Energy 284 (2023) 129335

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Multi-agent reinforcement learning for electric vehicle decarbonized routing and scheduling

Yi Wang [a], Dawei Qiu [a,*], Yinglong He [b], Quan Zhou [c], Goran Strbac [a]

[a] Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, UK
[b] Advanced Resilient Transport Systems, University of Surrey, Guildford, GU2 7XH, UK
[c] Birmingham C.A.S.E Automotive Research Centre, University of Birmingham, Birmingham, B15 2TT, UK

ARTICLE INFO

Keywords:
Electric vehicles
Carbon emissions
Carbon intensity
Routing and scheduling
Transport and power networks
Multi-agent reinforcement learning

ABSTRACT

Low-carbon transitions require joint efforts from electricity grid and transport network, where electric vehicles (EVs) play a key role. Particularly, EVs can reduce the carbon emissions of transport networks through eco-routing while providing the carbon intensity service for power networks via vehicle-to-grid technique. Distinguishing from previous research that focused on EV routing and scheduling problems separately, this paper studies their coordinated effect with the objective of carbon emission reduction on both sides. To solve this problem, we propose a multi-agent reinforcement learning method that does not rely on prior knowledge of the system and can adapt to various uncertainties and dynamics. The proposed method learns a hierarchical structure for the mutually exclusive discrete routing and continuous scheduling decisions via a hybrid policy. Extensive case studies based on a virtual 7-node 10-edge transport and 15-bus power network as well as a coupled real-world central London transport and 33-bus power network are developed to demonstrate the effectiveness of the proposed MARL method on reducing carbon emissions in transport network and providing carbon intensity service in power network.

1. Introduction

Over the last decades, the power and transport systems have undergone major changes in various aspects due to a number of technical, economic and environmental factors. One of the most remarkable things is associated with the climate change, which has altered our energy policy and energy mix [1]. Committee on Climate Change (CCC), the UK's independent climate advisory body, claims that by setting an ambitious new target to reduce greenhouse gas emissions to net zero by 2050, the UK can halt its contribution to global warming within 30 years [2]. As many countries have passed regulations to restrict fossil fuel consumption of traditional vehicles promising to a low-carbon future, a rapid increase in the use of electric vehicles (EVs) has been witnessed by both power and transport systems [3].

On the one hand, EVs can provide various ancillary services (e.g., demand–supply balance, frequency/voltage regulation, carbon intensity service, etc.) for the power system due to their significant advantages on mobility and flexibility, which boosts the decentralization and decarbonization of power systems [4]. On the other hand, the transport system is increasingly regarded as a key barrier to the low-carbon transition due to the high cost of substituting energy-dense liquid fossil fuels [5]. A potential solution is in the transition to electro-mobility, i.e., a shift from conventional fossil vehicles to EVs; thus, governments and companies are continually improving the required infrastructures (e.g., charging stations) to promote the popularization of EVs.

It is worth noting that the large-scale deployment of EVs in both power and transport systems introduces challenges on the efficient and stable route selection and power scheduling, due to the potential privacy concerns and the difficulty to handle various system uncertainties and dynamics. To this end, it is urgent to develop an effective coordinated control algorithm for these large-scale and small-size decentralized EVs to fully exploit their flexibility and mobility in coupled power and transport systems towards the low-carbon transition in both sides [6]. However, the joint routing and scheduling strategies of large-scale EVs have not been well investigated in existing literature under the topic of reducing carbon emissions. As such, this paper proposes a real-time and automatic control policy for mult-EVs eco-routing and vehicle-to-grid (V2G) scheduling problem toward the provision of carbon service in the context of coupled transport-power networks.

* Corresponding author.
E-mail address: d.qiu15@imperial.ac.uk (D. Qiu).

https://doi.org/10.1016/j.energy.2023.129335
Received 29 June 2023; Received in revised form 29 August 2023; Accepted 9 October 2023
Available online 11 October 2023

Nomenclature

A. Indices and Sets

$t \in T$	Index and set of time steps
$i \in I$	Index and set of EVs
$r \in R$	Index and set of transport roads
$g \in DG$	Index and set of diesel generators (DGs)
$g \in RES$	Index and set of renewable energy sources (RESs)
$k \in CS$	Index and set of charging stations (CSs)
$b \in B$	Index and set of electrical buses
$l \in L$	Index and set of electrical lines

B. Parameters

Δt	Time resolution (15 min)
λ_t^g	Grid electricity price at time t (£/kWh)
λ_t^c	Carbon intensity at time t (gCO2/kWh)
λ^e	Carbon price (£/tCO2)
$T_r^{rd,0}$	Free driving time on road r (min)
$\overline{D}_r^{rd}$	Capacity of road r
α^{rd}, β^{rd}	Road retardation coefficients (%)
L_r	Length of road r (km)
$d_{r,t}^{rd}$	Base flow of road r at time t
η^w	Weight constant of EVs (m/s^2)
R^g	Carbon emissions per fuel consumption (gCO2/kg)
η^e	Engine constant of EVs (kJ/s)
η^s	Speed constant of EVs (kg/m)
$\overline{P}_i^{ev}$	Power capacity of EV i (kW)
$\overline{E}_i$	Energy capacity of EV i (kWh)
$\underline{S}_i$	Minimum battery SoC of EV i (%)
$\overline{S}_i$	Maximum battery SoC of EV i (%)
η_i^c	Charging efficiency of EV i (%)
η_i^d	Discharging efficiency of EV i (%)
$\overline{v}$	Maximum permissible voltage (p.u.)
$\underline{v}$	Minimum permissible voltage (p.u.)
r_{bp}, x_{bp}	Resistance/Reactance of line bp (p.u.)
x_{bp}	Reactance of line bp (p.u.)
$\overline{l}_{bp}$	Ampacity Limit of line bp (A)
$\overline{P}_g^{gd}$	Active power import limit of grid g (kW)
$\underline{P}_b^{gd}$	Active power export limit of grid g (kW)
$\tilde{P}_{g,t}^{res}$	Active power capacity of RES g at time t (kW)
$\overline{Q}_g^{gd}$	Reactive power import limit of grid g (kVAR)
$\underline{Q}_g^{gd}$	Reactive power export limit of grid g (kVAR)
$\overline{P}_k^{cs}$	Capacity of CS k (kW)

C. Variables

$F_{i,r,t}^{rd}$	Fuel usage of vehicle i on road r at time t (kg/km)
$C_{r,t}^{rd}$	Vehicle carbon emissions on road r at time t (kg)
$V_{r,t}^{rd}$	Vehicle average velocity on road r at time t (m/s)
$a_{r,t}$	Vehicle acceleration on road r at time t (m/s^2)
$T_{r,t}^{rd}$	Commuting time of road r at time t (h)
$u_{i,r,t}^{rd}$	Binary indicating whether EV i is traveling on road r ($u_{i,r,t}^{rd} = 1$) or not ($u_{i,r,t}^{rd} = 0$) at time t
$P_{i,t}^c$	Charging power of EV i at time t (kW)
$P_{i,t}^d$	Discharging power of EV i at time t (kW)
$S_{i,t}^{ev}$	Battery SoC of EV i at time t (kWh)
$u_{i,t}^{ev}$	Binary indicating whether EV i charge ($u_{i,t}^{ev} = 1$) or discharge ($u_{i,t}^{ev} = 0$) at time t
$A_{i,t}$	Binary indicating whether EV i is connected with grid ($A_{i,t} = 1$) or not ($A_{i,t} = 0$) at time t
$E_{i,r,t}^{rd}$	Energy usage of EV i on road r at time t (kWh)
$P_{g,t}^{dg}$	Active power output of DG g at time t (kW)
$Q_{g,t}^{dg}$	Reactive power output of DG g at time t (kVAR)
$P_{g,t}^{res}$	Active power output of RES g at time t (kW)
$v_{b,t}$	Squared voltage of bus b at time t (p.u.)
$l_{bp,t}$	Squared current of line bp at time t (p.u.)
$P_{bp,t}$	Active power flow of line bp at time t (kW)
$Q_{bp,t}$	Reactive power flow of line bp at time t (kVAR)
$P_{g,t}^{gd}$	Active power supply from grid at time t (kW)
$Q_{g,t}^{gd}$	Reactive power supply from grid at time t (kVAR)

1.1. Literature review on model-based optimization methods

Previous work has attempted to solve EV routing and/or scheduling problems, which can be classified into three categories and mainly solved by model-based optimization methods. First, regarding the route planning problem, a fuzzy EV routing problem with recharging stations and time windows [7] and a cost-optimal route planning problem with vehicle dynamics [8] are investigated for travel distance minimization. Second, regarding the power scheduling problem, a stochastic optimization model capturing uncertain electricity prices is developed in [9] to investigate EVs' flexibility in providing up-/down-regulation services, while the optimal EV charging behaviors are modeled in [10] to reduce carbon emissions and wind curtailment. It is noted that the above papers focus on EV problems of either routing [7,8] or scheduling [9,10], while ignoring their coordinated effect. This effect is, however, very important since an efficient routing saves battery energy consumption on road and a smart charging strategy ensures the sufficient energy requirement for traveling purposes. As a result, there have been research papers [11–14] of the third category considering the joint routing and scheduling problems of EVs that are co-optimized to reduce travel time and energy cost. In [11], a two-stage decomposition algorithm is proposed to solve the joint routing and charging problem of multi-EVs towards the reduction of travel time and energy cost. In [12], the joint route selection and power scheduling problem is optimized to improve the overall economic profits of EVs. In [13], the optimal routing and charging problem of an EV fleet is formulated for high-efficiency dynamic transit systems, taking into account energy efficiency and charging price. In [14], an approximate distributed algorithm is developed to tackle the routing and charging problem of EVs. However, the above works only focus on the transport network, while the power network operation is not considered.

To address this issue, there have been papers [15–20] considering the joint routing and scheduling problem of EVs in the context of coupled transport-power networks. In [15], a security-constrained transport-power model is developed to investigate the hourly routing and V2G scheduling behaviors of EVs. In [16], a multi-period optimal traffic and power flow model is developed for the time-varying traffic and electricity demands. In [17], a bi-level framework is proposed to identify the optimal location of EV charging stations, considering both route selection and charging cost optimization. In [18], a holistic modeling framework is introduced to describe the distribution of traffic and power flows for the coupled transport-power network. In [19], the dynamic network equilibrium encapsulating the choices of route and charging location is studied to capture the temporally-dynamic interactions between transport networks and power networks. In [20],

a bi-level formulation is proposed to investigate the optimal charging pricing problem of charging stations (CSs) in coupled transport-power networks. However, the above works focus on the minimization of travel time or charging costs rather than the decarbonization of transport and power sectors.

To capture the carbon benefits of EVs in the transport-power network, authors in [21,22] have developed carbon-aware EV routing and charging strategies. The main idea is to reduce the emissions of fossil power plants via effective routing and charging behaviors. However, these two papers [21,22] together with the above-mentioned papers [12–20] do not consider any system uncertainty, which might be impractical in the transport-power network characterized by vast numbers of uncertainties and dynamics. There have been papers [23,24] developing planning strategies for low-carbon CSs, considering various uncertainties, e.g., traffic volumes, demand, and RESs. However, these papers [21–24] together with [13,18–20] do not consider the V2G capabilities and assume a constant charging rate for all EVs. Furthermore, these two papers focus on decarbonizing the power sector, while the carbon emissions caused by the congestion impact on transport networks are not considered.

1.2. Literature review on model-free learning methods

Despite the extensive efforts made to investigate the EV routing and/or scheduling problems, the limitations of model-based optimization methods cannot be erased: (1) it is assumed that the complete knowledge of transport network, EV models, and power network are already known prior to the experience, which is normally impractical taking into account the highly dynamic and stochastic real-world environment; (2) system uncertainties are captured by scenario-based stochastic programming methods, which can be time-consuming [25]. As a model-free and data-driven method, reinforcement learning (RL) [26] is used to learn the optimal decisions for agents (EVs) in a dynamic process by utilizing their experiences obtained from repeated interactions with the environment (coupled transport-power network), without any *prior* knowledge. In this regard, RL does not require any knowledge of simulated problems, since they are integrated into the environment that is regarded as a black box for RL agents. Additionally, as an online learning method, RL learns the uncertainties directly from historical data and can adapt to various conditions in milliseconds [27]. Thus, RL is stated as a valid solution for the EV real-time control problem in a complex environment.

