electronics

MDPI

Article

A Multi-Objective Deep Reinforcement Learning Approach for EV Charging Optimization Considering Transformer Lifespan

Ji Wang [1,2] and Changyu Du [3,*]

[1] CSG Electric Power Research Institute, Guangzhou 510663, China; wangji@csg.cn
[2] Guangdong Provincial Key Laboratory of Intelligent Measurement and Advanced Metering of Power Grid, Guangzhou 510700, China
[3] The School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: duchangyu@std.uestc.edu.cn

Abstract

With the large-scale integration of electric vehicles (EVs) into power grids, massive concentrated EV charging accelerates the loss of life (LOL) of distribution transformers. There is a need to reduce transformer LOL while optimizing user experience, thus formulating the EV charge optimization concerning transformer LOL (ECOTL) problem. However, due to its characteristics, such as multi-objective tradeoffs, nonlinearity, uncertainty, and the requirement for time-sensitive online control, the ECOTL problem poses severe challenges in obtaining a high-quality Pareto front, acquiring robust solutions, and improving solving speed. To address these, we propose a multi-objective multi-agent deep reinforcement learning (MOMADRL) method that deploys multiple Critics and Actors for each agent. During training, Critics learn value functions with distinct preferences and guide the corresponding Actors that specialize in learning optimal strategies to update their policies. After sufficient training, each Actor can determine a preference-specific charging strategy in the test phase. Case studies based on real electricity price and load data demonstrate that compared with traditional multi-agent deep reinforcement learning (MADRL) methods, the proposed method achieves better convergence performance on the training set. Moreover, it obtains better Pareto-front approximation than the learning-based baselines while providing fast online inference under the tested simulation conditions.

Keywords: electric vehicle; multi-objective optimization; Pareto front analysis; transformer loss; deep reinforcement learning

Academic Editor: Jen-Hao Teng

Received: 3 June 2026
Revised: 30 June 2026
Accepted: 2 July 2026
Published: 14 July 2026

1. Introduction

In recent years, the continuous intensification of global warming and the growing depletion of fossil energy reserves have made environmental pollution control and energy structure transformation a global consensus, creating an urgent need for coordinated reforms in the transportation and energy sectors to mitigate this dual crisis. Electric vehicles (EVs), leveraging their core advantages of zero tailpipe emissions and low reliance on fossil fuels, have emerged as a pivotal technological pathway to drive decarbonization in the transportation sector. Their market penetration is accelerating at a remarkable pace. According to data from the International Energy Agency (IEA), global EV sales reached 14 million units in 2023, accounting for 20% of the total annual passenger vehicle sales. This milestone, together with the growing demand for intelligent EV charging management

highlighted in recent review studies, signifies that EVs have entered a market-driven phase of large-scale development [1,2].

However, the large-scale popularization of EVs also poses severe challenges to the safe and stable operation of distribution networks (DNs). On one hand, concentrated charging demand (such as peak charging periods at post-work hours) will significantly increase the grid load, potentially leading to regional transformer overloads, voltage sags at the end of low-voltage feeders, and even accelerated aging of power equipment [3,4]. On the other hand, uncertainties in EV users' travel behaviors (e.g., travel routes, parking durations) and the spatiotemporal distribution of charging activities (e.g., charging behavior, charging time slots) result in strong randomness and volatility in charging loads. This further complicates the complexity of grid optimization and control tasks [5].

The core goal of EV charging optimization is to improve owners' charging satisfaction, which is mainly reflected in two aspects: reducing charging costs and alleviating charging anxiety. The former requires making charging and discharging decisions based on real-time electricity prices, allowing owners to charge during low-price periods to reduce economic burdens [6,7]. The latter stems from owners' concerns about insufficient battery range, which requires charging strategies to ensure vehicles have enough power to meet travel needs before departure [8,9]. In the distribution network, transformers, a key equipment for power transmission and voltage conversion, become overloaded when many EVs connect to residential chargers at fixed times [10]. This overload sharply raises winding temperature, accelerating insulation aging and increasing transformer loss of life (LOL). Measures to reduce transformer LOL are necessary. Therefore, the EV charge optimization concerning transformer LOL (ECOTL) problem is established, which faces a typical multi-objective conflict: these objectives are mutually restrictive, as improving one often comes at the cost of another, so there is no single optimal solution to maximize all objectives. It is necessary to solve for the problem's Pareto front—a set of non-dominated solutions (NDSs) where no solution performs better in one objective without worsening at least one other. This frontier provides decision-makers with alternative solutions to choose from based on specific scenario priorities [11]. However, solving for the Pareto front is extremely difficult, due to the conflicts between multiple objectives and the drama tic increase in computational complexity.

Many existing studies adopt day-ahead scheduling methods [12,13], but with the growing demand for electric vehicles (EVs) and efficient control, time-sensitive online control has become increasingly important. Researchers have proposed various EV control approaches. Physics-based methods [14,15] rely on accurate mathematical modeling, yet the strictly nonlinear calculation of transformer LOL in ECOTL modeling leads to heavy computational burden or even insolvability with existing solvers, making online control difficult. Heuristic methods such as genetic algorithm (GA) [16,17], particle swarm optimization (PSO) [18] and hybrid strategies [19] optimize via population iteration, evaluation-based selection and gradual approximation to the global optimum, but they face slow convergence, poor stability and difficulty in achieving global optimality when tackling the nonlinear ECOTL problem. For such problems characterized by strong randomness and rapid response requirements, machine learning methods are widely used for their data-driven nature, no need for precise modeling, fast decision-making and high adaptability, such as ensemble machine learning [20] and deep reinforcement learning [21–30]. Among these methods, multi-agent deep reinforcement learning (MADRL) has become a recent research hotspot because it enables decentralized control through agents that learn adaptive strategies from historical data [22,25,27,28]. Representative extensions include Stackelberg MADRL [29] and evolutionary-curriculum-learning-based MADRL [30]. The latter is highly relevant because it incorporates transformer lifetime into EV charging optimization. However, it

is designed to improve the training stability and scalability of large-scale coordinated EV charging under a scalarized reward, rather than to generate a set of non-dominated solutions. Therefore, it cannot directly approximate the Pareto front required for multi-objective EV charging decision-making.

To solve the multi-objective EV scheduling problem, researchers have proposed heuristic multi-objective evolutionary algorithms (MOEAs) [31–34], such as non-dominated sorting genetic algorithm (NSGA) series [31,32] and multi-objective evolutionary algorithm based on decomposition (MOEA/D) [33]. However, MOEAs face key issues in solving ECOTL: poor accuracy–efficiency balance under EV randomness, slow convergence due to transformer LOL model nonlinearity, and uneven solutions from ineffective diversity maintenance in high-dimensional spaces. Therefore, based on previous analyses, there is an urgent need to propose a MOMADRL method for multi-objective problems to solve ECOTL. Existing methods [35,36] have gaps: after training, the agent can usually only determine a compromised NDS under one preference. While a State-of-the-Art preference-driven method [37] is utilized to ensure precise alignment between preferences and value estimations during training, it relies on a single policy network to learn optimal strategies across the entire preference space. The limited representational capacity of a single network can lead to degraded decision-making performance. To fill this gap, inspired by [38], we propose a MOMADRL method based on a multi-Actor–Critic framework. By deploying multiple Actor networks for each agent, this method individually learns Pareto-optimal strategies under different preferences, ultimately supporting Pareto-front approximation. The innovations of this paper are as follows:

1. The ECOTL problem is modeled as a multi-objective decentralized partially observable Markov decision process (MO-Dec-POMDP), where charging cost, charging anxiety, and transformer LOL are treated as conflicting objectives. This formulation provides the problem basis for applying MOMADRL to learn multiple preference-specific charging policies rather than a single scalarized policy.
2. A MOMADRL method based on a multi-Actor-Critic framework is proposed. By deploying multiple Actor and Critic networks for each agent, each Actor-Critic pair is associated with a specific preference and learns the corresponding charging strategy. This design avoids the representational burden of using one single Actor to fit the entire preference space and enables the learned policies to approximate the Pareto front.
3. Case studies on 8-EV and 32-EV residential systems are conducted to evaluate the proposed MOMADRL method. The results show that, under the tested residential EV charging cases, the proposed method achieves better Pareto-front approximation than the learning-based baselines while maintaining fast online inference.

The work is organized as follows. First, the ECOTL model is established. Second, the multi-objective decentralized partially observable Markov decision process and the MOMADRL framework are introduced. Third, comparative experiments and sensitivity studies are conducted. Finally, conclusions and future research directions are presented.

2. Problem Statement

Consider a community distribution network consisting of a distribution transformer and multiple residential loads, which are categorized into EV charging loads and other loads.

We assume that each household with one EV is equipped with a smart charging controller. The transformer-related information and electricity price are pre-acquired from the utility's asset management system and broadcast through communication infrastructures, including advanced metering infrastructure (AMI) and standard cellular networks (4G/5G). The objective is to improve residents' charging satisfaction while reducing transformer

lifespan loss. Our modeling will be divided into two parts: (1) EV Charging Model; (2) Transformer Model.

2.1. EV Charging Model

Suppose there are I households in the considered community. Meanwhile, we discretize the time scale into multiple time steps $t = 0, 1, \ldots, T$. The residential load L_t^{resid} in this distribution network system is calculated as follows:

$$L_t^{\text{resid}} = L_t^{\text{EV}} + L_t^{\text{other}}, \tag{1}$$

where L_t^{other} represents demand of other loads at time t and L_t^{EV} is the total EV charging load in the network, which is defined as:

$$L_t^{\text{EV}} = \sum_{i=1}^{I} a_t^i \Delta t, a_t^i \in [a_{\min}, a_{\max}], \tag{2}$$

where a_t^i denotes the charging capacity of EV i within the time interval Δt at time t. A positive value indicates that the EV is charging, while a negative value indicates that the EV is discharging. $a_{\min}$ and $a_{\max}$ are minimum and maximum charging capacity, respectively.

During charging, EV owners seek to minimize charging costs. The charging cost for EV i at time t is given by:

$$CC_t^i = p_t \cdot a_t^i \cdot \Delta t, \tag{3}$$

where p_t is the electricity price.

Meanwhile, EV owners are concerned that the post-charging state of charge (SOC) may fail to meet mileage requirements, or that charging may be prematurely terminated due to unforeseen factors necessitating early departure. They thus seek to reach the target SOC as quickly as possible. Such psychological states of EV users are referred to as "charging anxiety." It is quantified as:

$$CA_t^i = \frac{\Omega^i \left(E_{tar}^i - E_t^i\right)}{t_{dep}^i - t} t \in [t_{arr}^i, t_{dep}^i), \tag{4}$$

where t_{arr}^i and t_{dep}^i are arrival time and departure time of EV i, respectively; E_t^i is the SOC of EV i at t; E_{tar}^i is the target SOC; and Ω^i is the maximum battery capacity. Equation (4) quantifies charging anxiety by defining user dissatisfaction as a function of time urgency: as the current time approaches the user-defined departure time, sensitivity to residual energy deficit rises sharply. Meanwhile, it captures the impact of driving range and energy requirements by penalizing the gap between the current and target SOC [30].

