DuMES: deep reinforcement learning based EV charging scheduling with dual-layer safety modules

Zhang, A., Liu, C., Makantasis, K., Chen, X. ORCID: https://orcid.org/0000-0001-9267-355X, Ward, T. and Cheng, L. (2025) DuMES: deep reinforcement learning based EV charging scheduling with dual-layer safety modules. IET Smart Energy Systems. ISSN 3065-9655 doi: 10.1049/ses2.70017 Available at https://centaur.reading.ac.uk/127268/

IET Smart Energy Systems

电力规划设计总院 China Electric Power Planning & Engineering Institute
华北电力大学 NORTH CHINA ELECTRIC POWER UNIVERSITY
IET The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

DuMES: Deep Reinforcement Learning-Based EV Charging Scheduling With Dual-Layer Safety Modules

Ao Zhang[1,2] | Cong Liu[3] | Konstantinos Makantasis[4] | Xiaomin Chen[5] | Tomas Ward[6] | Long Cheng[2]

[1]State Grid Electric Power Space Technology Company Limited, Beijing, China | [2]State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Beijing, China | [3]NOVA Information Management School, Nova University of Lisbon, Lisbon, Portugal | [4]Department of Artificial Intelligence, University of Malta, Msida, Malta | [5]Department of Computer Science, University of Reading, Reading, UK | [6]Insight Research Ireland Centre for Data Analytics, Dublin City University, Dublin, Ireland

Correspondence: Long Cheng (lcheng@ncepu.edu.cn)

Received: 11 June 2025 | **Revised:** 28 October 2025 | **Accepted:** 4 November 2025

ABSTRACT

Deep reinforcement learning (DRL) has become a promising approach for electric vehicle (EV) charging scheduling. However, its practical deployment poses potential risks to power infrastructure. DRL relies on trial-and-error interactions during training to approximate optimal policies, which may lead to unsafe decisions. To address this, a novel framework called dual-layer safety modules for EV charging scheduling (DuMES) is proposed. This framework introduces a decision-level safety layer into the conventional DRL architecture that adaptively detects and replaces unsafe actions. Furthermore, by integrating dual safety layers with reward shaping, the framework promotes convergence between raw and safe actions. This enhances training efficiency while ensuring power system stability during both training and deployment phases. The method was evaluated through simulation experiments on a charging station equipped with renewable energy and energy storage system (ESS). Comparative analyses with baseline methods demonstrate that DuMES effectively satisfies user charging demands, reduces operational costs and ensures compliance with safety constraints.

1 | Introduction

Due to the capability of connecting to the power grid and replenishing clean energy, electric vehicles (EVs) offer significant advantages over conventional fuel-powered vehicles in reducing carbon emissions. By 2024, the global fleet of EVs has reached 64 million, and it is projected to increase nearly fourfold to 250 million by 2030. Under this trend, EVs are expected to account for over 10% of all road vehicles by 2030 [1].

The widespread adoption of EVs has introduced significant challenges to the power system. The uncontrolled integration of mobile loads is likely to cause regional peak-valley fluctuations. Without effective countermeasures, this could lead to risks such as overloads of distribution equipment and increased load demands [2]. The upgrading of power infrastructure aims to improve the load-bearing capacity of the distribution network. However, it often faces high expansion costs and extended construction periods [3]. In contrast, charging scheduling decisions, through the dynamic adjustment of charging periods or charging power, leverage the flexible storage capabilities of EVs. This approach not only alleviates the load demand on the grid but also minimises electricity costs for users, presenting a more economical and efficient solution [4].

However, the charging behaviour of EVs exhibits significant spatiotemporal randomness. Influenced by factors such as time-of-use (TOU) electricity pricing and user range anxiety, the charging time and energy demand vary substantially across different vehicles [5]. Additionally, some EVs operate under a vehicle-to-grid (V2G) model, allowing them to feed energy back into the power grid, further increasing the complexity of the system. Traditional scheduling methods, such as rule-based methods and dynamic programming, rely on predefined rules

or precise system models. Their effectiveness largely depends on the accuracy of environmental assumptions and predictions, rendering them less adaptable to complex and dynamic real-world conditions [6]. To address these limitations, a class of approaches has emerged based on day-ahead scheduling, incorporating robust optimisation or stochastic optimisation techniques to mitigate the influence of uncertainties on decision-making [7]. Although these methods enhance scheduling robustness to some extent, they remain insufficient for real-time decision-making due to the high-dimensional and dynamic nature of the EV charging scene [8]. Therefore, existing scheduling methods exhibit inherent limitations and struggle to effectively address the increasingly complex problem of EV charging problems.

In recent years, inspired by the successful applications of DRL in domains such as autonomous driving, cloud computing and robot controlling [9–11], the integration of DRL into EV charging scheduling has emerged as a promising solution. Compared to conventional approaches, the primary advantage of reinforcement learning lies in its independence from prior knowledge, achieving optimisation solely through the interaction between the agent and the environment. Owing to this characteristic, reinforcement learning can effectively capture the stochastic features of charging scenarios, including TOU price, renewable energy generation and user behaviour, thereby enabling the learning of optimal scheduling strategies [12]. However, the deployment of DRL in real-world energy systems necessitates consideration of safety constraints. As a model-free method, DRL often struggles to accurately represent the physical constraints inherent in EV charging scenarios, potentially leading to constraint-violating decisions, which may result in energy inefficiencies or damage to power infrastructure.

In response to the aforementioned challenges, safe deep reinforcement learning (SDRL) has emerged as a promising approach. Several studies model the problem as a constrained markov decision process (CMDP), where the objective function incorporates safety constraints to guide the agent away from infeasible actions, thereby enhancing training efficiency. Other approaches introduce independent safety modules to correct constraint-violating behaviours without altering the underlying DRL algorithm, enabling seamless integration into the training process and improving overall safety. Recent studies in safe RL for EV charging have primarily addressed specific operational concerns within charging stations, such as battery degradation [13] and safe command allocation to prevent charging power deviations [14]. However, limited attention has been paid to the potential adverse impacts of high-power peaks on the electrical grid.

This study proposes DuMES, a DRL-based scheduling method that achieves multi-objective optimisation under safety constraints. The method integrates safety layer and reward shaping techniques into the DRL framework, introducing a dual-layer safety module while designing the reward function with incentives that encourage alignment with safe actions. Specifically, the constraints focus on mitigating power surges caused by instantaneous load spikes, aiming to ensure battery health and peak power safety. Through an iterative learning process, the safety layer in DuMES guarantees decision-making security during both training and deployment phases, whereas the reward design facilitates the agent towards more efficient learning of a feasible action space. DuMES is applied to schedule EV charging at a novel charging station with integrated renewables and energy storage. Experimental results demonstrate that the proposed DuMES effectively reduces operational costs while fulfiling user charging demands.

The main contributions of this paper are as follows:

- To optimise energy scheduling in V2G systems, we propose DuMES, a data-driven method based on DRL. The charging scheduling problem is formulated as an MDP, with the goal of meeting charging demands while minimising the operating costs of charging station.
- To address safety risks from the stochastic nature of DRL exploration, DuMES incorporates a dual-layer safety module. The first layer ensures battery safety during charging and discharging through iterative verification. The second layers smooths power peaks by adaptively delaying charging surges, taking into account factors such as rated power and flexibility margins.
- Simulations experiments are conducted on a charging station integrated with renewable energy and storage. The results show that DuMES consistently achieves optimal or near-optimal performance in operational cost and service reliability. Moreover, it enhances power safety, accelerates training and ensures safe agent behaviour during exploration.

The remainder of this paper is organised as follows. Section 2 reviews relevant work in the field of safe RL. Section 3 introduces the system model and formally defines the problem. Section 4 presents the proposed DuMES approach in detail. Section 5 reports experimental results and performance analysis. Finally, Section 6 concludes the paper with a summary of the main contributions.

2 | Related Work

In the field of energy management, DRL has emerged as an effective approach for addressing dynamic and real-time decision-making problems [26]. By balancing stochastic exploration with the exploitation of accumulated experience, DRL is capable of learning optimal scheduling strategies without relying on prior knowledge. However, during the exploration phase, DRL may generate actions that violate system constraints, potentially compromising the stable operation of energy systems. Ensuring the safety and feasibility of learnt policies in practical power system applications thus represents a critical research challenge.

The safe RL approach aims to learn a policy that maximises cumulative rewards while adhering to predefined safety constraints. Although existing studies vary in the formulation of objective functions and constraint conditions, they consistently emphasise the necessity for safe RL to ensure power system stability through the safe execution of actions. Table 1 presents a comparative analysis of relevant safe RL research in the energy sector from multiple perspectives.

TABLE 1 | Comparison of safe RL methods for energy management.

