Collaborative Stability and Benefit Optimization of V2G Based on Multi-agent Graph Reinforcement Learning with Privacy Protection

Qingwei Tang, Wei Sun, *Senior Member, IEEE,* Zhi Liu, *Senior Member, IEEE,* Yang Xiao, *Fellow, IEEE,* Xiaohui Yuan, *Senior Member, IEEE,* Chanjuan Zhao

***Abstract*—The adoption of renewable energy sources (RESs) has reshaped power systems, causing voltage violations and power imbalances. Electric vehicles (EVs) add to this burden but also offer potential active and reactive power support. Distribution system operators (DSOs) need effective distributed control to manage uncertainties and address privacy concerns. Existing research has mainly relied on model-based methods to solve active and reactive power control problems. However, these studies often neglect the privacy issues of EV owners and lack effective incentives to encourage users to participate in power regulation. This paper proposes a supply-side and demand-side cooperative control framework based on Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MATD3). This framework uses Graph Convolutional Networks (GCN) as the state encoder and combines a regional-global joint supervision strategy to construct the safety control policy, while the Federated-EVs control scheme is employed to protect user privacy and enhance user satisfaction. Simulation results show that the proposed algorithm demonstrates superior performance and advantages in peak voltage control, reducing peak power loss, meeting user demands, and improving user benefits.**

***Index Terms*—Active and Reactive Power Control, Electric Vehicles, EV, Multi-agent Reinforcement Learning, Federated Learning, Graph Convolutional Networks, Multi-objective optimization**

I. Introduction

With the rapid growth in the number of electric vehicles (EVs) [1], it is expected that their market share will exceed 20% of the total number of vehicles within the next decade [2]. However, the large-scale integration of EVs into active distribution networks (ADN) poses significant challenges to systems. On the one hand, EVs' charging demand significantly affects power systems' load characteristics; on the other hand, charging behaviors exhibit high uncertainty due to variations in users' driving patterns, potentially leading to system instability [3]. Moreover, global energy decarbonization has accelerated in recent years, driving the transition from fossil fuels to clean renewable energy sources (RESs). In this context, RESs such as photovoltaic (PV) and wind power (WT) have been widely integrated into the ADN. However, due to the variability and intermittency of renewable energy, their rapid deployment may lead to voltage violations [4].

Integrating EVs and RESs increases grid operational complexity, but their flexibility offers new scheduling opportunities for Distribution System Operators (DSOs) [5]. PV inverters and WT can help maintain voltage stability through dynamic reactive power regulation [6]. While EVs can be flexible scheduling units, their full potential depends on user participation. It is critical to incentivize users to allow their EVs to participate in grid scheduling tasks. This requires addressing two key challenges: (1) protecting user privacy and (2) providing sufficient economic incentives. Many regulation strategies necessitate real-time monitoring of users' electricity consumption and charging behavior, raising privacy concerns [7]. Additionally, users are more likely to cooperate if they receive significant economic benefits [8]. Efficiently scheduling power resources for ADN stability while safeguarding privacy remains a critical and challenging problem, highlighting the need for intelligent coordinated control algorithms.

A. Literature Review

1) Traditional Control and Optimization Approaches: Traditional voltage regulation in ADNs relies on On-Load Tap Changers (OLTCs) and Circuit Breakers (CBs), which suffer from mechanical inertia and limited responsiveness [9]. Optimization-based methods, such as Optimal Power Flow (OPF) [10] and droop control, require accurate system modeling and deterministic parameters, making them difficult to adapt to increasingly dynamic and uncertain ADN conditions. These limitations significantly reduce their practical applicability in real-time voltage regulation scenarios.

2) DRL-Based Voltage Control: To overcome the reliance on accurate models, deep reinforcement learning (DRL) has emerged as a data-driven, model-free approach for voltage control [12]. [13] applies a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm for autonomous voltage regulation, but lacks policy constraint mechanisms, risking

This work was supported in part by the National Natural Science Foundation of China under Grant 52277087, 62173120, and 52077049, in part by the Natural Science Foundation of Anhui Province under Grant 2108085UD07, 2108085UD11, 2008085UD04. *(Corresponding author: Wei Sun and Zhi Liu.)*

Qingwei Tang, Wei Sun are with the School of Electrical and Automation Engineering, Hefei University of Technology, Hefei, Anhui, 230009, China (e-mail: tangqingwei@mail.hfut.edu.cn, wsun@hfut.edu.cn).

Zhi Liu is with Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, 182-8585, Japan (e-mail: liu@ieee.org).

Yang Xiao is with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: yangxiao@ieee.org).

Xiaohui Yuan is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76207 USA (e-mail: xiaohui.yuan@unt.edu).

Chanjuan Zhao is with Anhui University, Hefei 230039, China (e-mail:chanjuanzhao@ahu.edu.cn).

unsafe actions. [14] introduces a Lipschitz-constrained neural network controller to improve safety, yet overlooks interactions across ADN subregions. [15] proposes a constrained Soft Actor-Critic (SAC) method to limit agent actions within safe zones, enhancing local security but lacking inter-zone coordination. Similarly, [16] introduces a consensus-based MARL framework to align neighboring agents' policies, yet its localized design hinders scalability in ADNs with complex topologies and high EV penetration.

However, most of these methods are limited by their inability to jointly optimize global and local control. As EVs proliferate in ADNs, this becomes more critical, since traditional DRL struggles with the high-dimensional, time-varying interactions between distributed EV agents and the grid [17].

3) Vehicle-to-Grid Coordination: EVs, through Vehicle-to-Grid (V2G) and Grid-to-Vehicle (G2V) technologies, can regulate both active and reactive power, enabling load flattening and improving energy utilization [18], [19]. For example, [20] designs an onboard V2G charger with flexible power factor support to improve bidirectional power flow feasibility, but it focuses only on hardware, lacking system-level coordination. [21] shows V2G's economic potential but assumes static pricing and full user compliance, overlooking behavioral uncertainties and incentives. [22] proposes a real-time coordination strategy for ADNs, yet relies on centralized control, ignoring privacy, latency, and scalability challenges under high EV integration.

Nonetheless, effective EV participation requires access to users' charging behaviors, locations, and power needs. Without proper handling, this dependency can lead to significant privacy concerns and user reluctance. Existing schemes often lack mechanisms to balance user benefits with grid regulation demands.

4) Federated Learning in Power Systems: Federated learning (FL) enables local model training without sharing raw data [23]. It has been applied to residential demand response [24], [25], multi-microgrid coordination [26], and combined with transfer learning to address data scarcity [27]. For V2G, studies [28]–[30] integrate FL with MARL to protect charging behavior privacy. The above studies assume static, homogeneous environments and struggle with dynamic EV scenarios. They also neglect the joint optimization of privacy, user incentives, and grid performance, limiting practical applicability.

However, these methods face limitations in dynamic EV-integrated ADNs, with challenges in real-time agent cooperation and maintaining learning performance amid mobility and heterogeneity. Coordination among user privacy, federated training, and grid stability control also remains underexplored.

5) Graph Neural Networks for Power Systems: The graph-structured nature of ADNs makes Graph Neural Networks (GNNs) particularly suitable for modeling spatial dependencies. Graph Convolutional Networks (GCNs), in particular, have shown promise in voltage control tasks by accurately capturing the relationships among buses and branches [31]. However, existing applications of GCNs are typically restricted to static or homogeneous settings. They often struggle with inconsistent feature dimensions across different ADN zones and fail to explicitly bridge global and local observations, which is crucial in distributed DRL applications.

It's worth noting that [42] proposes a DRL-based algorithm to optimize charging station revenue, user comfort, and waiting costs, enabling real-time scheduling of EVs. In [43], a GCN-based MARL algorithm is introduced for voltage and reactive power control in ADN. By employing regional partitioning and distributed control under partial observability, it achieves localized optimization. [44] develops a V2G scheduling framework that integrates edge computing with FL, effectively enhancing model aggregation efficiency and significantly reducing communication and computation overhead while preserving user privacy.

B. Contributions

To address the aforementioned issues, this paper proposes an innovative MARL method for the active and reactive power control of scheduling devices and EVs in the context of widespread RES integration. The main contributions are summarized as follows:

1) We model the active and reactive power control of scheduling devices and EVs in ADN as a decentralized partially observable Markov decision process (Dec-POMDP). This allows supply-side devices and demand-side EVs to fully utilize their power for a stable active power supply and reactive power compensation.
2) We propose a GCN-based module for global and local state encoding in ADNs. Considering the coupling between ADN sub-regions and the global system, we employ the KL divergence loss function to enforce consistency between global and local policies, thereby preventing unsafe control strategies.
3) We propose an FL-based training mechanism that enables each EV agent to learn independently at the local level without sharing state data, thereby reducing the risk of data privacy leakage. Furthermore, by incorporating a user incentive mechanism, we enhance user satisfaction and encourage active participation in the stable control of the ADN.
4) Based on MATD3, we propose DS-FGMATD3, a joint supply-demand control framework that stabilizes power fluctuations from RES, optimizes EV owners' benefits and comfort, and enhances overall system stability, energy efficiency, and user experience.
5) In the appendix, we integrate a battery degradation cost model into the reward function to protect the long-term interests of EV users. Furthermore, we assess the robustness of two FL aggregation methods under scenarios involving malicious parameter injection.

C. Paper Organization

The rest of the paper is organized as follows. Section II presents the real-world ADN system model with heterogeneous DERs and the Markov decision process (MDP) we developed. Section III introduces the MARL algorithm used to address the problem. Section IV discusses the experimental setup and presents the results of the case studies. Section V concludes the paper.

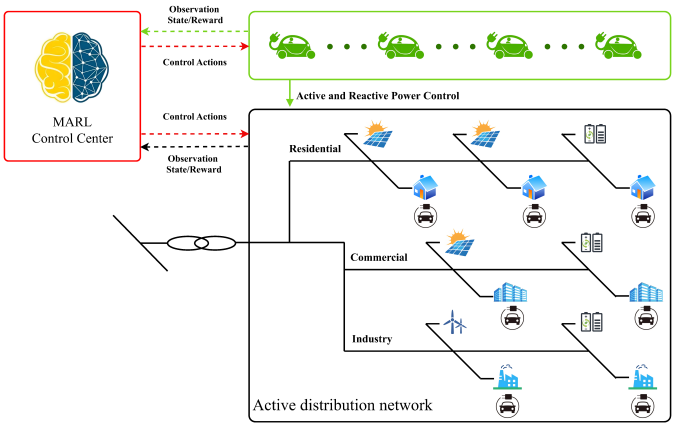


Fig. 1. Active distribution network with multiple renewable energy devices

II. Preliminaries

EVs are an emerging technology that helps DSOs manage the ADN, as shown in Figure 1. Modern ADN systems integrate various distributed RESs, such as PV and WT. Users drive EVs to work in the morning and return home in the evening, keeping them connected to charging stations at both locations. As controllable ADN units, these EVs are coordinated by the DSO through OPF to maintain grid stability.

This paper has two main objectives. On the supply side, we regulate controllable devices in the ADN, including inverters, static var compensators (SVC), energy storage systems (ESS), and EVs, to ensure voltage stability and minimize power losses. On the demand side, we optimize EV scheduling to maximize user benefits, protect privacy, and support DSOs in maintaining power system stability.

A. ADN stability control problem formulation

To simulate real electricity demand and the ADN environment, we divide the ADN into residential, commercial, and industrial zones, each equipped with PV, WT, ESS, and EV charging stations. The active power from renewable energy systems may lead to voltage violations, which can be mitigated by adjusting the reactive power of the inverter and SVC, thereby reducing power losses and maintaining voltage stability.