In addition to the dual objectives of minimizing charging costs and alleviating charging anxiety, the dynamic SOC changes of an EV during charging and discharging both directly influence these goals and are subject to physical constraints. The SOC evolution during charging and discharging is as follows:

$$E_{t+1}^i = \begin{cases} E_t^i + \frac{\eta_c a_t^i}{\Omega^i} \Delta t, if\ a_t^i \geq 0 \\ E_t^i + \frac{a_t^i}{\eta_d \Omega^i} \Delta t, if\ a_t^i < 0 \end{cases} \quad t \in \left[t_{arr}^i, t_{dep}^i\right), \tag{5}$$

$$E_t^i = E_{arr}^i\ if\ t = t_{arr}^i, \tag{6}$$

where η_c and η_d are charge efficiencies, and E_{arr}^i denotes the SOC of EV i upon arrival at the charging area.

2.2. Transformer Model

When an EV is connected to the distribution network, it can exchange energy with the main grid through the transformer. To accurately quantify transformer LOL, IEEE standard is adopted for modeling [39] based on the following motivations: First, it provides a standardized thermal model that explicitly maps electrical load variations to winding hotspot temperatures, establishing a calculable link between charging decisions and physical aging. Second, this deterministic model is computationally feasible for embedding into the reinforcement learning environment to generate real-time transformer LOL based on load data. The transformer LOL model is defined as follows:

$$\mathrm{LOL} = \frac{F_{\mathrm{EA}} \times \rho}{\mathrm{NI}}, \tag{7}$$

$$F_{\mathrm{EA}} = \frac{\sum_{n=1}^{N} F_{\mathrm{AA},n} \Delta t}{\sum_{n=1}^{N} \Delta t}, \tag{8}$$

$$F_{\mathrm{AA}} = \exp\left(\frac{15000}{383} - \frac{15000}{\theta_{\mathrm{H}} + 273}\right), \tag{9}$$

$$\theta_{\mathrm{H}} = \theta_A + \Delta\theta_{\mathrm{TO}} + \Delta\theta_{\mathrm{H}}, \tag{10}$$

$$\begin{aligned} \Delta\theta_{\mathrm{TO}} = \Delta\theta_{\mathrm{TO}}^{\mathrm{R}} \times \left(\left[\frac{K_{\mathrm{ul}}^2 R+1}{R+1}\right]^k - \left[\frac{K_{\mathrm{in}}^2 R+1}{R+1}\right]^k\right) \\ \times \left(1 - \mathrm{e}^{-\rho/\tau_{\mathrm{TO}}}\right) + \Delta\theta_{\mathrm{TO}}^{\mathrm{R}} \times \left[\frac{K_{\mathrm{in}}^2 R+1}{R+1},\right]^k \end{aligned} \tag{11}$$

$$\Delta\theta_{\mathrm{H}} = \Delta\theta_{\mathrm{H}}^{\mathrm{R}} \times \left(K_{\mathrm{ul}}^{2g} - K_{\mathrm{in}}^{2g}\right) \times \left(1 - \mathrm{e}^{-\rho/\tau_{\mathrm{W}}}\right) + \Delta\theta_{\mathrm{H}}^{\mathrm{R}} \times K_{\mathrm{in}}^{2g}, \tag{12}$$

where NI is the normal insulation life of the transformer; ρ is the duration of the load; F_{AA} is the aging acceleration factor; F_{EA} is the equivalent aging factor over the entire time period; N is the total number of time intervals; n is the time interval index; θ_{H} is the hot-spot temperature of the transformer winding calculated by (10); θ_A denotes the ambient temperature; $\Delta\theta_{\mathrm{TO}}$ denotes the temperature rise in the top oil above the ambient temperature calculated by (11); $\Delta\theta_{\mathrm{H}}$ denotes the winding hot-spot temperature rise above the top oil temperature calculated by (12); $\Delta\theta_{\mathrm{TO}}^{\mathrm{R}}$ denotes the top oil temperature rise above the ambient temperature under rated load; R is the ratio of rated load loss to no-load loss; k and g is the exponent for calculating, determined empirically; K_{ul} and K_{in} are the ratios of the final load and initial load to the rated load, respectively; and τ_{W} and τ_{TO} denote the winding time constant and top-oil time constant, respectively.

2.3. Multi-Objective EV Charge Optimization Model Concerning Transformer LOL

To measure the quality of solutions to multi-objective problems, we introduce the concepts of NDS, Pareto-optimal solution, and Pareto front:

Theorem 1. *Non-dominated Solution (NDS): Suppose there are m objective functions f. A policy π_1 is non-dominated if there exists no other policy π_2 such that $f_i(\pi_2) \leq f_i(\pi_1)$ for each $i \in [1, \ldots, m]$ and $f_j(\pi_2) < f_j(\pi_1)$ for at least one $j \in [1, \ldots, m]$. Policy π_1 is also called Pareto-optimal solution.*

Theorem 2. *Pareto Frontier: The Pareto set is the set of all Pareto-optimal solutions $\{\pi \mid \pi$ is Pareto optimal$\}$, and its image in the objective space is the Pareto front.*

Then, integrating the EV Charging Model and Transformer Model, the optimization problem under study is formulated with objectives of minimizing charging cost, charging anxiety, and transformer LOL.

The objective functions (overall charging cost f_t^{CC}, charging anxiety f_t^{CA}, and transformer LOL f_t^{LOL}) are calculated by:

$$f_t^{CC} = \sum_{i=1}^{I} CC_t^i, \tag{13}$$

$$f_t^{CA} = \sum_{i=1}^{I} CA_t^i, \tag{14}$$

$$f_t^{LOL} = \mathrm{LOL}_t^{\mathrm{resid}} - \mathrm{LOL}_t^{\mathrm{other}}, \tag{15}$$

where $\mathrm{LOL}_t^{\mathrm{resid}}$ and $\mathrm{LOL}_t^{\mathrm{other}}$ are transformer LOL determined by residential loads and the other load.

Therefore, the multi-objective EV charge optimization problem concerning transformer LOL is formulated as:

$$\min \boldsymbol{F} = [\sum_{t=1}^{T} f_t^{CC}, \sum_{t=1}^{T} f_t^{CA}, \sum_{t=1}^{T} f_t^{LOL}], \tag{16}$$
$$s.t.\ (1)\text{–}(15),$$

It is evident that this problem is a highly complex nonlinear multi-objective optimization problem, which is difficult for existing methods to solve:

Traditional physics-based methods typically rely on first establishing an accurate model of the problem and then using existing solvers to find solutions, yet they are usually unable to solve nonlinear problems effectively.

Heuristic methods, while having the advantage of not requiring precise mathematical modeling and being able to explore potential solution spaces, have limitations such as being prone to falling into local optima.

Current State-of-the-Art DRL-based methods usually perform well in single-objective scenarios, but when applied to multi-objective problems, agents must learn optimal decisions under multiple conflicting preferences to obtain NDSs and search for Pareto-optimal sets, which leads to high learning costs and ultimately results in poor training and convergence performance.

3. Proposed Approach

To address the challenges of nonlinearity, multi-objectivity, and time-sensitive online control in the ECOTL problem, we first model it as a decentralized partially observable Markov decision process (Dec-POMDP). Concurrently, we propose a multi-objective deep reinforcement learning-based framework.

3.1. Multi-Objective Decentralized Partially Observable Markov Decision Process

The ECOTL problem is modeled as a Dec-POMDP, where each EV acts as an agent. A Dec-POMDP framework fits well here: it captures the nature of multi-EV decentralized decision-making and the system's partial observability in the ECOTL problem, aligning with practical residential charging scenarios. Specifically, each agent can only access partial observations of the system state rather than the complete state, and agents make charging decisions independently during execution to collectively optimize the global objectives. To address the multi-objective characteristic of ECOTL, we propose a multi-objective Dec-POMDP (MO-Dec-POMDP), which consists of four components.

- **State**: the partial observation of agent i at t is:

$$s_t^i = \left(t, \Upsilon_t^i, \widetilde{p}_t, E_t^i, \widetilde{\theta}_{\mathrm{H},t}^{\mathrm{other}}, L_{t-1}^{\mathrm{EV}}\right) \tag{17}$$

  where $\widetilde{\theta}_{\mathrm{H},t}^{\mathrm{other}}$ is the winding hot-spot temperature at time t determined by the other load forecast value; Υ_t^i indicates whether EV i is charging (1 if it is charging, 0 otherwise); and $\widetilde{p}_t$ is the forecast electricity price. The benefit of this modeling is that agents can obtain more comprehensive objective-relevant information through partial observations, such as details $(\Upsilon_t^i, \widetilde{p}_t)$ related to charging costs, information (Υ_t^i, E_t^i, t) associated with charging anxiety, and data $(\widetilde{\theta}_{\mathrm{H},t}^{\mathrm{other}}, L_{t-1}^{\mathrm{EV}})$ concerning transformer LOL; the global state is $X_t = [s_t^1, s_t^2, \ldots, s_t^I]$.
- **Action**: action of agent i at time t is charging capacity $a_t^i \in [a_{\min}, a_{\max}]$, and the joint action of all agents is $A_t = [a_t^1, a_t^2, \ldots, a_t^I]$.
- **State transition**: when an agent takes an action a_t^i according to policy π^i, the state transforms from s_t^i to s_{t+1}^i at time $t+1$ with the transition function $s_{t+1}^i = P(s_t^i, a_t^i, \Upsilon_{t+1}^i, \widetilde{p}_{t+1}, E_{t+1}^i, \widetilde{\theta}_{\mathrm{H},t+1}^{\mathrm{other}})$, where E_{t+1}^i and $\widetilde{\theta}_{\mathrm{H},t+1}^{\mathrm{other}}$ is calculated by (5) and (10)–(12), respectively.
- **Reward:** For multi-objective optimization problems, multiple conflicting objectives should be traded off to obtain Pareto-optimal solutions. Therefore, an independent reward function is designed for each objective, which is a key distinction between the proposed MO-Dec-POMDP and the traditional Dec-POMDP. Since the original ECOTL problem is formulated as a minimization problem, where charging cost, charging anxiety, and transformer LOL are all expected to be minimized, these objective values should be converted into rewards that can be maximized. To make the reward values of different objectives comparable and avoid the dominance of objectives with larger numerical scales, each objective is normalized as follows:

$$r = -\frac{f - f_{\min}}{f_{\max} - f_{\min}}, \tag{18}$$

  where f is one of objectives f_t^{CC}, f_t^{CA}, and f_t^{LOL}. $f_{\max}$ and $f_{\min}$ are objective-specific normalization bounds. Specifically, $f_{\min}$ is obtained when only this objective is optimized and the other objectives are not considered. In contrast, $f_{\max}$ is obtained from the optimization results in which this objective is excluded. If multiple extreme-preference solutions are obtained, the largest value is used as $f_{\max}$. These bounds are determined before DRL training and then fixed during both training and testing. When $f = f_{\min}$, the normalized reward reaches its maximum value of 0; when $f = f_{\max}$, the normalized reward reaches its minimum value of -1. Therefore, Equation (18) provides a monotonically decreasing transformation from the original minimization objective to the normalized reward.

