

Reinforcement learning for vehicle-to-grid: A review

Hongbin Xie ^{a,b,1}, Ge Song ^{a,1}, Zhuoran Shi ^a, Jingyuan Zhang ^a, Zhenjia Lin ^c, Qing Yu ^b,
Hongdi Fu ^b, Xuan Song ^{a,*}, Haoran Zhang ^{b,d,**}

^a Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China

^b School of Urban Planning and Design, Peking University, Shenzhen, 508055, China

^c Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China

^d Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, 315201, China



ARTICLE INFO

Keywords:

Vehicle-to-grid
Reinforcement learning
Electric vehicle charging
Scheduling optimization
Markov decision process

ABSTRACT

The rapid development of Vehicle-to-Grid technology has played a crucial role in peak shaving and power scheduling within the power grid. However, with the random integration of a large number of electric vehicles into the grid, the uncertainty and complexity of the system have significantly increased, posing substantial challenges to traditional algorithms. Reinforcement learning has shown great potential in addressing these high-dimensional dynamic scheduling optimization problems. However, there is currently a lack of comprehensive analysis and systematic understanding of reinforcement learning applications in Vehicle-to-Grid, which limits the further development of this technology in the Vehicle-to-Grid domain. To this end, this review systematically analyzes the application of reinforcement learning in Vehicle-to-Grid from the perspective of different stakeholders, including the power grid, aggregators, and electric vehicle users, and clarifies the effectiveness and mechanisms of reinforcement learning in addressing the uncertainty in power scheduling. Based on a comprehensive review of the development trajectory of reinforcement learning in Vehicle-to-Grid applications, this paper proposes a structured framework for method classification and application analysis. It also highlights the major challenges currently faced by reinforcement learning in the Vehicle-to-Grid domain and provides targeted directions for future research. Through this systematic review of reinforcement learning applications in Vehicle-to-Grid, the paper aims to provide relevant references for subsequent studies.

1. Introduction

As the Fourth Industrial Revolution progresses, emerging technologies can help accelerate the energy transition, ensuring fair access to sustainable energy for everyone [1]. For example, using models to predict future power consumption [2,3], utilizing modern technology to achieve a balance between economic development and environmental protection [4–6], and enhancing the security of modern energy usage and distribution [7,8]. In energy management, with the proliferation of electric vehicles (EVs) and the rapid development of smart grids, Vehicle-to-Grid (V2G) technology is increasingly demonstrating its significant potential and value as an innovative energy management model [9–11]. V2G technology allows EVs users to charge from the grid at lower prices during off-peak hours, and during periods of high electricity demand, EVs can sell surplus stored idle electricity back to the grid at a higher price [12,13]. This bidirectional flow of electricity not only achieves peak shaving and valley filling, optimizing

the operational efficiency of the power system, but also enhances the flexibility and reliability of the grid, providing potential economic benefits to drivers, and effectively promoting the development of a clean, economical, and sustainable energy system [14–16].

1.1. Problems with growing V2G uncertainty and complexity

As user acceptance of V2G technology increases and policy incentives drive its adoption, V2G is rapidly developing while also facing a series of challenges brought about by the increasing complexity of the system [17–20]. As a large number of electric vehicles are randomly integrated into the grid, it will undoubtedly introduce numerous uncertainties across various aspects of the grid. Therefore, effectively managing the grid load fluctuations caused by the discharge of electric vehicles, optimizing the planning and layout of charging stations, and controlling costs have become critical issues that need

* Corresponding authors.[1]

** Corresponding authors.[2]

E-mail addresses: songx@sustech.edu.cn (X. Song), h.zhang@pku.edu.cn (H. Zhang).

¹ These authors contributed equally to this work.