An ADN is typically modeled as a tree structure, where $V = 0, 1, \ldots, N$ represents $N+1$ buses, and $E = 1, 2, \ldots, L$ represents L branches. For each $i \in N$, let v_i denote the voltage magnitude, θ_i represents the phase angle, and $s_i = p_i + jq_i$ represents the injected complex power. The injected active power is expressed as $p_i = p_i^g - p_i^l$, where p_i^g is the active power generated by the energy generation system, and p_i^l is the active power demanded by the load. The reactive power injection is $q_i = q_i^g - q_i^l$, where q_i^g is the reactive power from the inverters and SVCs, and q_i^l is the reactive power demanded by the load. The injected active and reactive power are then given by Equations 1 and 2.

$$p_i = v_i^2 \sum_{j \in V_i} g_{ij} - v_i \sum_{j \in V_i} v_j \left(g_{ij} \cos\theta_{ij} + b_{ij} \sin\theta_{ij}\right) \quad (1)$$

$$q_i = -v_i^2 \sum_{j \in V_i} b_{ij} + v_i \sum_{j \in V_i} v_j \left(g_{ij} \sin\theta_{ij} + b_{ij} \cos\theta_{ij}\right) \quad (2)$$

where $i \in V \setminus \{0\}$, and $V_i := \{j \mid (i,j) \in E\}$ represents the set of indices of buses connected to bus i. The parameters g_{ij} and b_{ij} represents the conductance and susceptance on branch (i,j), respectively, and $\theta_{ij} = \theta_i - \theta_j$ represents the phase angle difference between buses i and j. The objective of voltage control in the ADN is to mitigate severe voltage fluctuations and reduce power losses by regulating the reactive power of inverters or SVCs within the system. Equation 3 presents the objective function of voltage control.

$$\min_{Q_{i,t}} \sum_{t \in T} \left\{ \sum_{(i,j) \in \ell} P_{ij,t}^{\text{loss}} + \|\mathbf{v}(t) - \mathbf{v}_{\text{ref}}\| \right\} \quad (3)$$

$$P_{ij,t}^{\text{loss}} = r_{ij} l_{ij,t} \quad (4)$$

$$Q_{i,t}^{SVC/Inv,Min} \leq Q_{i,t}^{SVC} \leq Q_{i,t}^{SVC/Inv,Max}, \forall i,t \quad (5)$$

$$(Q_{i,t}^{PV/WT})^2 + (P_{i,t}^{PV/WT})^2 \leq (S_{i,\ max}^{PV/WT})^2, \forall i,t \quad (6)$$

$$0 \leq P_{i,t}^{esc} \leq \overline{P_i^{es}}, -\overline{P_i^{es}} \leq P_{i,t}^{esd} \leq 0, \underline{E_i^{es}} \leq E_{i,t}^{es} \leq \overline{E_i^{es}}, \forall i,t \quad (7)$$

$$E_{i,t+1}^{es} = E_{i,t}^{es} + P_{i,t}^{esc} \eta_i^{es} - \frac{P_{i,t}^{esd}}{\eta_i^{es}} \quad (8)$$

$$P_{i,t}^{esc} \cdot P_{i,t}^{esd} = 0 \quad (9)$$

$$\underline{\mathbf{v}} \leq \mathbf{v}(t) \leq \overline{\mathbf{v}}. \quad (10)$$

The power loss in the ADN is calculated using Equation 4, where r_{ij} represents the resistance of branch i, j, and l_{ij} represents the square of the current magnitude from node i to node j. Equation 5 constrains the reactive power of the compensation equipment within the lower bound $Q_{i,t}^{\text{SVC/Inv, Min}}$ and the upper bound $Q_{i,t}^{\text{SVC/Inv, Max}}$. The active and reactive power outputs of each PV or WT generation unit are constrained by the corresponding apparent power $S_{i,\text{max}}^{\text{PV/WT}}$. In Equation 7, $E_{i,t}^{es}$ represents the energy state of storage unit i at time step t, while $P_{i,t}^{esc}$ and $P_{i,t}^{esd}$ represents the charging and discharging powers, respectively. $E_{i,t}^{es}$ also represents the energy range constraint. Equation 8 models the time-coupling characteristics of the energy storage device, where $\eta_{i,k}^{es}$ indicates the charging efficiency. Equation 9 represents the mutually exclusive constraint for ESS, and Equation 10 defines the voltage operation constraints in the ADN, where the lower bound $\underline{\mathbf{v}}$ and upper bound $\bar{\mathbf{v}}$ specify the allowable voltage range for node i.

B. EVs regulation problem formulation

Users' travel patterns and electricity prices impact EV charging and discharging. V2G systems require adaptive control strategies to balance user needs while optimizing energy flow, reducing peak demand, and storing off-peak energy. This paper investigates the role of V2G in integrating renewables,

mitigating wind and solar intermittency, and improving RES utilization. We assume that all EVs regulate both active and reactive power, with the primary demand-side objective being to maximize the benefits for EV owners, as shown in Equation 11.

$$\begin{aligned}\max \sum_{t} \sum_{i=1}^{N} \left(p_{s,t} \cdot P_{i,t}^{evd} - p_{b,t} \cdot P_{i,t}^{evc}\right) \\ + \sum_{t} \sum_{i=1}^{N} \left(q_{s,t} \cdot Q_{i,t}^{evc} - q_{b,t} \cdot Q_{i,t}^{evd}\right)\end{aligned} \tag{11}$$

where $p_{s,t}$ represents the active power selling price at time t, and $P_{i,t}^{evd}$ represents the active power injected into the grid by EV i at time t. Similarly, $p_{b,t}$ represents the active power buying price at time t, and $P_{i,t}^{evc}$ represents the active power absorbed by the EV from the grid at time t. $q_{s,t}$ is the reactive power selling price at time t, and $Q_{i,t}^{evc}$ represents the reactive power injected into the grid by the EV at time t; $q_{b,t}$ represents the reactive power buying price at time t, and $Q_{i,t}^{evd}$ represents the reactive power absorbed by the EV from the grid at time t. The constraints for the V2G problem described in this paper are defined as follows.

$$0 \leq P_{i,t}^{evc} / P_{i,t}^{evd} \leq u_{i,t} \cdot \overline{P_{i,t}^{evc}} / \overline{P_{i,t}^{evd}} \tag{12}$$

$$\underline{Q_{i,t}^{ev}} \leq Q_{i,t}^{ev} \leq \overline{Q_{i,t}^{ev}} \tag{13}$$

$$\sqrt{\left(P_{i,t}^{ev}\right)^2 + \left(Q_{i,t}^{ev}\right)^2} \leq S_{i,max}^{ev} \tag{14}$$

$$P_{i,t}^{evc} \cdot P_{i,t}^{evd} = 0 \tag{15}$$

$$E_{i,t+1}^{ev} = E_{i,t}^{ev} + \eta_i^{evc} \cdot P_{i,t}^{evc} \cdot \Delta \mathrm{t} - \frac{P_{i,t}^{evd} \cdot \Delta \mathrm{t}}{\eta_i^{evd}} \tag{16}$$

where $P_{i,t}^{evc} / P_{i,t}^{evd}$ in Equation 12 represents the charging and discharging active power of EV i at time period t. In Equation 13, $Q_{i,t}^{ev}$ represents the reactive power regulation range constraint. In Equation 14, $S_{i,max}^{ev}$ represents the maximum apparent power of EV i. Equation 15 describes the mutual exclusivity constraint for charging and discharging EV i. In Equation 16, $E_{i,t+1}^{ev}$ represents the state of battery energy storage of EV i at time $t+1$, and η_i^{evc} and η_i^{evd} represents the charging and discharging efficiencies of EV i.

C. EV benefit-oriented ADN stability control formulation

The widespread adoption of PV systems, WT, and ESS complicates voltage control due to the intermittent nature of RESs. At the same time, the increasing deployment of EVs and charging stations presents new opportunities for power management. EVs can function as flexible load regulation devices, assisting DSOs in mitigating grid inefficiencies and fluctuations in renewable energy generation, while also providing a new revenue stream for EV owners.

This paper presents a dynamic ADN model with heterogeneous distributed energy resources (DERs) to simulate real-world scenarios. Reactive power regulation devices, such as inverters and SVCs, are deployed within the ADN to enhance stability. We assume that EV users' residences and businesses are distributed across various ADN nodes, with users connecting their EVs to charging stations according to commuting schedules, resulting in dynamic load characteristics. Due to privacy concerns, the DSO lacks critical information about EV states, complicating its role in ADN stability control. To address this, we propose an FL method, where local model parameters are uploaded and processed by the DSO, ensuring privacy and security while enabling efficient ADN regulation. Furthermore, we introduce an EV control algorithm based on FL and MARL to optimize power operation and management.

D. Decentralized Partially Observable Markov Decision Process

DSOs manage power demand by regulating macro-level ADN parameters. However, the increasing share of RESs and their intermittent nature complicate ADN management. We model the EV cooperative ADN stability control problem as a discrete-time step Dec-POMDP problem. The Dec-POMDP of this paper is defined as $\langle I, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$, where I represents the number of agents (ESS, PV/WT inverters, SVC, and EV), $\mathcal{S}$ is the global state, $s \in \mathcal{S}$, $o_i \in \mathcal{O}_{1:I}$ is the local observation of agent i, $\mathcal{A}_{1:I}$ is the action set, $a_i \in \mathcal{A}_{1:I}$, $r_i \in \mathcal{R}_{1:I}$ is the reward function, and $\mathcal{T}(s, o_{1:I}, a_{1:I}, \omega)$ is the state transition function, where ω represents uncertainty parameters in the environment. The time interval between two consecutive time steps is $\Delta t = 3$ minutes. Based on its control strategy $\mu(o)$, the agent i selects an action $a_{i,t}$ at a time step t based on its local observation $o_{i,t}$. The environment transitions to the next state s_{t+1} according to the state transition function $\mathcal{T}$. Agent i receives a reward $r_{i,t}$ and obtains a new local observation $o_{i,t+1}$. The interaction process continues, generating the observation, action, and reward trajectories for each agent: $\tau_i = o_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, \ldots, r_{i,T}$, where $\mathcal{O}_i \times \mathcal{A}_i \times \mathcal{O}_i \to \mathbb{R}$. The objective of each agent is to maximize its cumulative discounted reward $R_i = \sum_{t=1}^{T} \gamma^t r_{i,t}$. The experiments in this paper are conducted on the IEEE standard bus model.

Before modeling the RL task, we make the following assumptions:

1) Smart charging stations are deployed at every node, enabling EVs to exchange energy with the ADN through these stations.
2) The charging and discharging behavior of the EVs is controllable by the DSO, with the ability to dynamically adapt to system requirements.
3) Each charging station has access to information about the currently connected EVs, allowing the DSO to implement control strategies via the stations.
4) Agents in different ADN zones can receive state information through the distributed public communication network and take corresponding actions.
5) All EV observations and decisions remain local. Charging stations only execute commands and do not share personal data—such as SOC or travel patterns—with the DSO. EV users can disconnect from the system at any time.

The components of the Dec-POMDP are detailed as follows.

1) Local observation: The observation space of each agent is constructed based on the zoning results of the electricity usage type, and it is a subset of the global observation. In this paper, $o_{i,t} \in \mathcal{O}_i$ is defined as $\left\{\mathbf{p}_{e,k}^{L}, \mathbf{q}_{e,k}^{L}, \mathbf{p}_{e,k}^{RES}, \mathbf{q}_{e,k}^{RES}, \mathbf{v}_{e,k}, \theta_{e,k}, \mathbf{S}_{i,t}^{\mathrm{EV}}, \pi_{e,k}\right\}$, where $\mathbf{p}_{e,k}^{L}$ and $\mathbf{q}_{e,k}^{L}$ represent the active power and reactive power of the load at the e-th node in the k-th region, respectively. $\mathbf{p}_{e,k}^{RES}$ and $\mathbf{q}_{e,k}^{RES}$ represent the active power and reactive power generated by RESs at the e-th node in the k-th region, respectively. $\mathbf{v}_{e,k}$ and $\theta_{e,k}$ represent the voltage magnitude and phase angle at the e-th node in the k-th zone, respectively. $\mathbf{S}_{i,t}^{\mathrm{EV}}$ denotes the real-time battery capacity of the EV i. $\pi_{e,k}$ represents the real-time electricity price at the e-th node in the k-th zone.
2) Action: The agents are selected as power regulation devices such as ESS, PV/WT inverters, and EVs. For the agents deployed on PV/WT inverters and SVC, the continuous action space of each agent i is given by $\mathcal{A}_i = \{a_{i,t} : -c \le a_{i,t} \le c\}$, where $a_{i,t}$ represents the ratio of the maximum reactive power that can be generated at the time step t. The reactive power generated by the $i-th$ PV/WT inverter or SVC is calculated using $q_i^{PV} = a_{i,t}\sqrt{\left(s_i^{max}\right)^2 - \left(p_i^{max}\right)^2}$, where s_i^{max} is the maximum apparent power at the $i-th$ node, which depends on the inverter's physical capacity. When $a_{i,t} > 0$, the PV/WT inverter injects reactive power into ADN; otherwise, it absorbs reactive power from ADN. The load tolerance of the ADN determines the value of parameter c. For the agents deployed in ESS and smart charging station, each agent i at time step t controls its action $a_{i,t} = [a_{i,t}^p, a_{i,t}^q] \in \mathcal{A}_i$, adjusting both active and reactive power. The action consists of two parts: the active power action $a_{i,t}^p \in [-1, 1]$, representing the magnitude of active power for charging (positive) or discharging (negative), and its value is the percentage of the active power capacity $P_{i,t}^{ev} \in [-P_{iev}, P_{iev}]$. The reactive power action $a_{i,t}^q \in [-1, 1]$, represents the magnitude of reactive power absorption (positive) or injection (negative). Its value is a percentage of the reactive power limit $Q_{i,t}^{ev} \in \left[-\sqrt{\left(\overline{S_i^{ev}}\right)^2 - \left(\overline{P_i^{ev}}\right)^2}, \sqrt{\left(\overline{S_i^{ev}}\right)^2 - \left(P_i^{ev}\right)^2}\right]$, where D is the set of all agents.
3) State transfer: The state transition is described by the formula $s_{t+1} = \mathcal{T}(s_t, o_{i,t}, a_{1:I,t}, \omega_t)$, which is influenced by the current environmental state s_t, the local observations $o_{i,t}$ and actions $a_{i,t}$ of all agents, as well as the stochastic factors in the environment $\omega_t = [\pi_t^P, \pi_t^Q, P_{i,t}^L, Q_{i,t}^L, \overline{P_{g,t}^{RES}}]$. In this scenario, ω_t reflects the randomness of lighting, wind energy fluctuations, node electricity price variations, and the stochastic nature of user electricity consumption behavior. The ADN's active and reactive power prices, π_t^P and π_t^Q, depending on the power market. The active and reactive power demand of residential electricity, $P_{i,t}^L$ and $Q_{i,t}^L$, are influenced by user behavior. The renewable energy generation, $\overline{P_{g,t}^{RES}}$, is affected by fluctuations in solar irradiation and wind speed. We express the active power of EV as $P_{i,t}^{ev} = P_{i,t}^{evc} + P_{i,t}^{evd}$, where $P_{i,t}^{evc}$ and $P_{i,t}^{evd}$ represent the charging and discharging active power, respectively. We define a time-varying availability variable $u_{i,t}$, where $u_{i,t} = 1$ indicates that the EV is connected to a charging station at time t, and $u_{i,t} = 0$ indicates that it is not connected. Note that the EV cannot charge and discharge simultaneously, so they are described separately:

$$\begin{cases} P_{i,t}^{evc} = min\left(a_{i,t}^{evp}\overline{P_i^{ev}}, \frac{\overline{E_i^{ev}} - E_{i,t}^{ev}}{\eta_i^{esc}\Delta t}\right), & if P_{i,t}^{evc} => 0, \\ P_{i,t}^{evc} = 0, & otherwise. \end{cases}$$

$$\begin{cases} P_{i,t}^{evd} = max\left(a_{i,t}^{evp}\overline{P_i^{ev}}, -\frac{E_{i,t}^{e}\eta_i^{esd}}{\Delta t}\right), & if P_{i,t}^{evd} =< 0, \\ P_{i,t}^{evd} = 0, & otherwise. \end{cases}$$

where η_i^{esc} and η_i^{esd} represent the charging and discharging efficiencies, respectively. In addition, the energy change $E_{i,t}^{ev}$ of an EV from time step t to $t+1$ is shown in Equation 17.

$$E_{i,t+1}^{\mathrm{ev}} = \begin{cases} E_{i,t}^{\mathrm{ev}} + \left(P_{i,t}^{evc}\eta_i^{evc} - \frac{P_{i,t}^{evd}}{\eta_i^{evd}}\right)\Delta t, & \text{if } u_{i,t} = 1 \\ E_{i,t}^{\mathrm{ev}} - E_i^{\mathrm{tra}}, & \text{if } u_{i,t} = 0 \end{cases} \tag{17}$$

4) Reward function: The reward function consists of four components to optimize the coordination between EVs and the ADN: 1) penalizing voltage deviations to ensure stability, 2) minimizing power losses to enhance efficiency, 3) maximizing user benefits through optimal EV charging and discharging, and 4) applying a penalty to the battery state to ensure sufficient capacity for travel. These objectives and constraints facilitate efficient interaction between EVs and the grid, as demonstrated in Equation 18.

$$\begin{aligned} R_d = -\Bigg[& \frac{1}{|V|}\sum_{i\in V} l_v(v_{i,t}) + \alpha \cdot l_q(q^{\mathrm{RES}}) \\ & - \beta \cdot \left(\pi_t^P \cdot P_{i,t}^{\mathrm{ev}} + \pi_t^Q \cdot Q_{i,t}^{\mathrm{ev}}\right) - l_s(E_{i,t}^{\mathrm{ev}}) \Bigg] \end{aligned} \tag{18}$$

where $l_v(\cdot)$ is a voltage barrier function, the L3-shape function [12] is adopted as the form of the voltage barrier function, and $l_q(\mathbf{q}^{RES}) = \frac{1}{|D|}\|\mathbf{q}^{RES}\|_1$ represents the reactive power generation loss. Since it is difficult to obtain the power loss of the entire power grid in practice, $l_q(\cdot)$ is used in the simulation environment as an easily computable approximation of power loss. The barrier function $l_v(\cdot)$ describes the degree of deviation of the bus voltage from v_{ref}, where higher values indicate greater deviations. $\alpha, \beta \in (0, 1)$ is used to balance multiple targets in the reward function, which we empirically set to 0.1 and 0.5, respectively. Compared to constrained [45] or hierarchical [46] MARL approaches, the scalar reward design employed in this work achieves a better balance of simplicity, efficiency, and interpretability.

Our objectives—such as voltage regulation, power loss minimization, and user benefit—are quantifiable and can be effectively combined into a single reward signal. Meanwhile, key operational constraints (e.g., voltage and power limits) are incorporated into the environment or action space, ensuring physical feasibility during training. While hierarchical MARL is better suited for multi-timescale or multi-level decision problems, our setting involves agents operating on the same temporal and control scale. Therefore, a flat MARL structure with scalar rewards offers a practical and robust solution for our decentralized, privacy-aware framework. The mathematical expression of the L3-shape function is given in Equation 19.

$$l_v\left(v_{i,t}\right)=\begin{cases}a\cdot\left|v_{i,t}-v_{\mathrm{ref}}\right|-b, & \text{if } \left|v_{i,t}-v_{\mathrm{ref}}\right|>0.05\\ -c\cdot\mathcal{N}\left(v_{i,t}\mid v_{\mathrm{ref}},0.1\right)+d, & \text{otherwise}\end{cases} \tag{19}$$

where the hyperparameters a, b, c, and d are set to 2, 0.095, 0.01, and 0.04, respectively. $\mathcal{N}\left(v_{i,t}\mid v_{\mathrm{ref}},0.1\right)$ denotes the probability density function of a normal distribution with mean $v_{\mathrm{ref}}=1.0$ p.u. and standard deviation 0.1. The threshold of 0.05 defines the safe voltage deviation margin. This function combines the steep penalty of a Laplace-like form outside the safe range with Gaussian smoothness within it. It enforces strong penalties for large voltage deviations while offering gentle guidance in the safe zone, thus providing effective gradients for optimization. We approximate the total reactive power loss in the reward function to improve training efficiency, as its exact computation is complex and computationally intensive. Specifically, we define $l_q\left(q^{\mathrm{RES}}\right)=\frac{1}{|D|}\|q^{\mathrm{RES}}\|_1$, where $|D|$ is the number of regulating devices, and q^{RES} denotes their reactive power output vector. This ℓ_1-norm term quantifies the regulation extent and serves as a surrogate for reactive losses, thereby indirectly reflecting energy loss trends in the network. We use the level of range anxiety $l_s(\cdot)$ as a penalization term for the EV's battery capacity. It is a nonlinear mapping between the remaining and maximum battery capacity, which describes the dynamic characteristics of the driver's anxiety level as the battery level changes, as shown in Equation 20.

$$RA^g=\frac{\ell(E^g_{t_{dep}})+|\ell(E^g_{\max})|}{\ell(0)+|\ell(E^g_{\max})|} \tag{20}$$

where $E^g_{t_{dep}}$ represents the remaining battery capacity of the current EV user, $E^g_{\max}$ is the maximum battery capacity of the current EV, and $\ell(x)$ is a logarithmic function used to quantify the impact of battery capacity. The model normalizes the anxiety value by using the denominator term $(\ell(0)+\ell(E^g_{\max}))$, which keeps the anxiety value within a standardized range. Supply-side agents are responsible for maintaining voltage stability and enhancing user benefits within a multi-agent framework. At each time step t, they receive a shared global reward, defined as the average instantaneous reward of all EV agents: $R_s=\frac{1}{N}\sum_{i=1}^{N}R_d$. This shared reward mechanism aligns the objectives of the supply-side with demand-side performance, thereby promoting system stability and user satisfaction. It facilitates implicit coordination between the supply-side and demand-side.

E. User Incentive Index

We have designed a user incentive index (UII) to evaluate EV users' participation in the smart grid, considering both economic benefits and range anxiety. This index quantifies user participation in the grid's scheduling and optimization. The UII is defined as $I=\alpha\cdot P-\beta\cdot RA$, where P represents the user's economic benefit, reflecting the reward from participating in smart grid scheduling. A higher P increases participation enthusiasm. RA represents range anxiety due to insufficient battery capacity, with a lower RA enhancing participation willingness. The parameters α and β are the weight coefficients for economic benefit and range anxiety, respectively, and are set to $\alpha=0.5$ and $\beta=0.5$ in this paper.

III. Proposed method

This section presents the methodology proposed in this paper. To effectively address the stability control issue of EVs coordinating with ADN, we introduce a novel MARL method called demand and supply side - federated graph MATD3 (DS-FGMATD3), as depicted in Figure 2. The three key implementation details of this method are essential for solving the problem:

First, we propose a dynamic framework for joint demand-supply regulation, where EVs and the ADN collaborate using multi-level strategies based on local data and global feedback.

Second, to address demand-side privacy concerns, we introduce a FL-based training mechanism, allowing EVs to update the global model without sharing local data. This approach incentivizes user participation in ADN stability control and mitigates RES power fluctuations.

Finally, we enhance the MATD3 method to optimize renewable energy variability and EV charging uncertainty. Leveraging global supervision and distributed strategies, the algorithm balances demand-side charging with supply-side energy smoothing, guided by ADN voltage levels and user benefits.

A. GCN-based Safety Strategies Learning for Distribution Networks

We assume an electrical zoning scheme based on urban planning, consisting of residential, commercial, and industrial zones, as shown in Figure 2. Electricity consumption varies across these zones: residential areas peak at night, commercial zones during working hours, and industrial zones remain high throughout the day. EV users commute between zones to simulate daily travel patterns.

The ADN is modeled as a graph, where nodes represent buses and edges denote electrical connections. In this non-Euclidean space, nodes exhibit voltage coupling and power flow correlations. Network partitioning leads to subgraphs with varying node counts and inconsistent feature dimensions.

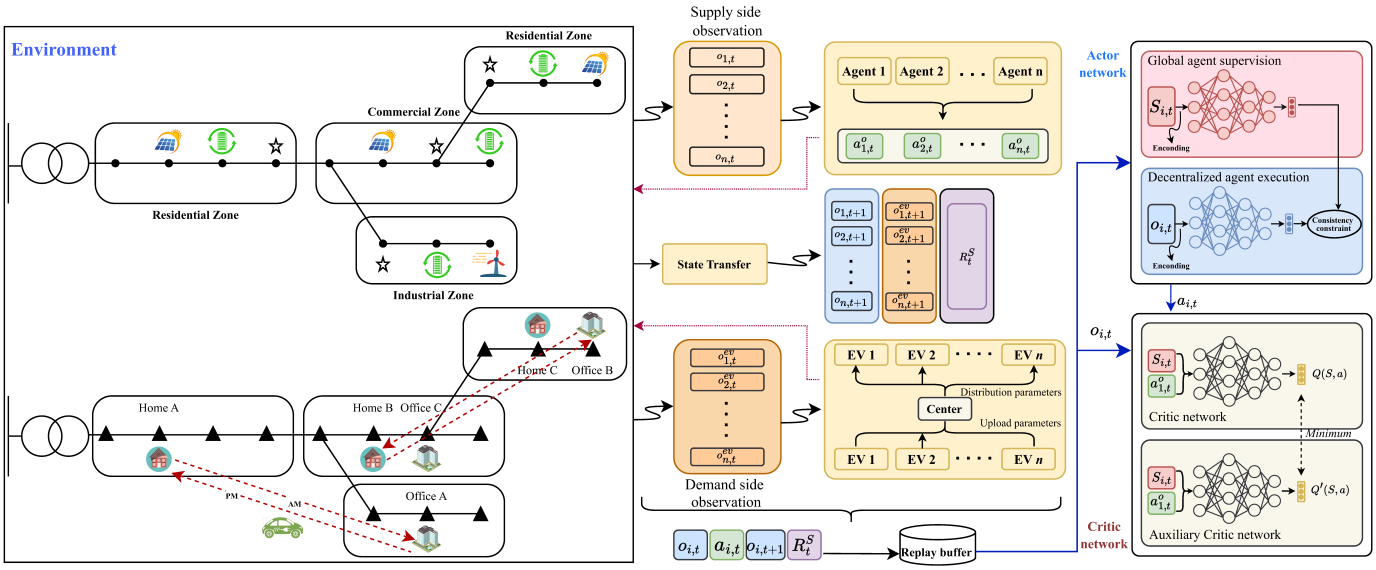


Fig. 2. DS-FGMATD3 framework for joint supply-side and demand-side regulation

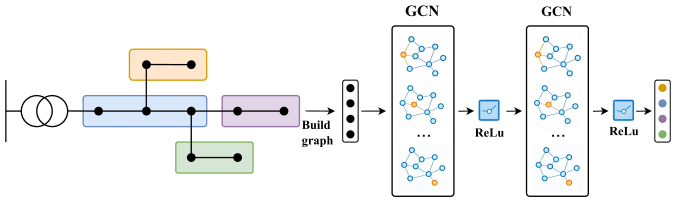


Fig. 3. GCN-based state feature encoder module

While zero-padding is a common workaround, it may introduce noise and distort data. In contrast, GCNs leverage local topology to aggregate node features, ensuring structural integrity and consistent representation. This enables agents to infer both local states and global dynamics for coordinated and safe control. Thus, we employ GCNs to encode local and global features, which are then input into the policy network, as shown in Figure 3. The ADN is represented as $\mathcal{G} = (V, E)$, where V is the set of nodes and E is the set of edges. Each node $v_i \in V$ corresponds to a local observation o_i, while the global observation S represents the overall features of the ADN. GCN performs both global and local encoding of the observations, as shown in Equation 21.