Methods	Constraint Type	Safe Exploration	Expert Knowledge	Action Alignment	Training efficiency	
					Param sensitivity	Convergence
DuMES	Soft-hard	Yes	Initial	Yes	Low	Fast
Reward shaping [13, 15]	Soft	No	No	Yes	High	Fast
Lagrangian relaxation [16–19]	Soft	No	No	Yes	Low	Slow
Trust region [20, 21]	Hard	Yes	Initial	No	High	Slow
Safety layer [14, 22, 23]	Hard	Yes	Initial	No	Low	Unstable
Shielding [24, 25]	Hard	Yes	All	No	Low	Unstable

Based on the extent of expert knowledge utilised, energy management methods based on safe RL can generally be categorised into two types. One category independently of any expert knowledge, embedding safety feedback directly into the objective function for optimisation. The reward shaping technique incorporates static penalty terms into the immediate rewards to guide the agent in actively avoiding unsafe or infeasible actions [13]. In contrast, the Lagrangian relaxation method introduces adaptive multipliers λ, which can be updated online within the overall objective function, thus reducing the sensitivity to hyperparameter settings. For example, the study [16] proposes a distributional soft actor–critic conservative augmented Lagrangian algorithm to address the battery temperature and health issues during fast charging of EVs, which was validated in real-world charging scenarios. Similarly, the work [17] introduced an actor–critic–Lagrangian (ACL) algorithm to resolve voltage violation problems in EV charging stations. Overall, these approaches contribute to more efficient learning of safety constraint boundaries by the agent, thereby enhancing training efficiency. However, due to the absence of explicit safety constraint mechanisms, it is challenging to ensure the system remains in a safe operating state throughout the training process.

Another category of methods constructs monitors based on expert knowledge to facilitate safety decision-making. The trust region method dynamically updates the trust region to ensure that policy at each step is projected onto the safety set. For example, the work [21] addresses the operational safety issues in distribution networks by proposing a projection-based embedded multi-agent DRL algorithm. This method effectively limits the decision space of the agent through action smoothing, achieving a 100% safety rate in experimental evaluations. Additionally, the safety layer and shielding mechanism are commonly used hard constraint techniques, although their implementations differ. The safety layer is typically embedded within the DRL framework for additional checks, whereas shielding mechanisms act as external components that intervene and provide corrections only when necessary. Regarding the issue of grid stability under large-scale EV integration, the research [23] proposes a dual-layer safety layer design based on the steady-state voltage security region (SVSR). This design utilises load margin indicators (LMI) to dynamically address uncertainties in charging and discharging strategies. The work [25] constructs an action feasibility space based on expert knowledge that adapts to V2G characteristics. When the current action may lead to overcharging or deep discharging of the battery energy storage system (BESS), it utilises the shielding mechanism to clip and replace the action, ensuring system stability. These methods strictly guarantee system safety during both the training and deployment phases. However, due to their reliance on constraints rather than incentives, the agent often struggles to fully comprehend the alignment logic of safe actions. This can result in slower policy convergence and even training instability in complex scenarios.

In contrast, we propose a novel safe RL-based method, termed DuMES, which incorporates a dual-layer safety module and a safety-driven reward function to guide the agent towards efficient energy scheduling in complex EV charging scenarios. Distinct from the aforementioned approaches, DuMES not only ensures strict safety in decision-making but also accelerates policy convergence by aligning raw actions with safety-refined actions. Furthermore, DuMES constructs safety constraint boundaries using a small amount of expert knowledge solely during the initial learning phase, thereby eliminating the reliance on explicit constraints and reducing both the complexity and dependency associated with penalty coefficient tuning. Simulation results demonstrate that DuMES significantly enhances decision-making safety while achieving optimal or near-optimal performance across various scheduling optimisation metrics.

3 | System Model for DUMES

In this section, we first outline the general system model employed in the DuMES framework, followed by a formal definition of the safety constraint and optimisation problem in charging scheduling. For better reference, the important notation is listed in Table 2.

3.1 | System Architecture

The architecture of the proposed DuMES method is shown in Figure 1. In the considered charging scenario, a V2G charging station capable of bidirectional energy transfer is employed. To further enhance the flexibility of regulation, renewable energy generation and ESS units are integrated into the framework. The entire framework is centred around a scheduling centre, where the DuMES strategy is deployed to perform real-time energy scheduling. This strategy dynamically determines the power allocation for both chargers and ESS at any given time. Operating under the V2G paradigm, the chargers not only supply energy to EVs but also enable energy feedback to the grid when appropriate. The main grid, in conjunction with renewable sources,

constitutes the energy supply for the charging station, thereby facilitating bidirectional energy flow. On the power market side, there is a continuous exchange of information with the scheduling centre, including dynamic electricity prices, repurchase signals and other relevant data.

Upon arrival at the charging station, the scheduling centre collects relevant charging information of EVs, including battery capacity, target state-of-charge (SoC) and estimated departure time. Subsequently, the DuMES framework integrates this information with environmental parameters, such as TOU prices and renewable energy output. Based on this integration, it formulates an energy scheduling policy π, which determines the optimal power allocation for each device at the current time step. According to this policy, chargers coordinates with the onboard EV batteries with the ESS to execute the specified charging and discharging operations. It is important to note that, as this study primarily focuses on the potential of EVs to contribute to grid stability and energy complementarity, EVs are typically assumed to remain parked for extended durations. Therefore, it is assumed that the charging station provides service only when chargers are available; otherwise, newly arrived EVs will depart immediately rather than waiting in a queue.

TABLE 2 | The used notation.

Notation	Meaning
Cid_i	The id of the ith EV
SoC_i	The SoC demand by ith EV
C_i	The battery capacity of the ith EV
$arrivalT_i$	The arrival time of the ith EV
DDL_i	The departure time of the ith EV
Loc_i	The area type of the ith EV
Pid_j	The id of the jth charger
P_j^{char}	Charge power of jth charger
η_j^{char}	Charge coefficient of jth charger
P_j^{dis}	Discharge power of jth charger
η_j^{dis}	Discharge coefficient of jth charger
P_j^{max}	Maximum rated power of jth charger

3.2 | Problem Formulation

To model the optimisation problem in this work, we provide mathematical definitions for the arrival EVs, chargers and other components, along with definitions related to constraint scheduling optimisation in the DuMES framework.

Arrival EVs: In this study, the arrival vehicle of the charging and discharge station is regarded as the key flexible energy storage. For the EV reaching the charging station at any time, it is defined as $Car_i = \{Cid_i, SoC_i, C_i, arrivalT_i, DDL_i, Loc_i\}$. Here, Cid_i denotes the ID uniformly assigned by the scheduling centre. SoC_i represents the expected battery state of the user to be achieved during this charging service; it is a battery energy ratio that reflects the desired level of charge. C_i is the onboard battery capacity of the EV. $arrivalT_i$ indicates the arrival time of the EV, and DDL_i is the latest estimated departure time provided by the EV user, which is submitted to the scheduling centre immediately upon arrival at the charging station. In addition, location plays an important role in charging scheduling, as it directly affects V2G participation [27]. Therefore, the variable Loc_i is introduced in the definition of arrival EVs to represent the type of the current region, which is closely associated with the arrival time $arrivalT_i$. Based on a statistical analysis of the National Household Travel Survey (NHTS) [28], real-world distributions of driver behaviour were extracted. These distributions reveal the temporal characteristics of travel patterns for four typical parking location types, as illustrated in Figure 2.

As shown in the figure, EV travel patterns differ markedly across scenarios. In residential areas, users primarily charge their

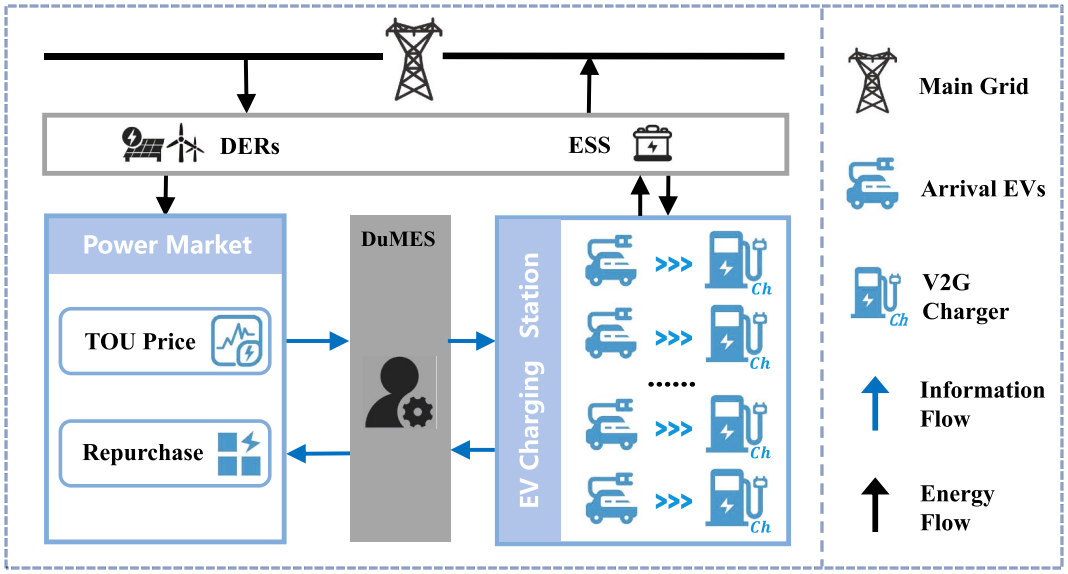

FIGURE 1 | The general system architecture of DuMES.

vehicles in the evening and overnight. During daytime hours, EVs are mostly parked in public locations such as office or commercial districts. The design also considers an all-day scenario, characterised by the absence of distinct traffic peaks; instead, the arrival probability remains approximately constant throughout the day, as observed in locations like hospitals and highway service areas. These variations highlight the need for a scheduling strategy with sufficient flexibility to adapt to diverse mobility patterns.