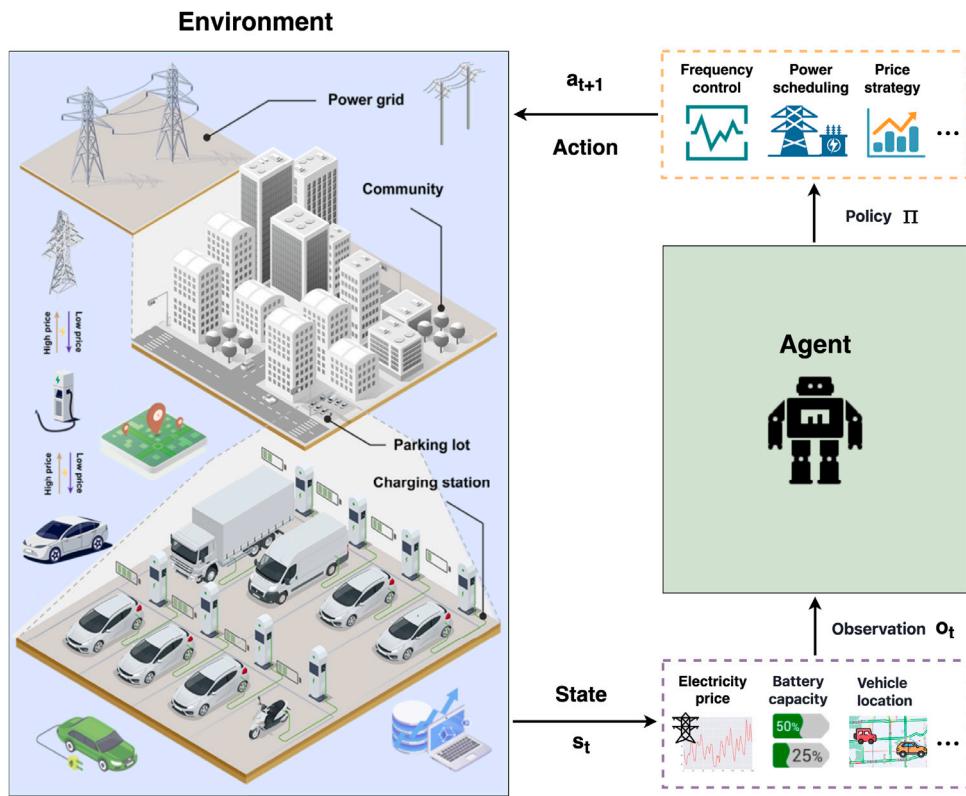


Fig. 1. Schematic of V2G with reinforcement learning.

to be addressed [21,22]. Secondly, an important task is to efficiently schedule a large number of electric vehicles for charging and discharging during peak and off-peak hours [23]. In addition, it is necessary to establish optimal charging strategies to extend the service life of batteries [24]. Lastly, ensuring the safety and stability of the entire charging network is equally crucial. The optimization of these issues is imperative, otherwise, it may severely hinder the development and application of V2G technology.

1.2. Why reinforcement learning for V2G?

Reinforcement Learning (RL), as a cutting-edge optimization decision-making technology, possesses significant advantages over traditional optimization methods. Traditional optimization techniques often rely on preset models and fixed rules, which exhibit obvious limitations when dealing with complex and dynamic environments. They struggle to adapt flexibly to environmental changes and unforeseen events. In contrast, RL mimics the human learning process by continuously interacting with the environment through agents. It improves decision-making strategies iteratively through trial-and-error and feedback loops, enabling adaptation to ever-changing environments [25–27].

The application of RL in V2G systems is particularly noteworthy. As shown in Fig. 1, RL can flexibly adjust charging and discharging strategies based on variable factors such as grid load, electricity price fluctuations, and the battery status of electric vehicles, ensuring efficient system operation [28,29]. In contrast, traditional methods require extensive presets and manual intervention when dealing with these variables, making it challenging to achieve automation and efficient optimization [30]. As the volume of data increases, it becomes imperative to incorporate other relevant data to improve the efficiency of V2G applications.

Moreover, RL continuously learns and optimizes its strategies through sustained interaction with the environment, effectively handling grid disturbances and other unexpected events [31]. This enables the system to maintain high adaptability and ensures robustness and stability when facing unknown or random events [32]. Traditional methods fall short in this regard, struggling to handle complex and volatile real-world environments flexibly.

Therefore, RL not only theoretically holds significant advantages but also demonstrates outstanding effectiveness in practical applications. It can efficiently meet the optimization and scheduling needs of V2G systems, providing strong support for the stable operation of the power grid and the energy management of electric vehicles.