$$H^{(l+1)} = \phi\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{21}$$

where $H^{(l)}$ and $W^{(l)}$ are the feature and weight matrices of the l-th layer, ϕ is the activation function, $\widetilde{A} = A + I$ is the adjacency matrix with self-connections, and $\widetilde{D}$ is its diagonal matrix. Since the global observation dimension is slightly larger than the local one after feature mapping, MLP modules are used to align them.

The processing flow of the local encoder mirrors that of the global encoder. However, the local encoder comprises four independent GCN modules, each dedicated to processing local observations of a specific dimension. Each module learns the corresponding feature representations. After encoding the states with the GCN, we obtain the global and local features, represented as $H = H^{\mathcal{G}}, H^{\ell}$. Each agent generates corresponding decisions based on the encoded observations. The $H^{\mathcal{G}}$ and H^{ℓ} both use the same encoding method, which first processes the observations through an Observation Encoder. We chose the Gated Recurrent Unit (GRU) to construct this module. It converts $H^{\mathcal{G}}$ and H^{ℓ} into an embedding vector and obtains the global module decision $a^{\mathcal{G}}_{i,t}$ and the local module decision $a^{\ell}_{i,t}$ after a multi-layer perceptron feature mapping. To ensure that the agent adheres to safety guidelines while performing control tasks, we introduce additional constraints that fine-tune the decisions made by the agent and reduce the occurrence of dangerous decisions.

After obtaining the global action $a^{\mathcal{G}}_n$ and local action a^{ℓ}_n, we aim to go beyond a single observational perspective in decision-making. Since PV and WT are affected by local sunlight and wind speed, they exhibit regionality and randomness. Moreover, ADN's power flow cannot be fully decoupled, making coordination between global and local actions crucial for system safety. To enforce consistency, we apply a KL divergence loss, termed consistency loss, as shown in Equation 22. Note that $a^{\mathcal{G}}_n$ only supervises a^{ℓ}_n and is not executed in practice. Consistent losses enhance coordination between global and local decisions, ensuring the agent considers electricity demand and sub-zone interactions.

$$\mathcal{L}_{kl} = D_{KL}\left\{\mathcal{N}\left(a^{\mathcal{G}}, \sigma_a^2\right) || \mathcal{N}\left(a^{\ell}, \sigma_a^2\right)\right\} \tag{22}$$

B. Federated Learning-based Learning for Demand-Side Strategies

We employ a FL-based mechanism to safeguard data privacy. In this approach, the EV agent processes its state information and makes decisions locally, without transmitting raw data to the central server. This decentralized strategy improves privacy by enabling agents to use their data for model training while minimizing the risk of data leakage.

Each EV agent i inputs its state information $\mathcal{O}_{i,t}^{\mathrm{EV}}$ and the state of the ADN into the actor-network. The local model parameters θ_i are updated by minimizing the Actor and Critic losses. As the EV's location changes in real-time, its local observations vary accordingly. After training, each EV agent uploads its updated model parameters θ_i^{new} to the central server, which ensures privacy and reduces communication overhead, as described in Equation 23. The DSO-managed central server aggregates these updates using a weighted average to update the global model, facilitating collaborative learning.

$$\theta^{global} = \frac{1}{N}\sum_{i=1}^{N}\theta_i^{new} \tag{23}$$

After updating the DSO central server, the global model parameters θ^{global} are redistributed to all agents for the next local training round. This iterative process continues until convergence, enabling the DSO to intervene in real-time strategies via federated averaging. The FL process transmits only local model parameters (i.e., network weights) from EV agents to the DSO, without exposing raw inputs or action trajectories. The DSO aggregates these to update the global model, ensuring no EV-specific behavior or status is revealed during training or execution.

C. MATD3-based strategy learning and objective optimization

We will not further elaborate on the MATD3 algorithm [32]. This paper combines the maximum entropy strategy with deep deterministic policy gradient to improve policy stability and sample efficiency. The state-value function aims to minimize the mean squared error between the critic-network output and the target critic-network output, as shown in Equation 24.

$$\mathcal{L}_Q\left(\theta^Q\right) = \mathbb{E}_{\tau(s,a,r,s',d)\sim D}\left[\left(Q\left(s,a\middle|\theta^Q\right) - \left(R+\gamma\left(1-d\right)\min_{a'}Q'\left(s',a'\middle|\theta^{-Q}\right)\right)\right)^2\right] \tag{24}$$

where $\tau(s,a,r,s',d) \sim D$ consists of states $\mathbf{S}$, actions $\mathcal{A}$, rewards R, next states $\mathbf{S}'$, and termination flags d sampled from the experience replay buffer D. $Q(s,a|\theta^Q)$ is the critic-network evaluating the state-action value after executing action a in state $\mathbf{S}$. R is the reward from the environment, and γ is the discount factor for future rewards. d is the termination flag. The MATD3 algorithm trains two critic-networks and uses the smaller Q-value to avoid overestimation, with $\min_{a'} Q'(s',a'|\theta^{-Q})$ representing the minimum of the two Q-values.

We employ the maximum entropy learning mechanism and entropy function to enhance the agent's exploration of the environment and improve policy discovery. The actor-network seeks to balance expected benefits and entropy, minimizing Equation 25.

$$\mathcal{L}_\pi\left(\theta^\pi\right) = -\mathbb{E}_{\tau(s,a,r,s',d)\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t\left(r(s_t,a_t) - \alpha H\left(\pi(\cdot|s_t)\right)\right)\right] \tag{25}$$

where $R(s_t,a_t)$ is the immediate reward, and α is the entropy regularization coefficient. In this paper, we set it to 0.1. $H\left(\pi(\cdot|s_t)\right)$ is the entropy of the policy at state s_t, defined in Equation 26, which measures the uncertainty of the policy.

$$H\left(\pi\left(\cdot|s_t\right)\right) = -\sum_a \pi(a|s_t)\log\pi(a|s_t) \tag{26}$$

The parameters of the target actor-network and target critic-network are updated by using a more stable soft update method as shown in Equation 27.

$$\begin{cases}\theta^{Q'} \leftarrow \lambda\theta^Q + (1-\lambda)\theta^{Q'},\\ \theta^{\pi'} \leftarrow \lambda\theta^\pi + (1-\lambda)\theta^{\pi'}\end{cases} \tag{27}$$

We use delayed policy updates, as shown in Equation 28, where ϵ is the exploration noise and c is used to limit the noise size.

$$a = \pi_{\theta'}(s) + clip\left(\epsilon,-c,c\right),\ \epsilon\sim\mathcal{N}\left(0,\sigma^2\right) \tag{28}$$

When we optimized the GCN to extract the observation features of the ADN, we used the loss function in Equation 29 to train the neural network, which is designed to improve the representation of the state features extracted by the GCN.

$$\mathcal{L}_\mathcal{G} = \eta_1\mathcal{L}_\pi\left(\theta^\pi\right) + \eta_2\mathcal{L}_Q\left(\theta^Q\right) + \eta_3 e^{-r} \tag{29}$$

where e^{-r} is a reward-based regularization term, r represents the sum of rewards. By jointly optimizing this loss function, the GCN can more effectively learn the state features of ADN. Algorithm 1 provides the pseudocode for the MAGRL algorithm of AVC. We set η_1, η_2 and η_3 to 0.4, 0.3, and 0.3, respectively.

IV. Experimental Examples

A. Case Study Setup

We conduct a case study on a modified IEEE 14-bus distribution network, where each bus is equipped with an EV charging station. The network includes three PV systems, one WT, four ESSs, and four SVCs, with the topology depicted in Figure 2.

The dataset is split 80% for training and 20% for testing to ensure independent evaluation. The training set is used for model fitting and tuning, while the testing set evaluates generalization on unseen data. It includes three years of user active/reactive power and renewable generation data from Wallonia's Elia Group [33], with load data from [40]. Samples are collected every three minutes, totaling about 525,600 points. The first 33 months (481,440 samples) form the training set, and the last 3 months (44,160 samples) form the testing set. Splitting occurs before feature extraction and normalization. More details appear in [12] and [41]. The four price parameters in Equation 11—active power selling price $p_{s,t}$, active power buying price $p_{b,t}$, reactive power selling price $q_{s,t}$, and reactive power buying price $q_{b,t}$—are based on real-time pricing data from the State Grid Anhui Electric

Algorithm 1 DS-FGMATD3

Initialize critic-networks $Q_{(\theta_1)}, Q_{(\theta_2)}$ and GCN network $\mathcal{F}_\mathcal{G}$ with random parameters $\theta_1, \theta_2, \mathcal{F}_\mathcal{G}$.
Initialize target networks $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\mathcal{F'}_\mathcal{G} \leftarrow \mathcal{F}_\mathcal{G}$.
Initialize replay buffer $\mathcal{D}$.
for agent $i = 1, 2, \ldots, N$ **do**
  Initialize actor-network π_ϕ with random parameters ϕ
  Initialize target networks $\phi' \leftarrow \phi$.
end for
Set global time step $T \leftarrow 0$.
for episode=1 to T **do**
  for time step $t = 1, 2, \ldots N$ **do**
    Interact with environment and obtain a transition (s, a, r, s') in $\mathcal{D}$.
    Sample mini-batch of N transitions (s, a, r, s') from $\mathcal{D}$.
  end for
  Update global time step $T \leftarrow T + 1$.
  for agent $i = 1, 2, \ldots N$ **do**
    Extract global and local observation features $H^\mathcal{G}, H^\ell$, using $\mathcal{F}_\mathcal{G}$.
    Select action with exploration noise using 28, and observe reward r and new state s'.
    Store transition tuple (s, a, r, s') in $\mathcal{D}$.
    for $j = 1 \rightarrow N_c$ **do**
      Update the parameters θ_1, θ_2 of critic using 24 and generated transitions.
    end for
    for $j = 1 \rightarrow N_a$ **do**
      Update the parameters ϕ of actor using 25 and 26 and generated transitions.
    end for
    Soft-update the parameters of the target critic-network and target actor-network using 27
    Update the parameters $\mathcal{F}_\mathcal{G}$ of GCN using 29
  end for
end for

Power Company. The reactive power price is set at 10% of the active power price. Figure 4 shows the time-varying electricity prices across different sub-regions.

The PV and WT systems have apparent power capacities of 50 kVA and 400 kVA, respectively. The power factor is set to 0.95 for residential nodes and 0.85 for commercial/industrial nodes. Three users with distinct EV usage patterns are considered, as detailed in Table I, and the technical parameters for the Tesla Model S are provided in Table II. All EVs' initial SOCs are set to 100% only at the start of the first training episode. Thereafter, SOC values evolve dynamically through interactions with the environment, accurately reflecting EVs' energy changes over time.

