinated scheduling and revenue allocation framework for the DNO and the EVAs. By placing the loadability directly into the upper-level objective rather than treating it as a passive constraint, the framework actively steers EV charging and discharging to relieve stress during margin-tight periods.

The improved Shapley value allocation mechanism further transforms user participation and loadability contribution into concrete economic incentives, so that users who provide stronger support precisely when the system needs it are systematically better compensated. Combined with the SAC–MATD3 dual-agent alternating training framework and the proposed SD–PER mechanism—which extends prioritized experience replay with opponent strategy variation awareness—the approach coordinates grid-side security, aggregator-side economic returns, and user-side behavioral incentives within a single learning-based architecture. Case studies on the IEEE 33-bus system show that the pro-

posed framework lifts the average peak-period loadability from 1.05 to 1.24, recovers the evening-peak minimum nodal voltage from 0.937 to 0.953 p.u. while lowering the average voltage-deviation index by about 20.5%, reduces the net user cost on the EVA side, and attains a scheduling quality within 1% of the KKT/MPEC bi-level benchmark, while the SD-PER mechanism accelerates convergence to about 1440 episodes and raises the final reward by about 40% over the base SAC. On the real 72-bus network, the average peak-period loadability is correspondingly raised from 1.05 to 1.20 with the minimum margin kept above the safety threshold, confirming the effectiveness, practical applicability, and scalability of the method in complex game-theoretic environments.

It should be noted that the present framework is still established under deterministic profiles without explicitly modeling the uncertainty of renewables, load and user behavior, and relies on offline training validated by simulation, which leaves room for further improvement. Future work will therefore proceed along three directions. First, the deterministic setting will be extended with explicit uncertainty modeling of renewable generation, load, and EV mobility, for instance through scenario-based or distributionally robust formulations and risk-aware reward design, so as to enhance the robustness of the scheduling policy. Second, the framework will be generalized to competitive scenarios with multiple interacting aggregators, in which the strategic behavior and possibly conflicting interests among EVAs are captured within the game, and its scalability to larger multi-aggregator distribution systems will be examined further. Third, the practical deployment challenges will be studied, including online or transfer learning to adapt the offline-trained policy to real operating conditions and the influence of communication and measurement imperfections, with the aim of promoting the deeper integration of flexible distributed resources into intelligent distribution networks.

Author Contributions: Conceptualization, X.L.; methodology, X.L. and N.T.; validation, X.M., N.T. and J.L.; investigation, X.M.; writing—original draft, X.L.; supervision, X.M., Y.Z. and J.L.; software, Y.Z.; writing—review and editing, N.T., Y.Z. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by State Grid Jiangsu Electric Power Co., Ltd. Science and Technology Project, grant number J2025109.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Xiangchao Li, Xiaobo Mao, and Ni Tang are employed by State Grid Wuxi Power Supply Company. The remaining authors declare that this research was conducted without any commercial or financial relationships that could be considered a potential conflict of interest. The authors declare that this study received funding from the State Grid Jiangsu Electric Power Co., Ltd. Science and Technology Project. The funder was involved in the study design, data collection, the writing of this article, and the decision to submit it for publication. The funder was not involved in the simulation and data analysis.

References

1. Impram, S.; Nese, S.V.; Oral, B. Challenges of renewable energy penetration on power system flexibility: A survey. *Energy Strategy Rev.* **2020**, *31*, 100539. [\[CrossRef\]](#)
2. Li, B.; Chen, M.; Zhong, H.; Ma, Z.; Liu, D.; He, G. A review of long-term planning of new power systems with large share of renewable energy. *Proc. CSEE* **2023**, *43*, 555–581. [\[CrossRef\]](#)
3. Sun, Y.; Xing, H.; Shi, Y.; Quan, W.; Yang, X. Multi-time scale adequacy assessment and quantitative analysis of new-type electric power systems. *Smart Power* **2025**, *53*, 35–42. [\[CrossRef\]](#)
4. Nimalsiri, N.I.; Mediawathe, C.P.; Ratnam, E.L.; Shaw, M.; Smith, D.B.; Halgamuge, S.K. A survey of algorithms for distributed charging control of electric vehicles in smart grid. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4497–4515. [\[CrossRef\]](#)