Similar to above studies based on model-based optimization methods, the existing RL literature on EV routing and/or scheduling problems can also be classified into three categories: (1) RL for EV route selection [28,29]; (2) RL for EV power scheduling [30,31]; and (3) RL for joint route selection and power scheduling [32–35]. However, some limitations to the transport network and electricity grid have not been explicitly modeled. On one hand, the routing decisions in most papers [28,29,32,33] are characterized by selecting the potential path to the designated CSs, while the real-time road dynamics are not captured. To address this issue, the real-time routing behaviors characterized by different road directions are modeled in [34,35]. However, the road congestion impact has not been investigated. Additionally, none of the above papers studied the eco-routing problem of EVs in the transport network. On the other hand, the scheduling decisions in most papers are characterized by less flexibility with strong assumptions, e.g., the charging rate is constant in [32] which is not controllable by EVs; only three modes (charging, discharging, and idle) [31,33,34] or discrete power rate [30] can be selected by EVs. In such simple settings, EVs are unable to adequately display their charging and discharging behaviors, resulting in sub-optimal solutions.

Previous work has successfully applied various single-agent reinforcement learning (SARL) methods to handle the single EV routing and/or scheduling problems, e.g., policy gradient (PG) for the routing problem [28], deep Q-network (DQN) for the scheduling problem [30], and joint routing and scheduling problem [32,33]. Nevertheless, our paper considers multiple EVs that operate in the coupled transport-power network, rendering the problem to a multi-agent setting that can be solved via multi-agent RL (MARL) methods. The most straightforward method in the MARL family is independent learning; for example, each EV agent in [31] deploys a Q-learning method to train its charging policy in a distribution network; and each EV agent in [34] deploys a DQN method to train its individual routing and scheduling control policy in electricity grid. Such an independent manner is fast, however, often ineffective to solve a large-scale multi-agent problem, because it focuses on local information only and ignores the correlative effect with other EV agents and thus causing the learning non-stationary issue [36]. To overcome this problem, authors in [35] introduce a central controller (e.g., an EV aggregator) that assists local EV agents to train their joint routing and scheduling policy in a centralized manner. In contrast to independent learning, the central controller requires knowledge from all local EV agents that may destroy their privacy. In addition, it is worth noting that EV routing and scheduling decisions are mutually exclusive and in different domains, i.e., they cannot simultaneously make decisions on routing in the transport network and scheduling in the power network. However, the RL methods employed in all the above papers [32–35] consider routing and scheduling decisions simultaneously, which can be inefficient. Furthermore, considering the algorithm's scalability, it can be infeasible to learn specialized policies for each EV agent under the MARL setup, leading to high computational costs.

1.3. Paper contribution

Inspired by the aforementioned issues, this paper focuses on investigating the benefits of multi-EVs in reducing carbon emissions on both the transport and power networks. A hierarchical MARL method is proposed to assist EVs in learning a two-level framework that can make effective routing and scheduling decisions in a sequential manner without prior knowledge. The main contributions are summarized below:

1. Propose a carbon-aware EVs joint eco-routing and V2G scheduling problem in a coupled transport-power network. In contrast to [11–24] that employ model-based optimization methods, this paper proposes a model-free learning method. In contrast to [21–24] that only consider the carbon emissions in power sectors, this paper focuses on reducing carbon emissions in both transport and power sectors. In contrast to [11–14] that ignore power network operations, this paper captures the coupled transport-power network.
2. Formulate the multi-EVs joint eco-routing and V2G scheduling problem as a *Partially Observable Markov Game* (POMG). EVs' mobility and flexibility are fully exploited to reduce carbon emissions in the transport network associated with road congestion and the power network via carbon intensity service provision. In contrast to [28,30,32,33] that employ SARL methods, this paper formulates the studied problem in a multi-agent setup.
3. Develop a novel MARL algorithm to effectively solve this POMG by (a) learning a hierarchical architecture to choose between making decisions either in the transport network or the power network; (b) employing a hybrid policy to handle both discrete routing and continuous scheduling action domains; and (c) adopting a parameter-sharing (PS) framework to learn a shared control policy, accelerating the training speed and improving the training performance.
4. Learn a generalized and real-time automatic MARL control policy that can adapt to the transport and power dynamics, including a real-world transport network in central London with large-scale EV penetration.

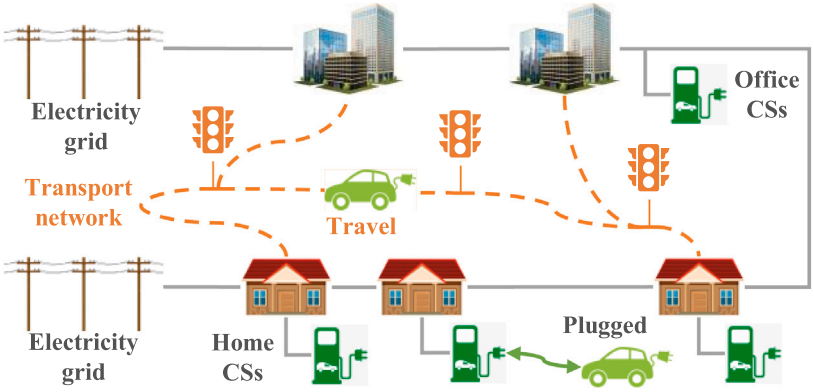


Fig. 1. The illustration of multi-EVs joint eco-routing and power scheduling problems in the transport network and the electricity grid.

1.4. Paper organization

The rest of the paper is organized as follows. Section 2 presents the formulations of the utilized EV model and the coupled power-transport network. Section 3 introduces the POMG formulation of the EV joint routing and scheduling problem. Section 4 details the proposal MARL algorithm to solve the POMG. The experiment setup and the case studies are presented in Sections 5 and 6, respectively. Finally, Section 7 concludes this paper.

2. Problem formulations

To fully explore the benefits of EVs in reducing carbon emissions, we focus on a multi-EVs joint eco-routing and V2G scheduling problem in the context of the coupled transport and power networks, as depicted in Fig. 1. Overall, the problem has two perspectives: (1) EVs make appropriate routing decisions in the transport network to alleviate road congestion and consequently reduce carbon emissions while ensuring the completion of daily journeys, i.e., travel from home to office in the morning and return home from office in the evening; (2) EVs make reasonable charging/discharging behaviors in response to electricity prices and carbon intensity signals when they are connected to the power network via CSs, e.g., charge power during the night for the next day's traveling purpose and discharge power to the grid for carbon service provision.

In detail, when EVs are traveling in the transport network, they can observe the real-time local transport information of vehicle locations and traffic volumes, and then optimally select moving directions via a smart routing algorithm. When EVs are plugged into the CSs, they can observe the real-time grid information of price and carbon signals and the EV battery state-of-charge (SoC), and then optimally manage the charging/discharging power via a smart scheduling algorithm. It is worth noting that the above two decisions are coupled, since efficient transport routing can allow more time for EVs to utilize their V2G flexibility, while reasonable power scheduling can provide sufficient battery energy for EVs to exploit their transport mobility.

2.1. Eco-routing in transport network

A transport network can be represented as a directed graph $G = (N, R)$, where N denotes the set of traffic nodes and R is the set of traffic roads. The transport requirement is modeled by origin–destination (O-D) demand pairs (e.g., home-office and office-home). Each O-D pair is connected by a set of routes, while each route consists of a set of transport roads in the transport network [15].

Specifically, as shown in Fig. 1, EVs commute through the routes of O-D pairs to connect with certain nodes in the transport network for their two daily journeys: (1) home to office (e.g., 6:00–9:00) and (2) office to home (e.g., 17:00–20:00). Consequently, there are several potential routes $k \in Ro$ between residential homes and commercial offices, where each home-office route is connected by a set of roads $r \in R$ in the transport network. The traffic volume $D^{rd}_{r,t}$ at time step t on road r can be expressed as

$$D^{rd}_{r,t} = d^{rd}_{r,t} + \sum_{i \in I} u^{rd}_{i,r,t}, \forall r \in R, \forall t \in T, \tag{2.1}$$

which includes the base flow $d^{rd}_{r,t}$ (represented by the total number of non-EV vehicles) on road r and the EV flow (represented by the sum of $u^{rd}_{i,r,t}$) [15], where the binary $u^{rd}_{i,r,t} \in \{0, 1\}$ indicates if EV i is traveling at time step t on road r ($u^{rd}_{i,r,t} = 1$) or not ($u^{rd}_{i,r,t} = 0$). The traffic volume $D^{rd}_{r,t}$ is time-varying and can lead to different commuting time $T^{rd}_{r,t}$ at time step t on road r, which is expressed as

$$T^{rd}_{r,t} = T^{rd,0}_{r}[1 + \alpha^{rd}(\frac{D^{rd}_{r,t}}{\overline{D}^{rd}_{r}})^{\beta^{rd}}], \ \forall r \in R, \forall t \in T, \tag{2.2}$$

where $\overline{D}^{rd}_{r}$, $T^{rd,0}_{r}$, α^{rd}, and β^{rd} correspond to the capacity of road r, the free-flow driving time (i.e., without road congestion), and two retardation coefficients, respectively [37]. It is expected from (2.2) that large traffic volume $D^{rd}_{r,t}$ may cause serious road congestion and longer commuting time $T^{rd}_{r,t}$. As a result, the commuting time of an O-D pair (route) k can be expressed as

$$T^{rd}_{k,t} = \sum_{r \in R} T^{rd}_{r,t} \Theta_{r,k}, \forall k \in Ro, \forall t \in T, \tag{2.3}$$

where $\Theta_{r,k}$ indicates whether road r is part of route k. If road congestion occurs on road r that is part of route k, the commuting time of route k will increase accordingly.

Given the real-time traveling time $T^{rd}_{r,t}$, the average vehicle speed $V^{rd}_{r,t}$ at time step t on road r can be calculated as

$$V^{rd}_{r,t} = \frac{L_r}{T^{rd}_{r,t}}, \ \forall r \in R, \forall t \in T, \tag{2.4}$$

where L_r is the length of road r. In addition, each journey is required to finish within a specified time interval T^{trl}_{i} (e.g., 1 h), which can be constrained as

$$\sum_{r \in R} \sum_{t=t^{dep}_i}^{t^{arr}_i} u^{rd}_{i,r,t} T^{rd}_{r,t} \le T^{trl}_{i}, \ \forall i \in I, \tag{2.5}$$

where t^{dep}_i and t^{arr}_i respectively indicate the departure (e.g., 7:00) and arrival (e.g., 7:45) time of EV agent i per journey.

To model the real-time fuel consumption and carbon emissions caused by the traveling activities in the transport network, we introduce the generic model for computing fuel consumption and carbon emissions [38], which are based on the theory of vehicle dynamics and can capture the influence of various factors (e.g., the angle of inclination, rolling resistance, engine power, etc.). Since it is difficult to obtain all these factors in practice, to simplify the fuel consumption

model and make it applicable to the studied eco-routing problem, only the easy-to-measure factors (vehicle speed, acceleration, and angle of inclination) are regarded as variables in the model, while the difficult-to-measure factors can be substituted by parameters that are accurately estimated [38]. Specifically, the fuel consumption per meter of non-EV i at time step t on road r can be represented as

$$F^{rd}_{i,r,t} = \beta_1 \cos\theta + \beta_2 \sin(\theta) + \beta_3 (V^{rd}_{r,t})^2 + \beta_4 a_{r,t} + \beta_5 a_{r,t}/V^{rd}_{r,t} + \beta_6/V^{rd}_{r,t} + \beta_7, \ \forall i \in I, \forall r \in R, \forall t \in T, \tag{2.6}$$

where the parameters $\beta_1 - \beta_7$ can be estimated by the maximum likelihood estimation (MLE) method and evaluated by the squared correlation coefficient and mean absolute percentage error [38], $a_{r,t}$ is the acceleration of a vehicle at time step t on road r. Finally, the carbon emissions related to fuel consumption at time step t on road r can be expressed as

$$C^{rd}_{r,t} = R^g \sum_{i=1}^{d^{rd}_{r,t}} F^{rd}_{i,r,t}, \ \forall r \in R, \forall t \in T, \tag{2.7}$$

where R^g is carbon emissions per kilogram fuel consumption.