TABLE I
BEHAVIORAL CHARACTERISTICS OF EV USERS

	Home/Office	Work	Home	Commuting	Power Consumption
Person A	Node 2/Node 10	AM 9:00- PM 16:00	PM 17:00-AM 8:00	1 Hour	25%
Person B	Node 6/Node 14	AM 8:00- PM 17:00	PM 18:00-AM 7:00	1 Hour	20%
Person C	Node 13/Node 7	AM 8:30- PM 17:30	PM 18:00-AM 8:00	0.5 Hour	10%

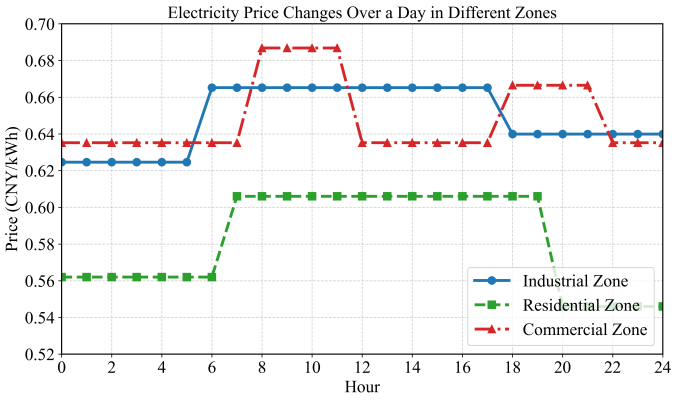


Fig. 4. Electricity price trends in different sub-regions

TABLE II
TECHNICAL PARAMETERS OF EV MODELS

Parameter	Value
Apparent Power Capacity S^{ev} (kVA)	16.7 kVA
Maximum Active Power P^{ev} (kW)	16.7 kW
Battery Capacity E^{ev} (kWh)	100 kWh
Charging/Discharging Efficiency η^{evc}/η^{evd} (%)	90%

We compare our algorithm with three baseline MARL methods: 1) MADDPG [34], which employs centralized training and decentralized execution for collaboration in multi-agent environments; 2) MAPPO [35], a probabilistic algorithm that enhances policy stability and sample efficiency through importance sampling; and 3) MAAC [36], which utilizes centralized training to improve policy learning efficiency. The hyperparameters of these methods are provided in Table III. While these methods are widely used in general MARL tasks, their integration into cooperative power control enhances scalability for distributed EV-based regulation.

TABLE III
HYPERPARAMETERS FOR THE MARL ALGORITHM

Parameter	Value	Parameter	Value
Actor learning rate (α_ϕ)	1×10^{-4}	Soft update rate (τ)	1×10^{-2}
Critic learning rate (α_θ)	1×10^{-4}	Discount factor (γ)	0.99
Actor/Critic architecture	2 layers, 128 nodes	Minibatch size (J)	32
Replay buffer size (D)	1×10^6	Optimizer for RL/GCN	Adam
Number of episodes	600	Random seeds	(100, 200, 300)
Target strategy update steps	2	Exploration noise std. (σ)	0.2
Clip noise limit (c)	0.2	GCN layers	3
GCN hidden size	128	GCN learning rate	1×10^{-4}
Activation	ReLU	Dropout rate	0.2
Weight decay	5×10^{-4}	GRU encoder hidden size	64

B. Reward Comparison

We compare the strategy quality and convergence speed of our algorithm with three baseline MARL algorithms (MADDPG, MAPPO, and MAAC) from the demand-side perspective, as shown in Figure 5. The proposed DS-FGMATD3 algorithm outperforms these algorithms in both performance and convergence. It reaches stable convergence and the highest reward within 200 iterations, whereas the comparison algorithms exhibit significant fluctuations. DS-FGMATD3 achieves higher average rewards for all users (Person A, B, and C),

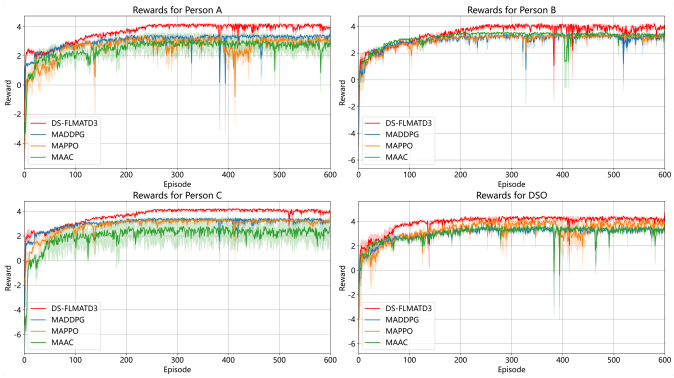


Fig. 5. Comparison of user training rewards

while MADDPG, MAPPO, and MAAC show lower rewards with larger fluctuations. Table IV presents the average cumulative rewards over 30 test days. To verify the response characteristics of the supply-side reward and the convergence process of its strategy at different regulation stages, we present its time-series evolution curve in the bottom-right subplot of Figure 5.

TABLE IV
TEST CUMULATIVE REWARDS FOR DIFFERENT ALGORITHMS

Algorithm	Person A	Person B	Person C
DS-FGMATD3	118.98	115.32	116.88
MADDPG	100.69	97.97	89.28
MAPPO	99.86	95.14	90.12
MAAC	89.89	83.29	79.66

C. Analysis of user benefits

To evaluate the impact on user benefits, we compare the proposed algorithm with three mainstream ones, as shown in Figures 6 and 7. The results reveal significant differences in average daily benefits. Our algorithm consistently outperforms the others. During training, it provides a daily benefit for Person A, B, and C that is approximately 50% higher than the second-best algorithm. For example, Person A's maximum benefit reaches ¥10, while the comparison algorithm's is ¥5; Person B's is ¥11 versus ¥5, and Person C's is ¥12 compared to ¥7.

During testing, both our algorithm and the comparison algorithms demonstrate high benefit stability. For Persons A, B, and C, the proposed algorithm outperforms the others by 15% to 25%, with peak benefits around ¥12, while the suboptimal algorithm achieves a maximum of ¥9. These results confirm that DS-FGMATD3 maintains strong benefit stability throughout both the training and testing stages.

D. Analysis of voltage regulation and power loss

To assess the performance of different algorithms in optimizing power losses and voltage, we compare the DS-FGMATD3 algorithm with three widely used algorithms. Figures 8 (a) and (b) present the power loss results during training and testing, respectively. All algorithms converge effectively during training, with DS-FGMATD3 achieving the lowest power loss—50% lower than MADDPG, 60% lower than MAPPO, and 63% lower than MAAC. Figure 8 (b) shows the power loss over a 24-hour period, where peak consumption in the morning and afternoon impacts power loss. DS-FGMATD3 consistently outperforms the others, maintaining the lowest power loss throughout the day.

Figures 9 and 10 present bus voltage results for the four algorithms during both training and testing. As shown in Figure 9, MADDPG and MAPPO maintain stable voltage at certain nodes, demonstrating some robustness to environmental changes, while MAAC shows significant voltage violations. In contrast, DS-FGMATD3 maintains smoother and more stable voltage across all nodes, exhibiting strong robustness in dynamic environments. The safe voltage range is indicated by a red dotted line during the testing phase. During peak periods of electricity consumption and renewable generation, grid voltage fluctuations increase. Figure 10 demonstrates that DS-FGMATD3 effectively keeps the voltage within the safe range with minimal fluctuations, even amid shocks from renewable energy, load demands, and EV charging. In comparison, MAPPO and MAAC experience large fluctuations, particularly during the afternoon peak, struggling to adapt to rapid changes in load and generation, resulting in local voltage instability.

E. Analysis of active and reactive power support of EVs for ADNs

Figure 11 shows the active and reactive power control results for three EVs over 24 hours, alongside their SOC variations. The active power profiles indicate that each EV charges during low-demand periods and discharges during high-demand periods to maximize benefits. The reactive power profiles show that all EVs support grid stability by contributing to reactive power compensation. These results validate the proposed method's effectiveness in providing reliable reactive power support. Integrating EVs into the ADN offers controllable resources for the DSO, but excessive use of EV batteries for grid regulation may disrupt commuting and reduce user satisfaction. The DSO's regulation causes noticeable SOC fluctuations, with lower battery levels during peak commuting hours and active participation in grid management during working hours and evenings. The experimental results show that the proposed algorithm effectively keeps battery levels optimal, ensuring user comfort and economic benefits. Figure 11 illustrates the total hourly power of EV charging and discharging aggregated statistically. It does not imply that the battery undergoes charging and discharging simultaneously.

F. Comparison of user incentive index among different algorithms

As shown in Figure 12, DS-FGMATD3 outperforms all other algorithms in the UII for all users, both during training and testing, highlighting its superior ability to balance economic benefits and range anxiety. In contrast, other algorithms show significantly lower UII values, confirming that

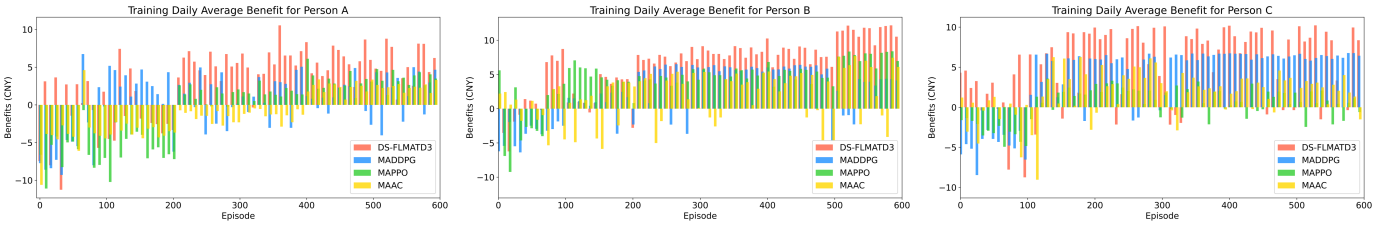


Fig. 6. User training benefits (¥) comparison

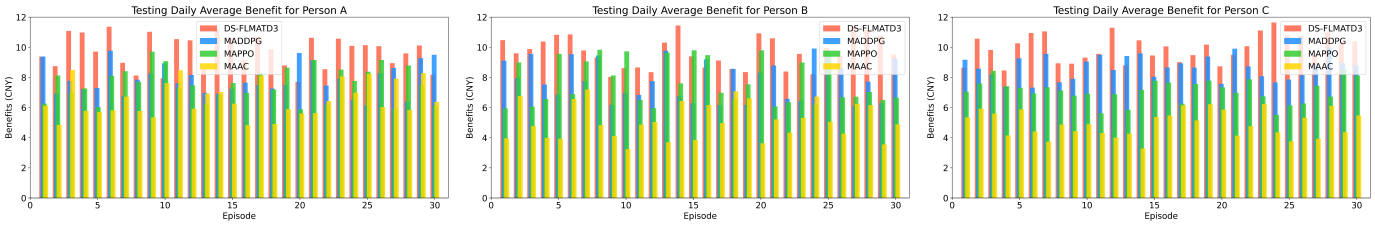


Fig. 7. User testing benefits (¥) comparison

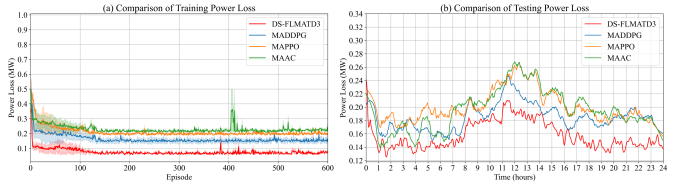


Fig. 8. Training/Testing ADN power loss (MW) comparison

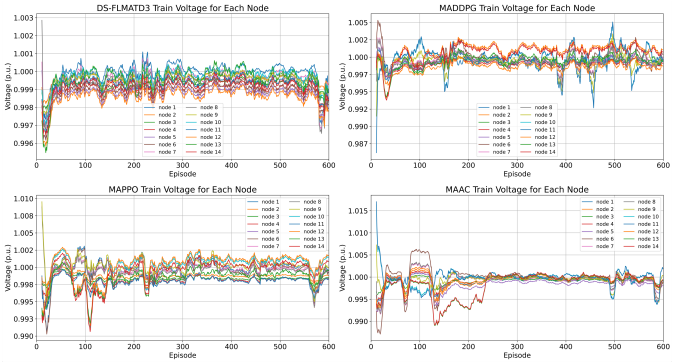


Fig. 9. Training ADN voltage (p.u.) comparison

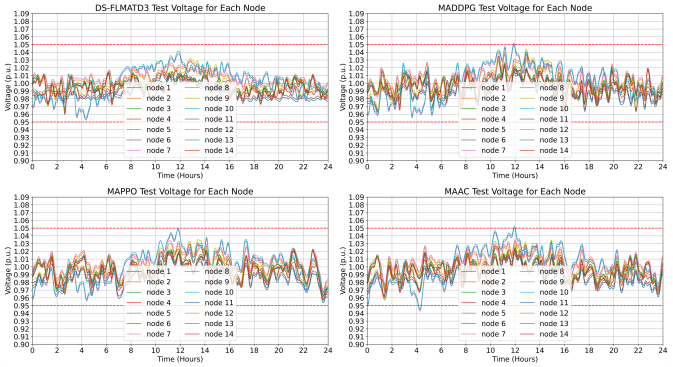


Fig. 10. Testing ADN voltage (p.u.) comparison

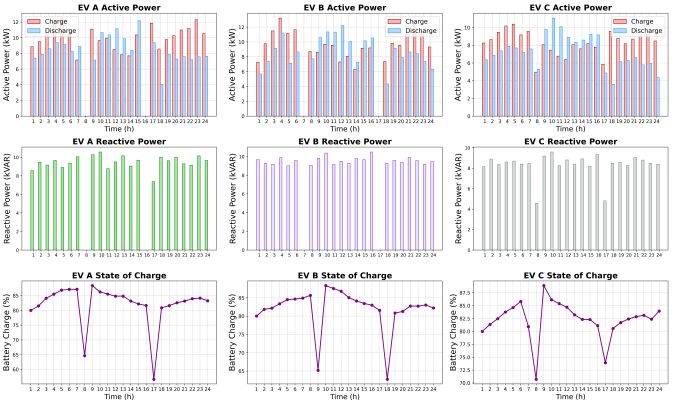


Fig. 11. The active and reactive power support provided by different EVs to the ADN

DS-FGMATD3 not only boosts economic benefits but also reduces range anxiety, increasing user engagement in smart grid scheduling. Its consistent performance across different users further emphasizes its robustness and adaptability, setting it apart from other algorithms.

