www.rericjournal.ait.ac.th

Coordinated Scheduling of Flexible Resources in Active Distribution Networks Using Multi-Agent Reinforcement Learning and Safety-Constrained Control

Zhishuang Wang [+], Chao Pang [+,1], Jian Chen [#], and Shiqian Ma [+]

ARTICLE INFO

Article history:
Received 14 October 2025
Received in revised form
10 March 2026
Accepted 17 April 2026

Keywords:
Active distribution network
Flexible resource scheduling
Multi-agent reinforcement learning
Multi-objective coordinated control
Quadratic programming safety layer

ABSTRACT

The rapid popularization of controllable loads such as electric vehicles (EVs), battery energy storage systems (BESS), and heating, ventilating and air conditioning (HVAC) has greatly improved the flexibility of active distribution networks, but it also brings challenges to safe and efficient real-time scheduling. However, most existing control strategies rely on either static optimization or single-agent reinforcement learning, which have shortcomings in scalability, adaptability to dynamic operating conditions, and ensuring network constraint satisfaction. To remedy these deficiencies, we propose a multi-agent reinforcement learning (MARL) -based scheduling framework combined with quadratic programming (QP) safety projection, which unifies flexible resource modeling, learns adaptive cost-aware and comfort-aware control policies, and guarantees real-time physical feasibility. The case study on the IEEE 123-node test system shows the superior performance of the proposed method compared to the existing methods. Specifically, in a typical scenario with 40% PV penetration, the proposed framework achieves 19.3% peak demand reduction and 13.4% operational cost savings. Moreover, it maintains a high EV user satisfaction of 96.5% and strictly guarantees network security with a zero voltage violation rate, which indicates its potential for reliable deployment in modern active distribution networks.

1. INTRODUCTION

The global transition to low-carbon sustainable energy systems and the rapid spread of distributed energy sources (DER) such as photovoltaic power generation, battery energy storage systems, and electric vehicles are fundamentally reshaping the structure and operation of distribution networks [1]. At the same time, the electrification of transportation and the increasing prevalence of flexible building loads, including heating, ventilation and air conditioning (HVAC) systems [2], are transforming demand into a more controlled but highly uncertain component of the power grid [3]. These developments have created unprecedented opportunities for demand side management, peak shaving [4], and integration of renewable [5], while also introducing new operational risks. However, the stochastic and dynamic nature of distributed energy sources and flexible loads, coupled with stringent cybersecurity constraints such as voltage limits [6] and line thermal capacities [7], make it increasingly challenging to maintain reliable and cost-efficient operation [8].

Faced with these dynamic changes, the traditional operation mode [9] based on day-ahead scheduling and static rule control is no longer applicable. Predefined charging and discharging curves cannot adapt to sudden changes in PV output, unexpected takeout of EVs [10], or unanticipated load peaks [11]. As a result, the distribution network often suffers from voltage fluctuations [12], thermal overloads [13], increased system losses, and rising operating costs, which reduce reliability [14] and economic benefits. This environment highlights the urgent need for data-driven [15], adaptive, and computationally efficient scheduling frameworks to fully exploit the flexibility of distributed energy sources while guaranteeing real-time network security [16].

Despite this pressing need, a rigorous review of existing scheduling strategies reveals a persistent dichotomy between computational scalability and strict operational security in the management of flexible resources [17]. Centralized optimization techniques, including mixed integer linear programming (MILP) and model predictive control (MPC) [18], offer mathematically rigorous solutions but suffer from severe scalability bottlenecks. When applied to large-scale networks with hundreds of heterogeneous DERs [19], their computational burden becomes unsustainable for real-time applications, particularly when dealing with rolling updates and non-convex constraints [20]. In contrast, heuristic and price-based schemes are fast but

DOI: https://doi.org/10.64289/iej.26.0206.5530729

[+]State Grid Tianjin Electric Power Research Institute, Tianjin, China 300380

[#]State Grid Tianjin Electric Power Company, Tianjin, China 300380.

[1]Corresponding author:
Email: pangchao1111@outlook.com

lack theoretical guarantees and often produce control actions that are locally advantageous but globally detrimental to network security [21].

To circumvent the online computational limitations of classical optimization, reinforcement learning (RL) has emerged as a compelling data-driven alternative that can bypass the online computational limitations of classical optimization [22]. While multi-agent reinforcement learning (MARL) with centralized training and distributed execution (CTDE) has successfully mitigated the curse of dimensionality to achieve scalable coordination [23], [24], key analyses of current MARL applications in power systems reveal a fundamental flaw: the lack of deterministic physical feasibility. The vast majority of existing MARL frameworks formulate distribution network scheduling as a standard, unconstrained Markov decision process (MDP). To handle stringent network constraints (e.g., voltage boundaries [25] and line thermal capacity), the vast majority of current literature relies on penalty-based reward shaping or Lagrange relaxation techniques. However, these soft-constraint approaches only encourage the security of statistical expectations during the offline training phase. Their attempts to map complex algebraic power flow constraints using purely stochastic policy gradients are inherently unable to exclude severe transient physical violations when the policy encounters unseen out-of-distribution scenarios during real-time execution. In addition, while some security exploration and shielding methods offer stringent guarantees, they are primarily designed for robot kinematics and are computationally difficult to handle the continuous, high-dimensional power flow manifolds of active distribution networks. Furthermore, existing MARL objectives remain largely one-dimensional and fail to systematically balance competing multi-objective criteria such as accurate grid support [26] and uncompromising end-user comfort.

To bridge the gap between the adaptability of soft learning strategies and the absolute deterministic reliability required for real-world grid operations, this paper proposes a MARL-based scheduling framework that integrates a quadratic programming (QP) secure projection layer for real-time [27], network-constrained heterogeneous flexible resource coordination in active distribution systems [28]. The main contributions are as follows:

1) Comprehensive Distributed Modeling: A unified multi-agent representation of EV, BESS, and HVAC loads is constructed, featuring a comprehensive system state space including node voltage, electricity price, state of charge (SoC) level, indoor temperature, and user travel information. This design supports scalable collaborative decision-making while strictly preserving the heterogeneous characteristics of flexible resources.
2) Multi-Objective Reward Design: A novel reward function is designed to jointly optimize grid-side objectives (cost minimization, peak demand reduction, and voltage stability) and user-side preferences (EV SoC implementation and comfort maintenance). This mechanism successfully links system-level economic efficiency with end-user experience.
3) Real-time control with security guarantees: A QP-based projection layer is introduced to ensure that the actions generated by the MARL agent strictly satisfy physical network constraints, including voltage and current limits, line capacity, SoC boundaries, and temperature comfort thresholds. This layer acts as a mathematically hard constraint filter, guaranteeing safe real-time deployment and completely preventing unsafe control behavior under unseen operating conditions.

The remainder of this paper is organized as follows: Section 2 introduces the system modeling and problem description of heterogeneous flexible resources and network constraints. Section 3 introduces the proposed MARL framework, reward design, and QP-based secure projection mechanism. Section 4 verifies the effectiveness and scalability of the proposed method through detailed numerical studies and comparative analyses. Section 5 concludes the paper and discusses future research directions.

2. PROBLEM FORMULATION

This section mathematically formulates the problem of coordinated scheduling of EVs, BESS, and HVAC systems in an active distribution network. The formulation proceeds by defining the system modeling and notations, then present the mathematical models of flexible resources, and finally define the optimization objective.

2.1 System Model and Notations

We model the distribution network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$, where $\mathcal{V}$ is the set of buses (nodes), with $|\mathcal{V}| = N_b$ denoting the number of buses. $\mathcal{E}$ is the set of distribution lines (edges), with $|\mathcal{E}| = N_l$ denoting the number of lines. $\boldsymbol{A}$ is the network adjacency matrix, with $A_{ij} = 1$ if buses i and j are connected, otherwise $A_{ij} = 0$. Each bus $i \in \mathcal{V}$ at time t is characterized by a feature vector (See Equation 1):

$$x_{i,t} = \left[v_{i,t}, p_{i,t}^{\text{base}}, p_{i,t}^{\text{RES}}, \overline{SoC}_{i,t}^{EV}, \text{SoC}_{i,t}^{\text{BESS}}, \bar{T}_{i,t}^{\text{in}}\right] \quad (1)$$

where $v_{i,t}$ is nodal voltage magnitude, $p_{i,t}^{\text{base}}$ is the base load, $p_{i,t}^{\text{RES}}$ is renewable injection.