Overall, compared to microscopic CO2 emission models such as the comprehensive modal emission model (CMEM) and vehicle specific power (VSP) based model, the introduced vehicle dynamic based carbon emission model is more applicable to the studied eco-routing problem [38]. It is because the input variables are average speed, average acceleration, and angle of inclination, which are easy to measure from current transportation information systems.

2.2. V2G scheduling in power network

There are many CSs in the power network, as shown in Fig. 1, that are located at both home and office areas for EVs to charge and discharge their batteries. Specifically, the charging/discharging characteristics of EV i in the power network can be formulated as a set of constraints (2.8)–(2.11). For each EV i, the charging/discharging power at time step t can be bounded by

$$0 \le P^c_{i,t} \le u^{ev}_{i,t} A_{i,t} \overline{P}^{ev}_i, \ \forall i \in I, \forall t \in T, \tag{2.8}$$

$$(u^{ev}_{i,t} - 1) A_{i,t} \overline{P}^{ev}_i \le P^d_{i,t} \le 0, \ \forall i \in I, \forall t \in T, \tag{2.9}$$

in which the binary $u^{ev}_{i,t} \in \{0,1\}$ corresponds to the charging ($u^{ev}_{i,t} = 1$) and discharging ($u^{ev}_{i,t} = 0$) characteristic of EV i. The other binary $A_{i,t} \in \{0,1\}$ indicates whether EV i is connected to the grid ($A_{i,t} = 1$) or not ($A_{i,t} = 0$), led by its routing behaviors $\sum_{r \in R} u^{rd}_{i,r,t} = 0$ for $A_{i,t} = 1$ and vice versa. The state-of-charge (SoC) $S^{ev}_{i,t}$ of EV i is limited by

$$\underline{S}_i \le S^{ev}_{i,t} \le \overline{S}_i, \forall i \in I, \forall t \in T, \tag{2.10}$$

where $\underline{S}_i$ and $\overline{S}_i$ are the upper and lower bounds of battery SoC, respectively. Additionally, the battery SoC dynamics during a daily scheduling cycle can be expressed as

$$S^{ev}_{i,t+1} = \begin{cases} S^{ev}_{i,t} + \frac{(P^c_{i,t}\eta^c_i + P^d_{i,t}/\eta^d_i)\Delta t}{\overline{E}_i} & \text{if } A_{i,t} = 1 \\ S^{ev}_{i,t} - \frac{E^{rd}_{i,r,t}}{\overline{E}_i} & \text{if } A_{i,t} = 0 \end{cases} \tag{2.11}$$

where η^c_i, η^d_i are battery charging and discharging efficiencies, while $\overline{E}_i$ is the battery energy capacity. $E^{rd}_{i,r,t}$ corresponds to the energy consumption of EV i at time step t on road r:

$$E^{rd}_{i,r,t} = \eta^w M_i L_r + \eta^e \frac{L_r}{V^{rd}_{r,t}} + \eta^s L_r (V^{rd}_{r,t})^2, \forall i \in I, \forall r \in R, \forall t \in T, \tag{2.12}$$

which is formulated as a function of road average speed $V^{rd}_{r,t}$, vehicle weight M_i, and road distance L_r, while the coefficients η^w, η^e, and η^s correspond to the constants of weight, engine, and speed, respectively [39]. Furthermore, to ensure sufficient energy upon departure, an inequality constraint is imposed as

$$S^{ev}_{i,t^{dep}} \overline{E}_i \ge \sum_{r \in R} \sum_{t=t^{dep}_i}^{t^{arr}_i} u^{rd}_{i,r,t} E^{rd}_{i,r,t}, \ \forall i \in I. \tag{2.13}$$

Finally, EVs can also contribute to the carbon emission reduction in the power network through providing the carbon intensity service offered by the National Grid. More specifically, the carbon intensity of electricity is a measure to calculate CO_2 emissions produced per kWh of electricity, which can be timely forecast and estimated by the National Grid's Carbon Intensity API [40]. This carbon intensity forecast includes CO_2 emissions related to electricity generation only, where the emissions are from all large metered power stations, interconnector imports, transmission and distribution losses, and also account for the national electricity demand, embedded wind, and solar generation [40]. In Great Britain (GB), the carbon intensity λ^c_t at time step t can be expressed as

$$\lambda^c_t = \frac{\sum_{g \in G} P^{pn}_{g,t} \lambda^{pn}_g}{D^{pn}_t}, \forall t \in T, \tag{2.14}$$

where λ^{pn}_g, $P^{pn}_{g,t}$, and D^{pn}_t correspond to the carbon intensity of fuel type g, the generation of fuel type g, and the national demand at time step t, respectively. As a result, the National Grid can publish the carbon intensity λ^c_t in real-time that activates local EVs to optimize their power scheduling behaviors (i.e., $P^c_{i,t}, P^d_{i,t}$) to minimize the CO_2 emissions.

2.3. Power network model

The power network operation is fully modeled by a branch flow algorithm [41], which can be solved by the distribution network operator (DNO) for each time step t.

$$\min_{\Xi^{grd}} \sum_{g \in GD} \lambda^g_t P^{gd}_{g,t} + \sum_{g \in DG} \lambda^{dg}_g P^{dg}_{g,t}, \tag{2.15}$$

where

$$\Xi^{grd} = \{P^{gd}_{g,t}, Q^{gd}_{g,t}, P^{dg}_{g,t}, Q^{dg}_{g,t}, P^{res}_{g,t}, P_{bp,t}, Q_{bp,t}, \nu_{b,t}\}, \tag{2.16}$$

subject to

$$\sum_{g \in B_{gd}} P^{gd}_{g,t} + \sum_{g \in B_{dg}} P^{dg}_{g,t} + \sum_{g \in B_{res}} P^{res}_{g,t} - \sum_{k \in B_{cs}} P^{cs}_{k,t} = \sum_{d \in B_{ed}} P^{ed}_{d,t} - \sum_{(p,b) \in L} P_{pb,t} + \sum_{(b,p) \in L} P_{bp,t}, \forall b \in B, \tag{2.17}$$

$$\sum_{g \in B_{gd}} Q^{gd}_{g,t} + \sum_{g \in B_{dg}} Q^{dg}_{g,t} = \sum_{d \in B_{ed}} Q^{ed}_{d,t} - \sum_{(p,b) \in L} Q_{pb,t} + \sum_{(b,p) \in L} Q_{bp,t}, \forall b \in B, \tag{2.18}$$

$$\underline{P}^{dg}_g \le P^{dg}_{g,t} \le \overline{P}^{dg}_g, \forall g \in DG, \tag{2.19}$$

$$\underline{Q}^{dg}_g \le Q^{dg}_{g,t} \le \overline{Q}^{dg}_g, \forall g \in DG, \tag{2.20}$$

$$\underline{P}^{gd}_g \le P^{gd}_{g,t} \le \overline{P}^{gd}_g, \forall g \in GD, \tag{2.21}$$

$$\underline{Q}^{gd}_g \le Q^{gd}_{g,t} \le \overline{Q}^{gd}_g, \forall g \in GD, \tag{2.22}$$

$$P^{res}_{g,t} \le \tilde{P}^{res}_{g,t}, \forall g \in RES, \tag{2.23}$$

$$\underline{\nu} \le \nu_{b,t} \le \overline{\nu}, \forall b \in B, \tag{2.24}$$

$$P^2_{bp,t} + Q^2_{bp,t} \le l_{bp,t} \nu_{b,t}, \forall (b,p) \in L, \tag{2.25}$$

$$\nu_{b,t} - \nu_{p,t} = 2 \cdot (r_{bp} P_{bp,t} + x_{bp} Q_{bp,t}) + l_{bp}(r^2_{bp} + x^2_{bp}), \forall (b,p) \in L, \tag{2.26}$$

$$P^{cs}_{k,t} = \sum_{i \in K_{ev}} (P^c_{i,t} + P^d_{i,t}) \le \overline{P}^{cs}_k, \forall k \in CS, \tag{2.27}$$

where the objective (2.15) is to minimize the total operation costs, including (1) the procurement cost from main grid at electricity price

λ_t^g and (2) the generation cost of diesel generators (DGs) λ_g^{dg}. The decision variables are collected in set Ξ^{grd}. The power flow constraints correspond to the active and reactive power balances (2.17)–(2.18) at bus b, where B_{gd}, B_{ed}, B_{dg}, B_{pv}, and B_{cs} correspond to the sets of main grid g, static load d, DG g, renewable energy source (RES) g, and CS k located at bus b, respectively. The output limits of active and reactive power of DG g and main grid GD g as well as the active power limit of RES g are constrained in (2.19)–(2.20), (2.21)–(2.22) and (2.23), respectively [42]. The nodal voltage and power flow limits are constrained in (2.24) and (2.25), respectively, while the power flow constraints are expressed in (2.26) capturing the relationship between voltage profiles and power flows. Eq. (2.27) expresses the charging/discharging power of CS k, which is the aggregated power of all EVs plugged into the CS k, limited by its power capacity $\overline{P}_k^{cs}$.

2.4. Problem challenges

Solving the above coupled eco-routing (Section 2.1) and V2G scheduling (Section 2.2) optimization problem faces several challenges: (1) it may raise privacy issues to acquire explicit system models and technical parameters for the construction of optimization models; (2) it is difficult to approximate system uncertainties (e.g., traffic volumes, price and carbon signals), the decision-making process thus cannot exactly capture the stochastic and dynamic characteristics; (3) it is time-consuming to solve such a complex optimization problem including both transport and power networks; (4) it does not generalize to the system dynamics, since the optimal decisions need to be re-optimized for any new state condition.

To this end, we employ a data-driven and model-free MARL method for EVs operating in a decentralized fashion with privacy protection. Additionally, the uncertainty features can be learned during the training process via deep learning techniques. Finally, once the MARL method is properly trained, the learned policy can be directly implemented into the real-world decision-making process in milliseconds.

3. Partially observable Markov game

Since each EV is operated as a sequential decision-making process in a decentralized manner and can only observe partial information, the studied multi-EVs joint eco-routing and V2G scheduling problem within a transport-power network can be modeled as a Partially Observable Markov Game (POMG) with discrete time steps, where a group of agents (EVs) interact with the partially observable environment (transport-power network) with imperfect information. In other words, each EV is regarded as an agent who can manage its routing and scheduling decisions, the transport network (Section 2.1) and the power network (Section 2.3) are the environment. In this context, EVs can observe the information of the transport-power environment and then make real-time routing and scheduling decisions without performing an optimization. Specifically, at each time step, all EV agents simultaneously make an action (e.g., routing or scheduling) and then receive a reward (e.g., revenues from carbon service provision) and an observation (e.g., electricity price, carbon intensity signals, traffic volumes, etc.). The objective for each agent is to maximize its cumulative rewards it receives over the day. The interaction process between EV agents and the transport-power environment is illustrated in Fig. 2. In general, the POMG can be defined by a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$, where the detailed components are specified as below.

3.1. State and observation

The environment state $s_t \in \mathcal{S}$ describes the configurations $\{o_{1,t}, \dots, o_{I,t}\}$ of all agents I at time step t and the environment's partial information (e.g., the mathematical models and technical parameters of the coupled transport-power network), where the local observation $o_{i,t}$ of each EV agent i at time step t can be specified as

$$o_{i,t} = [R_{i,t}^{rd}, N_{i,t}^{rd}, \vec{D}_{i,t}^{rd}, S_{i,t}^{ev}, \lambda_t^g, \lambda_t^c], \forall i \in I, \tag{3.1}$$

which are divided into two parts: (1) the transport information of the road index $R_{i,t}^{rd}$, the terminated node index $N_{i,t}^{rd}$, the road traffic volumes $\vec{D}_{i,t}^{rd}$ the EV i is potentially moving to; and (2) the electricity information of EV's battery SoC $S_{i,t}^{ev}$, grid electricity price λ_t^g, and carbon intensity signal λ_t^c. Note that the behaviors of gasoline vehicles in the transport network follow a preset daily pattern [15] and are represented as part of the traffic volume $\vec{D}_{i,t}^{rd}$ in the transport environment.