Chargers: In this study, chargers are defined by their bidirectional energy scheduling capabilities. For each charger j in a charging station, the configuration is denoted as $CP_j = \{Pid_j, P_j^{char}, \eta_j^{char}, P_j^{dis}, \eta_j^{dis}, P_j^{\max}\}$, where Pid_j is the unique identifier. The model incorporates both charging power P_j^{char} and discharging power P_j^{dis}, along with their corresponding efficiencies η_j^{char} and η_j^{dis}. A maximum power limit $P_j^{\max}$ is imposed to ensure operational safety by constraining agent decisions and preventing excessive instantaneous power output.

Optimisation model: The DuMES aims to minimise energy costs incurred by the charging station while ensuring full compliance with user charging requirements. Within this design, the optimal power allocation for the charger set $J = [j_1, j_2, ..., j_n]$ is determined to achieve cost minimisation, expressed as follows:

$$\omega = \min \sum_{i=1}^{I} Cost_i \tag{1}$$

where $Cost_i$ represents the cost of EV i. This represents the cost incurred by the charging station when acquiring energy from the main grid to facilitate energy supply. It is calculated as follows:

$$Cost_i = \rho(t) * P_t^{Grid} \tag{2}$$

The instantaneous cost, defined as the product of electricity price $\rho(t)$ and main grid power P_t^{Grid}, can be minimised by shifting demand to off-peak TOU periods to lower electricity prices and maximising renewable energy utilisation to reduce dependence on grid power.

Furthermore, cost minimisation must be based on the successful response to charging service. Given that EVs arrive at and depart from charging stations randomly with uncertain behaviours, it is crucial to ensure that they reach the desired state of charge before departure. The scheduling success of DuMES is defined as follows:

$$\text{success}\left(\text{Cid}_i, Pid_j\right) = \begin{cases} 1, & \text{if } SoC_i^{dep} \geq SoC_i^{\exp} \\ 0, & \text{else} \end{cases} \tag{3}$$

where SoC_i^{dep} and $SoC_i^{\exp}$ represent the state of charge at departure and the expected level of EV i, respectively. This design incorporates user-centric constraints into the energy scheduling method to improve service quality.

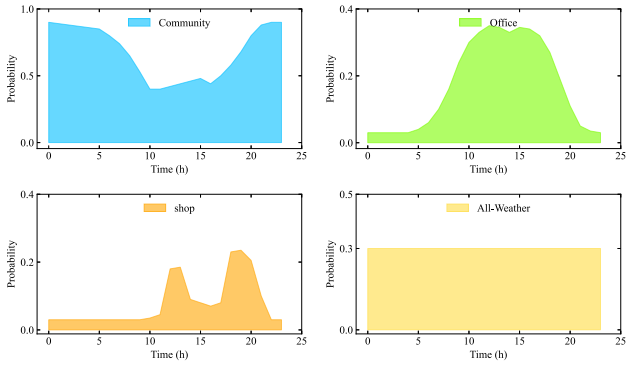

FIGURE 2 | Probability distributions of arrival time by scenarios.

3.3 | Safety Modules

In power systems, which represent highly sensitive environments, it is essential to implement the safety layer to prevent the trial-and-error behaviour inherent in DRL. The DuMES method incorporates two primary safety constraints: (1) the real-time SoC of EVs must remain within their capacity limits, as either overcharging or undercharging may result in battery degradation; (2) the total power output must not exceed a predefined threshold, since excessive peak power during high-demand periods may pose significant risks to the stability of the power grid. The dual-layer safety model designed in the DuMES framework is illustrated in Figure 3. Serving as an intermediate layer between the agent and the charging environment, the security module is responsible for transforming potentially hazardous actions into safe values through two levels of mapping, before these actions are executed within the environment.

3.3.1 | Battery Safety Module

To address the issue of battery damage during the charging and discharging processes, this work presents a battery safety module. The exploration of safe decision-making actions within this module can be iteratively conducted as follows:

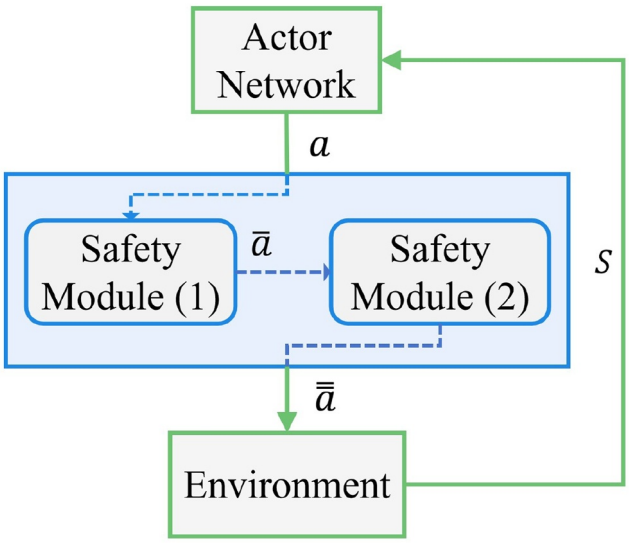

FIGURE 3 | Dual-layer safety model in DuMES.

$$P_{n,t}^{agent} = a_{n,t} * E \tag{4}$$

Specifically, Equation (4) is designed to compute the preliminary decision power $P_{n,t}^{agent}$. By multiplying the power coefficient $a_{n,t}$ with the battery capacity E, the initial value for safe power iteration can be obtained.

$$P_{n,t}(k+1) = P_{n,t}(k) + \rho\left(P_{n,t}^{agent} - P_{n,t}(k)\right) \tag{5}$$

As the number of iterations k increases, the decision power $P_{n,t}^{agent}$ gradually converges to the safe power $P_{n,t}$. The convergence rate is primarily governed by the exploration parameter ρ, which is adaptively calculated based on the variation in battery energy during iteration, as defined in Equation (6). A larger value of ρ leads to faster convergence, whereas smaller values, including zero, slows the process or causes it to terminate.

$$\begin{aligned}\rho = \Gamma\Big(&[s_n]^+\left[\left(\overline{E} - \epsilon\right) - SoC_n(k)\right]^+ \\ &+ [-s_n]^+\left[SoC_n(k) - \left(\underline{E} + \epsilon\right)\right]^+\Big)\end{aligned} \tag{6}$$

Here, Γ is a positive constant, and $[s_n]^+$ is used to extract the charging or discharging signal at the current time, where $[\bullet]^+ = \max(\bullet, 0)$ and $s_n = \mathrm{sgn}\left(P_{n,t}^{agent}\right)$; $\overline{E}$ and $\underline{E}$ denote the upper and lower bounds of the battery capacity; ϵ is a small positive constant used to control the threshold of the safety margin; and $SoC_n(k)$ represents the expected battery energy, which evolves iteratively with the safe power $P_{n,t}$.

This iterative formulation is introduced to ensure that the adjustment of charging power is adaptive, rather than relying on abrupt truncation at the safety boundaries. By gradually steering the safe power $P_{n,t}$ towards the agent decision $P_{n,t}^{agent}$, the mechanism allows the learning process to preserve continuity in high-dimensional action spaces, thereby facilitating stable policy optimisation. At the same time, the adaptive design guarantees that once the state approaches the safety margin, the update is immediately suppressed, ensuring strict compliance with operational limits.

Overall, based on Equations (4–6), when the predicted battery energy $SoC_n(k)$ remains within the safe range $\left[\overline{E} - \epsilon, \underline{E} + \epsilon\right]$, the safe power $P_{n,t}$ asymptotically converges to the decision power $P_{n,t}^{agent}$ at a rate governed by the adaptive parameter $\rho \neq 0$. If $SoC_n(k)$ approaches the safety boundary, that is, $\overline{E} - \epsilon < SoC_n(k)$ or $\underline{E} + \epsilon > SoC_n(k)$, the parameter is set to $\rho = 0$, immediately halting the update of $P_{n,t}$ and forcing convergence. This mechanism enables fast convergence when operating safely away from the boundaries while effectively freezing updates near the limits to prevent overcharge or undercharge risks.

3.3.2 | Power Safety Module

Since public charging stations are typically located in commercial or office areas, their peak loads tend to coincide with traffic peaks, which can lead to overloads when the distribution capacity is insufficient. DuMES incorporates a power safety module to enforce this constraint by adjusting the power allocated to each EV based on charging anxiety and power limit margins, thereby smoothing the overall peak load. This process is implemented through a multi-step exploratory mapping as follows:

Step 1: Initialise the power safety threshold ξ, then compute the safety margin $\triangle\Phi$ as follows:

$$\triangle\Phi_t = \sum_{n \in N}\left(P_{\max} - \overline{P}_{n,t}\right) \tag{7}$$

The safety margin represents the difference between the total power $\overline{P}_{n,t}$ and the rated power $P_{\max}$. The safety threshold ξ is a constant used to regulate the operational range of the power safety module. If the safety margin $\triangle\Phi_t$ is below the threshold ξ, the process proceeds to Step 2; otherwise, it terminates.

Step 2: Compute the maximum allowable increase in charging power for EV $n \in N$ at time t; the charging power $\overline{P}_{n,t}$ is determined as follows:

$$\begin{aligned}P_{n,t}^{+} = \min(&\triangle\Phi_t, \quad P_{\max} - \overline{P}_{n,t}, \\ &E \cdot (SOC^{\max} - SoC_n(t))/\eta^{ch})\end{aligned} \tag{8}$$

where $P_{n,t}{}^{+}$ is determined as the minimum value among three factors: the safety margin; the difference from the rated power; and the remaining capacity available for charging, which represents the maximum power that can be accepted without causing overcharging of the onboard battery.