To appropriately represent scenarios where multiple controllable resources are aggregated at a single bus i $\overline{SoC}_{i,t}^{EV}$ and $\bar{T}_{i,t}^{\text{in}}$ denote the averaged state-of-charge of all connected EVs and the averaged indoor temperature of all HVAC-equipped buildings at that node, respectively. $\text{SoC}_{i,t}^{\text{BESS}}$ represents state-of-charge of the stationary BESS. $\boldsymbol{x}_{i,t}$ defines the aggregated bus-level states primarily used for network-level power flow calculations and safety evaluations. In contrast, the microscopic agent-level observations contain the specific, unaggregated individual states (e.g., exact SoC and temperature parameters of a specific EV or HVAC unit) required for decentralized MARL decision-making.

To guarantee practical deployability during the decentralized execution phase, it is important to clarify how agents obtain these state variables. Specifically, the nodal voltage magnitude $v_{i,t}$ does not require global broadcast from a central coordinator. Instead, it is directly obtained through real-time local measurements by advanced metering infrastructure or the local controllers of the flexible resources (e.g., EV station, BESS inverters, and building energy management systems) deployed at bus i. Similarly, local power injections ($p_{i,t}^{\text{base}}$ and $p_{i,t}^{\text{RES}}$) are also monitored locally. This purely local data collection mechanism fits perfectly with the decentralized execution paradigm, significantly reducing communication overhead and latency.

2.2 Nodal Power Balance

The net active power injection at bus i at time t is governed by (See Equation 2):

$$P_{i,t} = p_{i,t}^{\text{base}} + \sum_{k\in\mathcal{E}_i} u_{k,t}^{\text{EV}} + \sum_{b\in\mathcal{B}_i} u_{b,t}^{\text{BESS}} + \sum_{h\in\mathcal{H}_i} u_{h,t}^{\text{HVAC}} - p_{i,t}^{\text{RES}} \tag{2}$$

where $u_{k,t}^{\text{EV}}$, $u_{b,t}^{\text{BESS}}$, and $u_{h,t}^{\text{HVAC}}$ represent the control actions of EV k, BESS b, and HVAC h at bus i, respectively. For the HVAC system, $u_{h,t}^{\text{HVAC}}$ directly represents the actual electrical active power consumed from the distribution network. The corresponding thermal power delivered to the building to change the indoor temperature is determined by multiplying this input electrical power by the performance factor, which is included in the energy conversion efficiency parameter η_h^{HVAC} in the subsequent thermodynamical model.

2.3 Flexible Resource Models

2.3.1 EV Charging Model

SoC dynamics of EV k are described by:

$$\text{SoC}_{t+1}^{k} = \text{SoC}_{t}^{k} + \frac{\eta_{\text{ch}}^{k}[u_{k,t}^{\text{EV}}]_+ - \frac{1}{\eta_{\text{dis}}^{k}}[u_{k,t}^{\text{EV}}]_-}{E_k^{\max}}\Delta t \tag{3}$$

subject to:

$$0 \le \text{SoC}_t^k \le 1, \quad |u_{k,t}^{\text{EV}}| \le \bar{u}_k^{\text{EV}} \tag{4}$$

where $E_k^{\max}$ represents the battery capacity, $\bar{u}_k^{\text{EV}}$ is max charging/discharging power, and $[x]_+$, $[x]_-$ denote positive/negative parts of the bidirectional action, respectively. To maximize operational flexibility, it is assumed that electric vehicle discharge (V2G) is physically allowed during the entire plug-in period. However, the actual discharge behavior is regulated by a multi-objective reward mechanism, detailed in Section 3.2, in order to strictly meet the travel needs of users and reduce battery loss. Specifically, V2G actions are severely penalized if they fail to meet the user's required departure SoC or result in excessive battery depletion costs, thus establishing a practical balance between grid support and EV user satisfaction (See Equations 3 and 4).

2.3.2 BESS Model

Similarly, the dynamics and operational constraints of the BESS b are modeled as follows (See Equations 5 and 6):

$$\text{SoC}_{t+1}^{b} = \text{SoC}_{t}^{b} + \frac{\eta_{\text{ch}}^{b}[u_{b,t}^{\text{BESS}}]_+ - \frac{1}{\eta_{\text{dis}}^{b}}[u_{b,t}^{\text{BESS}}]_-}{E_b^{\max}}\Delta t \tag{5}$$

with:

$$\text{SoC}_{\text{b}}^{\min} \le \text{SoC}_{\text{t}}^{\text{b}} \le \text{SoC}_{\text{b}}^{\max}, |u_{b,t}^{BESS}| \le \bar{u}_{b,t}^{BESS} \tag{6}$$

2.3.3 HVAC Model

The indoor temperature dynamics of building h follow a first-order thermal RC model:

$$T_{h,t+1}^{\text{in}} = T_{h,t}^{\text{in}} + \frac{\Delta t}{C_h}\left(\frac{T_t^{\text{out}} - T_{h,t}^{\text{in}}}{R_h} - \eta_h^{\text{hvac}} u_{h,t}^{\text{HVAC}}\right) \tag{7}$$

with comfort constraints:

$$T_h^{\min} \le T_{h,t}^{\text{in}} \le T_h^{\max} \tag{8}$$

Although the first-order RC equivalent circuit is a simplified approximation to the thermal dynamics of complex commercial buildings, this form is deliberately adopted to maintain computational tractability under the continuous state space of the MARL framework and the linear constraints of the QP projection layer. More importantly, the proposed closed-loop rolling execution mechanism is able to effectively compensate for the potential impact of these unmodeled higher-order effects. By continuously updating state observations using real-time temperature feedback, the framework can dynamically correct HVAC control actions, thus ensuring strict adherence to user comfort bounds even when the physical model is simplified (See Equations 7 and 8).

2.4 Power Flow and Network Constraints

We adopt a simplified Lin*DistFlow* model:

$$P_{ij,t} = \frac{v_{i,t} - v_{j,t}}{R_{ij}}, \quad |P_{ij,t}| \le \bar{P}_{ij} \tag{9}$$

And nodal voltage limits:

$$v^{\min} \le v_{i,t} \le v^{\max} \tag{10}$$

The power flow calculation method adopted in Equations (9) and (10) is based on the following assumptions: node voltages are close to the nominal value of 1.0 p.u., line losses are negligible, and reactive power has been locally compensated. Therefore, the proposed model is mainly suitable for upper active power scheduling and energy management. Although ignoring the exact R/X ratio and reactive power dynamics introduces some unmodeled AC characteristics, this intentional simplification is fundamentally necessary. It ensures that the subsequent QP-based safe projection layer remains a strictly convex optimization problem, thus guaranteeing the millisecond level of computational efficiency required for real-time MARL execution.

2.5 Objective Function

The overall objective is to minimize the weighted sum of operational costs and peak load by penalizing comfort violation behaviors:

$$\min_{\{u^{\mathrm{EV}},u^{\mathrm{BESS}},u^{\mathrm{HVAC}}\}} \sum_{t=1}^{T} \lambda_t P_t^{\mathrm{sys}} + \alpha \max_t P_t^{\mathrm{sys}} + \beta \sum_{h,t} \left(T_{h,t}^{\mathrm{in}} - T_{h,t}^{\mathrm{set}}\right)^2 \quad (11)$$

where $P_t^{\mathrm{sys}} = \sum_i P_{i,t}$ is total system demand at time t, λ_t is time-of-use (TOU) price, α, β are weighting coefficients for peak shaving and comfort penalty (See Equation 11).

3. METHODOLOGY

Building upon the physical models established in Section 2, this section develops a closed-loop scheduling framework that seamlessly integrates MARL with a QP-based safety projection. This architecture ensures real-time adaptability while mathematically guaranteeing the operational feasibility of heterogeneous flexible resources in active distribution networks. The algorithmic pipeline encompasses a CTDE-based MARL formulation, a multi-objective reward mechanism, proximal policy optimization (PPO) -driven policy updates, and a strict safety projection layer.

3.1. MARL Formulation

We denote the set of agents as $\mathcal{I} = \mathcal{E} \cup \mathcal{B} \cup \mathcal{H}$, where $\mathcal{E}$, $\mathcal{B}$, and $\mathcal{H}$ represent EVs, BESS units, and HVAC devices, respectively. At time step t, each agent i observes a local state o_t^i, which inherently comprises its individual physical status (e.g., SoC for EVs/BESS, indoor temperature for HVACs), the local bus voltage magnitude $v_{i,t}$, and the temporal feature t. Conditioned on this local observation, agent i selects an action a_t^i, via its stochastic policy π_{θ^i}, and interacts with the environment.