1.3. Differences with existing reviews

Here we first compared this review with existing reviews on V2G technology [17,29,33–41], as shown in Table 1. Through this comparison, we found that there is currently a lack of comprehensive analysis on the application of reinforcement learning in V2G optimization and scheduling. Furthermore, in the more mature G2V field, we compared previous research reviews, which mainly focus on the exploration and application of reinforcement learning techniques in electric vehicle charging scheduling [42–47]. These reviews provide an in-depth investigation of the critical role of reinforcement learning in advancing electric vehicle charging optimization technologies and offer valuable insights for future research. However, these reviews generally do not address discharge scheduling technologies, and there is still a lack of systematic investigation into the application of reinforcement learning in the V2G field. To the best of our knowledge, our review is the first to concentrate on discussing and summarizing the application of reinforcement learning in the V2G field, clearly elucidating the effectiveness of reinforcement learning in V2G systems and its positive impact on promoting the development of V2G technology.

Table 1

Comparison of existing review literature. The size of the orange area in the circle represents the relevance of the review topic to optimization scheduling.

Year	Review	Title	Main focus	Scheduling optimization theme	RL algorithm
2020	BK Sovacool, et al. [33]	Actors, business models, and innovation activity systems for vehicle-to-grid (V2G) technology: A comprehensive review	Investigating the stakeholders and business models for promoting V2G technology in decarbonizing European passenger transport.	○	✗
2021	B Bibak, et al. [34]	A comprehensive analysis of Vehicle to Grid (V2G) systems and scholarly literature on the application of such systems	Providing a comprehensive review of the integration of electric vehicles and renewable energy in smart grids through V2G systems.	○	✗
2022	M İnci, et al. [35]	Integrating electric vehicles as virtual power plants: A comprehensive review on vehicle-to-grid (V2G) concepts, interface topologies, marketing and future prospects	Investigating the process of integrating electric vehicles into virtual power plants through V2G technology, discussing system configuration, interface topology, and market potential.	●	✗
2022	B Viswanath, et al. [36]	Vehicle-to-Grid (V2G) Optimization for Grid Peak Demand Reduction and Decarbonization: A State-of-the-Art Review	Reviewing the current research progress on V2G optimization technologies , and discussing its impact on reducing peak power demand and promoting environmental sustainability.	●	✗
2023	S Panchanathan, et al. [37]	A Comprehensive Review of the Bidirectional Converter Topologies for the Vehicle-to-Grid System	Investigating the bidirectional converter topologies and charging systems for V2G technology, aiming to facilitate active power transfer between electric vehicles and the power grid.	○	✗
2023	G Vishnu, et al. [38]	Review of Challenges and Opportunities in the Integration of Electric Vehicles to the Grid	Reviewing the benefits and challenges of EV-power grid integration through V2G technology, focusing on its potential to revolutionize both the transportation and electric power sectors.	●	✗
2023	Qiu Dawei, et al. [29]	Reinforcement learning for electric vehicle applications in power systems: A critical review	Reviewed a large number of potential power applications under power systems , extensively discussing the applications of reinforcement learning in power management .	●	✓
2024	N Uribe-Pérez, et al. [39]	Communications and Data Science for the Success of Vehicle-to-Grid Technologies: Current State and Future Trends	Comprehensively reviewing the V2G system, with a focus on the role of communication technologies and the challenges faced, to guide future research and formulate policies.	○	✗
2024	M Wan, et al. [17]	Feasibility and Challenges for Vehicle-to-Grid in Electricity Market: A Review	Discussing the business models related to V2G technology and assess their feasibility and challenges in the electricity market.	●	✗
2024	A Gonçearuc, et al. [40]	The barriers to widespread adoption of vehicle-to-grid: A comprehensive review	Analyzing barriers to the adoption of V2G technology, aiming to understand their relative importance and facilitate stakeholder efforts towards its widespread adoption.	○	✗
2024	G Chen, et al. [41]	Control Strategies, Economic Benefits, and Challenges of Vehicle-to-Grid Applications: Recent Trends Research	Exploring recent trends in research on control strategies for V2G applications, assessing their economic benefits, and discussing the challenges associated with their implementation.	●	✗
2024	Ours	Reinforcement learning for vehicle-to-grid (V2G): A review	Systematically investigating the application trajectory of reinforcement learning algorithms in the field of V2G scheduling optimization , the latest research findings, the technical challenges faced, and the future development trends, aims to point the direction for subsequent research and provide practical references.	●	✓