3.2. Multi-Objective Multi-Agent Deep Reinforcement Learning

3.2.1. Multi-Objective Value Function

After MO-Dec-POMDP is established, a Pareto set $\Pi = \{\pi \mid \pi \textit{ is Pareto optimal}\}$ should be obtained to solve the ECOTL problem; we need to evaluate whether the policy is Pareto-optimal under different objective preferences. To this end, we propose a multi-objective value function to assess the quality of an agent's action in the current state under varying objective preferences.

First, the vector of rewards under varying objective preferences is defined as:

$$\hat{\boldsymbol{r}}_t = W^T \boldsymbol{r}_t = [r_{t,1}, r_{t,2}, \ldots, r_{t,M}] \tag{19}$$

where $W = [\boldsymbol{\omega}_{CC}; \boldsymbol{\omega}_{CA}; \boldsymbol{\omega}_{LOL}]$ is a weighting matrix; $\boldsymbol{\omega}_{\cdot} = [\omega_{\cdot,1}, \omega_{\cdot,2}, \ldots, \omega_{\cdot,M}]$ is the set of weighting factors for the corresponding objective, which satisfies $\boldsymbol{\omega}_{CC} + \boldsymbol{\omega}_{CA} + \boldsymbol{\omega}_{LOL} = \mathbf{1}$; and M is the number of weighting factors.

Second, for each timestep t, all agents decide charging actions after observing partial current information, and charging terminals compute reward value vectors $\hat{\boldsymbol{r}}_t$. In this problem, the agents aim to learn an optimal set of policies $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_M]$ to maximize the discounted cumulative reward value vector $\sum_{t'=0}^{\infty} \gamma^{t'} (\hat{\boldsymbol{r}}_{t+t'})$, where $\gamma \in [0, 1]$ is the discount factor. To achieve this, we introduce a multi-objective value function:

$$\boldsymbol{Q} = [Q_1, Q_2, \ldots, Q_M] = \mathbb{E}[\textstyle\sum_{t'=0}^{\infty} \gamma^{t'} \hat{\boldsymbol{r}}_{t+t'} \Big| X_t = X, A_t = A], \tag{20}$$

3.2.2. Multi-Objective Actor–Critic Framework

Existing DRL methods face significant limitations when applied to the complex ECOTL problem. First, basic frameworks struggle with the action space and efficiency: DQN [20,22] is restricted to discrete actions, leading to the "curse of dimensionality" in power control; PPO [24] exhibits low sample efficiency due to its on-policy nature; and TD3's [29] deterministic policy often leads to premature convergence and entrapment in local optima. Moreover, regarding multi-objective optimization, traditional Actor–Critic methods typically rely on a single Actor to learn strategies across different preferences. This design often fails to balance conflicting objectives effectively, resulting in high training costs.

To address these gaps, we propose the MOMADRL method based on the soft Actor–Critic framework [40]. The underlying soft Actor–Critic framework offers distinct advantages suited for the ECOTL problem. It operates directly within continuous action spaces and leverages maximum entropy to significantly enhance exploration. This prevents entrapment in local optima and ensures a diverse search for Pareto-optimal strategies, while its off-policy nature robustly improves sample efficiency. Structurally, we deploy M Actor and M Critic networks for each agent to individually learn the optimal policies and value functions under each single preference. This avoids the aliasing of multi-objective information and preference imbalance caused by a single network design, making the optimization of each objective more targeted. The establishment process of the proposed method is as follows:

Multi-objective value function is estimated by multiple Critic networks:

$$\boldsymbol{Q} = [\phi_{\delta_1}(X_t, A_{t,1}), \phi_{\delta_2}(X_t, A_{t,2}), \ldots, \phi_{\delta_M}(X_t, A_{t,M})], \tag{21}$$

where $\phi_{\delta}(\cdot)$ is the parameterized neural network.

To obtain the Pareto front, agents need to learn optimal Pareto solution sets Π under multiple preferences. For this, we configure M Actor networks $\boldsymbol{\pi}_{\mu}^{i} = [\pi_{\mu_1}^{i}, \pi_{\mu_2}^{i}, \ldots, \pi_{\mu_M}^{i}]$ for each agent i, where μ is the network parameter. Additionally, to output continuous charging action quantities, we adopt a Gaussian sampling method: specifically, each Actor network outputs the mean $\lambda(s)$ and variance $\sigma^2(s)$ of a Gaussian distribution $\mathcal{N}$, with actions sampled accordingly. However, this sampling process is non-differentiable, hindering policy network updates via backpropagation. We thus introduce the reparameterization trick: first sample ϵ from a unit Gaussian distribution $\mathcal{N}_0$, then generate actions through $a = \lambda(s) + \sigma(s) \cdot \epsilon$.

3.2.3. Multi-Actor–Critic Training Scheme

The parameter optimization method for Critic network m of agent i is to minimize the loss function and perform a gradient update:

$$y^i_{t,m} = r^i_{t,m} + \gamma(\begin{matrix}\min\limits_{j=1,2} Q^j_{\delta^i_{m,tar}}(X_{t+1,m}, A_{t+1,m}) \\ +\alpha\sum_{i=1}^{I} H(\pi^i_{\mu_m}(\cdot/s^i_{t+1,m}))\end{matrix}), a^i_{t+1,m} \sim \pi^i_{\mu_m}\left(\cdot/s^i_{t+1,m}\right), r_{t,m} \in \hat{r}_t, \tag{22}$$

$$\mathcal{L}^{\,i}_m(\delta^i_m) = \mathbb{E}_{X_t, A_{t,m}, r_{t,m}, X_{t+1,m} \sim B}\left[\left(Q_{\delta^i_m}(X_t, A_{t,m}) - y_{t,m}\right)^2\right], \tag{23}$$

$$\delta^i_m \leftarrow \delta^i_m + \eta\,\nabla_{\delta^i_m}\mathcal{L}^{\,i}_m\left(\delta^i_m\right) \tag{24}$$

where α is a coefficient of an entropy regularization term, which is used to balance the weights between entropy $H(\pi^i_{\mu_m}(\cdot/s^i_{t+1,m})) = -\log(\pi^i_{\mu_m}(\cdot/s^i_{t+1,m}))$ and Q-value; $Q^j_{\delta^i_{m,tar}}(\cdot)$ is the target value function that is estimated through the target Critic network $\phi_{\delta^i_{m,tar}}$, which is soft-updated by:

$$\delta_{m,tar} \leftarrow \upsilon \cdot \delta_m + (1-\upsilon) \cdot \delta_{m,tar}, \tag{25}$$

where υ is the soft update coefficient.

The parameter optimization process for Actor network m of agent i is:

$$\mathcal{L}^{\,i}_m(\mu^i_m) = \mathbb{E}_{X_t, A^{i\cdot}_{t,m} \sim B}\left[\begin{matrix} -\alpha H(\pi^i_{\mu_m}(\cdot|s^i_t)) \\ -\min\limits_{j=1,2} Q^j_{\delta^i_m}\left(X_t, a^i_{t,m}, A^{i\cdot}_{t,m}\right)\end{matrix}\right], \tag{26}$$

$$\mu^i_m \leftarrow \mu^i_m + \eta_\pi\,\nabla_{\mu^i_m}\mathcal{L}^{\,i}_m\left(\mu^i_m\right), \tag{27}$$

where $A^{i\cdot}_{t,m} = [a^1_{t,m}, a^2_{t,m}, \ldots, a^{i-1}_{t,m}, a^{i+1}_{t,m}, \ldots]$ represents the joint action of other agents except the current agent i.

The entire training process of all agents is detailed in Algorithm 1, and the flowchart of the entire training process is shown in Figure 1. For each episode e, the system first sequentially samples the weight factor vector $\boldsymbol{\omega} = [\omega_{CC,m}, \omega_{CA,m}, \omega_{LOL,m}]$ from set W; at each time step t, if in the exploration phase (the current episode number is less than the exploration episode number), agents randomly choose charging actions that meet physical constraints, while in the exploitation phase, each agent i observes the current state s^i_t and selects charging actions $a^i_{t,m}$ via its Actor network; after executing these actions in the community system, the charging station terminal calculates the reward vector r and the next time step's global state $X_{t+1,m}$ based on all agents' joint actions $A_{t,m}$ then store the state transition vector H in the experience replay buffer B; once the buffer is full, a batch D of state transition vectors is randomly sampled from it, the next time step's joint actions A_{t+1} are determined using all agents' policy networks, and all Actor networks and Critic networks are updated in line with Equations (22)–(27).

The proposed MOMADRL method adopts a centralized-training-with-decentralized-execution (CTDE) structure. During training, the Critic networks use global state information and joint actions from the replay buffer to stabilize multi-agent learning. During online execution in the simulation environment, each Actor makes decisions in parallel using only local EV information, including EV commuting behavior, charging behavior, and SOC, together with public system-level signals broadcast by the aggregator or AMI, such as electricity price and transformer-related information. Therefore, private information from other EVs is not required during decentralized execution. Through continuous interaction between agents and the residential community, all Actors can make decisions parallelly and Pareto-optimal solutions can eventually be obtained, with core reasons as follows: First, a

complete set of weight combinations for charging cost, anxiety cost, and LOL is generated before training, ensuring coverage of all preference dimensions in multi-objective optimization and laying the foundation for the diversity of Pareto-optimal solutions. Second, each weight corresponds to an independent Actor–Critic network; after sufficient training, the strategy under each weight can stably output the optimal solution for that preference, forming multiple groups of candidate solutions with different objective trade-off relationships. Finally, these candidate solutions undergo Pareto dominance screening, and the remaining ones are Pareto optimal solutions.

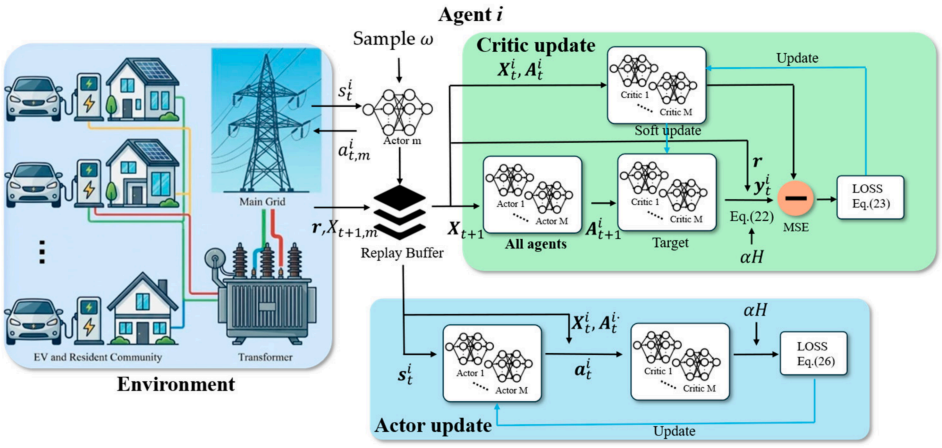

Figure 1. Flowchart of the proposed MOMADRL method.