The joint observation vector is defined as $\boldsymbol{o}_t = [o_t^i]_{i\in\mathcal{I}}$, and the global action vector is $\boldsymbol{a}_t = [a_t^i]_{i\in\mathcal{I}}$. For the centralized training phase, the global state s_t evaluated by the central critic is constructed by concatenating all local observations (i.e., $s_t \supseteq \boldsymbol{o}_t$) and augmenting them with global exogenous variables such as the system-wide base load and aggregate PV generation profiles. The decentralized joint policy is factorized as (See Equation 12):

$$\pi_\Theta(\boldsymbol{a}_t|\boldsymbol{o}_t) = \prod_{i\in\mathcal{I}} \pi_{\theta^i}(a_t^i|o_t^i) \quad (12)$$

where π_{θ^i} is the stochastic policy of agent i, parameterized by θ^i, and $\Theta = \{\theta^i\}_{i\in\mathcal{I}}$ is the set of all policy parameters. Both the individual actor policies and the centralized critic value function are represented by multi-layer perceptron (MLP).

3.2 Multi-Objective Reward Design

To align MARL training with this objective, we design a composite reward function (See Equation 13):

$$r_t = -\alpha C_t - \beta \Phi_{\mathrm{peak}}(t) - \gamma \sum_{h\in\mathcal{H}} \ell_{h,t}^{\mathrm{HVAC}} - \mu \sum_{k\in\mathcal{E}} \ell_k^{\mathrm{EV}} - \nu \ell_t^{\mathrm{grid}} \quad (13)$$

where the individual penalty components (operating cost, peak demand surrogate, EV satisfaction, HVAC comfort, and grid voltage/thermal penalties) map directly to the physical operational constraints. Operating cost:

$$C_t = \sum_{i\in\mathcal{V}} \lambda_t P_{i,t}^{\mathrm{grid}} + \sum_{b\in\mathcal{B}} c_b^{\mathrm{deg}} |u_{b,t}^{\mathrm{BESS}}| + \sum_{k\in\mathcal{E}} c_k^{\mathrm{deg}} |u_{k,t}^{\mathrm{EV}}| \quad (14)$$

Peak demand surrogate:

$$\Phi_{\mathrm{peak}}(t) = \left(\sum_{i\in\mathcal{V}} P_{i,t}^{\mathrm{grid}}\right)^2 \quad (15)$$

EV satisfaction penalty:

$$\ell_k^{\mathrm{EV}} = \kappa_k \max\{0,\ \mathrm{SoC}_{k,t^{\mathrm{dep}}}^{\mathrm{req}} - \mathrm{SoC}_{k,t^{\mathrm{dep}}}\} \quad (16)$$

HVAC comfort penalty:

$$\ell_{h,t}^{\mathrm{HVAC}} = \xi_h (T_{h,t}^{\mathrm{in}} - T_{h,t}^{\mathrm{set}})^2 \quad (17)$$

Grid penalty:

$$\ell_t^{grid} = \sum_{i\in\mathcal{V}} \left[\psi_i^+ \max\left(0, v_{i,t} - v^{\max}\right) + \psi_i^- \max\left(0, v^{\min} - v_{i,t}\right)\right] + \sum_{(i,j)\in\mathcal{E}} \chi_{ij} \max(0, |S_{ij,t}| - S_{ij}^{\max}) \quad (18)$$

This scalarization process ensures that the learned policies simultaneously achieve grid-level economic efficiency and user satisfaction.

The selection of reward weights $(\alpha, \beta, \gamma, \mu, \nu)$ is essential to balance the multi-objective operational goals. These coefficients are calibrated through a two-step process: (1) dimensional normalization to align disparate physical scales, and (2) empirical tuning to stabilize convergence.

A fundamental theoretical challenge arises when mapping the macroscopic peak demand objective to the step-wise MARL reward. In the original formulation, peak demand is evaluated using a max operator over the entire horizon. However, using the exact max operator in MARL would yield a sparse, episode-level reward signal, severely hindering the agents' temporal credit assignment. To circumvent this sparsity, we adopt the squared sum of grid power, $\Phi_{\mathrm{peak}}(t) = \left(\sum_{i\in\mathcal{V}} P_{i,t}^{\mathrm{grid}}\right)^2$, as a rigorous, step-wise surrogate. Mathematically, penalizing the squared power continuously discourages large load deviations and minimizes the variance of the net load profile. This variance-minimization effect intrinsically aligns with the max operator's goal of load flattening and peak shaving, while providing a dense and differentiable learning signal at every time step t (See Equations 14-18).

3.3 PPO under CTDE

Driven by the imperative for training stability in multi-agent environments and robustness against the severe non-stationarity issues commonly faced by off-policy algorithms using experience replay, we adopt PPO for policy updates, with a centralized critic $V_\phi(s_t)$.

Importance ratio (See Equation 19):

$$\rho_t^i(\theta^i) = \frac{\pi_{\theta^i}(a_t^i|o_t^i)}{\pi_{\theta_{\text{old}}^i}(a_t^i|o_t^i)} \tag{19}$$

Clipped objective:

$$\mathcal{L}_{\text{actor},i}^{\text{PPO}} = \mathbb{E}_t[min(\rho_t^i \hat{A}_t,\ \text{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon)\hat{A}_t) - \beta_{\text{ent}}\mathcal{H}(\pi_{\theta^i}(\cdot|o_t^i))] \tag{20}$$

GAE advantage estimator

$$\hat{A}_t = \sum_{\ell=0}^{L-1}(\gamma\lambda)^\ell \delta_{t+\ell},\ \delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{21}$$

Critic loss:

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E}_t[(V_\phi(s_t) - \hat{V}_t)^2] \tag{22}$$

Total objective:

$$\mathcal{L}_{\text{total}} = \sum_{i\in\mathcal{I}}(-\mathcal{L}_{\text{actor},i}^{\text{PPO}}) + \beta_V \mathcal{L}_{\text{critic}} \tag{23}$$

Equation (20) stabilizes policy updates by clipping; Equation (21) balances bias and variance; Equation (22) combines actor and critic objectives into the final training loss (See Equation 23).

3.4 QP-Based Safety Projection

To mathematically guarantee the feasibility of the RL-proposed actions under the stringent network and device constraints, we design a secondary projection layer. This layer solves the following strictly convex QP problem at every execution step (See Equation 24):

$$u_t^* = \text{argmin}_{\boldsymbol{u_t}} \|u_t - a_t\|_2^2 \tag{24}$$

Subject to:

$$\underline{u} \le u_t \le \overline{u} \tag{25}$$

$$\text{SoC}^{\min} \le \text{SoC}_t + Bu_t \le \text{SoC}^{\max} \tag{26}$$

$$T^{\min} \le A_T T_t^{in} + B_T u_t^{\text{HVAC}} + d_t \le T^{\max} \tag{27}$$

$$v^{\min} \le v_0 + MP_t(u_t) \le v^{\max} \tag{28}$$

$$HP_t(u_t)| \le \overline{F} \tag{29}$$

Eqs. (25)-(29) enforce device power bounds, SoC feasibility, HVAC comfort, and network security. The QP ensures that the executed action u_t^* is feasible, even if the RL proposal a_t is not. The matrices B, A_T, B_T, M, and H represent the linear mappings derived from the physical device dynamics and the network topology, which are comprehensively detailed in Appendix.

3.5. Closed-Loop Execution

The entire operation follows a rolling cycle: (i) perceiving global and local states; (ii) sampling tentative actions from π_{θ^i}; (iii) applying QP projection; (iv) executing feasible control; (v) updating the critic and actor via PPO.

To clearly illustrate the interaction among these components, Figure 1 depicts the overall framework and information flow. As shown, architecture is primarily divided into an offline centralized training phase (Phase 1) and an online decentralized execution phase (Phase 2). During real-time execution, heterogeneous agents map their local observations to tentative joint actions, which are then intercepted and corrected by a QP-based safety projection layer to ensure strict physical feasibility before interacting with the active distribution network environment. This guarantees adaptability to uncertainties and strict feasibility at each step.