Step 3: When reducing power, user charging demands should be considered. Therefore, we refer to the least laxity first (LLF) method, which accounts for EV charging time anxiety [29]. The laxity is defined as follows:

$$\psi_{n,t} = T_{n,t}^{Leave} - t - \left(\frac{\left(SOC^{\max} - SoC_{n,t}\right) * E}{\eta * P^{\max} * \triangle t}\right) \tag{9}$$

which represents the remaining schedulable time for EV n at the current time t, calculated as the expected departure time minus the current time and the estimated battery full charge time. A smaller $\psi_{n,t}$ indicates a higher priority for ensuring charging power.

Step 4: Using the maximum allowable increase charging power $P_{n,t}{}^{+}$ and the relaxation factor $\psi_{n,t}$, the priority of each EV $n \in N$ can be determined as follows:

$$S_{n,t} = \alpha \cdot \frac{1}{P_{n,t}^{+}} - \beta \cdot \psi_{n,t} \tag{10}$$

The objective is to prioritise the reduction of charging power for EVs with lower anxiety levels and higher load conditions. The above equation reflects this relationship: for any EV $n \in N$, the smaller its additional achievable power and relaxation factor, the higher its priority.

Step 5: Based on the obtained priority $S_{n,t}$, the EV set N is reordered to form $\widetilde{N}$. From this reordered set, the top X vehicles

are selected to constitute the adjustment set $\tilde{N}_X$, enabling the following safe power control:

$$\overline{\overline{P}}_{n,t} = \begin{cases} \overline{P}_{n,t} - P^{adjust}, & \text{if } n \in \tilde{N}_X \\ \overline{P}_{n,t}, & \text{if } n \in N \backslash \tilde{N}_X \end{cases} \tag{11}$$

For the EV set N, the top X vehicles with higher priority will have their power reduced by an adjustment value, forming the new safe power level. The remaining vehicles will retain the safe power output determined by the previous module.

Step 6: Finally, the safety margin $\triangle\Phi_t$ is updated using the new safe power $\overline{\overline{P}}_{n,t}$.

$$\triangle\Phi_t = \sum_{n \in N} \left(P_{\max} - \overline{\overline{P}}_{n,t} \right) \tag{12}$$

If the safety margin satisfies the threshold condition $(\triangle\ \Phi_t \geq \xi)$, the safety exploration concludes; otherwise, return to Step 2.

Through the dual mapping in the aforementioned modules, the DuMES framework establishes a hard constraint on charging scheduling safety, ensuring strict security during the training phase. In the following sections, we present the detailed implementation of the method, along with the soft constraints by reward shaping.

4 | Method Implementation

The proposed DuMES framework is implemented by the DRL method developed in this section. This approach enables adaptive online learning in complex environments and achieves efficient energy allocation during the execution phase. This section first introduces the MDP formulation of the scheduling problem and then presents the implementation process based on the proximal policy optimisation (PPO) algorithm, which demonstrates high solving efficiency in high-dimensional continuous spaces.

4.1 | Markov Decision Process for DuMES

As a machine learning technique, DRL learns a policy for maximising cumulative rewards by continuously interacting with the environment, extracting state information and generating actions. This section formulates the MDP for multi-objective scheduling optimisation, including the state space, action space and the design of the reward function.

4.1.1 | State Space

The state space is an abstract representation of the scheduling environment. For the EV charging station considered in this study, the state at any given moment is described as follows:

$$\begin{aligned} S(t) = \{ & \rho(t), G(t), \\ & SoC_1(t) \ldots SoC_n(t), T_1^{Leave}(t) \ldots T_n^{Leave}(t) \} \end{aligned} \tag{13}$$

The state space can be divided into two components based on whether the corresponding unit performs charging or discharging actions. The global state mainly includes system-wide information such as the real-time electricity price $\rho(t)$ and renewable generation output $G(t)$. In contrast, the local state mainly comprises the charging state $SoC_n(t)$ of the n-th charger (or ESS) at time t and its estimated departure time $T_n^{Leave}(t)$. It is worth noting that for ESSs installed as stationary energy buffers within the station, its expected departure time is assumed to be a fixed negative value.

4.1.2 | Action Space

The action space is the set of all possible actions that the agent can execute within a given environment. In the proposed DuMES, the action at any time t represents the charging/discharging power coefficient a for each charger or ESS, where $a \in [-1, 1]$, and can be expressed as follows:

$$a_t = \left(\alpha_{1,t}, \quad \alpha_{2,t} \ldots \quad \alpha_{n,t} \right), \alpha_{n,t} \in [-1, 1] \tag{14}$$

The size of the action space depends on the number of charging stations n. Each continuous variable $\alpha_{n,t}$ represents the magnitude of power, with positive values indicating charging and negative values indicating discharging.

4.1.3 | Reward Function

As a scheduling approach from the perspective of the charging station operator, DuMES primarily aims to minimise the cost of purchasing power from the main grid. Meanwhile, to reflect more realistic considerations, the reward function is additionally designed with two other components to achieve multi-objective optimisation, including meeting user charging demands, ensuring safety constraints and reducing operational costs.

This section defines the power purchase cost from the main grid as $R_t^{(1)}$, serving as the primary objective for V2G scheduling optimisation at charging stations. Specifically, the aim is to minimise this cost by guiding the agent to fully utilise renewable energy output and leverage the flexibility of onboard batteries to respond to real-time electricity price fluctuations.

$$R_t^{(1)} = \rho(t) * P_t^{Grid} \tag{15}$$

where $\rho(t)$ denotes the real-time TOU price at time t, and P_t^{Grid} represents the instantaneous power from the main grid. Their product corresponds to the electricity purchase cost at the given time. Furthermore, this section defines the insecurity level of decision-making power as $R_t^{(2)}$, expressed by the following formula:

$$R_t^{(2)} = \sum_{n \in EV} |P_{n,t}^{agent} - \overline{P}_{n,t}| \tag{16}$$

Here, the difference between decision power $P_{n,t}^{agent}$ and safety power $\overline{P}_{n,t}$ is used as a penalty for unsafe raw outputs of the agent. This reward shaping component is designed to guide the agent to focus on the underlying patterns in the safety module outputs, ultimately leading its decisions to converge to the safety boundary.

Finally, this part defines the penalty for unmet charging demand as $R_t^{(3)}$, which represents the discrepancy between the expected battery departure state SoC_n^{tar} and the actual battery departure state $SoC_{n,t}$.

$$R_t^{(3)} = \sum_{n \in EV, t \in DDL} \max\left(SoC_n^{tar} - SoC_{n,t}, 0\right)^2 \tag{17}$$

This penalty term equals zero when the charging demand is met and increases gradually with the degree of deviation, namely when the battery charge level at the departure time does not reach the expected value. This design is intended to prevent the agent from excessively prioritising optimisation of cost and safety at the expense of user service.

In summary, the total reward R_t^{total} at any given t is defined as the weighted sum of the three optimisation objective reward terms, and the formulation is as follows:

$$R_t^{total} = \eta_1 R_t^{(1)} + \eta_2 R_t^{(2)} + \eta_3 R_t^{(3)} \tag{18}$$

The penalty factors $0 > \eta_1 > \eta_2 > \eta_3$ are introduced to balance the reward weights and the normalisation process. Specifically, beyond the primary objective of cost minimisation, the condition $0 > \eta_1 > \eta_2$ assigns a higher penalty priority to charging safety. Moreover, since the agent obtains feedback $R_t^{(3)}$ only at the departure time of each EV, the negative penalty weight η_3 is made more negative to facilitate more effective learning by the agent.

4.2 | DuMES Implementation

The scheduling framework DuMES, improved with PPO-based safety enhancements, is described as follows. Initially, the agent continuously interacts with the environment to sample trajectory data over step t. This data includes the state, action, reward, next state, as well as the action probability under the current policy. All collected trajectories are stored in an experience replay buffer for subsequent training. Subsequently, the agent randomly samples mini-batches from the buffer to update both the policy and value networks. During this process, the clip objective function is employed to constrain the extent of policy updates, thereby ensuring policy stability during optimisation. Algorithm 1 provides an overview of this process.

ALGORITHM 1 | The proposed DuMES.

Require: Initial condition of EVs and chargers, discount rate γ, exploration rate ϵ and clip threshold ε.

Initialise: the actor and critic with random parameters θ, φ; weights c_1, c_2; constants k, t_{end}, G.
for training episode $= 1$ to k **do**
  Initialise environment state s.
  while $t \neq t_{\text{end}}$ **do**
    Sample action: $raw\_a \sim \pi_\theta(a|s)$.
    if raw_a is safe **then**
      $\overline{\overline{a}} \leftarrow raw\_a$.
    else
      $\overline{a} \leftarrow \text{Safe\_Layer}_1(raw\_a, s; \alpha)$.
      $\overline{\overline{a}} \leftarrow \text{Safe\_Layer}_2(\overline{a}, s; \beta)$.
    end if
    Execute $\overline{\overline{a}}$, observe s', r.
    Store transition (s, a, r, s') into replay buffer Δ.
    Update $s \leftarrow s'$.
  end while
  randomly select samples $S_\triangle$ from $\triangle$.
  Compute $L_{\text{clip}}(\theta)$.
  Update policy network θ via gradient descent.
  Update value network φ after G iterations.
end for

Training: During the training phase, the DuMES method iteratively optimises both the policy and value function based on the PPO algorithm, aiming to learn an optimal energy scheduling strategy within a highly dynamic and uncertain environment. The training process begins by initialising the parameters of the actor and critic networks, denoted respectively as θ and φ, along with setting several essential hyperparameters, including the discount factor γ, exploration rate ϵ and clipping threshold ε. At the start of each training episode, the environmental state s is initialised, after which the agent samples a raw action raw_a from the actor network according to the current policy $\pi_\theta(a|s)$. To ensure that the learnt policy consistently adheres to predefined system constraints throughout training, DuMES incorporates a dual safety layer mechanism. When raw action raw_a fails to satisfy these constraints, the action is sequentially refined by Safety Layer 1 and Safety Layer 2. These layers introduce parameterised control factors, α and β, respectively, to adaptively handle varying constraint conditions and generate a final action, denoted as $\overline{a}_{\text{safe}}$, that is both valid and safe for real-world execution.