○ Unrelated ● Partially related ■ Related

1.4. Our contributions

The purpose of this review is to fill the gap in the literature analysis of reinforcement learning in the V2G field. As shown in Fig. 2, this review provides a comprehensive organization, classification, and functional elaboration of reinforcement learning technology, focusing on the research and practical application of reinforcement learning in various V2G optimization tasks. It summarizes the core issues and challenges currently encountered by reinforcement learning in V2G applications and proposes potential key areas for future research. Overall, the main contributions of this review can be summarized as follows:

- To the best of our knowledge, this review is the first systematic review and summary of the application of reinforcement learning in the V2G field, providing a holistic perspective for both academia and industry. It reveals the potential and application prospects of reinforcement learning technology in V2G systems.
- This review systematically analyzes the application effects of reinforcement learning methods on various important optimization tasks in V2G from multiple dimensions. It clearly identifies the effectiveness of reinforcement learning in optimizing V2G and its role in promoting the advancement of V2G technology.
- This review delves into the core challenges faced by current reinforcement learning methods in the application of the V2G field. It systematically sorts out these key issues and, based on the cutting edge of current research, prospectively puts forward a series of methodologies worth further exploration and study. These strategies aim at overcoming current technological bottlenecks and provide valuable insights and directions for the future optimization of V2G systems and the development of smart grids.

2. Basic of V2G

2.1. Overview of V2G

The emergence of V2G technology is aimed at addressing two core challenges in the energy sector: enhancing the stability of the power grid and improving the utilization efficiency of electric vehicle energy storage resources. As electric vehicles become increasingly popular, the high-capacity battery packs they are equipped with are gradually turning into potential distributed energy storage units. V2G technology is the key to unlocking this potential.

Compared to the G2V (Grid-to-Vehicle) model, V2G technology enables electric vehicles to supply power back to the grid, establishing a two-way energy exchange between electric vehicles and the power grid [68]. With V2G technology, during peak hours of grid demand, electric vehicle owners can sell the electricity stored in their idle batteries back to the grid at a higher price while still ensuring their daily mobility needs. Conversely, during off-peak hours, owners can charge their vehicles at a lower cost from the grid, thus generating economic benefits [69]. This bidirectional interaction model not only reduces the purchase and usage costs of electric vehicles but also assists the grid in achieving “peak shaving and valley filling”, providing effective support for the optimization of the global energy structure and sustainable development.

2.2. Benefits of V2G

The main benefits of V2G include the following points:

- For the power grid, V2G technology effectively helps to balance supply and demand, significantly improving the stability and