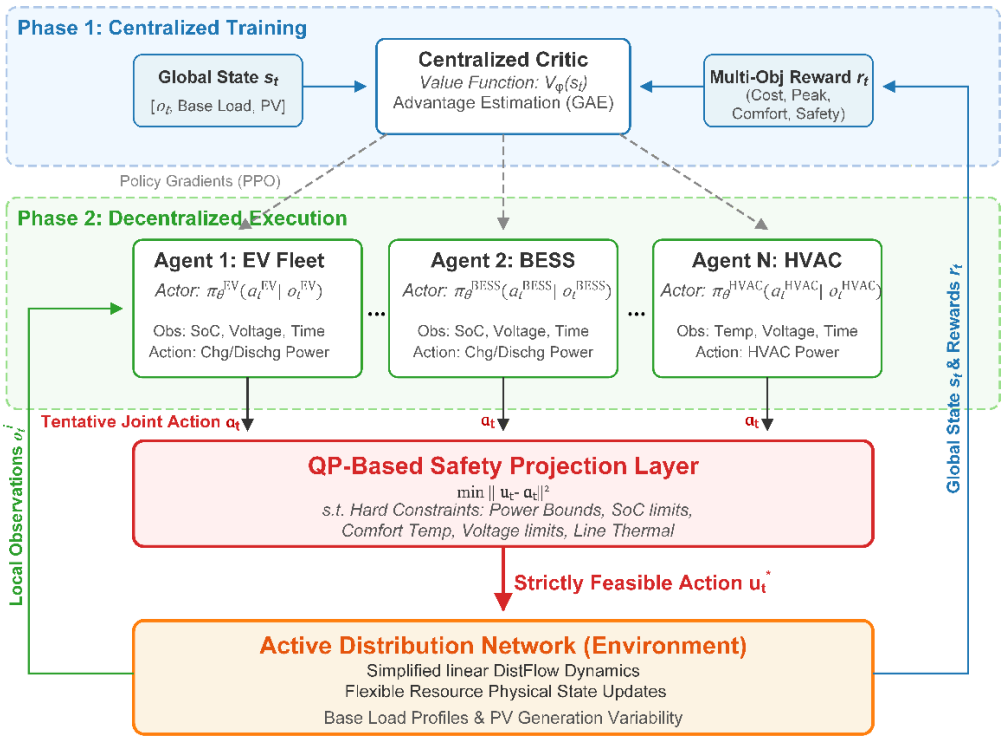

Fig. 1. Overall framework and information flow of the proposed CTDE-MARL with QP safety projection layer.

4. CASE STUDIES

To empirically validate the proposed CTDE-MARL with QP safety projection, comprehensive case studies are conducted on the IEEE 123-bus test system. We detail the test system and data, learning/solver configurations, baselines, evaluation metrics, main results, ablation studies, and sensitivity analyses. All modeling choices are consistent with Sections 2 and 3.

The experimental methodology is built upon a high-performance integrated simulation platform designed for complex power system coordination. All numerical experiments were conducted on a professional workstation equipped with Intel Core i9 CPU and 64 GB of memory to ensure computational efficiency. In terms of software ecosystem, both the MARL environment and the active distribution network's power flow dynamic model are developed using Python 3.9. Specifically, both MARL agents and training algorithms are implemented using the PyTorch framework.

In addition, the algorithm architecture and data processing have been rigorously designed to reflect real-world situations. The actor model and critic model in PPO algorithm both adopt MLP structure, including two fully connected hidden layers with 64 neurons each, and use ReLU activation function. In order to ensure the practicality of the simulation results, both the benchmark load and photovoltaic power generation curves were taken from real-world datasets. These profiles were subsequently normalized and scaled to match the specific capacity and voltage requirements of the IEEE 123-bus system.

4.1. Test System and Data

4.1.1 Network and timing

We adopt the IEEE 123-bus radial distribution system (depicted in Figure 2) with standard impedances and switch statuses (base voltage 4.16 kV; nominal per-unit voltage). The dispatch interval is $\Delta t = 15$ min over a 24 h horizon, i.e., T=96. This 15-minute dispatch interval aligns with standard distribution system operator scheduling practices. While local controllers for devices like HVACs operate on faster timescales (1–5 minutes) to track thermal setpoints, the proposed MARL framework provides the supervisory energy management signals. Voltage limits are $v^{\min} = 0.95$ p. u. and $v^{\max} = 1.05$ p. u. Line thermal limits follow feeder ratings; we aggregate them into $\overline{F}$ in the QP constraints.

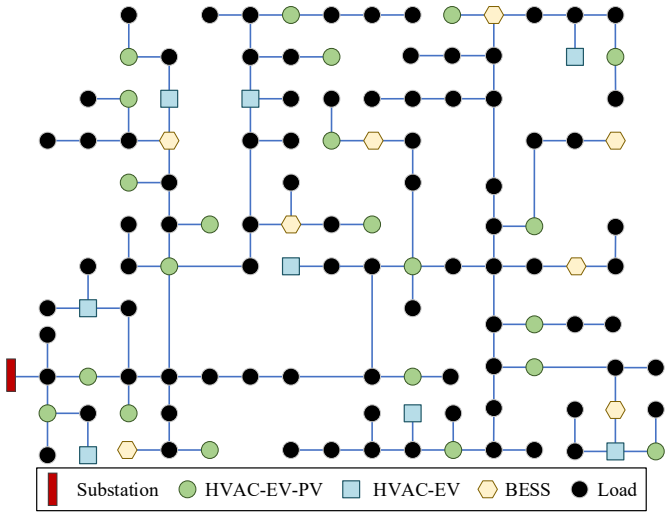

Fig. 2. IEEE 123-bus test system.

4.1.2 Demand, PV, and prices

The base load $p_{i,t}^{\text{base}}$ follows a typical normalized feeder profile comprising a mix of residential and commercial consumption patterns, scaled to a 6.0 MW daily maximum. Specifically, the daily load profile exhibits a morning peak between 07:00 and 09:00, and a more pronounced evening peak between 18:00 and 21:00. The PV injections $p_{i,t}^{\text{RES}}$ are placed on 22 selected nodes. The daily PV generation profile is derived from typical sunny-day solar irradiance data, with output gradually increasing after sunrise and peaking between 11:00 and 14:00. The installed PV capacities are distributed proportionally to the nodal base loads to achieve midday peak penetration levels $\rho_{\text{PV}} \in \{20\%, 40\%, 60\%\}$ of the local daytime demand. The TOU price λ_t adopts a three-tier structure with off-peak (e.g., 23:00–07:00), mid-peak (e.g., 07:00–10:00, 15:00–18:00), and peak (e.g., 10:00–15:00, 18:00–23:00) blocks, priced at {0.55, 0.85, 1.30} CNY/kWh, respectively.

4.1.3 Flexible resources and parameters

EV fleet: $N_{\text{EV}} = 500$ cars parked across 30 EVSE buses. Each EV has capacity $E_k^{\max} \in [45,75]$kWh , $\bar{u}_k^{\text{EV}} \in [7,11]$kW , efficiencies $(\eta_{\text{ch}}^k, \eta_{\text{dis}}^k) = (0.95,0.95)$, arrival/departure windows drawn from Gaussian mixtures (evening arrivals 17:00–21:00, morning departures 6:30–9:30). Required departure SoC $\text{SoC}_{k,t^{\text{dep}}}^{\text{req}} \in [0.8,0.95]$.

Stationary BESS: $N_{\text{BESS}} = 8$ units at feeder laterals; $E_b^{\max} \in [0.5,1.5]$MWh, $\bar{u}_b^{\text{BESS}} \in [0.25,0.5]$MW, efficiencies (0.96,0.96), SoC window [0.1,0.9].

HVAC buildings: $N_{\text{HVAC}} = 30$ commercial blocks, each modeled by the RC dynamics of (7). Comfort band $[E_h^{\min}, E_h^{\max}] = [22,26]$ ◦C, nominal setpoint $T_{h,t}^{\text{set}}$=24◦C,

thermal parameters C_h, R_h derived from floor area; COP absorbed into η_h^{hvac}.