As shown in Figure 4, the resulting safe action is then delivered to the environment, which returns the subsequent state s' and the immediate reward r. The transition tuple (s, a, r, s') is stored in a replay buffer Δ, and at the end of each episode, a batch of samples $S_\triangle$ is randomly drawn from it for updating the policy and value networks. The policy network is updated using the clipped loss function $L_{\text{clip}}(\theta)$, optimising the actor parameters θ through gradient descent in order to maintain the stability of the policy update and prevent significant deviation from the previous policy. In contrast, the critic network parameters φ are refined through multiple iterations (specified as G times) using the target value function, thereby enhancing the evaluation capability of the current policy.

Interaction: At the beginning of each training episode, EVs arrive at the charging station sequentially and connect to the grid. By collecting local information such as the battery SoC and departure times from each charger, the state information of the EVs can be obtained. The global information, such as electricity prices and renewable energy output, is then extracted and integrated. This

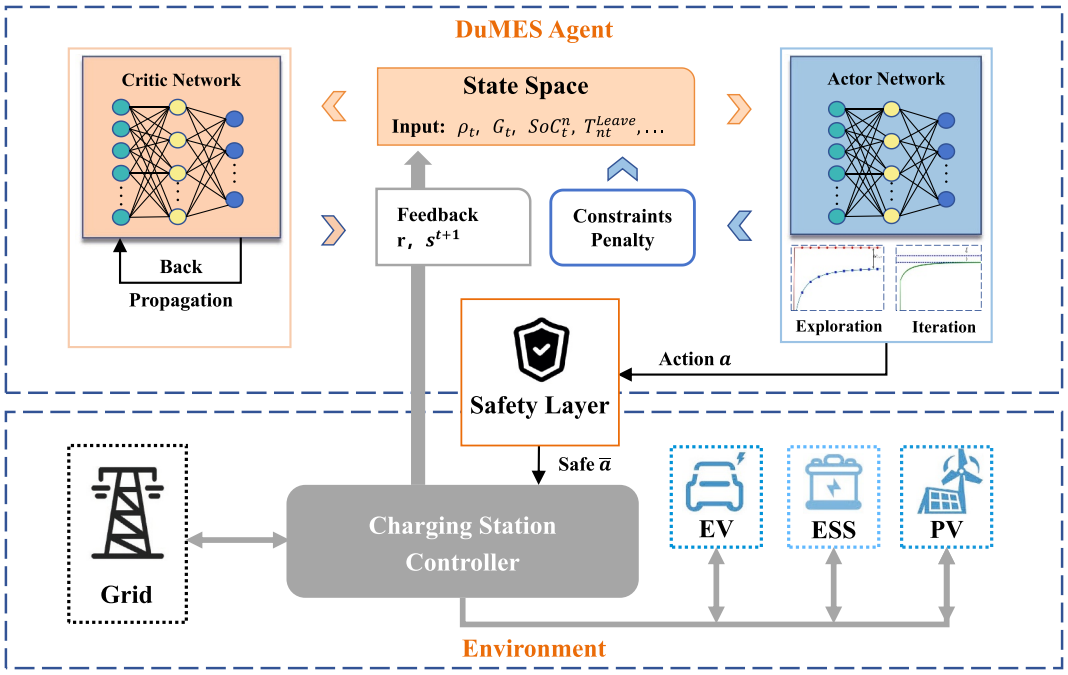

FIGURE 4 | The safe RL-based architecture for EV charging scheduling.

allows the initialisation of state space as $S(t) = \{\rho(t), G(t), SoC_1(t), ..., T_1^{Leave}(t), ...\}$. The agent, using a DNN to approximate the probability function, determines the appropriate actions to take. These actions are executed, and the agent continuously adjusts its strategy based on the rewards provided by the environment and the updates to the state. This interaction mechanism aims to guide the agent in optimising its strategy, ensuring that charging demands are met while minimising electricity costs and maintaining charging safety constraints.

Safety constraints: In this study, a major challenge lies in learning an optimal scheduling policy while strictly adhering to charging safety constraints. This difficulty arises from the model-free nature of DRL, which enables it to handle high-dimensional dynamic processes without any prior knowledge of system models. However, this also makes it difficult for the method to accurately capture physical constraints in the environment and take proactive safety measures. Furthermore, DRL-based approaches typically involve random exploration during the training phase, which involves extensive trial-and-error behaviour that poses potential risks in real-world applications. To address this issue, a safety layer design has been introduced, as shown in the safety layer component in Figure 4.

DuMES employs an actor–critic architecture for generating raw action decisions. In this framework, the actor network receives the current system state and outputs the final policy action, whereas the critic network estimates the state-action value function to evaluate the generated policy and facilitate decision-making optimisation. Upon generation of the preliminary action by the actor network, it is forwarded to the safety layer for constraint verification and correction. Control commands that may violate physical constraints are adjusted by the safety layer into constraint-satisfying safe actions $\bar{a}$, and then transmits these to the charging station controller within the system environment. The charging station controller subsequently schedules EVs and ESS based on the safe actions $\bar{a}$. In addition to the hard safety constraints enforced by the safety layer above, the environment also provides reward components related to the safe actions in its feedback after executing $\bar{a}$. These components serve to guide the agent in aligning its preliminary actions with the corresponding safe actions. By embedding the aforementioned safety constraints into the feedback learning process, the DuMES is able to optimise its scheduling strategy while ensuring compliance with safety requirements.

5 | Evaluation

This section presents an experimental evaluation of the proposed DuMES to verify its optimisation performance. First, the details of the experimental environment are described, followed by a performance assessment under various scenarios using several baseline methods, including two conventional scheduling algorithms, an advanced DRL algorithm and a safety-enhanced DRL variant. All methods are implemented in Python 3.9 using the PyTorch framework. The simulations are conducted on a local high-performance computing server equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM, 2.6 GHz) running Ubuntu 20.04 LTS.

5.1 | Simulation Setup

This section presents a series of evaluations to demonstrate the effectiveness of the DuMES framework in the context of charging scheduling. The simulation setup considers a charging station equipped with wind power generation and energy storage devices. The station includes 20 bidirectional chargers, each with a maximum charging power of 30 kW and a charging/discharging efficiency coefficient of 0.91, one energy

storage battery with a maximum charging power of 21 kW, a capacity of 50 kWh and a charging/discharging efficiency coefficient of 0.98. Additionally, the station includes 20 photovoltaic panels, each with dimensions of 2.279 m × 1.134 m and a power conversion efficiency coefficient of 0.21. Similar to ref. [30], a typical TOU tariff is adopted, as shown in Table 3.

According to data from the Federal Highway Administration of the United States Department of Transportation, the features of commuting EVs follow the normal distribution [28]. Therefore, the arrival time distribution of EVs conforms to the real-world data presented in the arrival EV model in Section 3.2, whereas the arrival SoC follows a normal distribution $N \sim \mathcal{N}(20, 2^2)$. Similarly, the expected parking time follows a normal distribution $N \sim \mathcal{N}(6, 1)$. In addition, each EV is assumed to be equipped with a 50-kWh onboard lithium battery, with the expected SoC set to 100%. The minimum capacity threshold is set to 2.4 kWh to protect battery lifespan [31]. The power safety threshold ξ is used to determine whether the current power consumption is approaching the permissible safety limit of the system. After evaluating multiple candidate values, ξ is set to 0.01 kW to balance the trade-off between detection accuracy and timely responses, thereby avoiding premature triggers or delayed recognition. It should be noted that the above settings serve only as simulation-specific parameters in this study. The proposed DuMES can effectively accommodate various travel patterns and electrical characteristics of EVs during the scheduling process while requiring only a limited set of prior knowledge such as battery capacity constraints during the initialisation phase of the safety module.

In terms of algorithm hyperparameter configuration, the PPO algorithm adopts a clipped strategy optimisation. The clipping ratio for policy updates is set to $\varepsilon = 0.2$, which serves to constrain the magnitude of policy changes and prevent abrupt alterations in the policy. Both the policy network and the value network are implemented as two-layer fully connected structures, with each layer containing 64 neurons. The activation function used is the hyperbolic tangent (Tanh). In each learning iteration, the generalised advantage estimation (GAE) coefficient is set to $\lambda = 0.95$, and the discount factor is defined as $\gamma = 0.99$, with the value network updated every five iterations and the policy network is updated in each iteration. More detailed configuration of the hyperparameters is presented in Table 4.

5.2 | Scheduling Performance

The performance of the DuMES method proposed in Algorithm 1 is validated using the following four baseline methods.

TABLE 3 | TOU tariff used in simulations.