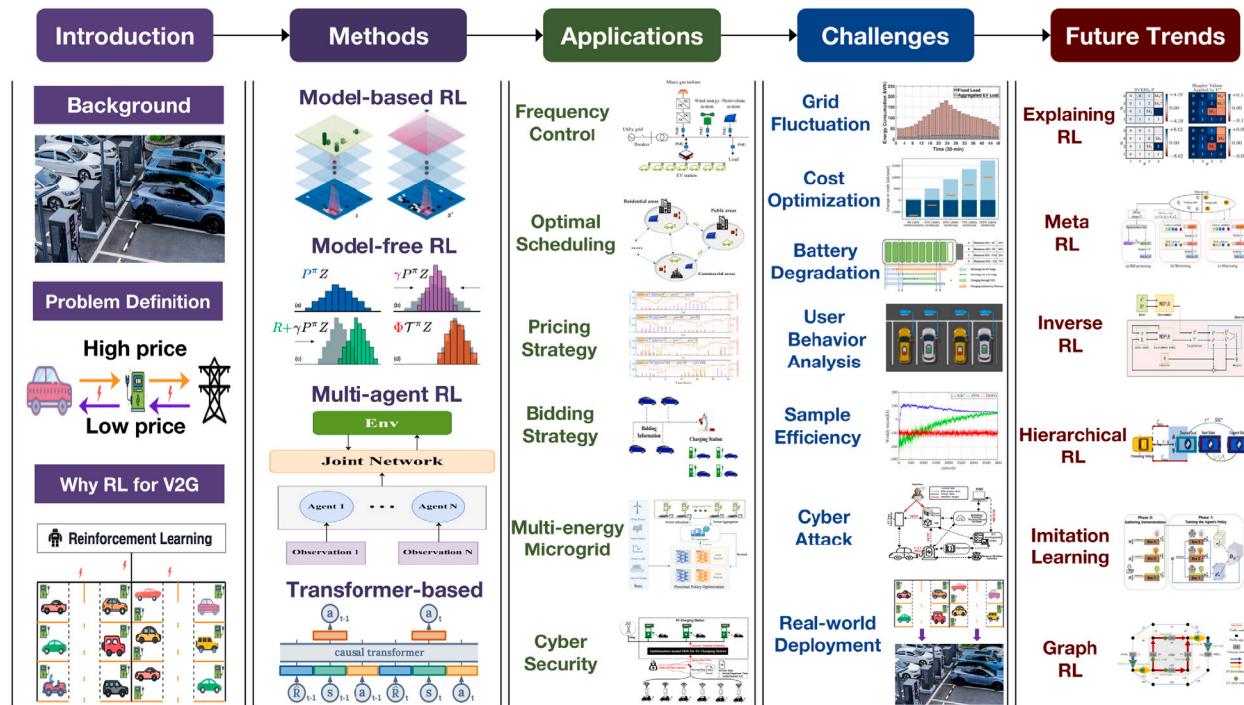


Fig. 2. Overview of the framework in this review [48–67].

flexibility of power supply, and enhancing the overall operation efficiency of the grid.

- For vehicle owners, the application of V2G technology allows them to sell electricity back to the grid during peak demand periods, thus opening up a new channel for earning additional income through electricity sales.
- For charging station aggregators, V2G technology greatly enhances the resource utilization efficiency of charging stations, which will undoubtedly vigorously promote the market development of charging infrastructure.
- For society as a whole, the promotion of V2G technology helps to facilitate the transition towards a cleaner, low-carbon energy system, reduces reliance on fossil fuels, and consequently lowers greenhouse gas emissions and environmental pollution, promoting sustainable development.

3. Reinforcement learning

In the upcoming sections, we will provide an overview of the background and classification of reinforcement learning (RL) in the context of V2G. In Section 3.1, we will introduce the fundamental concepts of RL, along with the methods and classifications that have evolved over time. In Section 3.2, we will explore the integration of RL with deep learning to tackle complex problems in high-dimensional state spaces. These methods will be categorized according to the classifications discussed in the previous section, with a focus on their development in deep reinforcement learning, recent research advancements, and the challenges they face. Section 3.3 will delve into Transformer-based reinforcement learning, highlighting its efficiency in handling complex temporal data in V2G, improving policy learning, adapting to dynamic changes, and optimizing multi-task objectives. In Section 3.4, we will examine multi-agent reinforcement learning, discussing various training and execution methods, and explaining how RL can address challenges in more complex scenarios. We will also emphasize the technological advancements and challenges that emerge as complexity increases. Finally, in Fig. 3, we will illustrate the evolution of RL algorithms in V2G over time, showing their increasing complexity and more specialized applications across various subfields.

3.1. Basic reinforcement learning

In this subsection, we will introduce some key concepts and ideas of reinforcement learning. First, we define a Markov decision process (MDP). Then, we will explain some commonly used learning methods in reinforcement learning. Finally, we will introduce some classic reinforcement learning algorithms based on common classification methods.

3.1.1. Markov decision process

MDP [98] describes a method for sequential decision-making, where actions not only affect the immediate reward but also influence subsequent states and thereby future rewards. MDP is an ideal mathematical framework for discussing reinforcement learning, allowing us to define and describe the problem conveniently [99]. An MDP is composed of a tuple of five elements: (S, A, P, r, γ) . S denotes the set of states ($s \in S$). A denotes the set of actions ($a \in A$). P denotes the transition model ($P(S_{t+1} = s' | S_t = s)$). r denotes the reward function $r(s_t, a_t)$. γ denotes the Discount factor $\gamma \in (0, 1]$. The purpose of reinforcement learning is to discover a policy π , where π denotes the policy $\pi(a|s) = P(A_t = a | S_t = s)$, such that the agent can achieve the maximum expected return starting from the initial state $\pi^* = \arg \min_{\pi} \mathbb{E}^{\pi} [\sum_{t=0}^{\infty} \gamma r(s_t, a_t)]$, where π^* denotes the optimal policy.

In V2G applications, it is essential to construct appropriate MDP models for different problems. A suitable MDP model is the foundation for effective learning in reinforcement learning.