4.2. Learning and Solver Setup

The reinforcement learning algorithm adopts PPO under the CTDE framework described in Section 3, with discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, entropy regularization coefficient $\beta_{\text{ent}} = 0.01$, clipping width $\epsilon = 0.2$, and critic loss weight $\beta_V = 0.5 (\gamma = 0.99, \lambda = 0.95, \beta_{\text{ent}} = 0.01, \epsilon = 0.2, \beta_V = 0$. Training is conducted using 5 random seeds of weather, price, and load scenarios for 1000 episodes of 24 h each, with a minibatch size of 8192 time steps and the Adam optimizer with learning rates $\eta_\theta = 3 \times 10^{-4}$ for the actor and $\eta_\phi = 1 \times 10^{-3}$ for the critic ($\eta_\theta = 3 \times 10^{-4}, \eta_\phi = 10^{-3}$). To fortify policy robustness against uncertainties, domain randomization is applied by injecting zero-mean Gaussian perturbations to the price signal (standard deviation $\sigma = 5\%$), outdoor temperature ($\sigma = 1.5$°C), and PV generation ($\sigma = 8\%$). Regarding the QP safety layer, a sparse interior-point solver with a warm-start mechanism is deployed. For the fully populated IEEE 123-bus test system, the QP problem at each time step involves exactly 538 continuous decision variables (representing the active power commands for 500 EVs, 8 BESS units, and 30 HVAC systems). Accordingly, there are approximately 2,640 linear inequality constraints strictly enforcing the power bounds, state-of-charge limits, temperature comfort bands, and the network-wide voltage and line thermal limits. Despite the large scale and high dimensionality of mathematics, typical solution times for small, medium, and large network configurations are strictly limited to approximately 42 milliseconds, 87 milliseconds, and 121 milliseconds per step, respectively. This computational efficiency fully meets the strict requirements.

4.3. Baselines

To comprehensively evaluate the proposed scheduling framework, this paper considers several representative benchmark methods. The rule-based (RB) approach adopts short-sighted control driven by time of use pricing, where electric vehicles strictly charge during low price periods, HVAC systems maintain strict set values, and energy storage systems greedily discharge during peak periods. The single agent reinforcement learning (SARL) benchmark method models the entire system as a global agent trained using an actor critic scheme but lacks explicit security guarantees. The MARL-NoQP variant adopts the same distributed PPO strategy under CTDE but omits the QP safety projection layer.

The fourth baseline method MPC uses the Lin*DistFlow* model to perform deterministic rolling optimization. It is worth noting that the prediction range of this method is limited to within two hours and has full predictive ability. It must be recognized that there are fundamental methodological differences between this traditional optimization baseline method and the proposed MARL framework. Although theoretically extending the prediction window of MPC can enhance its forward-looking optimality, for active distribution networks with hundreds of flexible resources, expanding the prediction range will exponentially increase the online computational burden, thereby violating the strict delay constraints required for real-time scheduling. In sharp contrast, the proposed MARL-QP method cleverly transfers this large amount of computational work to the offline training stage. By accumulating thousands of hours of operational experience offline, the MARL agents implicitly execute with a long-term global perspective online, requiring only lightweight neural network forward passes coupled with a millisecond-level QP projection. Consequently, the juxtaposition of MPC and MARL in these case studies extends beyond a mere numerical comparison; it explicitly highlights the critical trade-off between the online computational paralysis inherent to traditional optimization and the highly efficient, proactive execution of data-driven control.

4.4. Evaluation Metrics

To rigorously evaluate algorithmic stability against uncertainties, all quantitative performance metrics are computed across five independent random seeds and reported in the format of mean ± standard deviation. We consider system-level, user-level, and safety/feasibility metrics. All are averaged over test days unless specified.

(1) Peak reduction rate (PR) (See Equation 30)

$$\text{PR} = \frac{\max_t P_{\text{sys},t}^{\text{base}} - \max_t P_{\text{sys},t}^{\text{ctrl}}}{\max_t P_{\text{sys},t}^{\text{base}}} \tag{30}$$

$P_{\text{sys},t}^{\text{base}}$: system import without control;

$P_{\text{sys},t}^{\text{ctrl}}$: system import with the tested method.

(2) Cost saving (CS) (See Equation 31)

$$\text{CS} = \frac{\sum_t \lambda_t P_{\text{sys},t}^{\text{base}} - \sum_t \lambda_t P_{\text{sys},t}^{\text{ctrl}}}{\sum_t \lambda_t P_{\text{sys},t}^{\text{base}}} \times 100\% \tag{31}$$

(3) EV satisfaction (EV-Sat) (See Equation 32)

$$\text{EV-Sat} = \frac{1}{N_{\text{EV}}} \sum_{k=1}^{N_{\text{EV}}} 1\left(\text{SoC}_{k,t^{\text{dep}}} \geq \text{SoC}_{k,t^{\text{dep}}}^{\text{req}}\right) \times 100\% \tag{32}$$

where $1(\cdot)$ is the indicator function, which returns 1 if the condition inside is true and 0 otherwise. In this context, $1(\text{SoC}_{k,t^{\text{dep}}} \geq \text{SoC}_{k,t^{\text{dep}}}^{\text{req}})$ equals 1 when the SoC of the k-th EV at its departure time t^{dep} meets or exceeds the required SoC for user satisfaction $\text{SoC}_{k,t^{\text{dep}}}^{\text{req}}$, and 0 otherwise.

(4) HVAC comfort violation (MSE-Comfort) (See Equation 33)

$$\text{MSE-Comfort} = \frac{1}{N_{\text{HVAC}} T} \sum_{h,t} \left(T_{h,t}^{\text{in}} - T_{h,t}^{\text{set}}\right)^2 \tag{33}$$

(5) Voltage violation time (V-viol) (See Equation 34)

$$V - viol = \frac{\Delta t}{60} \sum_t \sum_{i \in \mathcal{V}} \mathbf{1}\left(v_{i,t} < v^{\min} \text{ or } v_{i,t} > v^{\max}\right) \tag{34}$$

(6) Feasibility rate (Feas-rate) (See Equation 35)

$$\text{Feas} - \text{rate} = \frac{1}{T}\sum_{t} 1\,(\text{all constraints satisfied at } t) \times 100\% \quad (35)$$

4.5. Main Results

As detailed in Table 1, the proposed MARL-QP framework demonstrates superior comparative performance under the 40% PV penetration scenario. Achieving a remarkable PR of 19.3% and CS of 13.4%, it comprehensively outperforms the RB, MPC, SARL, and MARL-NoQP baselines. Crucially, user-oriented performance is significantly elevated, with EV-Sat reaching 96.5% and the HVAC comfort error minimized to 0.49 °C².

From the network security perspective, the proposed methodology entirely eliminates voltage violations (V-viol = 0 min/day) and achieves a 100% feasibility rate. This outcome definitively proves that all physical constraints, including the voltage bounds v^min ≤v_(i,t)≤v^max , are rigorously enforced by the QP safety layer. By stark contrast, while the MARL-NoQP variant attains competitive economic gains, it exhibits severe network vulnerability (31.8 min/day of voltage violations), irrefutably confirming the absolute necessity of integrating a mathematical safety projection mechanism prior to real-world deployment.

As visually corroborated by Figure 3, the proposed framework elegantly circumvents the zero-sum trade-offs typically inherent in static or single-agent approaches, successfully maximizing grid-side economic objectives without compromising end-user satisfaction. And the narrow error bars illustrated in the figures mathematically confirm that the MARL-QP framework exhibits high statistical robustness and reproducibility, entirely dispelling concerns regarding random seed sensitivity.

Table 1. Comparative performance metrics of scheduling methods for active distribution networks with 40% PV penetration.

Method	PR (%) ↑	CS (%) ↑	EV-Sat (%) ↑	MSE-Comfort (°C²) ↓	V-viol (min/d) ↓	Feas-rate (%) ↑
RB	4.6 ± 0.6	2.1 ± 0.4	86.7 ± 1.2	0.82 ± 0.05	47.5 ± 4.5	71.3 ± 2.8
MPC	12.8 ± 0.0	8.9 ± 0.0	92.4 ± 0.0	0.54 ± 0.00	9.2 ± 0.0	96.1 ± 0.0
SARL	9.3 ± 0.8	6.0 ± 0.7	90.1 ± 1.5	0.67 ± 0.06	22.6 ± 3.8	89.4 ± 3.1
MARL-NoQP	14.1 ± 0.5	9.4 ± 0.6	92.8 ± 0.8	0.63 ± 0.04	31.8 ± 4.2	83.7 ± 2.5
Proposed	19.3 ± 0.4	13.4 ± 0.3	96.5 ± 0.5	0.49 ± 0.02	0.0 ± 0.0	100.0 ± 0.0

Table 2. Peak-hour power import and energy shifting contributions under different scheduling strategies.