Algorithm 1: Training process

1 : Initialize :
Replay buffer B, sampling batch D, training episode E, number of exploration episodes E_e, weighting matrix W, network parameters μ, δ, and hyperparameter γ, η, v.
2: **for** e = 1 to E **do**:
3 : Sequential sample the vector of weighting factors $\boldsymbol{\omega} = [\omega_{CC,m}, \omega_{CA,m}, \omega_{LOL,m}]$ from W;
4: **for** t = 1 to T **do**:
5 : **if** $e < E_e$:
6: All the agents randomly select charging behaviors that satisfy physical constraints;
7: **else**:
8 : observe the current state X_t, charging action selection is determined by m-th Actor network of all agents;
9: **end if**
10 : The joint action is carried out in the environment, and transition tuple $H = [X_t, A_{t,m}, \boldsymbol{r}, \boldsymbol{\omega}, X_{t+1,m}]$ stored in B;
11: **if** B is full:
12: Randomly sample D from the replay buffer;
13: Update all networks through the Formulas (22)–(27);
14: **end if**
15: **end for**
16: **end for**

4. Results

4.1. Experimental Setup

In this paper, a semi-synthetic case study is conducted on a distribution network system consisting of a 50 kVA transformer and eight households, each equipped with one EV. The EV-related parameters, including arrival time, departure time, initial SOC, battery capacity, and charging/discharging power limits, are set similarly to those used in existing EV charging studies [7,29,30]. The detailed parameters of the transformer and EV models are provided in Table 1. To better capture the fast thermal dynamics of the transformer, a 15 min control interval is adopted, resulting in 96 decision steps from 12:00 PM on the first day to 12:00 PM on the second day. The step size represents the uniform sampling interval in the objective-weight space and is set to 0.20, 0.15, and 0.10.

Table 1. Parameters of the ECOTL model.

Model	Parameter	Value
Transformer	θ_A	40 °C
	$\Delta\theta_{\mathrm{TO}}^{\mathrm{R}}$	53 °C
	$\Delta\theta_{\mathrm{H}}^{\mathrm{R}}$	27 °C
	NI	1.8×10^5 h
	R	4.1
	$\tau_{\mathrm{TO}}/\tau_{\mathrm{W}}$	6.86/0.08 h
	k, g	0.8
EV	E_{arr}	[0.2, 0.6]
	t_{arr}	Clip ($\mathcal{N}$ (6, 1^2), 5, 7)
	t_{dep}	Clip ($\mathcal{N}$ (20, 1^2), 18, 22)
	Ω	24 KWh
	$a_{\max}/a_{\min}$	6.3 KW/−6.3 KW

Both the Actor and Critic networks employ a four-layer fully connected architecture, with the ReLU activation function applied. The Actor network uses two separate output layers to predict the mean and standard deviation of the action distribution, respectively, and its final output is processed using the tanh activation function. The Adam optimizer is used. To validate the performance of the proposed method, real-world data are utilized for testing: The electricity price data are sourced from actual 2021 pricing in the Commonwealth Edison (ComEd; Chicago, IL, USA) service territory [41], whereas the base-load data are obtained from one year of smart-meter readings in a real distribution network [42]. Among the available dataset, the first 200 days are used as the training set, while the subsequent 40 days are used as the test set. All comparative experiments are repeated under five independent random seeds, and the final results are reported as mean ± standard deviation. The hyperparameters of the method proposed is shown in Table 2.

Table 2. Hyperparameters of the method proposed.

Hyperparameter	Value
Discount factor γ	0.99
Replay buffer size B	7200
Transition batch D	512
Learning rate η/η_π	$3 \times 10^{-3}/3 \times 10^{-3}/3 \times 10^{-4}$
Soft updated coefficient υ	5×10^{-3}
Number of neurons in each hidden layer	128

All simulation experiments are carried out on a computer configured with an Intel Core i9-10980XE CPU and an NVIDIA GeForce RTX 2060 GPU. The program is developed using Python 3.11, and the MADRL component was implemented using PyTorch 2.5.0 with CUDA 12.4 support.

Pareto-front quality is evaluated using three standard multi-objective metrics. First, the number of NDSs is computed for each test day and then averaged over all test days. Second, hypervolume (HV) is computed as:

$$HV = \frac{1}{D}\sum_{d=1}^{D} \lambda(\cup_{F\in P_d}[F, F_r]), \tag{28}$$

where D is the number of test days, P_d denotes the Pareto set obtained on day d, $\boldsymbol{F}_r = [6.0, 850.0, 10^{-6}]$ is the reference point, and $\lambda(\cdot)$ represents the Lebesgue measure of the dominated objective-space volume. HV is selected because it jointly measures convergence toward better objective values and the coverage of the Pareto front; a larger HV indicates a higher-quality Pareto set. Third, spacing (SPA) is calculated as the average, over all test days, of the standard deviation of the nearest-neighbor distances among the NDSs. SPA is selected to quantify the uniformity of the Pareto-front distribution; a smaller SPA indicates a more even distribution.

$$SPA = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{|P_d|-1}\sum_{i=1}^{|P_d|}(d_i - \overline{d})^2}, \tag{29}$$

To evaluate the performance of the proposed method, three benchmark methods are designed for comparative experiments:

1. Deterministic Information Scenario Optimization Based on NLopt [43]: This method assumes that information such as electricity prices, base loads, and EV users' commuting behaviors is known a priori. This assumption is unachievable in real-world scenarios. Since the ECOTL problem exhibits nonlinear characteristics after incorporating transformer LOL modeling, the nonlinear solver NLopt is employed to solve the deterministic optimization problem based on global information. This process yields theoretical Pareto-optimal solutions, which serve as a performance benchmark for subsequent method comparisons.
2. Stochastic Programming (SP): A benchmark for MOEA. First, apply random perturbations (with a deviation range of 20%) to the input load and electricity price data to generate 1000 perturbation scenarios. Then, use data compression technology to extract 20 representative scenarios from them, ensuring these scenarios retain the key statistical characteristics of the original samples. On this basis, utilize the Genetic Algorithm (GA) to solve the ECOTL problem for the 20 representative scenarios, and obtain the possible Pareto-optimal solution with the goal of minimizing the expected objective for each distinct weight.
3. Multi-agent Soft Actor-Critic (MASAC): This algorithm adopts a common framework in MADRL [24]. It is characterized by adding vectors with different preferences to state observations, while only deploying one pair of Actor–Critic to learn and evaluate the EV charging strategy. Its hyperparameter settings are consistent with the proposed method.
4. Preference-Driven Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (PD-MATD3) [37]: This algorithm's network updates are preference-driven. By integrating a directional angle term into the loss function of MATD3 [30], it forces the learned value function vector to precisely align with the user-specified preference vector in

geometric direction. Its hyperparameter and state observation settings are consistent with the MASAC method.

4.2. Training Process

Figure 2 presents the reward value curves of the MADRL-based methods under four preference settings with a step size of 0.2. The reward value represents the cumulative reward obtained in each training episode. For each preference weight, the agents of the three methods were trained independently, and the reward curves were recorded throughout the entire training process. During the exploration stage (Episodes 1–300), all methods mainly relied on random charging actions to explore the state-action space and accumulate interaction experiences. Consequently, the reward values fluctuated significantly and remained at relatively low levels. After approximately Episode 300, the agents gradually entered the exploitation stage, where the accumulated experiences were utilized to learn effective charging strategies, leading to a steady increase in reward values and eventual convergence. The results under all four preference settings show that the proposed method achieves better training performance than the learning-based baselines. As shown in Figure 2a–d, MASAC, PD-MATD3, and the proposed method converge at approximately Episodes 440, 410, and 370, respectively, indicating that the proposed method requires fewer training episodes to learn stable charging policies. Averaging the converged rewards over the four preference settings, MASAC achieves a final reward of approximately -4.85, PD-MATD3 reaches -4.55, while the proposed method attains the highest reward of -4.25. Consequently, the proposed method improves the final reward by approximately 12.4% and 6.6% compared with MASAC and PD-MATD3, respectively. The inferior performance of MASAC can be attributed to its single Actor–Critic architecture, which must simultaneously learn charging policies for multiple preference conditions, resulting in value estimation and unstable policy updates. Although PD-MATD3 introduces preference-aware learning and therefore achieves faster convergence and higher rewards than MASAC, a single policy network is still required to handle different preferences, limiting its ability to accurately capture diverse decision-making objectives. In contrast, the proposed method assigns multiple preference-oriented Critics to learn preference-specific value functions and provide more accurate guidance for policy optimization. This mechanism effectively reduces the learning complexity of individual networks, and enables the agents to identify higher-quality Pareto-optimal charging strategies. As a result, the proposed method not only converges 16% and 10% faster than MASAC and PD-MATD3, respectively, but also consistently achieves higher rewards and more stable convergence behavior across all preference settings.

4.3. Test Set Performance

4.3.1. Comparison of Experimental Results

To evaluate Pareto-front quality and online inference latency under the 15 min control interval, Table 3 reports NDS, HV, SP, and latency time, where latency time is the average time for each single decision-making charging action across all test days.

For SP, two limitations are evident. First, the stochastic evolutionary process is prone to premature convergence and local optima, preventing the discovery of a complete Pareto set. Consequently, its NDS values are consistently lower than those of the proposed method by 8.9%, 8.8%, and 9.0% under step sizes of 0.20, 0.15, and 0.10, respectively. Second, scenario generation and GA optimization introduce extremely high computational complexity. As the number of weighting factors increases from 6 to 36, the testing time rises dramatically from 185.42 s to 426.58 s, which is more than four orders of magnitude higher than that of the proposed method. Although SP achieves the second-best HV among the compared