G. Benefit-oriented cost analysis of battery degradation

To evaluate the impact of incorporating battery degradation costs on strategy learning and user benefits, this study monitors the average daily benefits of Users A, B, and C during both the training and testing phases. As illustrated in Figures 13 (a) and (b), the benefits exhibit considerable fluctuations and occasionally become negative. It is important to note that "negative benefits" are relative, as calendar aging of batteries causes unavoidable losses even in the absence of V2G participation. With continued training, the strategies converge, allowing agents to balance grid support and degradation costs optimally. During testing, all users consistently achieve positive benefits, demonstrating the stability of the proposed algorithm. These results confirm that the method effectively

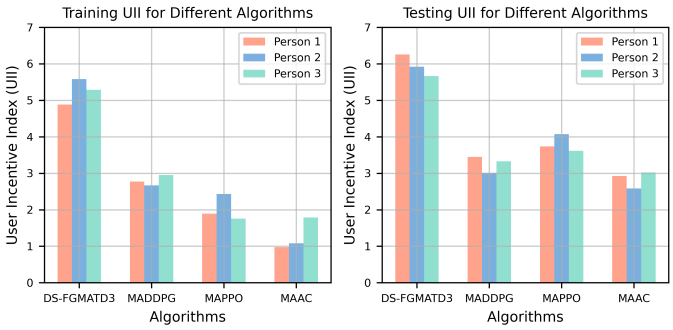


Fig. 12. Comparison of training and testing UII for different algorithms

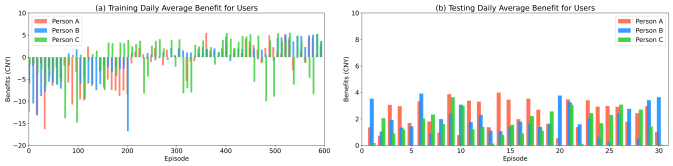


Fig. 13. Training/Testing daily average benefit for users

secures economic benefits while promoting sustainability and practical feasibility.

H. Performance evaluation under dynamic disconnection of electric vehicles

To evaluate the robustness of the proposed method against unexpected EV disconnections, a comparative experiment involving offline users was conducted. The reward curves are presented in Figure 14. Users A, B, and C were each offline during specific intervals (indicated by gray shaded areas). The results demonstrate that although temporary disconnections cause fluctuations in other users' rewards, the overall trend shows consistent improvement and convergence. The system rapidly recovers after each disconnection, indicating strong adaptability and robustness, thereby confirming the method's stability and effectiveness in dynamic and uncertain environments.

I. Impact of UII weighting coefficients on user incentive effectiveness

To evaluate the influence of the weighting coefficients α and β on the UII, ablation experiments were conducted using three weight combinations: $(\alpha, \beta) = (0.3, 0.7), (0.5, 0.5), (0.7, 0.3)$. These combinations were selected to investigate how the balance between economic benefits and range anxiety affects user incentive effectiveness and system stability. Figure 15 presents a comparison of the performance of four algorithms during training and testing across different UII weight settings. DS-FGMATD3 consistently outperforms the other algorithms under all tested configurations, demonstrating excellent user incentive capability and policy adaptability. Notably, the highest UII values occur at $(\alpha, \beta) = (0.5, 0.5)$ and $(0.3, 0.7)$, suggesting that user participation peaks when economic benefits are either dominant or moderately balanced with anxiety concerns. These findings confirm the robustness and generalization capacity of DS-FGMATD3 across varying user preference models.

J. Analysis of supply-side contributions

To evaluate the contribution of supply-side resources to voltage stability, Figure 16 presents the 24-hour active and reactive power outputs of WT, PV, ESS, and SVC. The left Figure indicates that WT and PV primarily supply active power, with PV output peaking during midday to meet the corresponding load demand. The ESS charges during periods of low load and discharges during peak demand, thereby flattening the load curve and reducing the pressure on system regulation. The right Figure illustrates the reactive power responses: as daytime load increases, reactive power demand also rises. During periods of high renewable generation, ESS and SVC jointly provide reactive power compensation. Notably, the SVC maintains a sustained response during high-risk intervals, thereby enhancing the voltage stability margin.

These supply-side resources demonstrate effective coordinated control: WT and PV supply clean active power, ESS mitigates load fluctuations via temporal energy shifting, and SVC delivers reactive support during peak demand periods. The results confirm the effectiveness of the proposed supply-demand coordination strategy in improving voltage stability and operational efficiency, especially under scenarios with high renewable energy penetration.

K. Supply/Demand-side-only voltage regulation

To evaluate the effectiveness and limitations of unilateral control methods, a comparison between (1) demand-side-only and (2) supply-side-only regulation strategies, as shown in Figure 17, the results demonstrate the inherent limitations of relying exclusively on single-sided resources for dynamic voltage control. Demand-side-only regulation results in delayed responses and inadequate support, particularly at the grid periphery. Conversely, supply-side-only control exhibits sluggish recovery and reduced precision when responding to dynamic load disturbances.

Experimental results show that single-sided strategies are insufficient to ensure the stability of distributed power systems, especially under nonlinear disturbances and spatial variability. The proposed coordinated supply–demand control mechanism fully leverages the complementary advantages of both sides, thereby enhancing the overall responsiveness, robustness, and effectiveness of voltage regulation.

L. Sensitivity analysis of reward function weights

To evaluate the impact of reward function (Equation 18) weights on policy performance and mitigate potential bias from improper configurations, we investigated two key parameters: α (the penalty weight for reactive power loss) and β (the weight for user profit). Figure 18 presents the results under various weight combinations. The findings indicate that changes in these weights significantly influence both user profit and voltage control. Specifically, increasing the system objective weight enhances grid-side regulation but reduces user

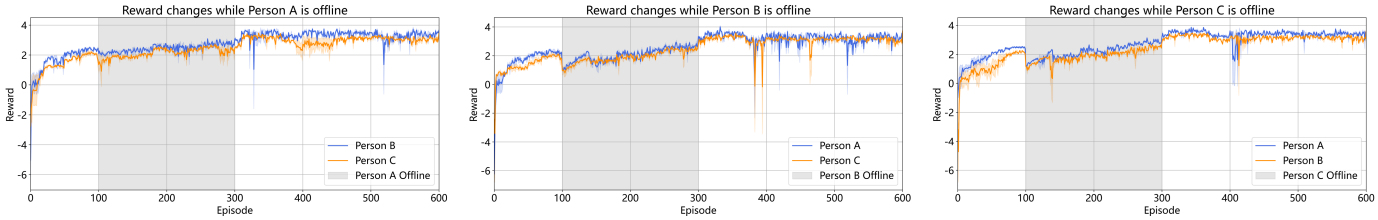


Fig. 14. Reward comparison under different EV offline

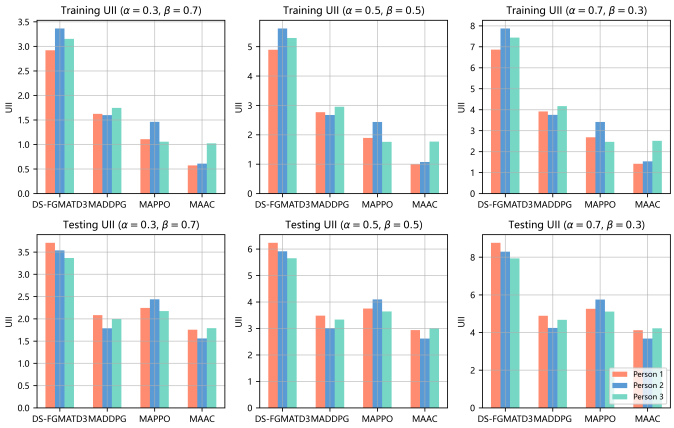


Fig. 15. Ablation study on UII parameter weights

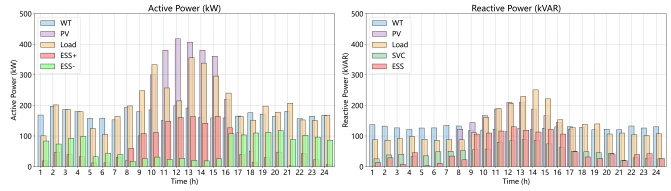


Fig. 16. Supply-side power support results

incentives. Overall, the agent demonstrates a robust and well-balanced performance without exhibiting bias toward any single objective. The selected weights $(\alpha = 0.1, \beta = 0.5)$ yield optimal results across all evaluation metrics, demonstrating effective multi-objective coordination.

M. Extended experiments on IEEE-33Bus and IEEE-141Bus systems

To evaluate the scalability of the proposed control strategy, extended experiments are conducted on the IEEE 33-bus and IEEE 141-bus systems. Figure 19 (a) presents the IEEE 33-bus system, where green nodes indicate deployed ESS, and

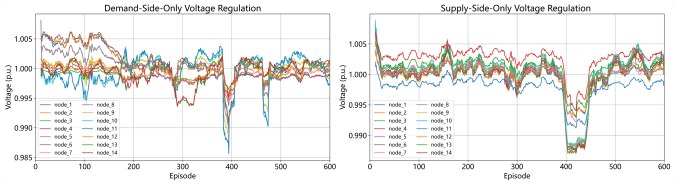


Fig. 17. Demand/Supply-side-only voltage regulation

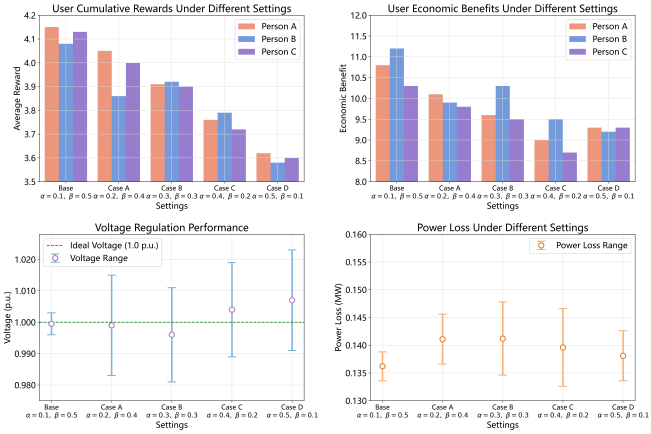


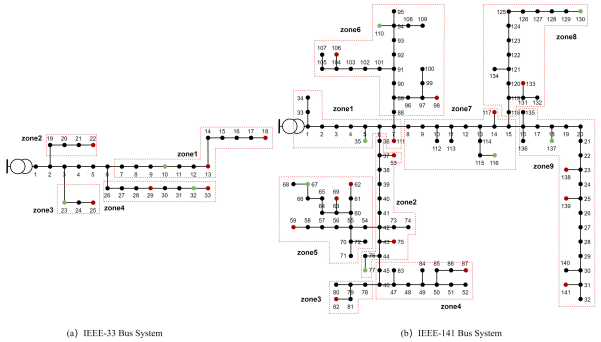
Fig. 18. Sensitivity analysis of reward function weights

red nodes represent reactive power devices such as SVCs and inverters. Zones 1–2 are residential, Zone 3 is commercial, and Zone 4 is industrial. Figure 19 (b) depicts the IEEE 141-bus system, which adopts the same zoning and device deployment strategy. Green nodes indicate ESS, while red nodes denote reactive power devices. This system is divided into nine zones: Zones 1–4 (residential), 5–7 (commercial), and 8–9 (industrial). Its larger scale and higher node density enable a more comprehensive evaluation of the control strategy's performance and scalability. Details of user behavior settings—including commuting paths, dwelling times, and energy usage across zones—are provided in Table V.