Method	Peak import (MW) ↓	EV V2G (MWh) ↑	BESS discharge (MWh) ↑	HVAC shift (MWh) ↑
RB	5.92	0.00	0.38	0.22
MPC	5.39	0.21	0.64	0.33
SARL	5.54	0.17	0.53	0.29
MARL-NoQP	5.26	0.28	0.71	0.36
Proposed	4.94	0.34	0.83	0.41

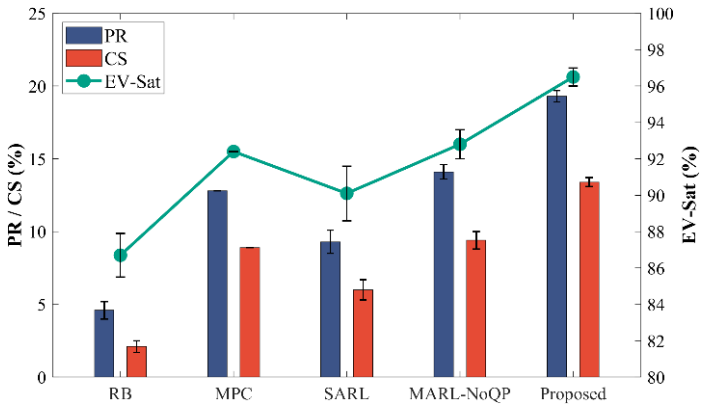

Fig. 3. Main results comparison of different scheduling methods.

Furthermore, Figure 4 and Table 2 deconstruct the exact mechanics of these improvements during the evening peak period. The proposed method increases the use of BESS discharging and EV vehicle-to-grid (V2G) services while coordinating HVAC pre-cooling, leading to the lowest peak power import among all methods. These coordinated actions enable effective peak shaving and cost reduction while respecting voltage and power flow limits. Overall, the results confirm that combining multi-agent reinforcement learning with a QP-based safety layer is a practical and reliable way to schedule heterogeneous flexible resources in active distribution networks.

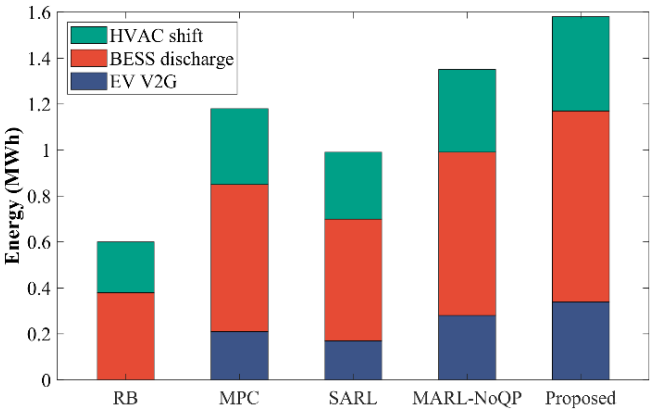

Fig. 4. Peak-hour energy contribution from flexible resources.

4.6. Ablation Studies

The necessity of the two core design architectures is strongly corroborated by the ablation studies presented in Table 3 and Table 4. The results demonstrate that the QP-based safety projection acts as a critical safeguard to prevent infeasible actions and ensure strict adherence to voltage and line limits while maintaining economic performance. Complementing this secure foundation, replacing a single global agent with a multi-agent structure under CTDE significantly improves both grid and user-side outcomes by facilitating scalable and cooperative scheduling across diverse DERs.

To profoundly understand the symbiotic interplay between the learned MARL policy and the deterministic QP safety layer, an analysis of their respective roles reveals a highly efficient division of labor. The MARL-NoQP variant organically achieves an 83.7% feasibility rate entirely on its own, a pivotal metric signifying that the foundational RL policy inherently shoulders the primary burden of system coordination. By successfully navigating the vast majority of operational constraints, the neural network allows the QP layer to intervene exclusively during the remaining 16.3% of the time steps. In these cases, the downstream projection activates specifically as an uncompromising safeguard against severe transient events or unexpected load spikes.

Because this mathematical filter operates purely as a non-intrusive safety net, the impressive economic and comfort-related outcomes are overwhelmingly driven by the learned MARL policy rather than the optimization layer. The objective function of the QP formulation is strictly confined to minimizing the Euclidean distance to the RL-proposed unconstrained actions; consequently, it does not inherently optimize for operational cost, peak reduction, or user satisfaction, leaving these global objectives entirely to the intelligent agents.

Beyond its functional non-intrusiveness, this architectural decoupling ensures that the computational overhead remains exceptionally practical for real-time deployment. Even for the fully populated 123-bus network, the maximum per-step solution time is strictly bounded at 121 ms. Relative to the standard 15-minute operational dispatch interval, this millisecond-level projection latency is virtually negligible, guaranteeing instantaneous control execution without disrupting the fluidity of the dynamic scheduling process.

Table 3. Effect of the safety projection layer on performance and feasibility.

Variant	PR (%)	CS (%)	V-viol (min/d)	Feas-rate (%)
MARL-NoQP	14.1 ± 0.5	9.4 ± 0.6	31.8 ± 4.2	83.7 ± 2.5
Proposed	19.3 ± 0.4	13.4 ± 0.3	0.0 ± 0.0	100.0 ± 0.0

Table 4. Effect of centralized training with CTDE vs. SARL.

Variant	PR (%)	CS (%)	EV-Sat (%)	MSE-Comfort (°C²)
SARL	9.3 ± 0.8	6.0 ± 0.7	90.1 ± 1.5	0.67 ± 0.06
Proposed	19.3 ± 0.4	13.4 ± 0.3	96.5 ± 0.5	0.49 ± 0.02

4.7. Sensitivity Analyses

4.7.1. PV penetration

Table 5 and Figure 5 show the performance trajectory of the proposed MARL-QP framework under gradually increasing photovoltaic penetration rates (20%, 40%, and 60%). With the increasing integration of renewable energy into the grid, both PR and CS indicators are showing a continuous upward trend. This improvement highlights the algorithm's ability to fully utilize the larger energy arbitrage space and more flexible load transfer capabilities brought by higher solar energy output. More importantly, although a photovoltaic penetration rate of 60% may bring more serious risks of reverse power flow and higher random fluctuations, QP based safety prediction can ensure that voltage deviation always remains at zero.

Table 5. Sensitivity of the proposed method to different PV penetration levels.

ρ_{PV}	Method	PR (%)	CS (%)	V-viol (min/d)
20%	Proposed	15.2 ± 0.3	9.6 ± 0.4	0.0
40%	Proposed	19.3 ± 0.4	13.4 ± 0.2	0.0
60%	Proposed	21.7 ± 0.4	15.8 ± 0.5	0.0

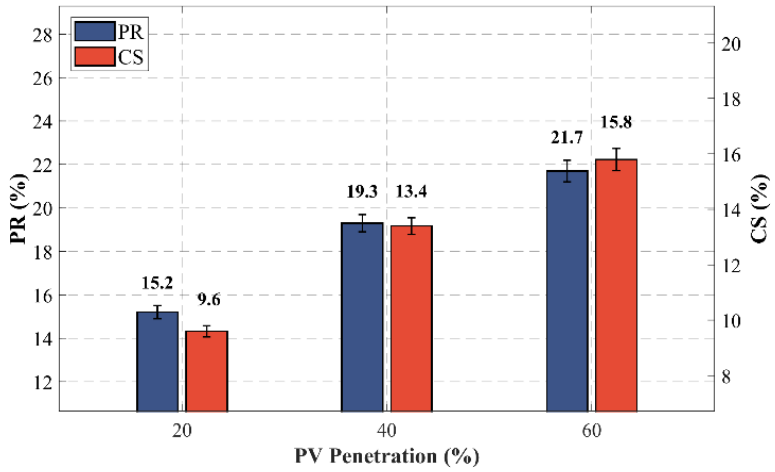

Fig. 5. Sensitivity of peak reduction and cost saving to PV penetration.

Table 6. Sensitivity of the proposed method to EV fleet size.