For optimal scheduling and pricing strategy, we can define:

- State: Grid load, vehicle information, current time, etc.
- Action: Different discharging strategies, discharging price, etc.
- Reward: Recovered energy, recovered price, etc.

For energy storage scheduling and optimization of the distribution network, we can define:

- State: Vehicle flow, storage cost, or construction cost.
- Action: Whether to store energy at a specific location or build a storage station, represented by a binary table.

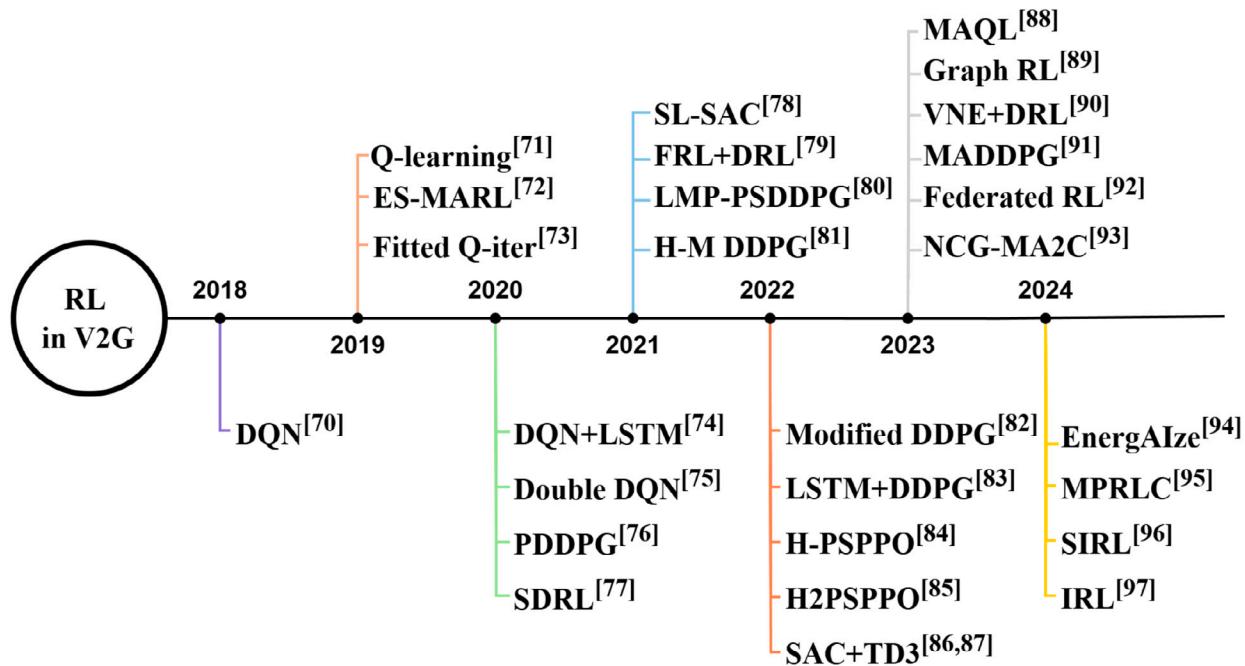


Fig. 3. Timeline of reinforcement learning in the V2G [70–97].

- Reward: Negative of the storage cost or construction cost.

By using this information, we can build a relevant simulation environment and learn policy (P) through this environment to achieve cost reduction and efficiency improvement.

Additionally, we can also construct different MDP models for various specific purposes in V2G. These models allow us to tailor the MDP models to the specific needs of different V2G applications, enhancing the effectiveness of reinforcement learning in achieving the desired outcomes. Such as multi-energy microgrid being more suitable for multi-agent MDP, optimization of distribution network being suitable for hierarchical MDP, and collaborative grid frequency control being more suitable for risk-sensitive MDP.

A good MDP model can help us better express the optimization goals and facilitate better optimization and decision-making by reinforcement learning algorithms. For V2G, the data is often very large and complex, making it even more important to choose the right data to build the MDP model.

Additionally, MDP has many variants that are used to discuss different model states under various conditions. For instance, the Partially Observable Markov Decision Process (POMDP) is used for cases where the entire state space cannot be observed. The Semi-Markov Decision Process (SMDP) is used for variable time steps. The Multi-Objective Markov Decision Process (MOMDP) is used to address different problems with multiple objectives. Among these, POMDP is often used in multi-agent reinforcement learning to express differences between agents. Based on the relationships between different agents, it has extended to frameworks like Partially Observable Multi-Agent Games (POMG), which must consider the strategies and actions of other agents. In practical applications, the appropriate MDP model should be chosen based on the conditions under obtained information. This can make the modeling results clearer and improve the effectiveness of reinforcement learning algorithms.