3.2. Action

The action $a_{i,t}$ of each EV agent i can be specified as

$$a_{i,t} = [a_{i,t}^{tsp}, a_{i,t}^{grd}], \forall i \in I, \tag{3.2}$$

which are also divided into two parts: (1) the discrete action $a_{i,t}^{tsp} \in \{0, 1, 2, \dots, N\}$ represents the potential directions of EV routing behaviors (e.g., straight, left, right, etc.) at node $N_{i,t}^{rd}$ in the transport network topology; and (2) the continuous action $a_{i,t}^{grd} \in [-1, 1]$ represents the magnitude of discharging (negative) and charging (positive) power of EV agent i as a percentage of its battery power capacity $[-\overline{P}_i^{ev}, \overline{P}_i^{ev}]$. It is notable that EV agent i cannot make routing action $a_{i,t}^{tsp}$ and scheduling action $a_{i,t}^{grd}$ simultaneously.

3.3. State transition

In the POMG, $\Delta t = 15$ min represents one time step. For each EV agent i at time step t, an action $a_{i,t}$ is computed using policy $\pi(a_{i,t}|o_{i,t})$ conditioned on its current local observation $o_{i,t}$. The environment then transits into the next state in accordance with the state transition function: $s_{t+1} = \mathcal{T}(s_t, a_{1:I,t}, \omega_t)$, which is influenced by the environment current state s_t, all agents' actions $a_{1:I,t}$, and the environment stochasticity ω_t. In this problem, $\omega_t = [\vec{D}_{i,t}^{rd}, \lambda_t^g, \lambda_t^c, P_{d,t}^{ed}, \tilde{P}_{g,t}^{res}]$ corresponds to the exogenous state features that are independent of agent actions and have intrinsic variability. Overall, there are five types of uncertainties considered in the studied EV routing and scheduling problem, which include electricity price signals, carbon intensity signals, electric demand and PV generation in the power network as well as traffic volumes in the transport network. Instead of using scenario-based stochastic programming to address such uncertainties in the conventional optimization methods, RL can manage these uncertainties by employing a data-driven method that does not depend on precise probability distributions for uncertainties but instead learns state features from the data set itself [26,43]. To better prove the effectiveness of RL in handling environment uncertainties, a test dataset (separate from training dataset) is normally used to evaluate the performance of the trained RL policy on generalization to different state conditions. Furthermore, once the RL policy is well trained, it can be directly deployed to the practical test process in milliseconds.

On the other hand, the state transition of endogenous state features $R_{i,t}^{rd}$, $N_{i,t}^{rd}$, and $S_{i,t}^{ev}$ are determined by action $a_{i,t}$ executed at time step t. In particular, the global positioning system (GPS) can automatically identify the map information $R_{i,t}^{rd}$ and $N_{i,t}^{rd}$ when EV agent i makes routing action $a_{i,t}^{tsp}$ in transport network. Furthermore, after EV agent i is parked to a CS and makes scheduling action $a_{i,t}^{grd}$, the charging and discharging power can be described as

$$P_{i,t}^c = [\min(a_{i,t}^p \overline{P}_i^{ev}, \frac{(\overline{S}_i - S_{i,t}^{ev})\overline{E}_i}{\eta_i^c \Delta t})]^+, \forall i \in I, \tag{3.3}$$

$$P_{i,t}^d = [\max(a_{i,t}^p \overline{P}_i^{ev}, \frac{(\underline{S}_{i,t}^{ev} - S_{i,t}^{ev})\overline{E}_i \eta_i^d}{\Delta t})]^-, \forall i \in I, \tag{3.4}$$

where operators $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$. Based on (3.3)–(3.4), we have the state transition $S_{i,t}^{ev}$ expressed as (2.11).

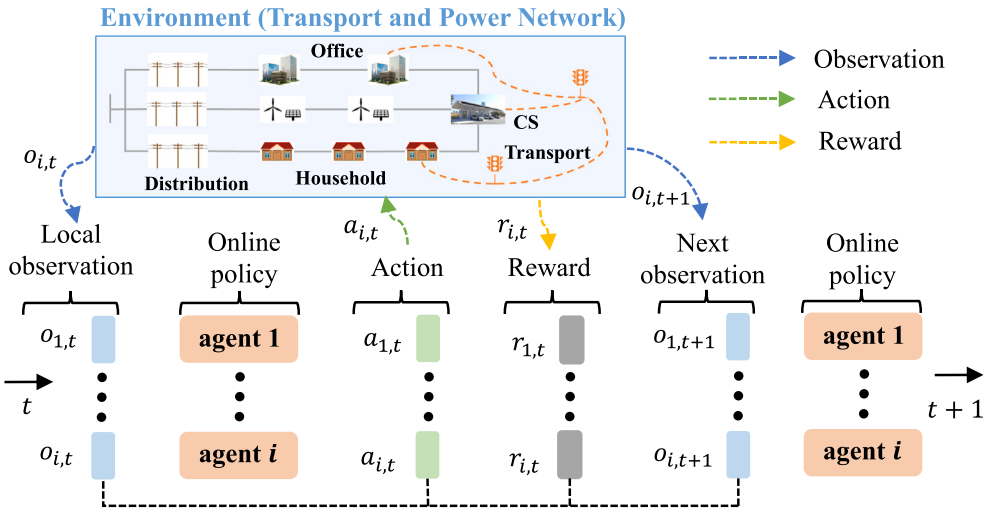


Fig. 2. The interaction process between EV agents and the transport-power environment.

3.4. Reward function

Each EV agent i receives its reward $r_{i,t}$ at the end of time step t. In transport network, the primary objective of EV i is to reduce road carbon emission cost given the carbon price c^e:

$$r_{i,t}^{co2} = -\sum_{r\in R} c^e C_{r,t}^{rd},\ \forall i \in I, \tag{3.5}$$

while also ensuring the completion of the journey at the required time. However, MARL is an unconstrained optimization that does not ensure the travel at the specified time, i.e., constraint (2.5) may not be satisfied. To properly account for this time-coupling constraint of EV daily journey, a penalty term $r_{i,t}^{trl}$ is introduced to penalize the extent of constraint violation with a penalty factor κ_1:

$$r_{i,t}^{trl} = -\kappa_1[\sum_{r\in R}\sum_{t=t_i^{dep}}^{t_i^{arr}} u_{i,r,t}^{rd} T_{r,t}^{rd} - T_i^{trl}]^+,\ \text{if } t = t_i^{arr}, \forall i \in I. \tag{3.6}$$

In the power network, the primary objective of EV agent i is to maximize carbon intensity service provision but minimize energy charging cost via V2G technique:

$$r_{i,t}^{v2g} = -\lambda_t^c c^e P_{i,t}^d - \lambda_t^g P_{i,t}^c, \forall i \in I, \tag{3.7}$$

while also ensuring the sufficient energy for each journey.

Similarly, since constraint (2.13) cannot be directly handled via the action domain, we introduce a penalty term $r_{i,t}^{soc}$ to avoid constraint violations with a penalty factor κ_2:

$$r_{i,t}^{soc} = \kappa_2[S_{i,t}^{ev}\overline{E}_i - \sum_{r\in R} u_{i,r,t}^{rd} E_{i,r,t}^{rd}]^-,\ \text{if } t \in \{t_i^{dep}, t_i^{arr}\}, \forall i \in I. \tag{3.8}$$

To further ensure the aggregate charging power within the CS power limit (2.27), we introduce another penalty for the CS overload that can be designed according to each EV's contribution to the aggregate charging power [44]:

$$r_{i,t}^{cs} = \begin{cases} -\kappa_3 \frac{|P_{i,t}^c + P_{i,t}^d|}{P_{k,t}^{cs}}(P_{k,t}^{cs} - \overline{P}_k^{cs})^2 & \text{if } P_{k,t}^{cs} \geq \overline{P}_k^{cs} \\ 0 & \text{else,} \end{cases} \forall i \in I. \tag{3.9}$$

Thus, the overall reward function can be designed as

$$r_{i,t} = \begin{cases} r_{i,t}^{co2} + r_{i,t}^{trl} & \text{if } A_{i,t} = 1 \\ r_{i,t}^{v2g} + r_{i,t}^{soc} + r_{i,t}^{cs} & \text{if } A_{i,t} = 0, \end{cases} \forall i \in I, \tag{3.10}$$

which depends on whether the EV agent i is in the traveling ($A_{i,t} = 1$) or connected to the grid ($A_{i,t} = 0$).

3.5. Objective

The above process continues for each time step until the episode (e.g., trading day) ends. In the POMG, each EV agent i seeks an optimal policy $\pi(a_{i,t}|o_{i,t})$ that maximizes its cumulative discounted reward

$$R_i = \sum_{t=0}^{T} \gamma^t r_{i,t},\ \forall i \in I, \tag{3.11}$$

where $\gamma \in [0, 1)$ is the discount factor that denotes the relative importance of future and immediate rewards. T = 24 h (96 time steps in 15 min resolution) is the daily horizon.

4. Multi-agent reinforcement learning

To solve the above POMG, we propose a novel MARL method called H2PSPPO with its general architecture being shown in Fig. 3, which can be divided into three parts: (1) constructing a hierarchical architecture with a two-level framework [45] to choose between transport network routing and power network scheduling, as these two decisions are mutually exclusive; (2) developing a hybrid policy [46] that can perform both discrete routing and continuous scheduling actions, because a single policy cannot output both discrete and continuous actions at the same time; (3) updating the control policy using the PPO algorithm in a PS framework [47] that exhibits stable and accelerated learning performance.

4.1. Two-level hierarchies

As discussed in Section 1.2, the routing action (discrete) and scheduling action (continuous) of EVs are mutually exclusive and in different domains. Therefore, it can be ineffective to directly apply MARL methods (e.g., [32–35]) to the studied problem, since either routing or scheduling action will be meaningless during the RL training process. In this section, a two-level hierarchical architecture is introduced to formulate the routing and scheduling characteristics of EVs as two separate RL policies for more effective learning performance.

Hierarchical reinforcement learning (HRL) refers to a type of RL method that can deal with several sub-policies working together in a hierarchical architecture. The two-level framework [45], one of the most common HRL techniques, is proposed as a temporal abstraction for RL actions, where the high-level (HL) action takes place over several time steps via the low-level (LL) actions. Specifically, for any EV agent i observing $o_{i,t}$ at time step t, based on the HL policy $\mu(x|o)$, an HL action is selected to make either transport network routing or power network

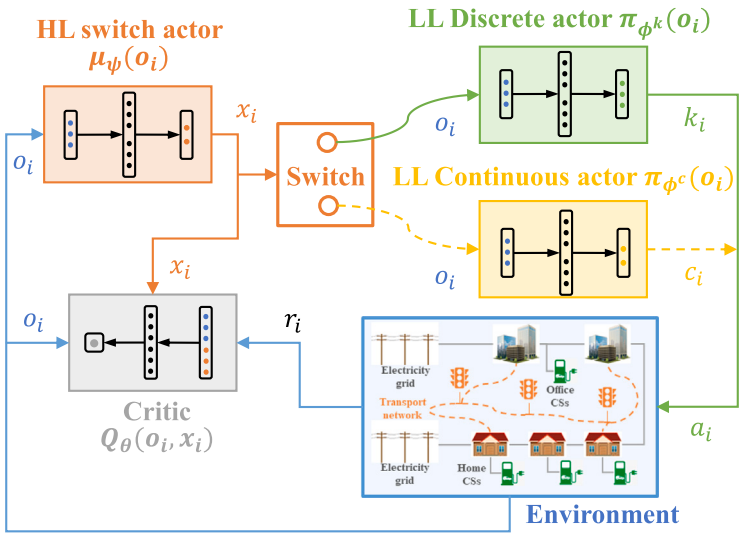


Fig. 3. Architecture of the proposed H2PSPPO method.

scheduling, i.e., $x_{i,t} = \mu_\psi(x|o) \to [0,1] \in \mathcal{X}_i$. Afterwards, one of the LL policies $\pi(a|o)$ is utilized to compute the LL action $a_{i,t}$. This process continues until the HL action switches to the other LL policy when probability $x_{i,t}$ crosses the 50% threshold. Similar to the vanilla RL, the reward over the two-level framework is given as $r_{i,t}$ in (3.10). Then, the objective of agent i over the proposed two-level framework within f time steps can be expressed as $R_i(o_{i,t}, x_{i,t}, o_{i,t+f}) = \mathbb{E}[\sum_{z=t}^{t+f} \gamma^{z-t} r_{i,z}]$. For each agent n, this process continues for T time steps, emitting a new trajectory of local observations, HL actions, LL actions, and rewards: $\tau_i = o_{i,1}, x_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, \ldots, r_{i,T}$ over $\mathcal{O}_i \times \mathcal{X}_i \times \mathcal{A}_i \times \mathcal{R}_i \to \mathbb{R}$.