N_{EV}	Method	PR (%)	EV-Sat (%)	V-viol (min/d)
300	Proposed	14.7 ± 0.3	97.2 ± 0.4	0.0
500	Proposed	19.3 ± 0.4	96.5 ± 0.5	0.0
800	Proposed	23.1 ± 0.6	95.9 ± 0.7	0.0

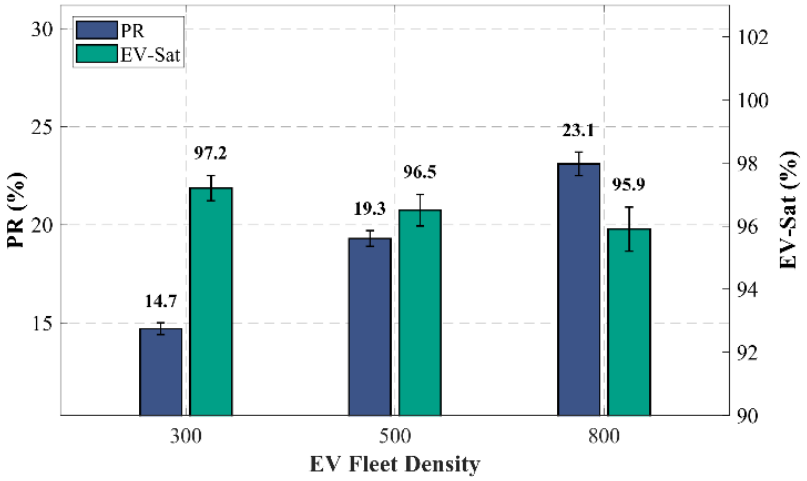

Fig. 6. Impact of EV fleet size on peak reduction and EV satisfaction.

4.7.2. Sensitivity to EV Fleet Density

Table 6 and Figure 6 investigate the algorithm scalability under different electric vehicle fleet densities (300, 500, and 800 vehicles). With the increase in the number of electric vehicles, overall flexibility will naturally improve PR (performance ratio). However, at the same time, EV satisfaction slightly decreases, reflecting the intensified competition for limited network charging capacity under strict line thermal constraints. Nevertheless, the security layer successfully achieved zero voltage violations in all scenarios. These results confirm the robustness and scalability of the framework, while also clearly revealing the inevitable physical trade-off between grid support and user experience in extreme electrification scenarios.

4.8. Feasibility Analyses

Figure 7 compares the ability of different scheduling strategies to maintain voltage feasibility in the distribution network. The left axis represents the daily cumulative voltage violation time (V-viol, in

minutes/day), and the right axis represents the percentage of feasible operating points (Feas rate). The proposed method completely eliminates voltage violations and achieves 100% feasibility, while MARL without QP layer still has obvious violations. The low feasibility of RB and SARL indicates that traditional control and single agent reinforcement learning are difficult to ensure strict grid safety.

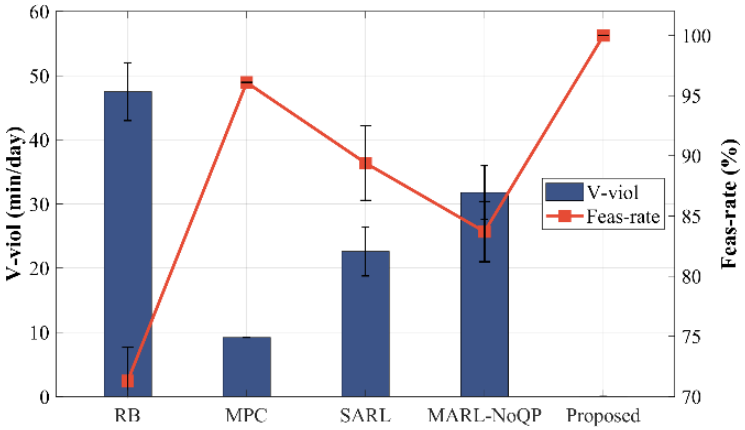

Fig. 7. Feasibility and voltage constraint satisfaction across control strategies.

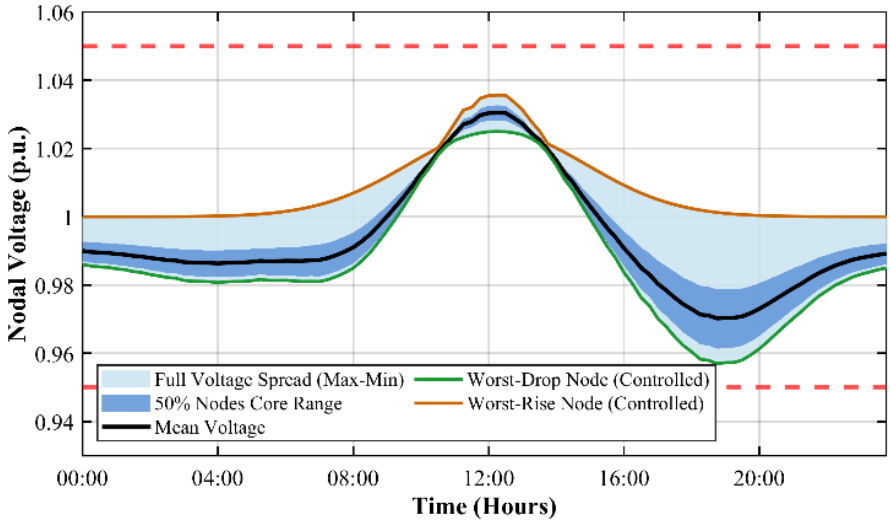

Fig. 8. Voltage profiles under the proposed framework.

Figure 8 verifies the overall system safety at the macro level by showing the spatiotemporal voltage profiles of all 123 buses for 24 hours. Although the system makes full use of the voltage margin to maximize the efficiency of PV grid connection at noon and cope with the EV charging peak in the evening, all node voltages do not exceed the legal limit. The worst-case voltage exactly remain stationary on the [0.95, 1.05] scale per unit boundary, which intuitively confirms that the downstream QP projection layer acts as a strict mathematical filter that effectively suppresses dangerous behavior, thus ensuring 100% operational feasibility.

4.9. Scalability Discussion

While this empirical study utilizes the IEEE 123-bus system, the proposed framework is engineered with intrinsic scalability for substantially larger distribution networks, fundamentally driven by its decentralized execution paradigm. Because decision-making is distributed among individual agents, the online inference complexity for each flexible resource remains a strict *O(1)* constant, functioning entirely independent of the macroscopic network expansion.

Beyond this localized neural network efficiency, the mathematical architecture of the safety layer further preserves system-wide scalability. By capitalizing on the highly sparse structural matrices inherent to radial distribution networks, the QP formulation avoids dimensional explosion. When processing these sparse algebraic constraints, modern interior-point solvers exhibit near-linear computational complexity, ensuring that the projection latency scales gracefully with grid size.

Validating these theoretical architectural advantages, the successful and simultaneous coordination of over 500 heterogeneous agents in this study serves as a definitive empirical proof-of-concept. This large-scale execution directly overcomes the high-dimensional action space bottlenecks that traditionally paralyze conventional multi-agent systems, confirming the framework's readiness for massive real-world deployment.

5. CONCLUSION

This paper presents a novel scheduling architecture that successfully bridges the fundamental gap between data-driven adaptability and deterministic operational safety in active distribution networks. By seamlessly integrating a CTDE-MARL framework with a rigorous QP projection layer, this methodology enables the secure and cooperative coordination of heterogeneous resources, including EVs, BESS, and HVAC systems While MARL agents autonomously learn highly cooperative, multi-objective dispatch policies, the downstream QP layer acts as an uncompromising mathematical filter to guarantee strict physical feasibility at every execution step.

Extensive case studies on the IEEE 123-bus system demonstrate that the proposed method achieves superior peak shaving and operational cost reductions, entirely eliminates network constraint violations, and maintains near-perfect end-user satisfaction.

Subsequent research will expand this framework toward hierarchical multi-level coordination between localized feeders and regional operators. Integrating probabilistic forecasting and transitioning into hardware-in-the-loop or real-world pilot deployments remain critical next steps to validate the cyber-physical resilience and practical deployability.

ACKNOWLEDGEMENT

This work is supported by Science and Technology Project of State Grid Tianjin Electric Power Company (Dianke - S&T 2025-02): Key Technologies and Applications of Transmission-Distribution Coordination for Risk Mitigation in Power Grids with High-Penetration Distributed Renewable Energy.

REFERENCES

[1] Shi J., Ma L., Li C. et al., 2022. A comprehensive review of standards for distributed energy resource grid-integration and microgrid. *Renewable and Sustainable Energy Reviews* 170: 112957.