	Period	TOU (¥/kWh)
Peak	12:00–15:00, 18:00–21:00	1.2710
Flat	6:00–12:00, 15:00–18:00. 21:00–23:00	0.7685
Valley	23:00–6:00	0.2576

1. Greedy method (GRD): A greedy implementation designed to minimise costs by prioritising the utilisation of renewable energy for power supply. Electricity is purchased from the main grid only when the real-time TOU price is less than or equal to the flat period.
2. Rule-based method (RBC): A rule-driven decision-making mechanism that operates with low latency. If any EV is scheduled to depart within the next 3 hours, RBC charges it at maximum power. Otherwise, the charging decisions are made by jointly considering TOU pricing and the output of renewable energy.
3. PPO [32]: An advanced real-time method based on the actor–critic framework that facilitates continuous action decision-making. However, it is founded upon the original DRL architecture and without the enhancements for ensuring operational safety.
4. DuMES: Introduced in Section 4.

This section evaluates the proposed method and baseline method under four charging scenarios, aiming to verify their scheduling performance on three optimisation objectives, as compared in Figure 5.

Scenes 1 to 4 correspond to the community, commercial, office area and all-day scenarios described in Section 3.2. Operational cost represents the energy expenditure incurred by the charging station to respond to charging demands over a complete testing day. This cost is minimised by fully utilising renewable energy output and storing energy during periods with off-peak electricity prices. In this context, GRD is a greedy strategy that aims solely at minimising costs. Under this strategy, it purchases electricity from the main grid for charging or energy storage only when the real-time electricity price does not exceed the flat rate, and it sells electricity otherwise. Consequently, GRD achieves the optimal reduction in operational costs by sacrificing other optimisation objectives. As shown in Figure 5a, GRD achieves the best performance in reducing operational costs. In comparison, the DRL-based PPO and DuMES methods, which incorporate multi-objective optimisation, obtain a cost efficiency that is second only to optimal and comparably close.

TABLE 4 | Hyperparameters in the DuMES method.

Discount factor γ	0.99
Buffer size	4000
Batch size	64
Learning rate	3e-4
Optimiser	Adam
Number of hidden layers	2
Number of hidden units	64
Activation function of hidden layers	Tahn
Clip parameter ε	0.2
GAE λ	0.95
Value function coefficient	0.5
Updating times	5

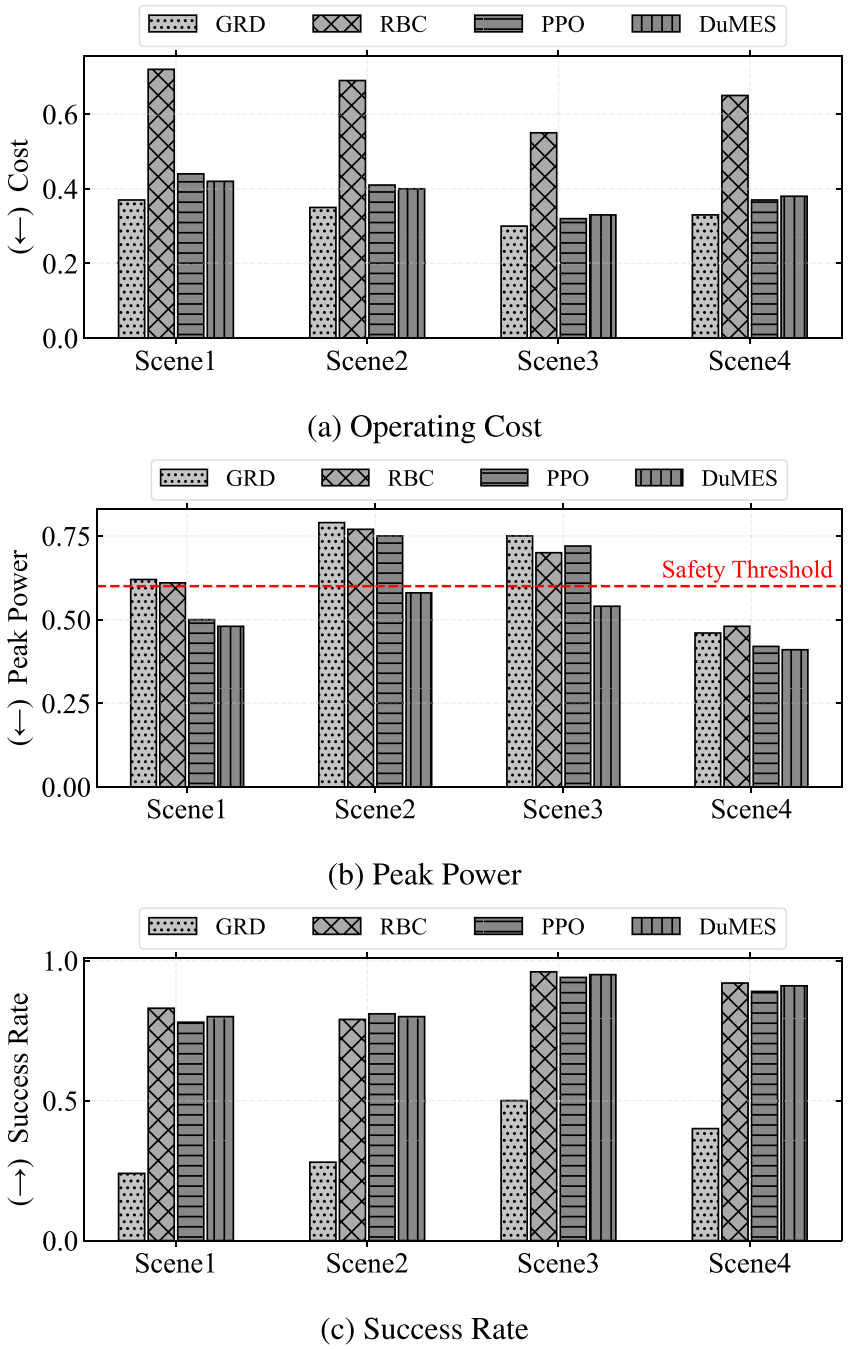

FIGURE 5 | Comparison across different scenes.

The peak power reflects the performance of the scheduling method in ensuring power safety. It is defined as the maximum total power exhibited by the charging station over a complete test day. As disordered charging may lead to power levels exceeding safe limits, the peak power is closely correlated with the safety constraints proposed in this study. Specifically, the power safety threshold is indicated by the dashed line in Figure 5b. All scheduling methods successfully avoid power violations in Scene 4, as this hypothetical scenario features continuous and stable vehicle arrival rates throughout the day, thereby preventing occurrences of uncoordinated and concentrated charging. Similarly, Scenario 1, representing a community with high traffic flow but without a pronounced peak in arrivals, results in only slight power violations for the GRD and RBC methods. However, in Scenes 2 and 3, the traffic flow exhibits evident peak and valley patterns, causing a large number of EVs to arrive within fixed time intervals. In such cases, the reward-driven mechanism of the DRL agent alone is insufficient to effectively mitigate peak power demands, and accordingly the GRD, RBC, and DRL methods all experience significant power violations. In contrast, DuMES integrates a dual-layer safety module and leverages the iterative relaxation mechanism of the power safety module and shifts low-priority charging demands to later time periods, thereby completely smoothing the power peak.

The service success rate indicates whether the charging service is provided in a timely manner throughout the test day. Due to its cost-greedy design, the GRD approach tends to pursue lower

real-time electricity prices, which may lead to missed opportunities for optimal charging before the user departs. As a result, it demonstrates the lowest success rate among the compared strategies. In contrast, RBC adopts a more conservative strategy by charging at maximum power during the 3 hours preceding the anticipated departure, which results in an excellent success rate. By comparison, DuMES achieves a suboptimal success rate, demonstrating its capability to balance multi-objective optimisation.

In summary, the proposed DuMES method achieves multi-objective optimisation comparable to PPO in scheduling performance. However, DuMES demonstrates a significant improvement in power safety without compromising scheduling effectiveness. The following section further compares the training efficiency and battery safety performance of DuMES and PPO.

5.3 | Training Efficiency and Safety

In EV charging scheduling, the computational efficiency of the policy considerably influences both real-time performance and optimisation effectiveness. Conventional DRL methods typically impose constraints through penalty terms in the reward, using hard boundary limitations without modifying the network architecture. However, when applied to high-dimensional continuous solution spaces, DRL often fails to capture the inherent safety patterns with this approach, thereby exacerbating sample and computational restrictions. In contrast, the proposed DuMES method adopts a more flexible safety module design that adaptively adjusts the constraint boundaries based on real-time states, which significantly reduces the number of trial-and-error iterations. Furthermore, by incorporating dynamic safety boundaries into the penalty terms of the reward function, the method guides the original network output towards alignment with safe actions, thereby accelerating the convergence rate.

To validate the improvement in training performance achieved by the proposed framework, this part compares PPO with its safety-enhanced variant, DuMES. Both methods are trained for a maximum of 5×10^5 steps with identical hyperparameter configurations. The only distinction lies in the formulation of the reward function: While the PPO employs hard boundary constraints within the penalty term, the DuMES utilises outputs from the safety module. Figure 6 presents the mean rewards obtained by the two methods.

It can be observed that both DuMES and PPO achieved effective convergence, as quantitatively compared in Table 5. Under identical hyperparameter settings, PPO converged after 309123 time steps, attaining an average reward of approximately −12.15. In contrast, DuMES converged after 254352 time steps, with an average reward of approximately −10.72. This represents an 11.77% improvement in reward and a 17.72% acceleration in convergence speed, thereby demonstrating the superiority of DuMES in terms of training efficiency.