Additionally, it is worth noting that the introduced penalty terms in (3.6), (3.8), and (3.9) in Section 3.4 will not be used in the same RL policy, since the proposed two-level hierarchical architecture allows the incorporation of different LL policies (the discrete LL policy for routing and the continuous LL policy for scheduling in the studied problem). More specifically, the penalty term $r_{i,t}^{trl}$ described in (3.6) corresponds to the transport network, which will be used to train the discrete LL policy for effective EV routing behaviors, while the penalty terms $r_{i,t}^{soc}$ and $r_{i,t}^{cs}$ described in (3.8) and (3.9) correspond to the power network, which will be used to train the continuous LL policy for effective EV scheduling behaviors.

4.2. Hybrid policy via PPO

In order to characterize the high-dimensional and continuous observation and action spaces of the HL and LL policies discussed above, an actor-critic architecture [26] is introduced for the hierarchical architecture. The actor module contains the HL policy $\mu(x|o)$ and the LL policy $\pi(a|o)$, while the critic module contains a state-value function $V(o,x)$ that specifies the expected value of selecting an HL selection $x_{i,t}$ in observation $o_{i,t}$. However, considering that the routing and scheduling actions of EV agents are in discrete and continuous spaces respectively, a hybrid LL policy $a_{i,t} = \{k_{i,t}, c_{i,t}\} \in \mathcal{A}_i$ with two actor branches (networks) [46] is developed to separately compute the discrete routing (when traveling in the transport network) and the continuous scheduling actions (when connected to the power network) for each EV agent i:

$$k_{i,t} = a_{i,t}^{tsp} \quad \text{and} \quad c_{i,t} = a_{i,t}^{grd}. \tag{4.1}$$

To model such action characteristics and inspired by the PPO algorithm [26], we generate (1) a probability distribution for the discrete actor network parameterized by ϕ_i^k to output the corresponding probabilities for all potential routing selections, this categorical policy $k_{i,t} = \pi_{\phi_i^k}(k|o)$ is then sampled for the optimal action $k_{i,t}$ in observation $o_{i,t}$; and (2) a Gaussian distribution for the continuous actor network parameterized by ϕ_i^c to output the corresponding mean and variance for scheduling behaviors, the stochastic policy $c_{i,t} = \pi_{\phi_i^c}(c|o)$ is then sampled for the optimal action $c_{i,t}$ in observation $o_{i,t}$. The discrete policy $\pi_{\phi_i^k}$ and continuous policy $\pi_{\phi_i^c}$ are then updated independently, which minimizes their own clipped surrogate objective to restrict the policy update:

$$L_{i,t}^{\text{CLIP}}(\phi_i^k) = \hat{\mathbb{E}}_t\left[\min(\zeta_{i,t}^k \hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^k, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\right], \tag{4.2}$$

$$L_{i,t}^{\text{CLIP}}(\phi_i^c) = \hat{\mathbb{E}}_t\left[\min(\zeta_{i,t}^c \hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^c, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\right], \tag{4.3}$$

where the approximated policy gradient and its pruning by clipping the probability ratio between $[1-\epsilon, 1+\epsilon]$ are represented by the first and second terms, respectively. As an on-policy method, PPO updates the policy network based on the transitions generated by the current policy network. The critic network can make a more accurate value prediction for the current policy network. However, the on-policy methods may suffer from poor sampling efficiency, since the prior transitions cannot be utilized frequently to update the policy network. To address this issue, the hyperparameter $\epsilon \in [0,1]$ is used to truncate the gradient update of the new policy π_ϕ from the old version π_ϕ^{old}, where the importance sampling technique [48] is used to improve the sample efficiency of PPO. The main idea is to sample the training data from a proposal distribution to approximate the expectation on average. In this case, the new policy π_ϕ can be evaluated with samples collected from the old policy π_ϕ^{old}.

In the hybrid policy, the probability ratio $\zeta_{i,t}^k$ takes into account the discrete policy, while the probability ratio $\zeta_{i,t}^c$ takes into account the continuous policy. Specifically,

$$\zeta_{i,t}^k = \frac{\pi_{\phi_i^k}(k_{i,t}|o_{i,t})}{\pi_{\phi_i^k}^{old}(k_{i,t}|o_{i,t})} \quad \text{and} \quad \zeta_{i,t}^c = \frac{\pi_{\phi_i^c}(c_{i,t}|o_{i,t})}{\pi_{\phi_i^c}^{old}(c_{i,t}|o_{i,t})}. \tag{4.4}$$

where the LL discrete policy $\pi_{\phi_i^k}(k|o)$ can be used to optimize the HL policy $\mu_{\psi_i}(x|o)$, which is characterized by discrete domain. Similarly, the probability ratio of HL policy $\zeta_{i,t}^x$ can be derived similarly as the LL discrete one $\zeta_{i,t}^k$ in (4.4).

The generalized advantage function $\hat{A}_{i,t}$ can be written as

$$\hat{A}_{i,t} = \delta_{i,t} + (\gamma\lambda)\delta_{i,t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{i,T-1}, \tag{4.5}$$

$$\delta_{i,t} = r_{i,t} + \gamma V_{\theta_i}(o_{i,t+1}, x_{i,t+1}) - V_{\theta_i}(o_{i,t}, x_{i,t}), \tag{4.6}$$

where $V_{\theta_i}(o,x)$ is the state-value function taking agent's local observations o_i and HL actions x_i into training [26], which is approximated by a critic network parameterized by θ_i, while $\gamma \in [0,1]$ and $\lambda \in [0,1]$.

4.3. Training process

As the proposed POMG comprises a group of EV agents with identical local observation, action, and reward function, the training performance of MARL policies can be improved by employing a PS framework. [47]. Particularly, PS enables all agents to share the utilized policy's parameters, which can be learned using the experiences accumulated by all agents. However, PS still allows for behavioral variation in the agents, because each agent receives different local observations. Specifically, H2PSPPO runs for all agents by their shared HL and LL policies $\mu_\psi(x|o), \pi_{\phi^k}(k|o), \pi_{\phi^c}(c|o)$ for T time steps, and collects the trajectories τ_i from the interactions with the environment. After a batch of trajectories are gathered from the buffer $\mathcal{J} = \{\tau_i\} \sim \mathcal{F}$, the EV agents then can utilize them to calculate the discounted reward-to-go $\hat{R}_{\iota,t} = \sum_{h=t}^{T} \gamma^{h-t} r_{\iota,h}$ and the advantage function $\hat{A}_{\iota,t}$ for each trajectory ι and time step t. Then, the three actor networks are trained by maximizing their objectives as follows:

$$\mathcal{L}(\psi) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^x_{\iota,t}\hat{A}_{\iota,t}, \text{clip}(\zeta^x_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{4.7}$$

$$\mathcal{L}(\phi^k) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^k_{\iota,t}\hat{A}_{\iota,t}, \text{clip}(\zeta^k_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{4.8}$$

$$\mathcal{L}(\phi^c) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^c_{\iota,t}\hat{A}_{\iota,t}, \text{clip}(\zeta^c_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{4.9}$$

where J indicates the training batch size. The critic network is trained by minimizing the loss function:

$$\mathcal{L}(\theta) = \frac{1}{J}\sum_{\iota=1}^{J} \left(\hat{R}_{\iota,t} - V_\theta(o_{\iota,t}, x_{\iota,t})\right)^2. \tag{4.10}$$

Given the above optimizations, the network weights of three actors and one critic can be updated as

$$\begin{aligned} \psi &\leftarrow \psi + \alpha^\psi \nabla_\psi \mathcal{L}(\psi), \\ \phi^k &\leftarrow \phi^k + \alpha^{\phi^k} \nabla_{\phi^k} \mathcal{L}(\phi^k), \\ \phi^c &\leftarrow \phi^c + \alpha^{\phi^c} \nabla_{\phi^c} \mathcal{L}(\phi^c), \\ \theta &\leftarrow \theta - \alpha^\theta \nabla_\theta \mathcal{L}(\theta), \end{aligned} \tag{4.11}$$

where $\alpha^\psi, \alpha^{\phi^k}, \alpha^{\phi^d}, \alpha^\theta$ indicate the learning rates of the gradient ascent/descent algorithm for actor/critic networks.

The pseudo-code of H2PSPPO is shown in Algorithm 1:

Algorithm 1 H2PSPPO for I agents

Initialize weights $\psi, \phi^k, \phi^c, \theta$ for actor and one critic networks
Set learning rates $\alpha^\psi, \alpha^{\phi^k}, \alpha^{\phi^c}, \alpha^\theta$
for episode (i.e., day) $epi = 1$ to E **do**
  Initialize the global state s_0 and local observation $o_{i,0}$
  Set an empty data buffer $\mathcal{F} = \{\}$
  For each agent i, sets an empty trajectory $\tau_i = []$
  For each agent i, selects HL action $x_{i,0}$ in observing $o_{i,0}$
  for time step (i.e., 15 min) $t = 1$ to T **do**
    repeat
      for agent (i.e., EV) $i = 1$ to I **do**
        if EV i is in transport network **then**
          Samples LL action $a_{i,t} = a^{tsp}_{i,t} = \pi_{\phi^k}(a|o)$
        else if EV i is in power network **then**
          Samples LL action $a_{i,t} = a^{grd}_{i,t} = \pi_{\phi^c}(a|o)$
        end if
      end for
      Execute all agents' actions $a_{1:I,t}$ to the environment
      DNO solves AC OPF (2.15)–(2.27) if HL action is for power network
      for agent (i.e., EV) $i = 1$ to I **do**
        Observes reward $r_{i,t}$ and next local observation $o_{i,t+1}$, stores one sample experience to trajectory τ_i += $[o_{i,t}, x_{i,t}, a_{i,t}, r_{i,t}]$
        while time step t % $J = 0$ **do**
          Agent i collects a set of trajectories τ_ι from buffer $\mathcal{F}$, then computes advantage function $\hat{A}_{\iota,t}$ and discounted reward-to-go $\hat{R}_{\iota,t}$
          Updates network weights $\psi, \phi^k, \phi^c, \theta$ in (4.11)
        end while
      end for
      Update state $s_t \leftarrow s_{t+1}$ and observation $o_{i,t} \leftarrow o_{i,t+1}$
    until HL action $x_{i,t} = \mu_\psi(x|o)$ is switched
    Update HL action $x_{i,t} \leftarrow x_{i,t+1}$
  end for
end for

5. Input data and experiment setup

5.1. Network topology and EV models

Case studies are conducted on a 7-node 10-edge transport network and a modified IEEE 15-bus power network with 10 EVs and 2 CSs, representing home and office areas, respectively. Detailed network structures are illustrated in Fig. 4, while the technical parameters of EVs are presented in Table 1 [49]. It can be observed from Fig. 4 that 1 CS is located at road node N_2 (green) for EV drivers parking at home, and 1 CS is located at road node N_4 (orange) for EV drivers parking at work places.