The data configuration for the new system is based on the settings in [12]. As illustrated in Figure 20, with the increasing number of nodes and the integration scale of distributed renewable energy sources, the algorithm exhibits some fluctuations during convergence. Nevertheless, it maintains an overall stable convergence trend. These results confirm the proposed algorithm's adaptability and scalability in complex scenarios, highlighting its potential for real-world large-scale applications.

TABLE V
BEHAVIORAL CHARACTERISTICS OF EV USERS ON IEEE-33BUS AND IEEE-141BUS

System	User	Home/Office	Work Time	Home Time	Commuting Time	Power Consumption
IEEE-33Bus	Person A	Node3/Node12	AM9:00–PM16:00	AM17:00–PM8:00	1 hour	25%
IEEE-33Bus	Person B	Node6/Node22	AM8:00–PM17:00	AM18:00–PM7:00	1 hour	20%
IEEE-33Bus	Person C	Node9/Node27	AM8:30–PM17:30	AM18:00–PM8:00	0.5 hour	15%
IEEE-141Bus	Person A	Node5/Node35	AM9:00–PM16:00	AM17:00–PM8:00	1 hour	25%
IEEE-141Bus	Person B	Node15/Node78	AM8:00–PM17:00	AM18:00–PM7:00	1 hour	20%
IEEE-141Bus	Person C	Node40/Node101	AM8:30–PM17:30	AM18:00–PM8:00	0.5 hour	15%



(a) IEEE-33 Bus System (b) IEEE-141 Bus System

Fig. 19. Example of zonal partitioning and device deployment in IEEE-33/141Bus systems

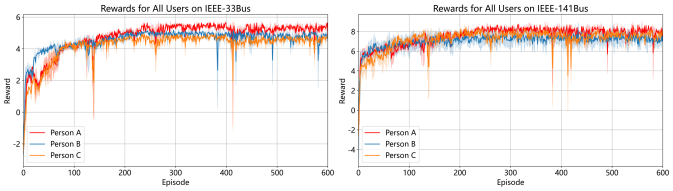


Fig. 20. DS-FGMATD3 rewards for all users on IEEE-33Bus and IEEE-141Bus

N. Comparative Robustness Evaluation of FedAvg and Median

Federated learning is susceptible to abnormal parameter uploads from malicious or faulty clients, which can substantially degrade the global model's performance. This study compares the FedAvg and median aggregation methods under scenarios involving malicious parameter injection. Specifically, starting from round 200, we simulate malicious clients uploading perturbed parameters, as illustrated in Figure 21. The results demonstrate that FedAvg suffers from notable performance degradation and convergence instability, while the median aggregation method maintains stable convergence, effectively mitigating the effects of abnormal updates and significantly improving training robustness.

V. Conclusion

This paper presents a DS-FGMATD3 control framework based on MARL to address stability challenges in ADNs caused by the integration of EVs and RESs. The proposed framework employs a GCN to encode both global and local system states, effectively capturing complex topological and operational interactions. A consistency loss function is introduced to enhance coordinated agent decision-making, while FL is adopted to preserve user privacy and enable collaborative model training. Simulation results indicate notable improvements in peak voltage regulation, power loss reduction, user demand satisfaction, and overall user welfare.

Future work will focus on enhancing the framework's scalability and generalization to accommodate larger and more complex network topologies, while simultaneously reducing model complexity to adapt to dynamic power fluctuations. Additionally, user preference heterogeneity will be addressed through personalized incentive parameters, and user profiling will be integrated with adaptive mechanisms for real-time strategy adjustment. Collaboration with behavioral scientists is also planned to incorporate psychological models such as Prospect Theory and to develop game-theoretic strategies that ensure incentive compatibility in V2G interactions.

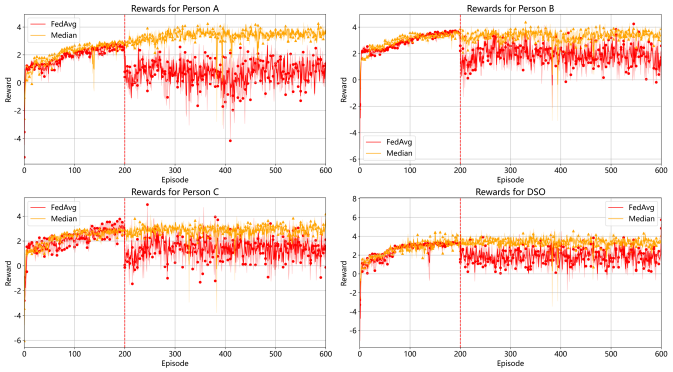


Fig. 21. Performance Comparison of FedAvg and Median

Appendix A
Electric Vehicle Commuting Energy Model

This paper defines the driving energy consumption E_{tra} in Equation 17 as the energy consumed by the traction battery during commuting, directly reflecting changes in the battery's SOC. Based on vehicle dynamics, the energy consumption is estimated using the following model:

$$E_{\text{tra}} = \alpha \cdot \left(\frac{1}{2\eta} m v^2 + c_r m g d + \frac{1}{2} \rho c_d A v^2 d \right) + E_{\text{aux}} \quad (30)$$

where the key parameters such as vehicle mass m, average speed v, and frontal area A are obtained from vehicle specifications; dynamic coefficients like c_r and c_d are calibrated using public road test data; physical constants such as efficiency η and air density ρ refer to [38], [39]. Detailed definitions and values are listed in Table VI.

To better reflect real-world conditions, such as frequent stop-and-go, slope driving, and auxiliary loads like air conditioning, an empirical coefficient α and an auxiliary term E_{aux} are introduced to correct the base model, enhancing its practical relevance. Although equivalent average speeds are used, factors like traffic congestion and signal delays are considered, with total commuting times set as 1/1/0.5 hours.

TABLE VI
Parameters of Commuting Energy Consumption Model

Parameter	Value	Parameter	Value
m	2200 kg	c_d	0.28
v	75/75/60 (km/h)	c_r	0.02
η	0.9	A	2.4 m^2
d	30/20/15 km	ρ	1.2 kg/m^3
E_{aux}	8 kWh	α	3

Appendix B
Anxiety Function

To better characterize user behavior in EV-grid interaction, especially under limited battery conditions, we introduce a

range anxiety function to quantify users' psychological discomfort due to low battery levels. This factor is critical in influencing participation in grid regulation tasks. The proposed anxiety function for the g-th EV user is defined as:

$$RA_3^g = \frac{\ell\left(E^g_{t_{\text{dep}}}\right) + |\ell\left(E^g_{\max}\right)}{\ell(0) + |\ell\left(E^g_{\max}\right)|} \tag{31}$$

where the intermediate function $\ell(\cdot)$ is given by:

$$\ell(E^g_{t_{\text{dep}}}) = \ln\left(1 \Big/ \left[1.01 - \left(|E^g_{\max} - E^g_{t_{\text{dep}}}|/E^g_{\max}\right)^2\right]\right) \tag{32}$$

where $E^g_{\max}$ denotes the maximum battery capacity of the g-th EV, and $E^g_{t_{\text{dep}}}$ represents its battery energy at the departure time. The term $\ell(0)$ corresponds to the function value when the battery is completely depleted and serves as the baseline for range anxiety. By combining a logarithmic function with a squared term, this formulation effectively captures the nonlinear relationship between battery level and user anxiety. When the battery level is low, anxiety rises sharply (i.e., steep slope); as the battery approaches full capacity, the increase in anxiety slows down (i.e., gentle slope), indicating reduced user concern. This design reflects the psychological traits of some EV users—especially short-distance commuters—who experience greater anxiety at low charge but feel at ease when the battery is sufficiently full.

APPENDIX C
BATTERY DEGRADATION AND COST MODELING

To quantify the impact of V2G behavior on battery lifespan, we introduce a battery degradation model D_h^{total}, which consists of two components: calendar aging and cycle aging, formulated as: $D_h^{\text{total}} = D_h^{\text{cal}} + D_h^{\text{cyc}}$.

A. Calendar Aging Modeling

Calendar aging is influenced by the SOC level S_h, temperature θ_h, and time d. It is modeled as:

$$D_h^{\text{cal}} = G(S_h) \cdot \exp\left(-\frac{E_a}{R\theta_h}\right) \cdot d^{0.5} \tag{33}$$

where $G(S_h)$ is a piecewise quadratic function of SOC, E_a is the activation energy, R is the gas constant, and θ_h is the battery temperature.

B. Cycle Aging Modeling

Cycle aging is primarily affected by the charge/discharge rate I_h^c, temperature θ_h, and the ampere-hour throughput A_h, and is modeled as:

$$\begin{cases} D_h^{\text{cyc}} = Z(\theta_h) \cdot e^{q_1 I_h^c} \cdot A_h \\ Z(\theta_h) = q_1\theta_h^2 + q_2\theta_h + q_3 \\ I_h^c = (P_h^{\text{ch}} + P_h^{\text{ds}})/Q_0, \quad A_h = (P_h^{\text{ch}} + P_h^{\text{ds}}) \cdot 10^3 \cdot \Delta t/V \end{cases} \tag{34}$$

where Q_0 is the rated battery capacity, V is the nominal voltage, and Δt is the time step, q_1, q_2, q_3 are fitting parameters related to temperature.

C. Degradation Cost Modeling

The degradation cost of the battery at time step h, denoted as π_h^{deg}, is expressed as:

$$\pi_h^{\text{deg}} = \pi_{\text{bes}} \cdot \frac{D_h^{\text{total}}}{1-\mu} \tag{35}$$

where μ is the remaining capacity ratio at the end of life (set to 0.9). Considering the time value of money, the current residual value of the battery, π_{bes}, is composed of three main components: (1) replacement cost, (2) operation and maintenance (O&M) cost, and (3) salvage value. The computation is given by: $\pi_{\text{bes}} = \pi_{\text{rep}} + \pi_{\text{om}} - \pi_{\text{sv}}$, where $\pi_{\text{rep}} = C_{\text{rep}}/(1+i)^{\phi}$ represents the present value of battery replacement cost, $\pi_{\text{om}} = C_{\text{om}} \cdot \left((1+i)^{\phi} - 1\right) / \left(i \cdot (1+i)^{\phi}\right)$ represents the present value of operation and maintenance (O&M) expenses over the battery's service life, and $\pi_{\text{sv}} = \gamma_{\text{sv}} \cdot \pi_{\text{rep}}$ represents the discounted salvage value at the end of life, where C_{rep} is the initial procurement cost of a new battery, C_{om} is the fixed periodic O&M cost, i is the annual discount rate, ϕ is the designed lifespan of the battery (in years), and γ_{sv} denotes the salvage value ratio, i.e., the proportion of residual value relative to the original cost. Detailed parameter settings are provided in Table VII, and more details can be found in [37].

To better reflect realistic scenarios, we incorporate the battery degradation cost $\pi_{i,t}^{\text{deg}}$ into the multi-agent reward function. The revised reward function is defined as:

$$R_d = -\Bigg[\frac{1}{|V|}\sum_{i\in V} l_v(v_{i,t}) + \alpha \cdot l_q(q^{\text{RES}}) - \beta \cdot (\pi_t^P \cdot P_{i,t}^{\text{ev}} + \pi_t^Q \cdot Q_{i,t}^{\text{ev}}) - l_s(E_{i,t}^{\text{ev}}) + \pi_{i,t}^{\text{deg}}\Bigg] \tag{36}$$

Based on this updated reward function, we re-train the algorithm to observe its impact on user benefits. The results of the experiment are shown in IV.G

TABLE VII
PARAMETERS IN BATTERY AGING AND COST MODELING

Parameter	Value	Parameter	Value
Q_0	100 kWh	π_{om}	100 ¥/year
V	360 V	i	10%
θ_h	298 K	ϕ	5 year
R	8.314 J/mol·K	q_1	1.85×10^{-6}
E_a	58.3 kJ/mol	q_2	-1.04×10^{-3}
π_{rep}	5000 ¥	q_3	0.2089
γ_{sv}	50%	q_4	0.65

REFERENCES

[1] Kabiri-Renani, Y., Arjomandi-Nezhad, A., Fotuhi-Firuzabad, M., et al. "Transactive-Based Day-Ahead Electric Vehicles Charging Scheduling," *IEEE Transactions on Transportation Electrification*, 2024.

[2] Jansen, M., Gross, R., Staffell, I., "Quantitative Evidence for Modelling Electric Vehicles," *Renewable and Sustainable Energy Reviews*, 2024, vol. 199, p. 114524.

[3] Zhao, A. P., et al., "Electric Vehicle Charging Planning: A Complex Systems Perspective," *IEEE Transactions on Smart Grid*, vol. 16, no. 1, pp. 754-772, 2024.