In addition, Figure 7 presents the SoC curves of 20 vehicles participating in the charging scheduling, comparing the DuMES method equipped with the safety module and the PPO method without it. It should be noted that the arrival and departure times in this figure were deliberately designed as an extreme case, where vehicles arrive within a concentrated time window and remain parked for an extended duration. This setup serves as a stress-test scenario to clearly demonstrate the robustness of the proposed safety module, and it does not represent the realistic travel patterns used in the performance evaluation.

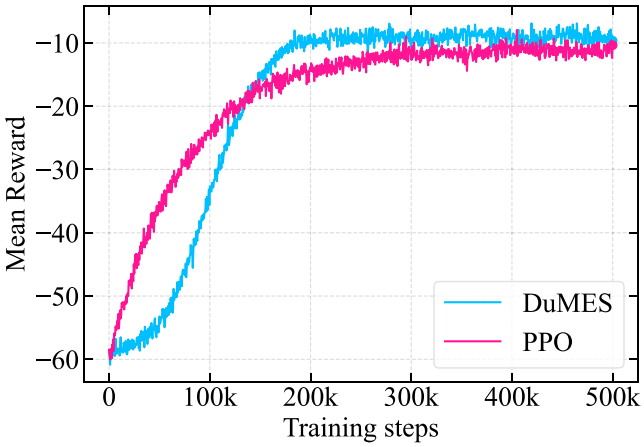

FIGURE 6 | Comparison of mean reward between PPO and DuMES.

TABLE 5 | Comparison of training efficiency.

Methods	Mean reward	Convergence timestep	Comparison of convergence (%)
DuMES	−10.72	254352	−17.72
PPO	−12.15	309123	100

As shown in Figure 7b, the absence of a safety module can lead to violations of the minimum SoC threshold during discharging operations, occasionally resulting in battery undercharge. This issue arises because the PPO algorithm relies solely on the reward function for safety guidance. Consequently, when the current battery state approaches the lower bound or zero, the agent may still select random exploratory actions with $a(t) < 0$, thereby causing the SoC below the permissible limit or even producing negative values. In contrast, the proposed DuMES method, as demonstrated in Figure 7a, ensures that the SoC of all EVs remains within the specified range, thereby guaranteeing battery safety.

To further illustrate the effectiveness of the battery safety module, Figure 8 presents the safety exploration process conducted by DuMES as the battery SoC increases. In Figure 8a, the comparative method is the DRL method without safety module. This method directly applies the decision power produced by the agent to the charger without evaluating the safety of the action. It is evident that the instantaneous increase in charging power under the conventional DRL method may expose the battery to an overcharge risk. In contrast, when the battery SoC is substantially lower than the threshold range, DuMES equipped with the safety module rapidly increases the corresponding safety power exploration value; subsequently, as the battery SoC approaches the limit, the increase becomes gradual until it stops, thereby ensuring that the current battery SoC in Figure 8b

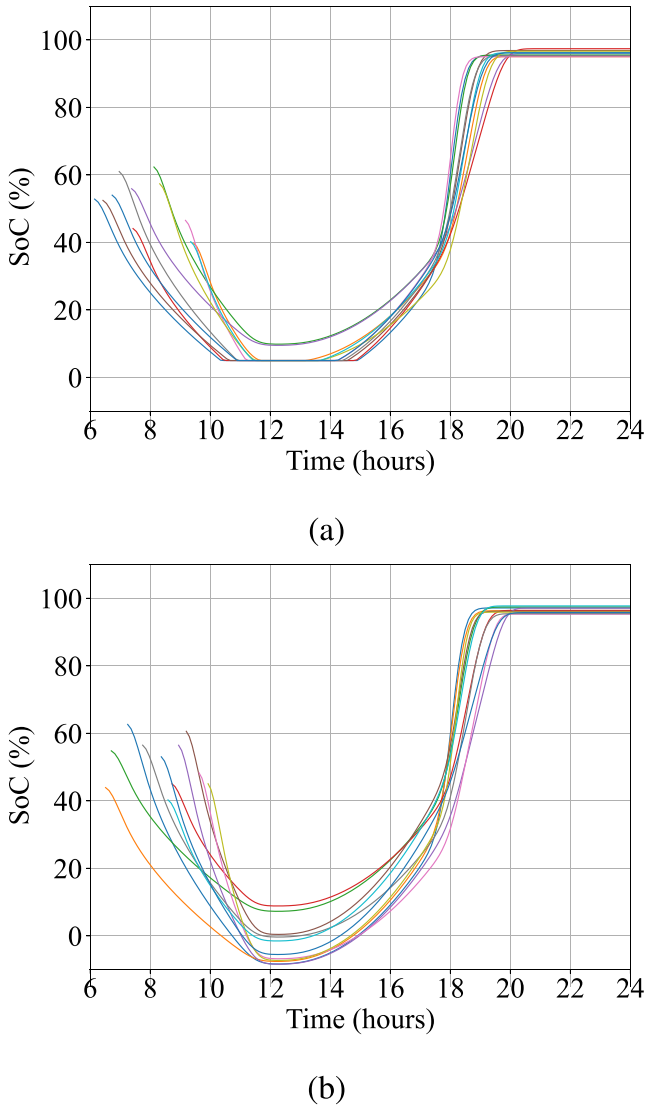

FIGURE 7 | SoC violation cases for 20 EVs: (a) with battery safety module and (b) without battery safety module.

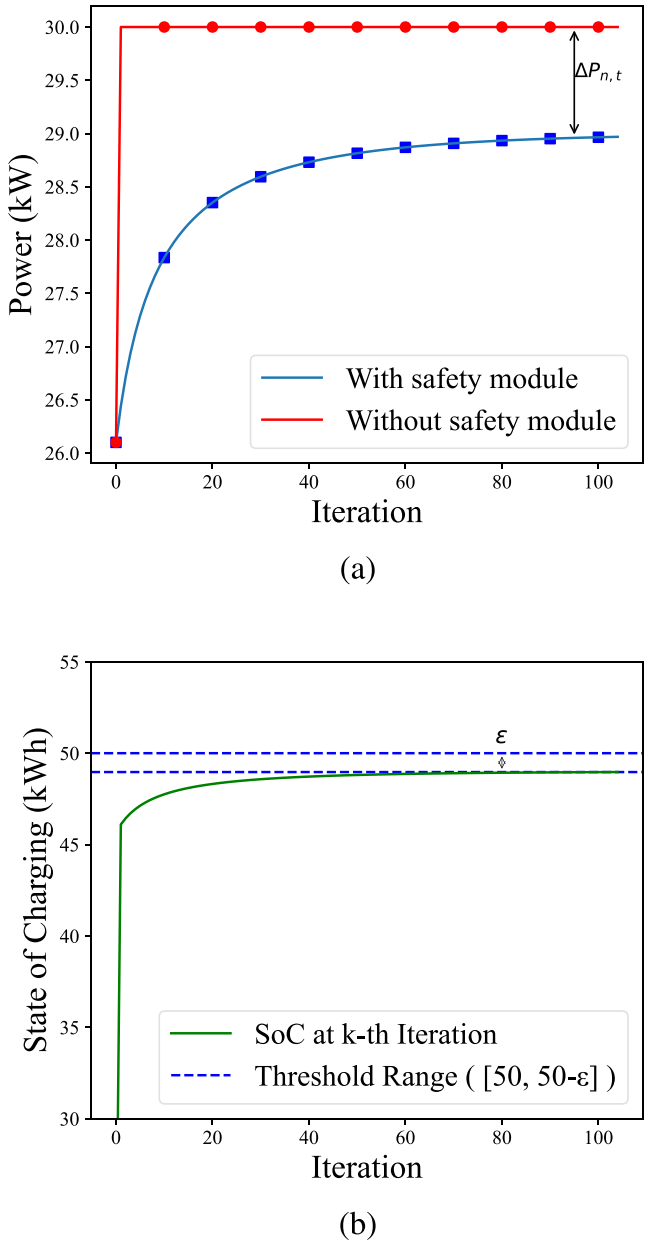

FIGURE 8 | Safe exploration for battery safety module.

always remains within the margin of the safety threshold. By defining the threshold range with an appropriately small positive number ϵ, the training process is ensured to converge normally.

6 | Conclusion

In this work, we propose DuMES, a dual-layer safety module framework designed to enhance standard DRL approaches for charging scheduling by introducing a decision-level safety layer. DuMES adaptively detects and replaces unsafe actions while incorporating reward shaping aligned with dual safety constraints, effectively mitigating potential grid risks during both training and deployment. Simulation studies on a charging station equipped with renewable generation and ESS demonstrate that DuMES not only satisfies user charging demands but also surpasses baseline methods in reducing operational costs and strictly adhering to safety limits. Furthermore, the reward shaping mechanism promotes convergence between the original and safe action policies, thereby accelerating training and reducing the overhead associated with trial-and-error learning.

Author Contributions

Ao Zhang: investigation, methodology, software, writing – original draft, writing – review and editing. **Cong Liu:** methodology, writing – review and editing, conceptualization. **Konstantinos Makantasis:** writing – review and editing, conceptualization. **Xiaomin Chen:** writing – review and editing, conceptualization. **Tomas Ward:** writing – review and editing, conceptualization. **Long Cheng:** investigation, conceptualization, methodology, writing – review and editing.

Funding

This research was supported by the Fundamental Research Funds for the Central Universities (2025JC002), and by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project—UIDB/

04152—Centro de Investigação em Gestão de Informação (MagIC)/ NOVA IMS.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. International Energy Agency, Global EV Outlook 2024, 2024, https://www.iea.org/reports/global-ev-outlook-2024.