The transport network data are presented in Table 2, expressing the transport time without congestion and the distance between two adjacent nodes. To capture the influence of charging limits of EVs, each CS has a charging power capacity of 40 kW. The EV routing and scheduling decisions are modeled for time resolutions of 15-min. Note that the choice of 15 min can be adjusted based on the scenario to be analyzed and computational needs.

Table 1
Technical parameters of utilized EVs.

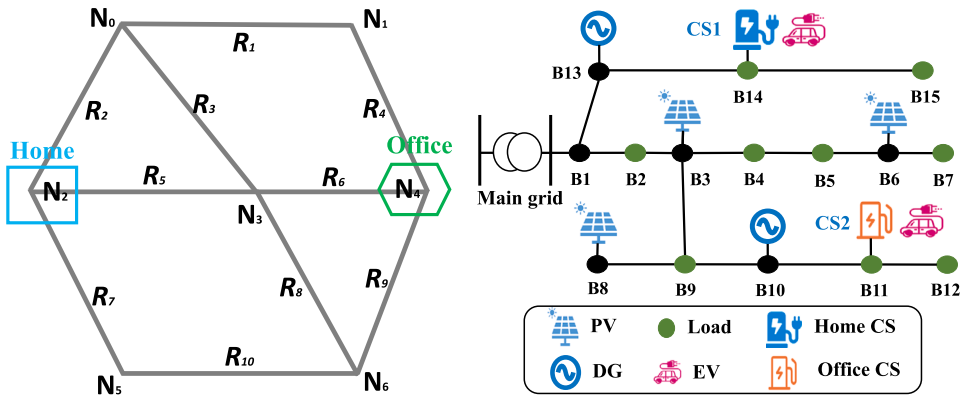
Parameters	$\overline{P}$ (kW)	$\overline{E}$ (kWh)	$\underline{S}, \overline{S}$ (%)	η^c_i, η^d_i (%)
Values	16.5	100	0,100	90,90

Table 2
Road data of 7-node 10-edge transport network.

Road	From node	To node	Travel time (h)	Distance (km)
0	0	1	0.20	19.4
1	0	2	0.18	17.1
2	0	3	0.19	18.6
3	1	4	0.22	19.8
4	2	3	0.13	11.6
5	3	4	0.13	10.5
6	2	5	0.19	17.6
7	3	6	0.24	17.6
8	4	6	0.15	15.0
9	5	6	0.16	16.4

5.2. Data descriptions

A real-world dataset is used to capture the uncertainties related to traffic volumes in the transport network [50]. To capture system uncertainties in the power network, we collect the yearly grid electricity prices from Nord-Pool group [51] and carbon intensity signals from the national grid [40], where their mean and standard deviation are illustrated in Fig. 5. The yearly residential load and PV generation data are collected from the real-world open-source dataset recorded by the Ausgrid [52]. In the experiment, we split the one-year price and carbon intensity data in the training (Jan.-Nov.) and the test (Dec.) sets for MARL methods. The carbon price is fixed at 300 £/tCO2 [53] over the year.



(a) The 7-node 10-edge transport network (b) The modified 15-bus power network

Fig. 4. The 7-node 10-edge transport and 15-bus power network.

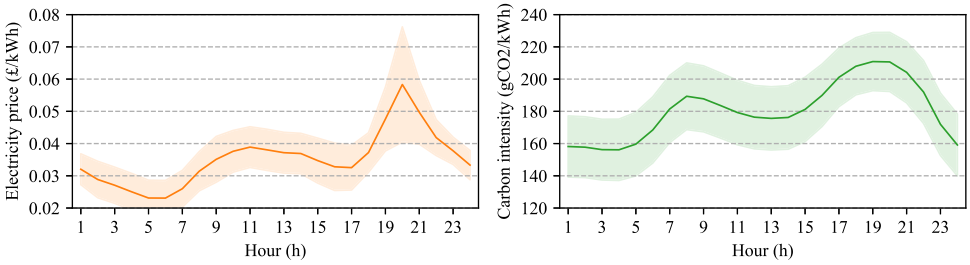


Fig. 5. Electricity price and carbon intensity signals.

5.3. Benchmarks

We compare the proposed H2PSPPO with two benchmark MARL methods: (1) **PPO**: each agent adopts a vanilla PPO algorithm, the actor network simultaneously computes routing and scheduling actions per time step, and no PS framework is integrated while each agent uses individual experiences to train its own policy; (2) **H2PPO**: based on PPO, each agent learns a hierarchical architecture and a hybrid policy for both routing and scheduling actions, but PS framework is not adopted. We run 5000 episodes of each MARL method with 10 random seeds for the initialization of weights and environment. Additionally, we compare our proposed model-free H2PSPPO method to a model-based optimization method **Central-Opt**: following the similar assumptions in [11,12,15], it is assumed that the coupled power-transport network can acquire all the information of EVs and the perfect knowledge of the system uncertainties.

5.4. RL network structure and hyperparameters

The proposed H2PSPPO contains three actor networks and one critic network. It is noted that all these four networks have 1 hidden layer with 64 units using the RELU as the activation function. `Tanh` and `Softplus` activation functions are used as the continuous actor outputs to construct the Gaussian policy with mean and std, respectively. `Softmax` activation function is used in the discrete actor output to construct the categorical policy with the corresponding probability for each discrete action dimension. The HL actor network inputs the local observation and outputs the probabilities of selecting making routing and scheduling decisions. The LL discrete actor network inputs the local observation and outputs the probabilities of selecting potential routing directions (e.g., straight, left, and right) in transport network. The LL continuous actor network inputs the local observation and outputs the magnitude of behaving discharging or charging power in the power network. For the critic network, it inputs the local observation and HL selection, linear is used without activate function for the output layer representing the state value.

We use Adam optimizer for all four networks with a learning rate $\alpha^{\psi}, \alpha^{\phi^k}, \alpha^{\phi^d} = 10^{-4}$ and $\alpha^{\theta} = 10^{-3}$, respectively. The batch size $N = 96$ refers to the number of environment steps (24 h). We employ a clip rate $\epsilon = 0.2$ and discount rates $\gamma = 0.99, \lambda = 0.98$.

Finally, the proposed H2PSPPO method has been implemented in Python with Tensorflow v2.6.0 [54], and the linearized AC-OPF algorithm (Section 2.3) has been implemented in Pyomo with Gurobi solver [55]. Case studies have been carried out on a computer with a 6-core 3.50 GHz Intel(R) Xeon(R) E5-1650 processor and 32 GB of RAM.

6. Case studies

6.1. Performance evaluation

This section assesses the training and test performance of the three examined MARL methods. Fig. 6 depicts the evolution of episodic total reward of 10 EVs over 5000 training episodes, where the solid lines and the shaded areas respectively depict the moving average over 100 episodes and the oscillations of 10 seeds. Furthermore, their corresponding averaged episodic training time as well as the averaged number of episodes and averaged total training time required to reach convergence are also collected in Table 3. Finally, we also plot the cumulative total reward of 10 EVs over the test month in Fig. 7 by deploying the well-trained three MARL policies and the optimization-based central-opt method.

The first observation we notice from Fig. 6 is that PPO (blue) has the most unstable and oscillatory training performance, resulting in the lowest reward level and failing to reach optimum. This is because the EV agents in PPO make both routing and scheduling actions at each time step, while one of them is meaningless (e.g., routing action does not help EV agents behave charging in the power network). In this case, H2PPO (orange) can effectively deal with this issue by constructing a

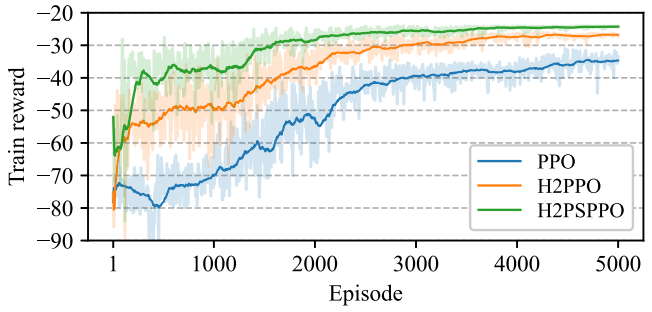


Fig. 6. Episodic training reward of 10 EVs for 3 MARL methods.

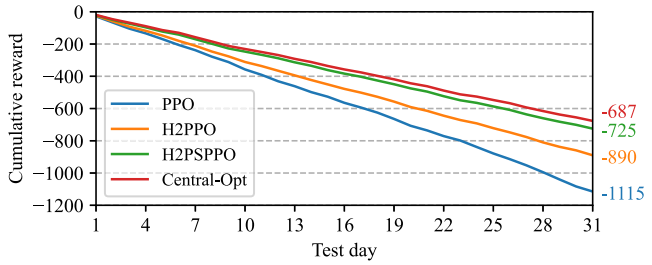


Fig. 7. Cumulative test reward of 10 EVs for 3 MARL and 1 optimization-based methods.

Table 3
Computational performance for 3 MARL methods.

Method	PPO	H2PPO	H2PSPPO
Episodic training time (s.)	1.55	1.98	1.86
Number of episodes (#)	5,000	4,000	3,000
Total training time (hr.)	2.15	2.20	1.55

hierarchical architecture and a hybrid policy, and consequently achieving a higher reward level. In addition, this performance can be further improved in terms of both higher reward level and faster training speed in our proposed H2PSPPO (green) via a PS framework. Similar to the training performance, the proposed H2PSPPO in the test performance (Fig. 7) outperforms PPO and H2PPO by 34.98% and 18.56% the cumulative reward over the test month (31 days), respectively. Note that the proposed H2PSPPO achieves the near optimal solution (5.24% gap) obtained from the optimization-based Central-Opt method over the test month.

We further assess the computational performance of three MARL methods during the training process. Table 3 shows that PPO has the shortest episodic training time of 1.55 s (since it only requires training one actor network to compute both routing and scheduling actions, eliminating the need for hierarchical architecture), followed by H2PPO and H2PSPPO (since they need to train three actor networks rather than the single actor network in PPO). Additionally, H2PSPPO (around 3000 episodes) demonstrates a faster convergence rate than H2PPO (around 4000 episodes) to reach convergence. This is because the PS technique can utilize all agents' experiences to update the single and shared policy. Finally, our proposed H2PSPPO (1.55 hr) costs the lowest computational time to reach convergence among three MARL methods.

6.2. Analysis of joint routing and scheduling behaviors

After evaluating the MARL training and test performance, this section aims to analyze the routing and scheduling behaviors of 10 EVs in the transport-power network, as illustrated in Fig. 8 (two typical journeys of H2 W and W2H) and Fig. 9 (charging and discharging power rate), respectively.

Table 4
Comparison of transport result and electricity cost for 3 routing and scheduling strategies.

	Travel distance (km)	Travel time (min)	Transport emissions (kg/100 km)	Electricity cost (£)
S1	474.14	403.55	**18.23**	**−26.47**
S2	**221.03**	461.26	23.38	41.24
S3	345.61	**343.73**	21.95	41.82

Considering that the relationship between vehicle speed and carbon emissions follows a polynomial curve [38], EVs in Fig. 8(a)–(b) make eco-routing behaviors with the objective of maintaining a reasonable vehicle speed range (e.g., 50–80 km/h) for the transport network, i.e., driving too slowly or too fast will cause high carbon emissions. On one hand, most EVs avoid the shortest route ($N_2-N_3-N_4$), since serious congestion can happen and cause large carbon emissions if these EVs choose this route. On the other hand, many EVs select roads $N_0 - N_2$, $N_2 - N_5$, and $N_4 - N_6$ as part of their commuting routes, since the corresponding traffic volumes are low. Increasing traffic volumes on these roads can effectively reduce the vehicle speed (e.g., around 60 km/h) and then lead to fewer carbon emissions.

EVs in Fig. 9 exhibit similar scheduling characteristics during the daily cycle. Different EVs can have different charging/discharging power at different time steps, which is influenced by many factors, e.g., the initial SoC, electricity prices, carbon intensity signals, road power consumption, RL policies, etc. In detail, they choose to charge power in the early morning (e.g., 5:00–7:00) and mid-day (e.g., 13:00–14:00) to minimize charging costs. This is because of the relatively low grid electricity prices and demand level in the early morning and the high PV generation level in the mid-day. On the other hand, EVs discharge power for carbon service provision in the late afternoon and evening (e.g., 17:00–22:00) in response to the high carbon intensity level and peak demand level. Furthermore, it can be found from Fig. 9 that the total charging and discharging power has reached the CS capacity at certain time steps, e.g., mid-day and evening, respectively.