baselines, its excessive computational burden severely limits its applicability in fast scheduling scenarios. MASAC adopts a single-Actor architecture to simultaneously learn policies corresponding to multiple preference vectors. This significantly increases the learning difficulty because a single policy network must approximate optimal actions for multiple conflicting objectives. Consequently, MASAC exhibits the worst Pareto-front quality among all methods. Its HV values are approximately 49.4%, 49.2%, and 48.9% lower than those of the proposed method under the three step sizes, respectively. Furthermore, its NDS values are 18.2%, 14.8%, and 8.0% lower than those achieved by the proposed method. Although MASAC obtains the smallest SPA values, indicating a relatively uniform distribution of discovered solutions, the corresponding HV and NDS values remain significantly inferior. This suggests that MASAC generates uniformly distributed solutions in a relatively narrow and suboptimal region of the objective space, resulting in poor convergence toward the true Pareto front. PD-MATD3 improves preference awareness by introducing azimuth-loss-guided learning. Compared with MASAC, it achieves substantial improvements in both HV and NDS, demonstrating that preference information is more effectively incorporated into value estimation. Nevertheless, PD-MATD3 still relies on a single Actor network, which fundamentally limits its ability to represent multiple optimal policies across different preference regions. Consequently, its HV values remain 10.2%, 11.0%, and 11.4% lower than those of the proposed method under the three step sizes, respectively. Similarly, its NDS values are lower than those of the proposed method by 22.8%. In contrast, the proposed method consistently achieves the best overall performance among all learning-based approaches. The key reason is that the multi-Actor architecture decomposes the original multi-preference optimization problem into several simpler sub-problems, allowing each Actor to specialize in a specific preference region. This substantially reduces policy-learning difficulty and improves the diversity of discovered Pareto-optimal solutions. As a result, the proposed method achieves the highest NDS values under all step sizes. Compared with SP, MASAC, and PD-MATD3, the average NDS improvement reaches 8.9%, 13.7%, and 22.8%, respectively. In terms of HV, the proposed method improves upon MASAC by approximately 95.0%, 96.8%, and 95.4%, and surpasses PD-MATD3 by 11.4%, 12.4%, and 12.8%, respectively. Meanwhile, the obtained SPA values remain on the order of 10^{-3}, indicating a well-distributed Pareto front. Furthermore, the Latency time remains nearly constant at approximately 0.013 s regardless of the number of weighting factors, demonstrating excellent scalability and fast response capability. Compared with NLopt, the proposed method produces almost identical Pareto-front diversity, with NDS differences of only 0.5%, 0.4%, and 0.4% under the three step sizes, respectively. Its online decision time is reduced from 5.84–8.73 s to only 0.012–0.014 s, corresponding to a speedup of approximately 450–650 times. Therefore, under the tested residential EV charging case study, the proposed method achieves better Pareto-front approximation than the learning-based baselines while maintaining fast online inference.

Table 3. Comparison results of various methods.

Step Size	No. of Weighting Factors	Methods	HV (1×10^{-5})	SPA	NDS	Latency Time (s)
0.20	6	NLopt	241.30 ± 7.24	0.0424 ± 0.0040	5.90 ± 0.08	5.84 ± 0.42
		SP	220.92 ± 10.21	0.0436 ± 0.0060	5.40 ± 0.10	185.42 ± 14.36
		MASAC	97.50 ± 9.75	0.0006 ± 0.0002	5.45 ± 0.30	0.012 ± 0.001
		PD-MATD3	170.68 ± 8.53	0.0001 ± 0.0001	4.83 ± 0.45	0.005 ± 0.001
		Proposed	190.11 ± 6.12	0.0007 ± 0.0002	5.93 ± 0.15	0.012 ± 0.001

Table 3. *Cont.*

Step Size	No. of Weighting Factors	Methods	HV (1×10^{-5})	SPA	NDS	Latency Time (s)
0.15	15	NLopt	262.42 ± 8.41	0.0392 ± 0.0035	14.75 ± 0.20	6.91 ± 0.53
		SP	239.70 ± 11.51	0.0395 ± 0.0050	13.50 ± 0.25	248.36 ± 19.74
		MASAC	106.28 ± 11.16	0.0007 ± 0.0002	13.63 ± 0.75	0.012 ± 0.001
		PD-MATD3	186.04 ± 9.86	0.0002 ± 0.0001	12.06 ± 1.13	0.005 ± 0.001
		Proposed	209.12 ± 7.14	0.0011 ± 0.0003	14.81 ± 0.38	0.013 ± 0.001
0.10	36	NLopt	294.39 ± 10.12	0.0366 ± 0.0030	35.40 ± 0.48	8.73 ± 0.68
		SP	267.31 ± 14.30	0.0378 ± 0.0040	32.40 ± 0.60	426.58 ± 31.27
		MASAC	119.18 ± 13.40	0.0008 ± 0.0002	32.70 ± 1.80	0.012 ± 0.001
		PD-MATD3	206.52 ± 12.20	0.0003 ± 0.0001	28.95 ± 2.70	0.005 ± 0.001
		Proposed	232.93 ± 8.90	0.0015 ± 0.0004	35.55 ± 0.90	0.014 ± 0.001

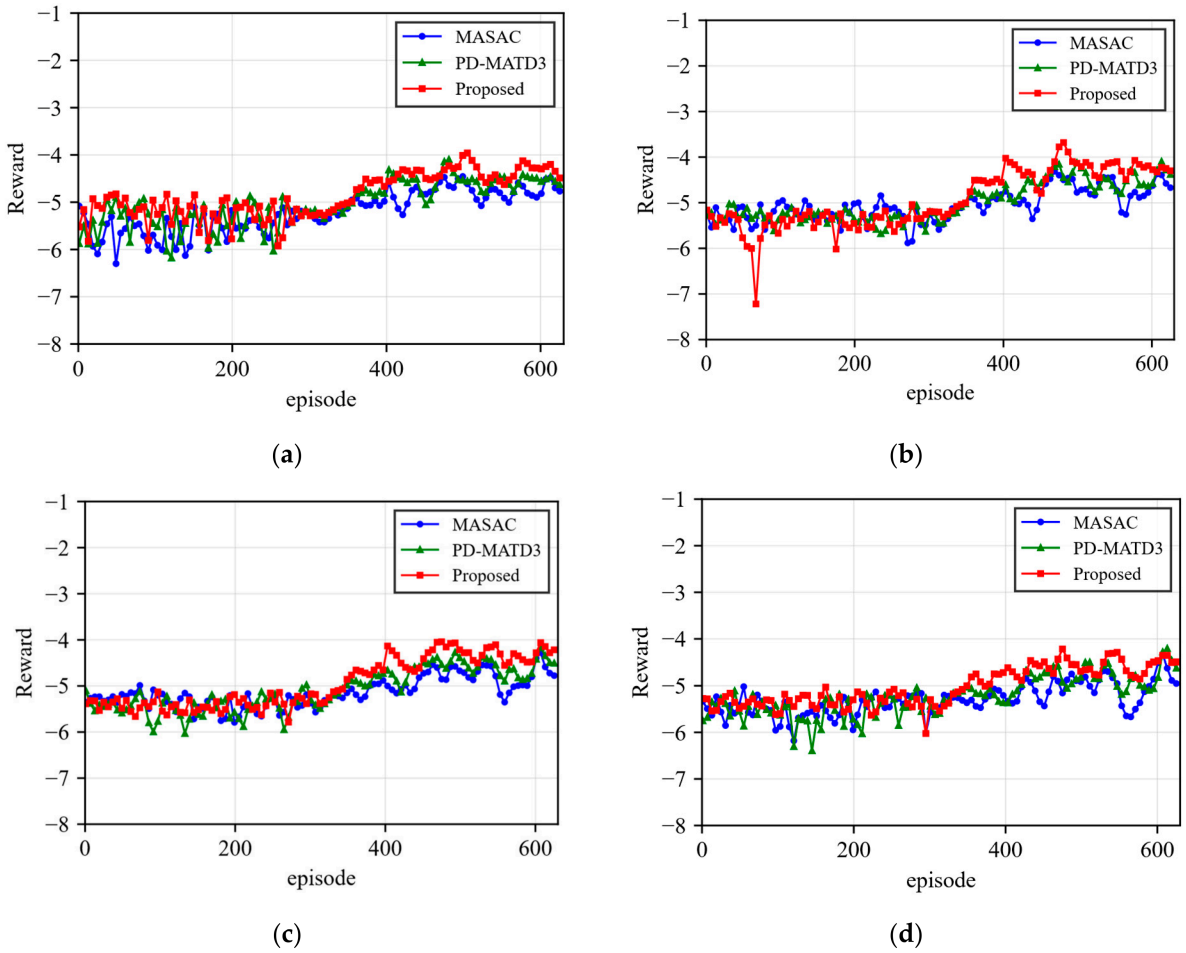

Figure 2. Reward value curves of MADRL-based methods under different preferences when step size is 0.2: (**a**) ω = [0.2, 0.2, 0.6]; (**b**) ω = [0.2, 0.6, 0.2]; (**c**) ω = [0.4, 0.2, 0.4]; (**d**) ω = [0.6, 0.2, 0.2].

4.3.2. Performance of Proposed Method Under a Single Day

To evaluate the performance of the proposed method, we visualize the Pareto front derived from the solutions obtained over a one-day period, as shown in Figure 3. The blue data points represent the NDSs. The red arrow indicates the possible moving directions of certain points. Crucially, the distribution of these solutions verifies the definition of NDSs stipulated in Theorem 1. As observed in the figure, any movement from a point

on this front to improve one objective inevitably leads to the deterioration of at least one other objective.

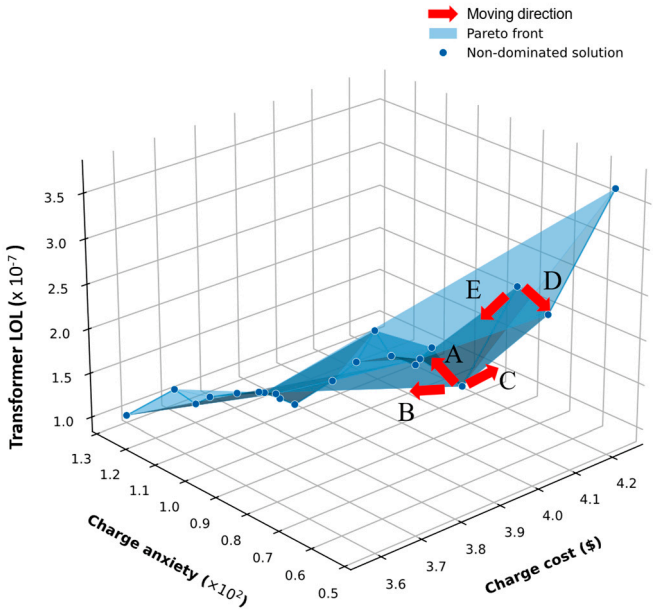

Figure 3. The Pareto front obtained by the proposed method on a certain day in the test set.

This phenomenon stems from the natural contradictions inherent in the optimization mechanisms of the three objectives. Specifically, charging cost is optimized by scheduling charging during low-electricity-price periods and increasing discharging revenues; user anxiety is minimized by maximizing charging volume and charging as early as possible; transformer LOL is reduced by limiting total charging power, increasing discharging support, and avoiding load concentration. Consequently, these conflicting mechanisms manifest as distinct trade-offs on the Pareto front. When charging cost is minimized, the system tends to concentrate charging during low-price valleys. This can cause sudden load spikes, leading to increased transformer LOL (moving direction A). Alternatively, aggressive discharging to lower costs may leave the battery with insufficient energy, thereby increasing user anxiety (moving direction B); conversely, when optimizing for user anxiety, the system prioritizes aggressive and immediate charging. This increase in power consumption and peak load naturally exacerbates transformer LOL and incurs higher charging costs (moving direction C); finally, when the priority shifts to minimizing transformer LOL, the system may shift EV loads to high-electricity-price periods or restrict high-power charging. This often forces charging into high-price intervals (increasing economic cost) or results in reduced total charged energy (increasing user anxiety), corresponding to moving direction D and E, respectively. Overall, the obtained NDSs form a clear trade-off surface among charging cost, charging anxiety, and transformer LOL, illustrating the effectiveness of the proposed method on the selected test day.