Considering the commonly used models in reinforcement learning, we will focus on discussing the differences, advantages, and disadvantages of MDP, POMDP, and POMG, as well as their application scenarios. When all states are known and finite, using MDP to describe the state-action transition relationship is undoubtedly the best choice,

as it better captures the characteristics of all states and generates results. However, when the state space is too large, using MDP will lead to a sharp increase in computational complexity, consuming more computational resources.

In cases where the state is not fully known, POMDP is often used to describe the state-action relationship. The agent needs to infer potential state information through current observations and make decisions. POMDP is more in line with the model state in real-world situations and also shows good representational capabilities in multi-agent reinforcement learning. However, it makes the learning process more complex and it is more difficult to reach the optimal solution.

POMG represents the process where different agents in the environment influence each other. The environment it describes is more complex, as it not only needs to infer the current potential state but also consider the actions and strategies of other agents to make decisions. POMG is suitable for describing multi-agent reinforcement learning problems where agents influence each other. In training, it needs to consider the strategy coordination between different agents and the non-stationarity exhibited as other agents' strategies change.

3.1.2. Monte Carlo

The Monte Carlo method approximates the solution to a problem through extensive random sampling, using random numbers to simulate and solve complex computational problems. In reinforcement learning, when dealing with large state-action spaces, we can use the Monte Carlo method to obtain approximate values for the state transition function and the reward function, and improve accuracy through multiple samples. The Monte Carlo method first evaluates the value function of the current policy by simulating multiple complete episodes. Through multiple samples, starting from the initial state, actions are taken according to the current policy until a terminal state is reached, recording each state, action, and reward. For each state, the cumulative return from that state to the end of the episode define as $V^\pi(s) = \mathbb{E}_\pi [G_t | S_t = s] \approx \frac{1}{N} \sum_{i=1}^N G_t^{(i)}$ which G denotes the return obtained each time state s is encountered. Finally, the average of all returns for that state is taken as the value estimate for that state. For each state, we choose the action that maximizes the expected return. The policy is updated to select the optimal action in each state. This process alternates until the policy converges.

3.1.3. Dynamic programming

Dynamic programming methods [100] involve breaking down a problem into several sub-problems, solving these sub-problems first, and then using their solutions to solve the original problem. The problems that dynamic programming can solve need to meet two conditions: first, the entire optimization problem can be decomposed into multiple sub-optimization problems; second, the solutions to the sub-optimization problems can be stored and reused. Therefore, it is used to solve model-based reinforcement learning problems [101], where the transition probabilities and rewards for each state and action are precisely known. There are two main dynamic programming-based reinforcement learning algorithms: policy iteration and value iteration. In policy iteration, we evaluate the current policy to obtain the state-value function, then improve the policy based on the state-value function $\pi^0 \rightarrow V^{\pi^0} \rightarrow \pi^1 \rightarrow V^{\pi^1} \rightarrow \dots \rightarrow \pi^*$, and continue evaluating and improving the policy until it converges to the optimal policy. In value iteration, we perform only one round of value updates during policy evaluation, and then directly improve the policy based on the updated values: $V^{k+1}(s) = \max_{a \in A} \{r(s, a) + \gamma \sum P(s'|s, a)V^k(s')\}$. Dynamic programming methods for solving reinforcement learning problems have high computational complexity and rely on environment models. They are not well applied in model-free dominant reinforcement learning. For V2G, we do not believe that a general and transparent environment model can be obtained. Therefore, although model-based reinforcement learning has successful algorithms like AlphaGo [50], it is challenging to apply them to V2G problems.

3.1.4. Temporal difference

Model-free reinforcement learning algorithms cannot know the environment's reward function and state transition function in advance. They need to learn directly from the data sampled during interactions with the environment. Model-free reinforcement learning cannot use dynamic programming algorithms for learning. Instead, it uses sampling methods and introduces temporal difference (TD) algorithms [102] to iterate and improve policies. TD methods combine the ideas of Monte Carlo and dynamic programming algorithms. When updating the value estimates of states, TD methods do not require the return G_t calculated after the entire sequence ends. Instead, they use the reward obtained at the current step plus the estimated value of the next state $V_\pi(s) = \mathbb{E}_\pi [R_t + \gamma V_\pi(S_{t+1}|S_t = s)]$.