6.3. Benefit of eco-routing and V2G scheduling

To investigate the benefits of eco-routing and V2G scheduling in reducing carbon emissions, a detailed comparison with respect to the proposed strategy (S1) is conducted, including two extra strategies: (S2) the routing decisions are made to solely minimize the traveling distance; (S3) the routing decisions are made to solely minimize the traveling time. In the power network, EV agents in both S2 and S3 make charging decisions until the battery capacity is full but do not behave discharge decisions for V2G services.

We summarize and compare the travel distance, travel time, transport emissions, and electricity cost (including carbon service provision revenue) of 10 EVs for three strategies in Table 4. It is expected that EVs under S2 and S3 receive the shortest traveling distance (221.03 km) and the least traveling time (343.73 min), respectively, while EVs under S1 contribute to the minimum transport emissions (18.23 kg/100 km) due to the eco-routing strategy. In the power network, EVs have a similar charging cost under S2 and S3, but are capable of making revenue (negative cost) under S1 when exhibiting V2G flexibility to provide carbon intensity service to the grid.

6.4. Application to a real-world transport network

To further analyze the scalability of the proposed MARL method, this section implements H2PSPPO to a real-world transport network that contains the area of central London in the UK, where four home areas (green) and three office areas (orange) are considered, and 300 EVs are simulated to capture the realistic traffic conditions. Regarding the electricity grid, a modified IEEE 33-bus power network including 7

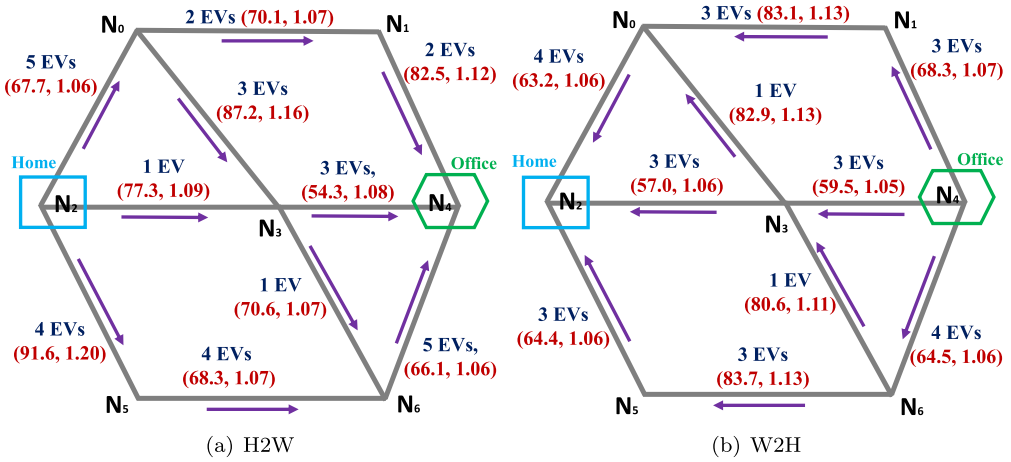


Fig. 8. EV eco-routing process in H2 W and W2H. The two red values in the bracket indicate average vehicle speed and carbon emissions per km.

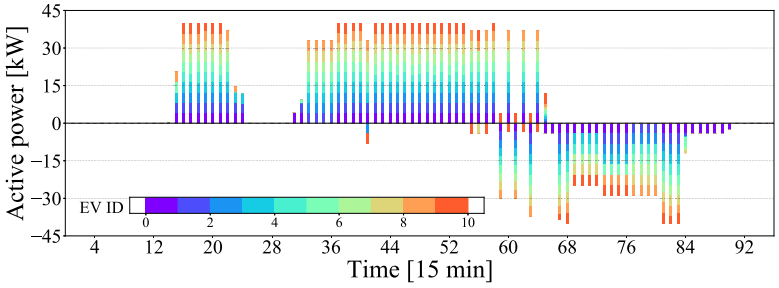


Fig. 9. Charging and discharging power of 10 EVs in the 15-bus power network. Each EV is represented by one color, while the height of each bar refers to its charging (positive) and discharging (negative) power.

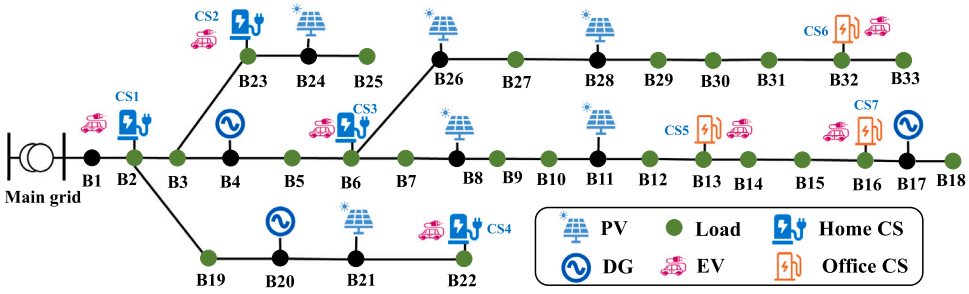


Fig. 10. The modified IEEE 33-bus power network.

CSs (4 home CSs and 3 office CSs) is utilized to simulate the EV charging and discharging behaviors, as illustrated in Fig. 10. Specifically, the charging stations installed at home and office areas have capacities of 300 kW and 400 kW, respectively. Fig. 11 depicts the evolution of episodic total reward of 300 EVs over 25,000 training episodes, where the solid lines and the shaded areas respectively depict the moving average over 100 episodes and the oscillations of 10 seeds. It can be observed that the proposed H2PSPPO can learn a steady and increasing reward for all 300 EV agents and gradually reaches convergence. The computing time over 25,000 episodes is around 20.83 h.

Similar to the virtual case, we assume that these utilized EVs have two typical journeys (H2W and W2H) per day, where their routing behaviors are presented in Figs. 12 and 13, respectively. Finally, EVs' aggregated charging and discharging behaviors are shown in Fig. 14 and their business cases for charging behaviors and carbon intensity service provision are recorded in Table 5. First of all, it can be observed from Figs. 12 and 13 that there are mainly four routes for 300 EVs to commute between home and office areas, of which the middle two exhibit the relatively shorter distance (i.e., the route passing

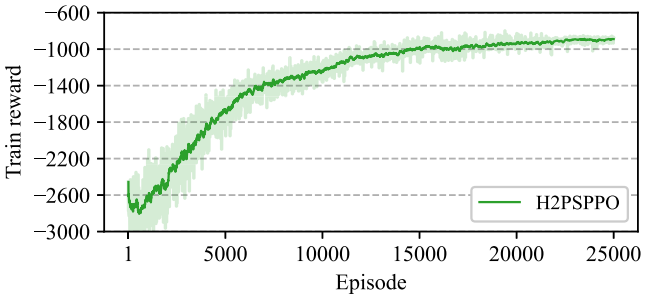


Fig. 11. Episodic training reward of 300 EVs for H2PSPPO method.

Shepherd's Bush/Marble Arch and the route passing Earls Court/Hyde Park). However, according to Figs. 12 and 13, most EVs prefer the top and bottom ones for daily commutes rather than the middle two, which have a shorter distance. This is because there are too many

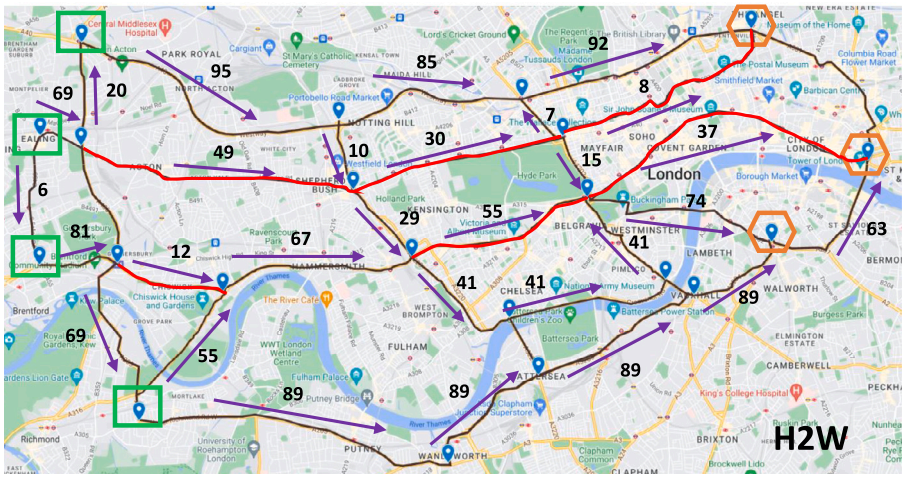


Fig. 12. Routing behaviors of 300 EVs for journey H2 W in central London.

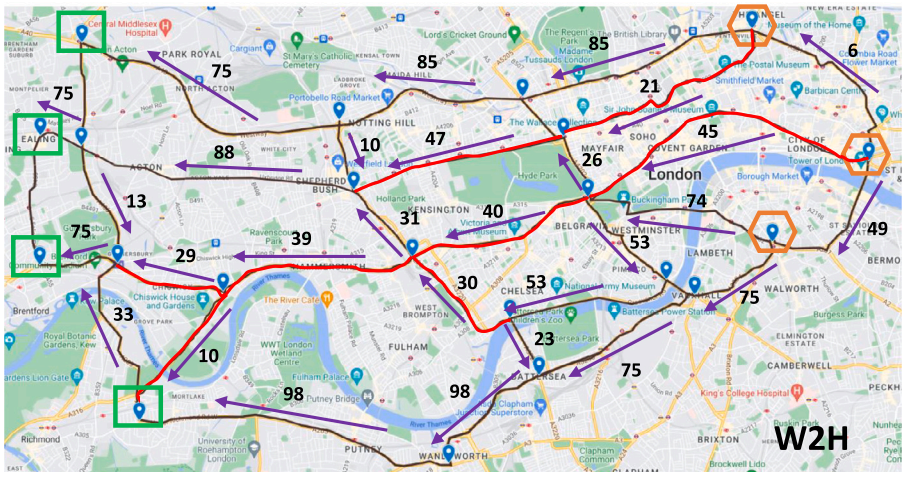


Fig. 13. Routing behaviors of 300 EVs for journey W2H in central London.

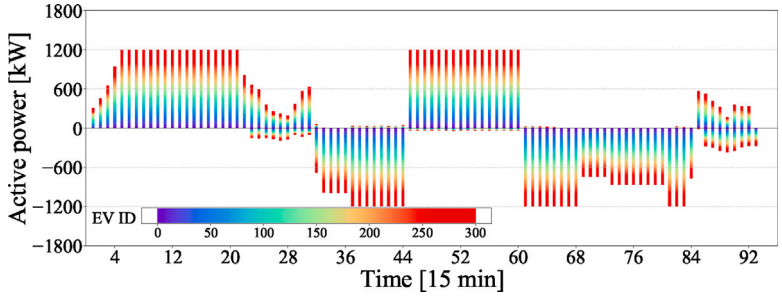


Fig. 14. Total power charging and discharging of 300 EVs in the 33-bus power network, where each EV is represented by one color in the color bar.

vehicles on the middle two, causing very serious congestion and much longer travel time as well as large carbon emissions (roads under serious congestion are indicated as the red ones in Figs. 12 and 13). Regarding the scheduling behaviors in the power network, Fig. 14 shows that these 300 EVs still behave with their charging power in the early morning and mid-day due to the low electricity prices and demand levels as well as the high PV penetration, respectively, while discharging power in the late afternoon and night in response to the high carbon intensity signals and peak demand levels. Additionally, EV power charging and discharging have reached the CS capacity in the mid-day and evening, reflecting the requirement for secure power system operations. Table 5 shows that EVs can obtain a certain amount of revenue (energy arbitrage) through charging from the grid at low electricity prices and discharging for carbon service provision at high carbon intensity signals.