We next analyze the detailed results of a specific day in the test set under a certain preference. Figure 4 presents the performance evaluation of the proposed method on this day when ω = [0.2, 0.6, 0.2], including the real-time electricity price, base load, and changes in charging power and energy of four EVs in the system across different time steps, with dashed lines indicating the upper and lower limits of charging power and energy. Since the optimization is performed with a 15 min time step, the charging power and energy values shown in Figure 4c,d are hourly averages calculated from the four 15 min intervals within each hour for clearer visualization.

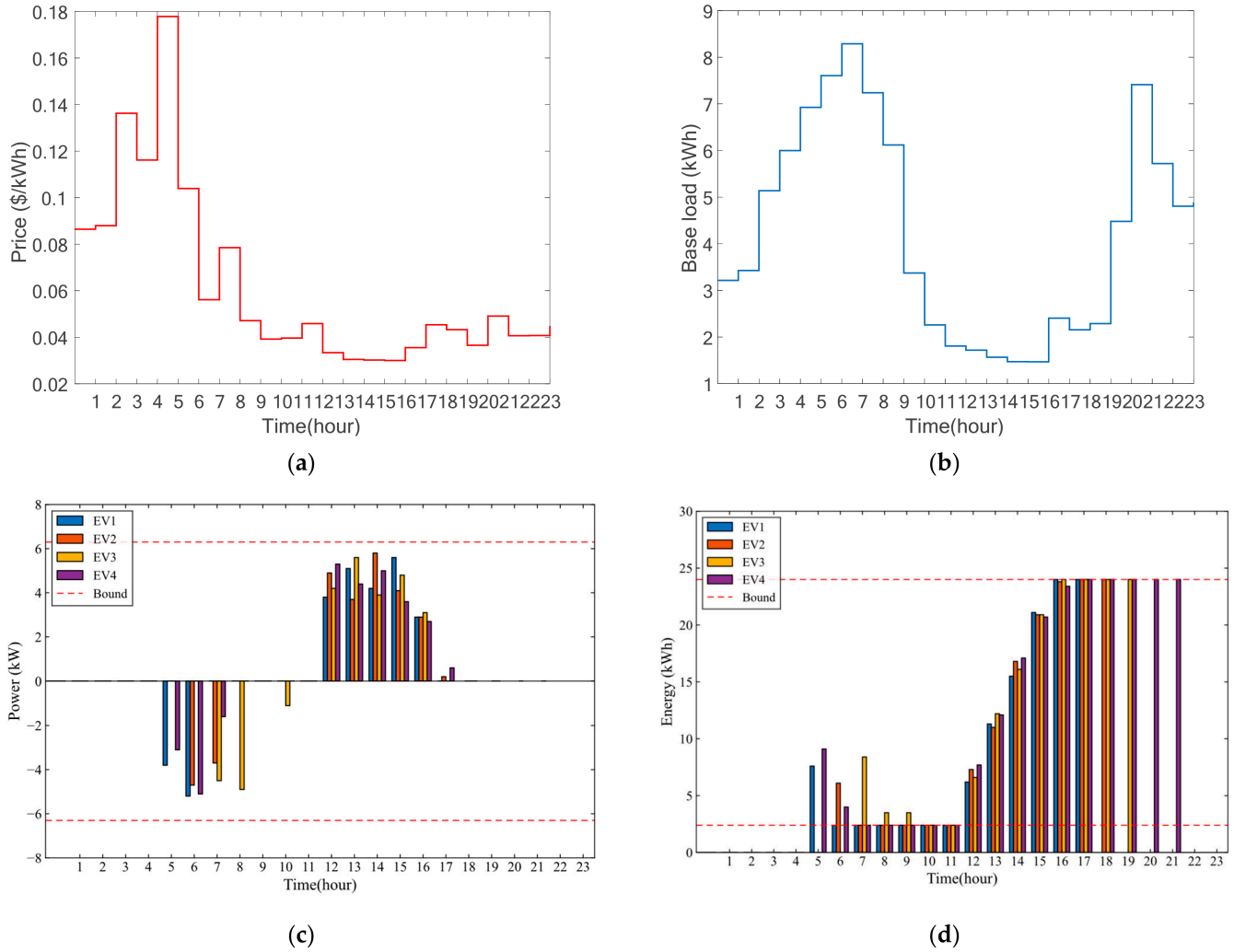

Figure 4. The performance evaluation of the proposed method on a specific day in the test set when ω = [0.2, 0.6, 0.2]: (**a**) real-time electricity price; (**b**) base load; (**c**) charging situation of 4 EV users; (**d**) energy of 4 EV users.

The results demonstrate that the proposed method can simultaneously consider charging cost, charging anxiety, transformer aging, and their corresponding preference weights when determining charging and discharging strategies. As shown in Figure 4a,b, electricity prices and base loads remain relatively high during hours 4–8. During this period, most EVs perform discharging or maintain low charging levels, with discharging power reaching approximately 4–5 kW. This behavior not only generates economic benefits through electricity selling but also alleviates transformer loading during peak-demand periods, thereby reducing the LOL of the transformer. After hour 11, both electricity prices and base loads decrease significantly. Consequently, all EVs gradually switch to charging mode and increase their charging power to approximately 4–6 kW, enabling energy replenishment at a lower charging cost while avoiding excessive load accumulation.

As illustrated in Figure 4d, the battery energy of all EVs increases steadily during low-price periods and reaches the battery capacity limit of 24 kWh by approximately hour 17, which is significantly earlier than most of their departure times. This behavior is mainly attributed to the relatively high anxiety preference under the current setting. To reduce the risk of insufficient energy and improve user satisfaction, the proposed method prioritizes charging completion once electricity prices and transformer loads become favorable. Consequently, all EVs achieve full charge well before departure while still taking advantage

of low-price periods, demonstrating that the proposed method can effectively balance charging anxiety, charging cost, and transformer aging considerations. Overall, by dynamically coordinating charging and discharging behaviors according to electricity prices, transformer loading conditions, and user preferences, the proposed method effectively balances economic benefits, equipment protection, and user satisfaction, thereby achieving a high-quality Pareto-optimal solution under the current preference setting.

4.3.3. Scalability Experiment

In order to verify the performance of the proposed method in a larger-scale EV system, the number of EVs is increased to 32, while the transformer capacity and base load are proportionally scaled by a factor of four. The step size is set as 0.2.

Table 4 summarizes the performance of different methods under the large-scale scenario. As the number of EVs increases, the dimensionality of the decision space grows substantially, making it more difficult to obtain high-quality Pareto-optimal solutions. Consequently, the SP method requires extensive evolutionary iterations to search for NDSs, resulting in the longest computation time of 4589.87 s.

Table 4. Comparison results for various methods in the scalability experiment.

Methods	HV (1×10^{-5})	SPA	NDS	Latency Time (s)
NLopt	4477.95 ± 10.92	0.17 ± 0.00	5.09 ± 0.01	144.25 ± 10.37
SP	2610.14 ± 14.52	0.20 ± 0.02	5.45 ± 0.08	4589.87 ± 354.69
MASAC	1566.63 ± 216.33	0.10 ± 0.04	4.44 ± 0.62	0.05 ± 0.01
PD-MATD3	2303.51 ± 202.88	0.12 ± 0.01	4.42 ± 0.89	0.02 ± 0.01
Proposed	3176.54 ± 152.77	0.04 ± 0.00	5.35 ± 0.20	0.05 ± 0.01

Compared with conventional MADRL methods, the proposed method demonstrates consistently superior performance. Specifically, it achieves an HV value of 3176.54×10^{-5}, which is 102.8% higher than MASAC and 37.9% higher than PD-MATD3, indicating a significantly better approximation of the Pareto front. Meanwhile, the proposed method obtains 5.35 NDSs, outperforming MASAC and PD-MATD3 by 20.5% and 21.0%, respectively, and remaining comparable to SP. In terms of Pareto-front distribution quality, the proposed method achieves the best SPA value of 0.04, representing improvements of 80.0%, 76.5%, 60.0%, and 66.7% over SP, Nlopt, MASAC, and PD-MATD3, respectively, which demonstrates a more uniform distribution of solutions along the Pareto front.

Furthermore, owing to the multi-Actor parallel decision-making mechanism, the proposed method requires only 0.05 s for online decision-making, reducing the computation time by approximately 99.99% compared with SP and 99.97% compared with Nlopt, while maintaining good Pareto-front quality. Overall, the above analysis results demonstrate its effectiveness and scalability for large-scale EV coordinated scheduling problems.

4.3.4. Study Without Vehicle-to-Grid

The previous experiments assume that EVs can both charge from and discharge to the grid. This setting is useful for evaluating the potential of bidirectional EV charging in reducing charging cost and transformer aging. However, in practical applications, vehicle-to-grid (V2G) operation may be limited by bidirectional charger availability, user willingness, and local grid regulations. Therefore, an additional study is conducted to evaluate the performance of the proposed MOMADRL method without V2G.

In this experiment, the discharging action is disabled by setting the lower charging-power bound to zero, i.e., $a_{\min} = 0$, and the other experimental settings are also kept the same. The comparison results are reported in Table 5. The step size is set to 0.2.

Table 5. Results of study without V2G.

Methods	HV (1×10^{-5})	SPA	NDS	Latency Time (s)
NLopt	207.72 ± 6.94	0.0442 ± 0.0038	5.68 ± 0.12	7.90 ± 0.46
SP	186.62 ± 9.54	0.0468 ± 0.0048	5.33 ± 0.11	250.80 ± 19.45
MASAC	60.69 ± 8.50	0.0005 ± 0.0002	4.40 ± 0.20	0.015 ± 0.001
PD-MATD3	129.48 ± 11.98	0.0003 ± 0.0001	4.36 ± 0.49	0.005 ± 0.001
Proposed	136.12 ± 6.51	0.0008 ± 0.0002	5.94 ± 0.22	0.009 ± 0.001

As shown in Table 5, the proposed MOMADRL method still maintains good performance. It achieves the highest HV among the learning-based methods, which is 124.3% higher than MASAC and 5.1% higher than PD-MATD3. Meanwhile, it obtains the best NDS among all compared methods, indicating that it can still generate more effective non-dominated solutions without EV discharging. Although its SPA is slightly higher than MASAC and PD-MATD3, it remains much smaller than those of NLopt and SP. In terms of online decision-making efficiency, the proposed method keeps a millisecond-level latency of 9 ms. Therefore, in the study without V2G, the proposed MOMADRL method still achieves a favorable Pareto-front quality with fast online inference.