5. Hussain, M.T.; Bin Sulaiman, N.; Hussain, M.S.; Jabir, M. Optimal management strategies to solve issues of grid having electric vehicles (EV): A review. *J. Energy Storage* **2021**, *33*, 102114. [\[CrossRef\]](#)
6. Zhang, M.; Yang, H.; Xu, Y.; Sun, H. Learning-based real-time aggregate flexibility provision and scheduling of electric vehicles. *IEEE Trans. Smart Grid* **2024**, *15*, 5840–5852. [\[CrossRef\]](#)
7. Wang, Q.; Huang, C.; Wang, C.; Li, K.; Xie, N. Joint optimization of bidding and pricing strategy for electric vehicle aggregator considering multi-agent interactions. *Appl. Energy* **2024**, *360*, 122810. [\[CrossRef\]](#)
8. Xu, X.; Zhan, Z.; Mi, Z.; Ji, L. An optimized decision model for electric vehicle aggregator participation in the electricity market based on the Stackelberg game. *Sustainability* **2023**, *15*, 15127. [\[CrossRef\]](#)
9. Qiao, J.; Mi, Y.; Ma, S.; Han, Y.; Wang, P. Trilayer Stackelberg game scheduling of active distribution network based on microgrid group leasing shared energy storage. *Appl. Energy* **2025**, *382*, 125157. [\[CrossRef\]](#)
10. Nie, Y.; Liu, J.; Liu, X.; Zhao, Y.; Ren, K.; Chen, C. Asynchronous multi-agent reinforcement learning-based framework for bi-level noncooperative game-theoretic demand response. *IEEE Trans. Smart Grid* **2024**, *15*, 5622–5637. [\[CrossRef\]](#)
11. Zhang, J.; Che, L.; Shahidehpour, M. Distributed training and distributed execution-based Stackelberg multi-agent reinforcement learning for EV charging scheduling. *IEEE Trans. Smart Grid* **2023**, *14*, 4976–4979. [\[CrossRef\]](#)
12. Sun, Q.; Li, Y.; Wu, Y.; Wu, K.; Zhang, M.; Wu, J. Game-theoretic and benefit equilibrium of V2G multilateral contracts: Mechanism design and coordination. *J. Energy Storage* **2026**, *159*, 121731. [\[CrossRef\]](#)
13. Wan, Y.; Wang, N.; Du, E.; Liu, X.; Wang, Y.; Chen, Z.; Kang, C. Coordinated operation of alternative fuel vehicle-integrated microgrid in a coupled power-transportation network: A Stackelberg–Nash game framework. *Appl. Energy* **2025**, *401*, 126800. [\[CrossRef\]](#)
14. Han, L.; Morstyn, T.; McCulloch, M. Incentivizing prosumer coalitions with energy management using cooperative game theory. *IEEE Trans. Power Syst.* **2019**, *34*, 303–313. [\[CrossRef\]](#)
15. Wang, Y.; Cao, R.; Luo, D.; Li, P.; Cheng, X. Benefit distribution of integrated regional energy systems under carbon trading mechanisms based on improved Shapley value methods. *Front. Energy Res.* **2023**, *11*, 1265924. [\[CrossRef\]](#)
16. Fang, F.; Yu, S.; Liu, M. An improved Shapley value-based profit allocation method for CHP-VPP. *Energy* **2020**, *213*, 118805. [\[CrossRef\]](#)
17. Glavic, M.; Fonteneau, R.; Ernst, D. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-Pap.* **2017**, *50*, 6918–6927. [\[CrossRef\]](#)
18. Li, Y.; Zhao, B.; Li, Y.; Long, C.; Li, S.; Dong, Z.; Shahidehpour, M. Safe-AutoSAC: AutoML-enhanced safe deep reinforcement learning for integrated energy system scheduling with multi-channel informer forecasting and electric vehicle demand response. *Appl. Energy* **2025**, *399*, 126468. [\[CrossRef\]](#)
19. Guo, G.; Gong, Y. Multi-microgrid energy management strategy based on multi-agent deep reinforcement learning with prioritized experience replay. *Appl. Sci.* **2023**, *13*, 2865. [\[CrossRef\]](#)
20. He, C.; Li, J. Electric Vehicle Charging and Discharging Scheduling Method Based on Clustering and Deep Reinforcement Learning. *Energies* **2026**, *19*, 2238. [\[CrossRef\]](#)
21. Yang, H.; Xu, Y.; Sun, H.; Guo, Q.; Liu, Q. Electric vehicles management in distribution network: A data-efficient bi-level safe deep reinforcement learning method. *IEEE Trans. Power Syst.* **2024**, *40*, 256–271. [\[CrossRef\]](#)
22. Wu, M.; Wen, Y.; Duan, Y.; Zhao, J.; Jin, Y.; Li, W.; Cai, Y. A collaborative optimal scheduling strategy for multiple virtual power plants based on multi-agent deep reinforcement learning. *Sustainability* **2026**, *18*, 5861. [\[CrossRef\]](#)
23. Zhao, Y.; Li, X.; Zheng, Y.; Guo, W. Federated Deep Reinforcement Learning for Energy Scheduling in Privacy-Sensitive PV-EV Charging Networks. *Electronics* **2026**, *15*, 1012. [\[CrossRef\]](#)
24. Ma, M.; Dong, P.; Dong, J.; Liu, M.; Ke, S.; Wang, C. A refined peak-shaving strategy for electric vehicles participating in urban grid zones based on master–slave games. *Power Syst. Technol.* **2025**, *49*, 2779–2788. [\[CrossRef\]](#)
25. Gao, J.; Gao, F.; Yang, Y.; Wu, H.; Zhang, Y.; Liang, P. Configuration optimization and benefit allocation model of multi-park integrated energy systems considering electric vehicle charging station to assist services of shared energy storage power station. *J. Clean. Prod.* **2022**, *336*, 130381. [\[CrossRef\]](#)
26. Wang, J.; Guo, C.; Yu, C.; Liang, Y. Virtual power plant containing electric vehicles scheduling strategies based on deep reinforcement learning. *Electr. Power Syst. Res.* **2022**, *205*, 107714. [\[CrossRef\]](#)
27. U.S. Department of Transportation; Federal Highway Administration. 2017 National Household Travel Survey (NHTS). Available online: <https://nhts.ornl.gov> (accessed on 15 April 2026).
28. Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; Dormann, N. Stable-Baselines3: Reliable reinforcement learning implementations. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.

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.