[2] Wang X. and Dong B., 2024. Long-term experimental evaluation and comparison of advanced controls for HVAC systems. Applied Energy 371: 123706.

[3] Nguyen A.T., Pham D.H., Oo B.L., Santamouris M., Ahn Y., and Lim B.T.H., 2024. Modelling building HVAC control strategies using a deep reinforcement learning approach. *Energy and Buildings* 310: 114065.

[4] Hojjatinia Z., Mezher A.M., Castillo-Guerra E., Cardenas-Barrera J.L., and Saleh S.A., 2024. Peak shaving impact on load forecasting: A strategy for mitigation. *IEEE Access* 12: 160846-160863.

[5] Barman P., Dutta L., Bordoloi S., Kalita A., Buragohain P., Bharali S. et al., 2023. Renewable energy integration with electric vehicle technology: A review of the existing smart charging approaches. *Renewable and Sustainable Energy Reviews* 183: 113518.

[6] Lu Y., Fu M., Gao J., Shi Q., and Liu W., 2023. Voltage limit violation risk evaluation method considering communication failures in distributed voltage regulation system. *International Journal of Electrical Power & Energy Systems* 147: 108886.

[7] Lai C.M. and Teh J., 2022. Network topology optimisation based on dynamic thermal rating and battery storage systems for improved wind penetration and reliability. *Applied Energy* 305: 117837.

[8] Xiong J., Sun Y., Wang J., Li Z., Xu Z., and Zhai S., 2022. Multi-stage equipment optimal configuration of park-level integrated energy system considering flexible loads. *International Journal of Electrical Power & Energy Systems* 140: 108050.

[9] Sun L., Xia J., She D., Ding W., Jiang J., Liu B. et al., 2024. A predictive fuzzy logic and rule-based control approach for practical real-time operation of urban stormwater storage system. *Water Research* 266: 122437.

[10] Wang H., Shi M., Xie P., Lai C.S., Li K., and Jia Y., 2023. Electric vehicle charging scheduling strategy for supporting load flattening under uncertain electric vehicle departures. *Journal of Modern Power Systems and Clean Energy* 11 (5): 1634-1645.

[11] Al-Othman A., Tawalbeh M., Martis R., Dhou S., Orhan M., Qasim M. et al., 2022. Artificial intelligence and numerical models in hybrid renewable energy systems with fuel cells: Advances and prospects. *Energy Conversion and Management* 253: 115154.

[12] Zhao C., Sun D., Zhang L., Nian H., and Hu B., 2023. Two-stage optimal re/active power control of large-scale wind farm to mitigate voltage dip induced frequency excursion. *IEEE Transactions on Sustainable Energy* 14 (1): 730-733.

[13] Al-Issa H.A., Qawaqzeh M., Kurashkin S., Halko S., Kvitka S., Vovk O. et al., 2022. Monitoring of power transformers using thermal model and permission time of overload. *International Journal of Electrical and Computer Engineering* 12 (3): 2323-2334.

[14] Alvarez-Alvarado M.S., Donaldson D.L., Recalde A.A., Noriega H.H., Khan Z.A., Velasquez W. et al., 2022. Power system reliability and maintenance evolution: A critical review and future perspectives. *IEEE Access* 10: 51922-51950.

[15] Li Q., Liu J., Cui B., Song W., and Ye J., 2024. Distribution system flexibility characterization: A network-informed data-driven approach. *IEEE Transactions on Smart Grid* 15 (1): 1188-1191.

[16] Pourghaderi N., Fotuhi-Firuzabad M., Moeini-Aghtaie M., Kabirifar M., and Lehtonen M., 2023. Exploiting DERs' flexibility provision in distribution and transmission systems interface. *IEEE Transactions on Power Systems* 38 (2): 1963-1977.

[17] Li Y., Wang J., Zhang Y., and Han Y., 2022. Day-ahead scheduling strategy for integrated heating and power system with high wind power

penetration and integrated demand response: A hybrid stochastic/interval approach. *Energy* 253: 124189.

[18] Sasaki K., Aki H. and Ikegami T., 2022. Application of model predictive control to grid flexibility provision by distributed energy resources in residential dwellings under uncertainty. *Energy* 239: 122183.

[19] Escobar F., Viquez J.M., Garcia J., Aristidou P., and Valverde G., 2022. Coordination of DERs and flexible loads to support transmission voltages in emergency conditions. *IEEE Transactions on Sustainable Energy* 13 (3): 1344-1355.

[20] Zhao Z., Guo J., Luo X., Lai C.S., Yang P., Lai L.L. et al., 2022. Distributed robust model predictive control-based energy management strategy for islanded multi-microgrids considering uncertainty. *IEEE Transactions on Smart Grid* 13 (3): 2107-2120.

[21] Quintana E. and Inga E., 2022. Optimal reconfiguration of electrical distribution system using heuristic methods with geopositioning constraints. *Energies* 15 (15): 5317.

[22] Lu Y., Xiang Y., Huang Y., Yu B., Weng L., and Liu J., 2023. Deep reinforcement learning based optimal scheduling of active distribution system considering distributed generation, energy storage and flexible load. *Energy* 271: 127087.

[23] Ye Y., Papadaskalopoulos D., Yuan Q., Tang Y., and Strbac G., 2023. Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets. *IEEE Transactions on Smart Grid* 14 (2): 1541-1554.

[24] Amato C., 2024. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv*:2409.03052.

[25] Lu Y., Fu M., Gao J., Shi Q., and Liu W., 2023. Voltage limit violation risk evaluation method considering communication failures in distributed voltage regulation system. *International Journal of Electrical Power & Energy Systems* 147: 108886.

[26] Zhang H., Xiao F., Zhang C., and Li R., 2023. A multi-agent system based coordinated multi-objective optimal load scheduling strategy using marginal emission factors for building cluster demand response. *Energy and Buildings* 281: 112765.

[27] Wu S., Liu T., Egerstedt M., and Jiang Z.-P., 2023. Quadratic programming for continuous control of safety-critical multiagent systems under uncertainty. *IEEE Transactions on Automatic Control* 68 (11): 6664-6679.

[28] Evangelopoulos V.A., Kontopoulos T.P. and Georgilakis P.S., 2022. Heterogeneous aggregators competing in a local flexibility market for active distribution system management: A bi-level programming approach. *International Journal of Electrical Power & Energy Systems* 136: 107639.

APPENDIX

Let $N = |\mathcal{I}|$ be the total number of flexible agents, N_b be the number of buses, and N_l be the number of branches. The global action vector $u_t \in \mathbb{R}^N$ concatenates the active power commands of all EVs, BESS, and HVACs.

SoC Mapping Matrix ($\boldsymbol{B}$): The matrix B maps the continuous active power actions to the state-of-charge variations over the interval Δt. Based on Eqs. (3) and (5), B is a highly sparse, block-diagonal matrix of dimension $\mathbb{R}^{(N_{EV}+N_{BESS})\times N}$. Its non-zero diagonal elements encapsulate the corresponding parameters $\frac{\eta_{ch}\Delta t}{E^{max}}$or charging) and $\frac{\Delta t}{\eta_{dis}E^{max}}$or discharging).

Thermal Mapping Matrices ($\boldsymbol{A_T, B_T}$): Derived from the first-order RC dynamics in Eq. (7), A_T is a diagonal matrix reflecting the thermal inertia $1-\frac{\Delta t}{c_h R_h}$, while B_T maps the HVAC electrical power actions to indoor temperature changes based on $\frac{\Delta t \eta_h^{\mathrm{HVAC}}}{c_h}$.

Voltage Sensitivity Matrix ($\boldsymbol{M}$): $M \in \mathbb{R}^{N_b \times N_b}$ linearly relates nodal active power injections to nodal voltage magnitude variations. Under the simplified LinDistFlow formulation (Eq. 9), M is analytically derived from the network resistance matrix R, constructed via the upstream/downstream incidence relationships of the radial topology.

Branch Flow Sensitivity Matrix ($\boldsymbol{H}$): $H \in \mathbb{R}^{N_l \times N_b}$ functions as the Power Transfer Distribution Factor (PTDF) matrix for the radial distribution network, linearly mapping the nodal net power injections $P_t(u_t)$ to the active power flows across each line to enforce the thermal limit $\overline{\boldsymbol{F}}$.