4.3.5. Study with Battery Degradation Cost

Frequent EV discharging may accelerate battery degradation. To further examine this practical factor, an additional study is conducted by incorporating the battery degradation cost into the charging-cost objective. Specifically, the degradation cost is formulated as [8]:

$$C^i_{\text{deg}} = C_E \left| \frac{m_k}{100} \right| \left| \frac{a^i_t \Delta t}{\Omega^i} \right|, \tag{30}$$

where C_E is the total battery cost of EVs, which is set to 300 \$/kWh, and m_k denotes the slope of the linear approximation of battery life with respect to cycle number, which is set to -0.015. Then, the charging-cost objective is replaced by the sum of the electricity cost and the battery degradation cost:

$$f^{CC}_t = \sum_{i=1}^{I} (CC^i_t + C^i_{\text{deg}}), \tag{31}$$

The other two objectives, namely charging anxiety and transformer LOL, remain unchanged, and the other experimental settings are also kept the same. The comparison results are shown in Table 6. The step size is set to 0.2.

Table 6. Results of study with battery degradation cost.

Methods	HV (1×10^{-5})	SPA	NDS	Latency Time (s)
NLopt	278.10 ± 8.48	0.0296 ± 0.0034	5.81 ± 0.10	6.98 ± 0.41
SP	273.20 ± 12.93	0.0543 ± 0.0039	3.56 ± 0.08	140.81 ± 19.90
MASAC	111.31 ± 7.27	0.0005 ± 0.0001	4.20 ± 0.32	0.014 ± 0.001
PD-MATD3	146.12 ± 10.54	0.0001 ± 0.0001	3.01 ± 0.61	0.007 ± 0.001
Proposed	240.45 ± 8.52	0.0006 ± 0.0003	5.39 ± 0.17	0.006 ± 0.001

As shown in Table 6, the proposed method still achieves the highest HV among the learning-based methods, which is 116.0% higher than MASAC and 64.6% higher than PD-MATD3. Meanwhile, its NDS value reaches 5.39, also outperforming MASAC and PD-MATD3; its SPA is slightly higher than those of MASAC and PD-MATD3, it remains much lower than those of NLopt and SP. The proposed method requires only 6 ms, maintaining

millisecond-level execution. Therefore, the proposed method maintains favorable Pareto-front quality and fast online inference under the tested battery-degradation-cost setting.

5. Discussion

In this paper, the IEEE standard [39] is assumed to quantify transformer LOL. It is worth noting that the core of this work is the MOMADRL framework. As a data-driven approach, the proposed agent optimizes its policy based on the feedback (reward signals) from the environment. Therefore, changes to the transformer LOL modeling approach are internal to the environment and do not affect the validity of the proposed method. Should more precise or realistic methods be developed in the future to better reflect actual aging conditions, they could be integrated by updating the reward calculation module, while the agent's learning mechanism would remain effective.

In addition, the influence of transformer parameter variations and system setup scalability has been further investigated. Variations in key thermal parameters, including ambient temperature, top-oil time constant, and winding time constant, may affect the transformer thermal response and consequently the calculated LOL. However, the sensitivity analysis results in Table 7 show that the proposed method maintains consistently stable Pareto-front quality under all tested parameter settings. Specifically, the obtained NDS remains highly comparable to that of the optimal Nlopt solver, while the SPA values remain within a narrow range of 0.0007–0.0009, indicating a uniformly distributed Pareto front. Moreover, the HV values exhibit only minor fluctuations across different parameter settings. These results demonstrate that the proposed method is not sensitive to variations in transformer thermal characteristics.

Table 7. Comparison of Pareto-quality metrics and deviations in sensitivity analysis.

	Parameter	**Ambient Temperature θ_A**			**Top-Oil Time Constant τ_{TO}**			**Winding Time Constant τ_W**		
		40 °C	**20 °C**	**10 °C**	**6.860**	**8.232**	**5.488**	**0.080**	**0.096**	**0.064**
HV	Nlopt	241.30	244.92	247.10	241.30	244.18	238.62	241.30	242.65	239.78
	Proposed	190.11	193.24	195.10	190.11	192.03	187.95	190.11	191.08	188.84
SPA	Nlopt	0.0424	0.0409	0.0398	0.0424	0.0413	0.0440	0.0424	0.0417	0.0432
	Proposed	0.0007	0.0008	0.0008	0.0007	0.0007	0.0009	0.0007	0.0007	0.0008
NDS	Nlopt	5.90	5.94	5.98	5.90	5.93	5.86	5.90	5.91	5.88
	Proposed	5.92	5.85	5.99	5.92	5.93	5.89	5.92	5.83	5.90

To further examine the scalability of the proposed method beyond the benchmark eight-EV case, an additional 32-EV residential charging scenario is tested. As the system scale increases, the number of agents, charging demand, and optimization complexity also increase, resulting in a more challenging scheduling problem. However, these changes do not affect the implementation framework of the proposed MOMADRL method. To verify its scalability, an additional experiment was conducted on a larger residential system with a 200 kVA transformer serving 32 households. As shown in Table 4, the proposed method continues to outperform the other learning-based baselines, achieving the highest NDS and the best SPA while maintaining a computation time of only 0.05 s. These results demonstrate that the proposed method can preserve fast online inference and high-quality Pareto-front approximation among learning-based methods under larger-scale residential scenarios.

Although the above results demonstrate the effectiveness of the proposed MOMADRL method under the tested simulation settings, several practical limitations should be acknowledged. First, the V2G operation considered in this study depends on the availability of bidirectional chargers, compatible EV battery-management systems, reliable commu-

nication and metering infrastructure, and appropriate market or regulatory mechanisms. Therefore, the V2G-based results should be interpreted as potential benefits under assumed bidirectional charging availability, rather than direct deployment-ready validation for all residential users.

Second, user acceptance remains an important practical factor. EV owners may be concerned about battery degradation, warranty coverage, privacy, insufficient compensation, and whether the desired SOC can be reached before departure. Practical implementation should therefore include SOC guarantees, opt-out mechanisms, and suitable incentive or compensation schemes.

Third, the current formulation mainly focuses on charging cost, charging anxiety, and transformer LOL, while other distribution-network constraints, such as AC power-flow feasibility, nodal voltage limits, feeder capacity limits, congestion, and phase imbalance, are not explicitly modeled. These factors may further restrict feasible charging and discharging actions in real networks. Future work will incorporate power-flow, voltage-security, and feeder-capacity constraints to improve the practical applicability of the proposed framework.

6. Conclusions

To address the strong randomness and multi-objective characteristics of the ECOTL problem, we propose a MOMADRL framework based on a multi-Actor–Critic design. The proposed method is intended to approximate a set of preference-specific charging policies for balancing charging cost, charging anxiety, and transformer LOL, while maintaining fast online inference. Comparative experiments with different benchmarks yield the following results:

The proposed MOMADRL method deploys multiple Critic networks for each agent, where each Critic learns value functions with distinct preferences to guide multiple Actors in personalized learning of Pareto-optimal strategies. In the training stage, the proposed method shows consistently better convergence performance than the learning-based baselines. Averaging the converged rewards over the four preference settings, MASAC achieves a final reward of approximately -4.85, PD-MATD3 reaches -4.55, while the proposed method attains the highest reward of approximately -4.25. Moreover, the proposed method converges approximately 16% and 10% faster than MASAC and PD-MATD3, respectively, verifying its improved training performance.

Meanwhile, through the multi-Actor design, the proposed MOMADRL method executes personalized charging decisions tailored to distinct preferences. During the testing phase of the semi-synthetic case study, it achieves the best HV among the learning-based methods, improving HV by 95.0–96.8% over MASAC and 11.4–12.8% over PD-MATD3 under different step sizes, while also obtaining the highest NDS values among the learning-based methods. Although an HV gap remains compared with the optimization-based benchmarks NLopt and SP, the proposed method achieves millisecond-level online decision-making and is at least 400 times faster than these benchmarks. These results demonstrate that the proposed method achieves a better Pareto-front approximation than the other learning-based baselines while maintaining fast online inference in the tested residential EV charging case study. In the scalability experiment with 32 EVs, the proposed method improves HV by 102.8% and 37.9% over MASAC and PD-MATD3, respectively, and maintains a low latency of 0.05 s, demonstrating its scalability for larger EV coordinated scheduling problems. In the study without V2G, it still achieves the highest HV among the learning-based methods, with improvements of 124.3% and 5.1% over MASAC and PD-MATD3, respectively, and obtains the best NDS value. In the study with battery degradation cost, it further improves HV by 116.0% and 64.6% over MASAC and PD-MATD3, respectively,

while maintaining millisecond-level execution. These results indicate that the proposed method shows promising scalability and adaptability under the tested 32-EV, no-V2G, and battery-degradation-cost scenarios.

Looking ahead, although the IEEE standard [38] provides a standardized theoretical basis for transformer LOL modeling, future work could validate the proposed framework using field data or more detailed aging models to further enhance its practical applicability. The current study is conducted under simulation-based residential EV charging scenarios. In practical applications, V2G deployment depends on the availability of bidirectional chargers, compatible EV battery-management systems, reliable communication and metering infrastructure, user willingness, and appropriate market or regulatory mechanisms. Moreover, several real-world constraints need to be further considered. User acceptance is particularly important, as EV owners may be concerned about battery warranty, privacy, and insufficient economic incentives. In addition, voltage limits and feeder capacity constraints are critical operational factors in real distribution networks. Therefore, future work will incorporate user participation behavior, AC power-flow constraints, nodal voltage limits, and feeder congestion constraints to improve the practical applicability of the proposed method. Meanwhile, the proposed MOMADRL framework provides a flexible approach for multi-objective problems, especially for scenarios with strong randomness and dynamics that require fast control. This framework can also be combined with uncertainty modeling methods to improve robustness. As the number of EVs increases, the network parameters and coordination complexity will grow significantly. Therefore, we will also design more scalable methods for large-scale multi-objective EV scheduling problems in future work.