To summarize, the scalability of the proposed H2PSPPO has been verified by the implementation of the coupled transport-power network including a real-world central London transport network and the modified IEEE 33-bus power network with 300 EVs, compared with existing literature [33–35] that deploy 24 EVs, 16 EVs, and 200 EVs, respectively. It is because the PS framework enables all agents to learn a shared policy rather than personalizing policies for each agent, enhancing training scalability and stability.

Table 5
Business cases of 300 EVs for charging cost and carbon service provision.

Number of EVs	Charging service cost (£)	Carbon service revenue (£)	Net revenue (£)
300	785.70	1792.45	1006.75

7. Conclusions

This paper proposes a novel MARL method for the multi-EVs joint eco-routing and V2G scheduling problem, aiming to simultaneously decarbonize the coupled transport and power networks through improved eco-routing behavior and the grid carbon intensity service. Specifically, this problem is first formulated as a POMG, while EV agents operate in a decentralized manner that does not require any prior knowledge of the joint environment (i.e., transport and power networks). To solve this problem, the proposed MARL method learns a hierarchical architecture for the mutually exclusive discrete routing and continuous scheduling behaviors via a hybrid policy. To enhance the scalability and speed up the training process, a PS framework is introduced for all EVs to train a shared control policy. Extensive case studies are performed in different settings, including a virtual 7-node 10-edge transport network and a real-world central London transport network. The results prove the effectiveness of the proposed method in addressing realistic EVs routing and scheduling problems, contributing to the low-carbon transport-power network.

CRediT authorship contribution statement

Yi Wang: Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Dawei Qiu:** Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Yinglong He:** Methodology, Writing – original draft, Writing – review & editing. **Quan Zhou:** Writing – original draft, Writing – review & editing. **Goran Strbac:** Conceptualization, Project administration, Methodology, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgment

This work was supported by the UK EPSRC project: 'Integrated Development of Low-Carbon Energy Systems (IDLES): A Whole-System Paradigm for Creating a National Strategy' (project code: EP/R0455 18/1) and the Horizon Europe project 'Reliability, Resilience and Defense technology for the griD' (Grant agreement ID: 101075714).

References

[1] Dowling P. The impact of climate change on the European energy system. Energy Policy 2013;60:406–17.
[2] Carmichael R. Behaviour change, public engagement and Net Zero, a report for the Committee on Climate Change. Centre for Energy Policy and Technology (ICEPT) and Centre for Environmental …; 2019.
[3] Hossain MS, Fang YR, Ma T, Huang C, Dai H. The role of electric vehicles in decarbonizing India's road passenger toward carbon neutrality and clean air: A state-level analysis. Energy 2023;273:127218.
[4] Heymann F, Milojevic T, Covatariu A, Verma P. Digitalization in decarbonizing electricity systems–Phenomena, regional aspects, stakeholders, use cases, challenges and policy options. Energy 2023;262:125521.
[5] Schwanen T. Achieving just transitions to low-carbon urban mobility. Nat Energy 2021;6(7):685–7.
[6] Colmenar-Santos A, Muñoz-Gómez A-M, Rosales-Asensio E, López-Rey Á. Electric vehicle charging strategy to support renewable energy sources in Europe 2050 low-carbon scenario. Energy 2019;183:61–74.
[7] Zhang S, Chen M, Zhang W, Zhuang X. Fuzzy optimization model for electric vehicle routing problem with time windows and recharging stations. Expert Syst Appl 2020;145:113123.
[8] Hulagu S, Celikoglu HB. An electric vehicle routing problem with intermediate nodes for shuttle fleets. IEEE Trans Intell Transp Syst 2020.
[9] Daryabari MK, Keypour R, Golmohamadi H. Stochastic energy management of responsive plug-in electric vehicles characterizing parking lot aggregators. Appl Energy 2020;279:115751.
[10] Dixon J, Bukhsh W, Edmunds C, Bell K. Scheduling electric vehicle charging to minimise carbon emissions and wind curtailment. Renew Energy 2020;161:1072–91.
[11] Yao C, Chen S, Yang Z. Joint routing and charging problem of multiple electric vehicles: A fast optimization algorithm. IEEE Trans Intell Transp Syst 2022;23(7):8184–93.
[12] Liu P, Wang C, Hu J, Fu T, Cheng N, Zhang N, et al. Joint route selection and charging discharging scheduling of EVs in V2G energy network. IEEE Trans Veh Technol 2020;69(10):10630–41.
[13] Chen T, Zhang B, Pourbabak H, Kavousi-Fard A, Su W. Optimal routing and charging of an electric vehicle fleet for high-efficiency dynamic transit systems. IEEE Trans Smart Grid 2018;9(4):3563–72.
[14] Tang X, Bi S, Zhang Y-JA. Distributed routing and charging scheduling optimization for internet of electric vehicles. IEEE Internet Things J 2019;6(1):136–48.
[15] Sun Y, Chen Z, Li Z, Tian W, Shahidehpour M. EV charging schedule in coupled constrained networks of transportation and power system. IEEE Trans Smart Grid 2019;10(5):4706–16.
[16] Lv S, Wei Z, Sun G, Chen S, Zang H. Optimal power and semi-dynamic traffic flow in urban electrified transportation networks. IEEE Trans Smart Grid 2020;11(3):1854–65.
[17] González S, Feijoo F, Basso F, Subramanian V, Sankaranarayanan S, Das TK. Routing and charging facility location for EVs under nodal pricing of electricity: A bilevel model solved using special ordered set. IEEE Trans Smart Grid 2022;13(4):3059–68.
[18] Wei W, Wu L, Wang J, Mei S. Network equilibrium of coupled transportation and power distribution systems. IEEE Trans Smart Grid 2018;9(6):6764–79.
[19] Xie S, Xu Y, Zheng X. On dynamic network equilibrium of a coupled power and transportation network. IEEE Trans Smart Grid 2022;13(2):1398–411.
[20] Cui Y, Hu Z, Duan X. Optimal pricing of public electric vehicle charging stations considering operations of coupled transportation and power systems. IEEE Trans Smart Grid 2021;12(4):3278–88.
[21] Lv S, Chen S, Wei Z, Zhang H. Power–Transportation coordination: Toward a hybrid economic-emission dispatch model. IEEE Trans Power Syst 2022;37(5):3969–81.
[22] Lv S, Chen S, Wei Z. Coordinating urban power-traffic networks: A subsidy-based Nash–Stackelberg–Nash game model. IEEE Trans Ind Inf 2022;19(2):1778–90.
[23] Wu T, Wei X, Zhang X, Wang G, Qiu J, Xia S. Carbon-oriented expansion planning of integrated electricity-natural gas systems with EV fast-charging stations. IEEE Trans Transp Electrif 2022;8(2):2797–809.
[24] Wu T, Li Z, Wang G, Zhang X, Qiu J. Low-carbon charging facilities planning for electric vehicles based on a novel travel route choice model. IEEE Trans Intell Transp Syst 2023.
[25] Ruan G, Kirschen DS, Zhong H, Xia Q, Kang C. Estimating demand flexibility using siamese LSTM neural networks. IEEE Trans Power Syst 2021.
[26] Qiu D, Wang Y, Hua W, Strbac G. Reinforcement learning for electric vehicle applications in power systems: A critical review. Renew Sust Energy Rev 2023;173:113052.
[27] 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 2023.
[28] Lin B, Ghaddar B, Nathwani J. Deep reinforcement learning for the electric vehicle routing problem with time windows. IEEE Trans Intell Transp Syst 2022;23(8):11528–38.
[29] Ren L, Fan X, Cui J, Shen Z, Lv Y, Xiong G. A multi-agent reinforcement learning method with route recorders for vehicle routing in supply chain management. IEEE Trans Intell Transp Syst 2022.
[30] Wan Z, Li H, He H, Prokhorov D. Model-free real-time EV charging scheduling based on deep reinforcement learning. IEEE Trans Smart Grid 2018;10(5):5246–57.
[31] Jiang C, Jing Z, Cui X, Ji T, Wu Q. Multiple agents and reinforcement learning for modelling charging loads of electric taxis. Appl Energy 2018;222:158–68.
[32] Qian T, Shao C, Wang X, Shahidehpour M. Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system. IEEE Trans Smart Grid 2019;11(2):1714–23.

[33] Zhang C, Liu Y, Wu F, Tang B, Fan W. Effective charging planning based on deep reinforcement learning for electric vehicles. IEEE Trans Intell Transp Syst 2020;22(1):542–54.
[34] Alqahtani M, Hu M. Dynamic energy scheduling and routing of multiple electric vehicles using deep reinforcement learning. Energy 2022;244:122626.
[35] Qiu D, Wang Y, Zhang T, Sun M, Strbac G. Hybrid multiagent reinforcement learning for electric vehicle resilience control towards a low-carbon transition. IEEE Trans Ind Inf 2022;18(11):8258–69.
[36] Lowe R, Wu YI, Tamar A, Harb J, Abbeel OP, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proc. 31st int. conf. neural inf. process. syst.. 2017, p. 6379–90.
[37] Yuanqing W, Wei Z, Lianen L. Theory and application study of the road traffic impedance function. J Highw Transp Res Dev 2004;21(9):82–5.
[38] Zeng W, Miwa T, Morikawa T. Prediction of vehicle CO2 emission and its application to eco-routing navigation. Transp Res Part C Emerg Technol 2016;68:194–214.
[39] Franceschetti A, Honhon D, Van Woensel T, Bektaş T, Laporte G. The time-dependent pollution-routing problem. Transp Res B Meth 2013;56:265–93.
[40] National Grid. Carbon intensity API - national data. 2021, URL https://carbonintensity.org.uk/.
[41] Wang Y, Qiu D, Strbac G, Gao Z. Coordinated electric vehicle active and reactive power control for active distribution networks. IEEE Trans Ind Inf 2023;19(2):1611–22.
[42] Wang Y, Rousis AO, Strbac G. A three-level planning model for optimal sizing of networked microgrids considering a trade-off between resilience and cost. IEEE Trans Power Syst 2021;36(6):5657–69.
[43] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT Press; 2018.
[44] Yan L, Chen X, Chen Y, Wen J. A cooperative charging control strategy for electric vehicles based on multiagent deep reinforcement learning. IEEE Trans Ind Inf 2022;18(12):8765–75.
[45] Frans K, Ho J, Chen X, Abbeel P, Schulman J. Meta learning shared hierarchies. 2017, arXiv preprint arXiv:1710.09767.
[46] Fan Z, Su R, Zhang W, Yu Y. Hybrid actor-critic reinforcement learning in parameterized action space. In: Proc. 28th int. joint conf. artif. intell.. 2019, p. 2279–85.
[47] Terry JK, Grammel N, Hari A, Santos L. Parameter sharing is surprisingly useful for multi-agent deep reinforcement learning. 2020, arXiv e-prints, arXiv–2005.
[48] Ruan G, Wang J, Zhong H, Xia Q, Kang C. Improving sample efficiency of deep learning models in electricity market. IEEE Trans Power Syst 2023;38(5):4761–73.
[49] Electric Vehicle Database. Tesla model S performance. 2019, URL https://ev-database.uk/car/1207/Tesla-Model-S-Performance.
[50] US Federal Highway Administration and Environmental Protection Agency (FHWA and EPA). National near road study. 2022, URL https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm.
[51] Nord Pool. Historical market data. 2021, URL https://www.nordpoolgroup.com/historical-market-data/.
[52] Ratnam EL, Weller SR, Kellett CM, Murray AT. Residential load and rooftop PV generation: an Australian distribution network dataset. Int J Sustain Energy 2017;36(8):787–806.
[53] Department of Energy & Climate Change. Guidance on estimating carbon values beyond 2050: an interim approach. 2021, URL https://www.gov.uk/government/publications/guidance-on-estimating-carbon-values-beyond-2050-an-interim-approach.
[54] Abadi M, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, Software available from tensorflow.org. URL https://www.tensorflow.org/.
[55] Gurobi Optimization. Gurobi optimizer reference manual. 2019, URL http://www.gurobi.com.