[4] Hou, Q., Dai, N., and Huang, Y., "Voltage Regulation Enhanced Hierarchical Coordinated Volt/Var and Volt/Watt Control for Active Distribution Networks With Soft Open Points," *IEEE Transactions on Sustainable Energy*, vol. 15, no. 3, pp. 2021-2037, 2024.

[5] Mosaddek Hossain Kamal Tushar, Adel W Zeineddine, and Chadi Assi. Demand-side management by regulating charging and discharging of the ev, ess, and utilizing renewable energy. *IEEE Transactions on Industrial Informatics*, 14(1):117–126, 2017.

[6] Daner Hu, Zhenhui Ye, Yuanqi Gao, Zuzhao Ye, Yonggang Peng, and Nanpeng Yu. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Transactions on Smart Grid*, 13(6):4873–4886, 2022.

[7] Shiva Raj Pokhrel and Mohammad Belayet Hossain. Data privacy of wireless charging vehicle to grid (v2g) networks with federated learning. *IEEE Transactions on Vehicular Technology*, 71(8):9032–9037, 2022.

[8] Michela Moschella, Pietro Ferraro, Emanuele Crisostomi, and Robert Shorten. Decentralized assignment of electric vehicles at charging stations based on personalized cost functions and distributed ledger technologies. *IEEE Internet of Things Journal*, 8(14):11112–11122, 2021.

[9] Young Jin Kim, James L Kirtley, and Leslie K Norford. Reactive power ancillary service of synchronous dgs in coordination with voltage control devices. *IEEE Transactions on Smart Grid*, 8(2):515–527, 2015.

[10] Yashodhan P Agalgaonkar, Bikash C Pal, and Rabih A Jabr. Distribution voltage control considering the impact of pv generation on tap changers and autonomous regulators. *IEEE Transactions on Power Systems*, 29(1):182–192, 2013.

[11] Pedram Jahangiri and Dionysios C Aliprantis. Distributed volt/var control by pv inverters. *IEEE Transactions on power systems*, 28(3):3429–3439, 2013.

[12] Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems*, 34:3271–3284, 2021.

[13] Shengyi Wang, Jiajun Duan, Di Shi, Chunlei Xu, Haifeng Li, Ruisheng Diao, and Zhiwei Wang. A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning. *IEEE Transactions on Power Systems*, 35(6):4644–4654, 2020.

[14] Wenqi Cui, Jiayi Li, and Baosen Zhang. Decentralized safe reinforcement learning for inverter-based voltage control. *Electric Power Systems Research*, 211:108609, 2022.

[15] Wei Wang, Nanpeng Yu, Yuanqi Gao, and Jie Shi. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Transactions on Smart Grid*, 11(4):3008–3018, 2019.

[16] Yuanqi Gao, Wei Wang, and Nanpeng Yu. Consensus multi-agent reinforcement learning for volt-var control in power distribution networks. *IEEE Transactions on Smart Grid*, 12(4):3594–3604, 2021.

[17] Zhichun Yang, Fan Yang, Huaidong Min, Hao Tian, Wei Hu, Jian Liu, and Nasrin Eghbalian. Energy management programming to reduce distribution network operating costs in the presence of electric vehicles and renewable energy sources. *Energy*, 263:125695, 2023.

[18] Xiaoying Tang, Suzhi Bi, and Ying-Jun Angela Zhang. Distributed routing and charging scheduling optimization for internet of electric vehicles. *IEEE Internet of Things Journal*, 6(1):136–148, 2018.

[19] Danilo Sbordone, I Bertini, B Di Pietra, Maria Carmen Falvo, A Genovese, and Luigi Martirano. Ev fast charging stations and energy storage technologies: A real implementation in the smart micro grid paradigm. *Electric Power Systems Research*, 120:96–108, 2015.

[20] Sepehr Semsar, Theodore Soong, and Peter W Lehn. On-board single-phase integrated electric vehicle charger with v2g functionality. *IEEE Transactions on Power Electronics*, 35(11):12072–12084, 2020.

[21] Jyotsna Singh and Rajive Tiwari. Cost benefit analysis for v2g implementation of electric vehicles in distribution system. *IEEE Transactions on Industry Applications*, 56(5):5963–5973, 2020.

[22] Sara Deilami, Amir S Masoum, Paul S Moses, and Mohammad AS Masoum. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *IEEE Transactions on smart grid*, 2(3):456–467, 2011.

[23] Badra Souhila Guendouzi, Samir Ouchani, Hiba EL Assaad, and Madeleine EL Zaher. A systematic review of federated learning: Challenges, aggregation methods, and development tools. *Journal of Network and Computer Applications*, page 103714, 2023.

[24] Shahab Bahrami, Yu Christine Chen, and Vincent WS Wong. Deep reinforcement learning for demand response in distribution networks. *IEEE Transactions on Smart Grid*, 12(2):1496–1506, 2020.

[25] Jae-Hoon Lee, Ji-Yoon Park, Hoon-Seong Sim, and Hyun-Suk Lee. Multi-residential energy scheduling under time-of-use and demand charge tariffs with federated reinforcement learning. *IEEE Transactions on Smart Grid*, 14(6):4360–4372, 2023.

[26] Ouns Bouachir, Moayad Aloqaily, Öznur Özkasap, and Faizan Ali. Federatedgrids: Federated learning and blockchain-assisted p2p energy sharing. *IEEE Transactions on Green Communications and Networking*, 6(1):424–436, 2022.

[27] Zhenyi Wang, Peipei Yu, and Hongcai Zhang. Privacy-preserving regulation capacity evaluation for hvac systems in heterogeneous buildings based on federated learning and transfer learning. *IEEE Transactions on Smart Grid*, 14(5):3535–3549, 2022.

[28] Linfang Yan, Xia Chen, Yin Chen, and Jinyu Wen. A cooperative charging control strategy for electric vehicles based on multiagent deep reinforcement learning. *IEEE Transactions on Industrial Informatics*, 18(12):8765–8775, 2022.

[29] Yuanming Shi, Shuhao Xia, Yong Zhou, Yijie Mao, Chunxiao Jiang, and Meixia Tao. Vertical federated learning over cloud-ran: Convergence analysis and system optimization. *IEEE Transactions on Wireless Communications*, 23(2):1327–1342, 2023.

[30] Khaled B Letaief, Yuanming Shi, Jianmin Lu, and Jianhua Lu. Edge artificial intelligence for 6g: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, 40(1):5–36, 2021.

[31] Chaoxu Mu, Zhaoyang Liu, Jun Yan, Hongjie Jia, and Xiaoyu Zhang. Graph multi-agent reinforcement learning for inverter-based active voltage control. *IEEE Transactions on Smart Grid*, 2023.

[32] Pengcheng Chen, Shichao Liu, Xiaozhe Wang, and Innocent 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(4):2656–2667, 2022.

[33] Elia. Solar pv power generation data. https://www.elia.be/en/grid-data/power-generation/solar-pv-power-generation-data.

[34] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[35] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[36] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, pages 2961–2970. PMLR, 2019.

[37] R. Khezri, D. Steen, E. Wikner, and L. A. Tuan, "Optimal V2G scheduling of an EV with calendar and cycle aging of battery: An MILP approach," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 4, pp. 10497–10507, 2024.

[38] Y. Wang, R. Lian, H. He, J. Betz, and H. Wei, "Auto-tuning dynamics parameters of intelligent electric vehicles via Bayesian optimization," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 3, pp. 6915–6927, 2023.

[39] A. Hamednia, N. Murgovski, J. Fredriksson, J. Forsman, M. Pourabdollah, and V. Larsson, "Optimal thermal management, charging, and eco-driving of battery electric vehicles," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7265–7278, 2023.

[40] Trindade, A., ElectricityLoadDiagrams20112014, UCI Machine Learning Repository, vol. 10, pp. C58–C86, 2015.

[41] Office of Energy Efficiency & Renewable Energy, Commercial and residential hourly load profiles for all TMY3 locations in the United States, 2014. [Online]. Available: https://openei.org/datasets/dataset/

[42] Li Y, Zhang Z, Xing Q. Real-time online charging control of electric vehicle charging station based on a multi-agent deep reinforcement learning. *Energy*, 2025, 319:135095.

[43] Hu D, Li Z, Ye Z, et al. Multi-agent graph reinforcement learning for decentralized Volt-VAR control in power distribution systems. *International Journal of Electrical Power & Energy Systems*, 2024, 155:109531.

[44] Shang Y, Li Z, Li S, et al. An information security solution for vehicle-to-grid scheduling by distributed edge computing and federated deep learning. *IEEE Transactions on Industry Applications*, 2024, 60(3):4381–4395.

[45] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 22–31. PMLR, 2017.

[46] Mingyu Yang, Yaodong Yang, Zhenbo Lu, Wengang Zhou, and Houqiang Li. Hierarchical multi-agent skill discovery. *Advances in Neural Information Processing Systems*, 36:61759–61776, 2023.

Qingwei Tang is currently pursuing the Ph.D. degree at the School of Electrical and Automation Engineering, Hefei University of Technology, China. His research interests focus on multi-agent reinforcement learning (MARL) and its applications in complex systems, including wireless communication networks, modern power systems. His work aims to enhance the intelligence, sustainability, and resilience of critical infrastructures through advanced learning and optimization techniques.

Wei Sun (Senior Member, IEEE) received his B.E. degree in Automation, M.S. degree in Detection Technology and Automatic Equipment, and Ph.D. in Electrical Engineering from Hefei University of Technology, China, in 2004, 2007, and 2012, respectively. He is currently a Professor at Hefei University of Technology. His research interests include wireless networks, networked control systems, and microgrids.

Zhi Liu (Senior Member, IEEE) received Ph.D. degree in informatics in National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interest includes video network transmission and mobile edge computing. He is now an editorial board member of Springer Wireless Networks and IEEE Transactions on Multimedia. He is a senior member of IEEE.

Yang Xiao (Fellow, IEEE) received the B.S. and M.S. degrees in computational mathematics from Jilin University, Changchun, China, in 1989 and 1991, respectively, and the M.S. and Ph.D. degrees in computer science and engineering from Wright State University, Dayton, OH, USA, in 2000 and 2001, respectively. He is currently a Full Professor with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL, USA. Dr. Xiao directed more than 20 doctoral dissertations and supervised over 20 M.S. theses/projects. He has authored or co-authored more than 300 Science Citation Index (SCI)-indexed journal papers (including over 70 IEEE/ACM Transactions) and 300 Engineering Index (EI)-indexed refereed conference papers and book chapters related to these research areas. His research interests include cyber-physical systems, the Internet of Things, security, wireless networks, smart grids, and telemedicine. Dr. Xiao was a Voting Member of the IEEE 802.11 Working Group from 2001 to 2004, involving the IEEE 802.11 (Wi-Fi) standardization work. He is a Fellow of IET, AAIA, AIIA, and ACIS. Dr. Xiao was a Guest Editor 37 times for different international journals, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) during 2022–2023, IEEE TRANSACTIONS ON NET-WORK SCIENCE AND ENGINEERING (TNSE) in 2021, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING in 2021, IEEE NETWORK in 2007, IEEE WIRELESS COMMUNICATIONS in 2006 and 2021, IEEE Communications Standards Magazine in 2021, and Mobile Networks and Applications (MONET)(ACM/Springer) in 2008. He is also the Editor-in-Chief of Cyber-Physical Systems Journal, International Journal of Sensor Networks (IJSNet), and International Journal of Security and Networks (IJSN). Dr. Xiao has been an Editorial Board Member or an Associate Editor for 20 international journals, including the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING since 2022, IEEE TRANSACTIONS ON CYBERNETICS since 2020, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS from 2014 to 2015, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2009, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2007 to 2014. He is/was a Member of Technical Program Committee for more than 300 conferences. He was the recipient of the IEEE TNSE Excellent Editor Award in 2022 and 2023.

Xiaohui Yuan (Senior Member, IEEE) is an Associate Professor and the Director of the Computer Vision and Intelligent Systems Lab at the University of North Texas, Denton, TX, USA. His research interests include artificial intelligence and machine learning. Dr. Yuan was a recipient of the Ralph E. Powe Professor Award in 2008 and the U.S. Air Force Visiting Professor Award in 2011, 2012, and 2013, respectively. He serves as an associate editor, the editorial board member, and a guest editor for several journals, and an organizing member for many international conferences.

Chanjuan Zhao received the B.S. degree in automation, the M.S. degree in control theory and control engineering, and the Ph.D. degree in electrical engineering from Hefei University of Technology, Hefei, China, in 2010, 2016, and 2020, respectively. She is currently a Lecturer with Anhui University, Hefei, China. Her research interests include networked control of microgrids, distributed optimization, and reinforcement learning.