Author Contributions: Methodology, J.W. and C.D.; software, J.W.; validation, J.W. and C.D.; investigation, J.W.; resources, C.D.; data curation, J.W. and C.D.; writing—original draft preparation, J.W.; writing—review and editing, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: Author Ji Wang was employed by the CSG Electric Power Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

EV	Electric vehicle
LOL	Loss of life
ECOTL	EV charge optimization concerning transformer LOL
MOMADRL	Multi-objective multi-agent deep reinforcement learning
MADRL	Multi-agent deep reinforcement learning
IEA	International Energy Agency
DN	Distribution network
GA	Genetic algorithms
PSO	Particle swarm optimization
MOEA	Multi-objective evolutionary algorithm
MOEA/D	Multi-objective evolutionary algorithm based on decomposition
DRL	Deep reinforcement learning
SOC	State of charge
Dec-POMDP	Decentralized partially observable Markov decision process
MO-Dec-POMDP	Multi-objective Dec-POMDP
AMI	Advanced metering infrastructure
SP	Stochastic programming

MASAC	Multi-agent soft Actor–Critic
PD-MATD3	Preference-driven multi-agent twin delayed deep deterministic policy gradient
NDS	Non-dominated Solution
HV	Hypervolume
SPA	Spacing
V2G	Vehicle-to-grid

References

1. Sandhiya, E.; Gajanand, M.S. Hybrid Charging Infrastructure for Electric Vehicles: A Comprehensive Review and Avenues for Future Research. *Appl. Energy* **2025**, *401*, 126749. [CrossRef]
2. Michailidis, P.; Michailidis, I.; Kosmatopoulos, E.B. Reinforcement Learning for Electric Vehicle Charging Management: Theory and Applications. *Energies* **2025**, *18*, 5225. [CrossRef]
3. Vagropoulos, S.I.; Kyriazidis, D.K.; Bakirtzis, A.G. Real-Time Charging Management Framework for Electric Vehicle Aggregators in a Market Environment. *IEEE Trans. Smart Grid* **2016**, *7*, 948–957.
4. Wan, D.; Mo, W.; Li, J.; Yang, C.; Wu, J.; Zhou, Q.; Gong, Y. Two-Layer Cooperative Optimization of Flexible Interconnected Distribution Networks Considering Electric Vehicle User Satisfaction Degree. *Electronics* **2023**, *12*, 4582. [CrossRef]
5. Wang, G.; Lu, H.; Yang, X.; Li, H.; Song, X.; Rong, J.; Wang, Y. Economic-Optimal Operation Strategy for Active Distribution Networks with Coordinated Scheduling of Electric Vehicle Clusters. *Electronics* **2025**, *14*, 3154. [CrossRef]
6. Papadopoulos, P.; Jenkins, N.; Cipcigan, L.M.; Grau, I.; Zabala, E. Coordination of the Charging of Electric Vehicles Using a Multi-Agent System. *IEEE Trans. Smart Grid* **2013**, *4*, 1802–1810. [CrossRef]
7. Namerikawa, T.; Okubo, N.; Sato, R.; Okawa, Y.; Ono, M. Realtime Pricing Mechanism for Electricity Market with Built-in Incentive for Participation. *IEEE Trans. Smart Grid* **2015**, *6*, 2714–2724. [CrossRef]
8. Wan, Z.; Li, H.; He, H.; Prokhorov, D. Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 5246–5257. [CrossRef]
9. Alsabbagh, A.; Wu, B.; Ma, C. Distributed Electric Vehicles Charging Management Considering Time Anxiety and Customer Behaviors. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2422–2431. [CrossRef]
10. Turker, H.; Bacha, S.; Chatroux, D.; Hably, A. Low-Voltage Transformer Loss-of-Life Assessments for a High Penetration of Plug-In Hybrid Electric Vehicles (PHEVs). *IEEE Trans. Power Deliv.* **2012**, *27*, 1323–1331. [CrossRef]
11. Lin, F.; Zeng, J.; Xiahou, J.; Wang, B.; Zeng, W.; Lv, H. Multiobjective Evolutionary Algorithm Based on Nondominated Sorting and Bidirectional Local Search for Big Data. *IEEE Trans. Ind. Inform.* **2017**, *13*, 1979–1988. [CrossRef]
12. Vandael, S.; Claessens, B.; Ernst, D.; Holvoet, T.; Deconinck, G. Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market. *IEEE Trans. Smart Grid* **2015**, *6*, 1795–1805. [CrossRef]
13. Liu, Z.; Wu, Q.; Huang, S.; Wang, L.; Shahidehpour, M.; Xue, Y. Optimal Day-Ahead Charging Scheduling of Electric Vehicles Through an Aggregative Game Model. *IEEE Trans. Smart Grid* **2018**, *9*, 5173–5184. [CrossRef]
14. Momber, I.; Siddiqui, A.; Román, T.G.S.; Söder, L. Risk Averse Scheduling by a PEV Aggregator Under Uncertainty. *IEEE Trans. Power Syst.* **2015**, *30*, 882–891. [CrossRef]
15. Karfopoulos, E.L.; Hatziargyriou, N.D. A Multi-Agent System for Controlled Charging of a Large Population of Electric Vehicles. *IEEE Trans. Power Syst.* **2013**, *28*, 1196–1204. [CrossRef]
16. Ghofrani, M.; Arabali, A.; Etezadi-Amoli, M.; Fadali, M.S. Smart Scheduling and Cost-Benefit Analysis of Grid-Enabled Electric Vehicles for Wind Power Integration. *IEEE Trans. Smart Grid* **2014**, *5*, 2306–2313. [CrossRef]
17. Chen, C.; Duan, S. Optimal Integration of Plug-In Hybrid Electric Vehicles in Microgrids. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1917–1926. [CrossRef]
18. Ali, Z.; Putrus, G.; Marzband, M.; Gholinejad, H.R.; Saleem, K.; Subudhi, B.D. Multiobjective Optimized Smart Charge Controller for Electric Vehicle Applications. *IEEE Trans. Ind. Appl.* **2022**, *58*, 5602–5615. [CrossRef]
19. Luo, D.; Ji, W.; Hu, X. Parameter Optimization and Control Strategy of Hybrid Electric Vehicle Transmission System Based on Improved GA Algorithm. *Processes* **2023**, *11*, 1554. [CrossRef]
20. Hecht, C.; Figgener, J.; Sauer, D.U. Predicting Electric Vehicle Charging Station Availability Using Ensemble Machine Learning. *Energies* **2021**, *14*, 7834. [CrossRef]
21. Du, C.; Hu, W.; Cao, D.; Wang, Y.; Wu, J.; Chen, Z. Bi-Level Distribution System Parallel Restoration by Integrating Deep Reinforcement Learning with Physics-Based Optimization. *J. Mod. Power Syst. Clean Energy* **2025**. [CrossRef]
22. Cao, D.; Hu, W.; Zhao, J.; Huang, Q.; Chen, Z.; Blaabjerg, F. A Multi-Agent Deep Reinforcement Learning Based Voltage Regulation Using Coordinated PV Inverters. *IEEE Trans. Power Syst.* **2020**, *35*, 4120–4123. [CrossRef]
23. Luo, L.; Chen, M.; Wang, R.; Gao, H.; Liu, J. Inter-Sub-District Operation Optimization Method in Distribution Network Based on Available Transfer Capacity of Adjacent Zones. *South. Power Syst. Technol.* **2025**, *19*, 99–110.

24. Cao, D.; Hu, J.; Liu, Y.; Hu, W. Decentralized Graphical-Representation-Enabled Multi-Agent Deep Reinforcement Learning for Robust Control of Cyber-Physical Systems. *IEEE Trans. Reliab.* **2024**, *73*, 1710–1720. [CrossRef]
25. Wu, Y.; Liu, W.; Liu, M.; Zhang, Y.; Wang, Y.; Tangwang, Q. Multi-Area Cooperative Algorithm Under Large-Scale New Energy Access to Power Grid. *South. Power Syst. Technol.* **2025**, *19*, 174–188.
26. Cao, D.; Liu, Y.; Wang, Y.; Zhang, Q.; Hu, W. Deep Reinforcement Learning in Power Systems Resilience: A Review. *IEEE Trans. Reliab.* **2025**, *74*, 5356–5370. [CrossRef]
27. Cao, D.; Zhao, J.; Hu, J.; Pei, Y.; Huang, Q.; Chen, Z.; Hu, W. Physics-Informed Graphical Representation-Enabled Deep Reinforcement Learning for Robust Distribution System Voltage Control. *IEEE Trans. Smart Grid* **2024**, *15*, 233–246. [CrossRef]
28. Wang, Y.; Hu, W.; Cao, D.; Zhao, P.; Abulanwar, S.; Chen, Z.; Blaabjerg, F. Local Distribution Voltage Control Using Large-Scale Coordinated PV Inverters: A Novel Multi-Agent Deep Reinforcement Learning-Based Approach. *IEEE Trans. Smart Grid* **2025**, *16*, 2683–2686. [CrossRef]
29. Zhang, J.; Che, L.; Shahidehpour, M. Distributed Training and Distributed Execution-Based Stackelberg Multi-Agent Reinforcement Learning for EV Charging Scheduling. *IEEE Trans. Smart Grid* **2023**, *14*, 4976–4979. [CrossRef]
30. Li, S.; Hu, W.; Cao, D.; Zhang, Z.; Huang, Q.; Chen, Z.; Blaabjerg, F. EV Charging Strategy Considering Transformer Lifetime via Evolutionary Curriculum Learning-Based Multiagent Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2022**, *13*, 2774–2787. [CrossRef]
31. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
32. Li, Y.; Cai, Y.; Zhao, T.; Liu, Y.; Wang, J.; Wu, L.; Zhao, Y. Multi-Objective Optimal Operation of Centralized Battery Swap Charging System with Photovoltaic. *J. Mod. Power Syst. Clean Energy* **2022**, *10*, 149–162. [CrossRef]
33. Yao, W.; Zhao, J.; Wen, F.; Dong, Z.; Xue, Y.; Xu, Y.; Meng, K. A Multi-Objective Collaborative Planning Strategy for Integrated Power Distribution and Electric Vehicle Charging Systems. *IEEE Trans. Power Syst.* **2014**, *29*, 1811–1821. [CrossRef]
34. Liemthong, R.; Srithapon, C.; Ghosh, P.K.; Chatthaworn, R. Home Energy Management Strategy-Based Meta-Heuristic Optimization for Electrical Energy Cost Minimization Considering TOU Tariffs. *Energies* **2022**, *15*, 537. [CrossRef]
35. Zhang, F.; Yang, Q.; An, D. Hamlet: Hierarchical Multi-Objective Reinforcement Learning Method for Charging Power Control of Electric Vehicles with Dynamic Quantity. *IEEE Trans. Smart Grid* **2024**, *15*, 783–794. [CrossRef]
36. Mu, C.; Shi, Y.; Xu, N.; Wang, X.; Tang, Z.; Jia, H.; Geng, H. Multi-Objective Interval Optimization Dispatch of Microgrid via Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2024**, *15*, 2957–2970. [CrossRef]
37. Basaklar, T.; Gumussoy, S.; Ogras, U.Y. PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
38. Wang, Y.; Qiu, D.; Teng, F.; Strbac, G. Towards Microgrid Resilience Enhancement via Mobile Power Sources and Repair Crews: A Multi-Agent Reinforcement Learning Approach. *IEEE Trans. Power Syst.* **2024**, *39*, 1329–1345. [CrossRef]
39. *IEEE C57.91-2011*; IEEE Guide for Loading Mineral-Oil-Immersed Transformers and Step Voltage Regulators. IEEE: New York, NY, USA, 2011.
40. Hao, L.; Jin, J.; Xu, Y. Laxity Differentiated Pricing and Deadline Differentiated Threshold Scheduling for a Public Electric Vehicle Charging Station. *IEEE Trans. Ind. Inform.* **2022**, *18*, 6192–6202. [CrossRef]
41. Historical Hourly Electricity Price from PJM. 2017. Available online: https://dataminer2.pjm.com/ (accessed on 10 April 2025).
42. A Real 240-Bus Distribution System with One-Year Smart Meter Data. Available online: http://wzy.ece.iastate.edu/Testsystem.html (accessed on 11 November 2025).
43. Library for Nonlinear Optimization. Available online: https://nlopt.readthedocs.io/en/latest/ (accessed on 11 November 2025).