2. S. P. Sone, J. J. Lehtomäki, Z. Khan, K. Umebayashi, and K. S. Kim, "Robust EV Scheduling in Charging Stations Under Uncertain Demands and Deadlines," *IEEE Transactions on Intelligent Transportation Systems* 25, no. 12 (2024): 21484–21499, https://doi.org/10.1109/tits.2024.3466514.

3. I. Diahovchenko, "Analyzing the Influence of Electric Vehicle Charging Scheduling on Distribution Transformer Lifespan," *Heliyon* 10, no. 21 (2024): e39904, https://doi.org/10.1016/j.heliyon.2024.e39904.

4. Z. Liu, Y. Chen, R. Zhuo, and H. Jia, "Energy Storage Capacity Optimization for Autonomy Microgrid Considering CHP and EV Scheduling," *Applied Energy* 210 (2018): 1113–1125, https://doi.org/10.1016/j.apenergy.2017.07.002.

5. A. Zhang, Q. Liu, J. Liu, and L. Cheng, "CASA: Cost-Effective EV Charging Scheduling Based on Deep Reinforcement Learning," *Neural Computing & Applications* 36, no. 15 (2024): 8355–8370, https://doi.org/10.1007/s00521-024-09530-3.

6. T. Long, Q.-S. Jia, G. Wang, and Y. Yang, "Efficient Real-Time EV Charging Scheduling via Ordinal Optimization," *IEEE Transactions on Smart Grid* 12, no. 5 (2021): 4029–4038, https://doi.org/10.1109/tsg.2021.3078445.

7. Y. Kabiri-Renani, A. Arjomandi-Nezhad, M. Fotuhi-Firuzabad, and M. Shahidehpour, "Transactive-Based day-ahead Electric Vehicles Charging Scheduling," *IEEE Transactions on Transportation Electrification* 10, no. 4 (2024): 8235–8245, https://doi.org/10.1109/tte.2023.3348490.

8. H. Li, Z. Wan, and H. He, "Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning," *IEEE Transactions on Smart Grid* 11, no. 3 (2019): 2427–2439, https://doi.org/10.1109/tsg.2019.2955437.

9. B. R. Kiran, I. Sobh, V. Talpaert, et al., "Deep Reinforcement Learning for Autonomous Driving: A Survey," *IEEE Transactions on Intelligent Transportation Systems* 23, no. 6 (2021): 4909–4926, https://doi.org/10.1109/tits.2021.3054625.

10. Y. Gu, Z. Liu, S. Dai, et al., "Deep Reinforcement Learning for Job Scheduling and Resource Management," in *Cloud Computing: An Algorithm-Level Review* (arXiv preprint arXiv:2501.01007, 2025).

11. D. Silver, A. Huang, C. J. Maddison, et al., "Mastering the Game of Go With Deep Neural Networks and Tree Search," *Nature* 529, no. 7587 (2016): 484–489, https://doi.org/10.1038/nature16961.

12. H. Xu, A. Zhang, Q. Wang, Y. Hu, F. Fang, and L. Cheng, "Quantum Reinforcement Learning for Real-Time Optimization in Electric Vehicle Charging Systems," *Applied Energy* 383 (2025): 125279, https://doi.org/10.1016/j.apenergy.2025.125279.

13. F. Ming, F. Gao, K. Liu, and X. Li, "A Constrained DRL-based bi-level Coordinated Method for Large-Scale EVs Charging," *Applied Energy* 331 (2023): 120381, https://doi.org/10.1016/j.apenergy.2022.120381.

14. J. Zhang, Y. Guan, L. Che, and M. Shahidehpour, "EV Charging Command Fast Allocation Approach Based on Deep Reinforcement Learning With Safety Modules," *IEEE Transactions on Smart Grid* 15, no. 1 (2023): 757–769, https://doi.org/10.1109/tsg.2023.3281782.

15. H. Ding, Y. Xu, B. C. S. Hao, Q. Li, and A. Lentzakis, "A Safe Reinforcement Learning Approach for Multi-Energy Management of Smart Home," *Electric Power Systems Research* 210 (2022): 108120, https://doi.org/10.1016/j.epsr.2022.108120.

16. L. Chen, Y. Tao, A. M. Lopes, M.-Y. Chow, and Y. Chen, "Safety-Optimized Fast Charging of Lithium-ion Battery Based on Distributional SAC-Conservative Augmented Lagrangian SDRL Algorithm," *IEEE Transactions on Vehicular Technology* 74, no. 8 (2025): 1–13, https://doi.org/10.1109/tvt.2025.3551661.

17. J. Fan, A. Liebman, and H. Wang, "Safety-Aware Reinforcement Learning for Electric Vehicle Charging Station Management in Distribution Network," in *2024 IEEE Power & Energy Society General Meeting*, (IEEE, 2024), 1–5.

18. X. Yang, H. He, Z. Wei, R. Wang, K. Xu, and D. Zhang, "Enabling Safety-Enhanced Fast Charging of Electric Vehicles via Soft Actor critic-Lagrange DRL Algorithm in a cyber-physical System," *Applied Energy* 329 (2023): 120272, https://doi.org/10.1016/j.apenergy.2022.120272.

19. T. Wu, A. Scaglione, and D. Arnold, "Constrained Reinforcement Learning for Predictive Control in Real-Time Stochastic Dynamic Optimal Power Flow," *IEEE Transactions on Power Systems* 39, no. 3 (2023): 5077–5090, https://doi.org/10.1109/tpwrs.2023.3326121.

20. Y. Ye, H. Wang, P. Chen, Y. Tang, and G. Strbac, "Safe Deep Reinforcement Learning for Microgrid Energy Management in Distribution Networks With Leveraged Spatial–Temporal Perception," *IEEE Transactions on Smart Grid* 14, no. 5 (2023): 3759–3775, https://doi.org/10.1109/tsg.2023.3243170.

21. M. Zhang, G. Guo, S. Magnússon, R. C. Pilawa-Podgurski, and Q. Xu, "Data Driven Decentralized Control of Inverter Based Renewable Energy Sources Using Safe Guaranteed multi-agent Deep Reinforcement Learning," *IEEE Transactions on Sustainable Energy* 15, no. 2 (2023): 1288–1299, https://doi.org/10.1109/TSTE.2023.3341632.

22. Z. Yi, X. Wang, C. Yang, C. Yang, M. Niu, and W. Yin, "Real-Time Sequential security-constrained Optimal Power Flow: A Hybrid Knowledge-Data-Driven Reinforcement Learning Approach," *IEEE Transactions on Power Systems* 39, no. 1 (2023): 1664–1680, https://doi.org/10.1109/tpwrs.2023.3262843.

23. Y. Jin, M. A. Acquah, M. Seo, S. Ghorbanpour, S. Han, and T. Jyung, "Optimal EV Scheduling and Voltage Security via an Online Bi-Layer Steady-State Assessment Method Considering Uncertainties," *Applied Energy* 347 (2023): 121356, https://doi.org/10.1016/j.apenergy.2023.121356.

24. A. Ajagekar and F. You, "Deep Reinforcement Learning Based Unit Commitment Scheduling Under Load and Wind Power Uncertainty," *IEEE Transactions on Sustainable Energy* 14, no. 2 (2022): 803–812, https://doi.org/10.1109/tste.2022.3226106.

25. P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-Shielded Multi-Agent Deep Reinforcement Learning for Safe Active Voltage Control With Photovoltaic/Battery Energy Storage Systems," *IEEE Transactions on Smart Grid* 14, no. 4 (2022): 2656–2667, https://doi.org/10.1109/tsg.2022.3228636.

26. B. Xiong, L. Zhang, Y. Hu, F. Fang, Q. Liu, and L. Cheng, "Deep Reinforcement Learning for Optimal Microgrid Energy Management With Renewable Energy and Electric Vehicle Integration," *Applied Soft Computing* 176 (2025): 113180, https://doi.org/10.1016/j.asoc.2025.113180.

27. K. Chaudhari, N. K. Kandasamy, A. Krishnan, A. Ukil, and H. B. Gooi, "Agent-Based Aggregated Behavior Modeling for Electric Vehicle Charging Load," *IEEE Transactions on Industrial Informatics* 15, no. 2 (2018): 856–868, https://doi.org/10.1109/tii.2018.2823321.

28. S. Bricka, T. Reuscher, P. Schroeder, et al., "Summary of Travel Trends: 2022 National Household Travel Survey," *Federal Highway Administration* tech. rep., (2022).

29. A. Subramanian, M. J. Garcia, D. S. Callaway, K. Poolla, and P. Varaiya, "Real-Time Scheduling of Distributed Resources," *IEEE*

Transactions on Smart Grid 4, no. 4 (2013): 2122–2130, https://doi.org/10.1109/tsg.2013.2262508.

30. Y. Yang, S. Yang, H. Moon, and J. Woo, "Analyzing Heterogeneous Electric Vehicle Charging Preferences for Strategic time-of-use Tariff Design and Infrastructure Development: A Latent Class Approach," *Applied Energy* 374 (2024): 124074, https://doi.org/10.1016/j.apenergy.2024.124074.

31. L. Lin, K. Ou, Q. Lin, J. Xing, and Y.-X. Wang, "Two-Stage Multi-Strategy Decision-Making Framework for Capacity Configuration Optimization of Grid-Connected PV/Battery/Hydrogen Integrated Energy System," *Journal of Energy Storage* 97 (2024): 112862, https://doi.org/10.1016/j.est.2024.112862.

32. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.-06347* (2017).