3.2. Deep reinforcement learning

Deep reinforcement learning (DRL) combines deep learning and reinforcement learning. It uses deep learning to address the challenges of storing and computing high-dimensional states and actions in reinforcement learning. In this section, we will specifically introduce some commonly used model-free deep reinforcement learning models. Based on different decision-making methods, we generally divide them into value-based reinforcement learning methods and policy-based reinforcement learning methods. Some algorithms combine both approaches, and we will also provide a detailed introduction to these.

3.2.1. Value-based

Value-based methods make decisions by learning the value of states or state-action pairs (such as Q-values). Common algorithms like Deep Q-Networks (DQN) output the value of each possible action given a state and then select the action with the highest value. Value-based methods are typically used in environments with discrete action spaces. Value-based methods are intuitively expressive, allowing us to quickly learn a stable policy for their application in V2G. As shown in [103], these methods have already been applied to energy management, and we believe they have similar application scenarios in V2G.

DQN [104]. DQN simulates the Q-value table using a neural network, where the neural network outputs the Q-values for each action. During training, it interacts with the environment using an epsilon-greedy strategy to explore and generate experiences, which are stored in a replay buffer. During training, a batch of experiences is randomly sampled from the replay buffer for updates $Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$. DQN introduces a target network $Q_{tar}(s, a)$, whose parameters θ are periodically updated. The pseudocode is shown as algorithm 1.

Algorithm 1 Deep Q-Network (DQN) Algorithm

```

1: Initialize replay buffer  $D$  to capacity  $N$ 
2: Initialize action-value function  $Q$  with random weights  $\theta$ 
3: Initialize target action-value function  $\hat{Q}$  with weights  $\theta^- = \theta$ 
4: for each episode do
5:   Initialize state  $s_0$ 
6:   for  $t = 0, 1, 2, \dots, T$  do
7:     With probability  $\epsilon$  select a random action  $a_t$ 
8:     Otherwise select  $a_t = \arg \max_a Q(s_t, a; \theta)$ 
9:     Execute action  $a_t$  and observe reward  $r_t$  and next state  $s_{t+1}$ 
10:    Save the transition  $(s_t, a_t, r_t, s_{t+1})$  into the replay buffer  $D$ 
11:    From  $D$  randomly select a mini-batch of transitions
         $(s_j, a_j, r_j, s_{j+1})$ 
12:     $y_j = \begin{cases} r_j & \text{if done}_{j+1} \\ r_j + \gamma \max_{a'} \hat{Q}(s_{j+1}, a'; \theta^-) & \text{otherwise} \end{cases}$ 
13:    Perform a gradient descent step on  $(y_j - Q(s_j, a_j; \theta))^2$  with
        respect to the network parameters  $\theta$ 
14:    Every  $C$  steps, reset  $\hat{Q} = Q$ 
15:  end for
16: end for

```

Rainbow DQN [105]. Due to the popularity of the DQN algorithm, many different algorithms have been developed based on it. Rainbow DQN summarizes and integrates these algorithms. Therefore, by introducing Rainbow DQN, we can summarize the DQN improvement algorithms that preceded Rainbow.

- **Double Q-learning** [106]: Reduces overestimation bias in Q-learning by decoupling action selection and action evaluation.
- **Prioritized Experience** [107]: Improves data efficiency by sampling experiences based on their importance and replaying transitions with higher learning potential more frequently.
- **Dueling Network Architectures** [108]: Helps generalize between actions by separating the representation of state value and action advantage.
- **Multi-step Learning**: Alters the bias-variance trade-off by using multi-step bootstrap targets, helping to propagate newly observed rewards more quickly to earlier visited states.
- **Distributional RL** [109]: Better handles uncertainty by learning a distribution of discounted returns instead of estimating the mean.
- **Noisy Nets**: Enables effective exploration without additional information by using stochastic network layers for exploration.

C51 [49]. If the range of the distribution is from V_{min} to V_{max} , and it is evenly divided into N discrete points. Each equal subdivision is: $\{z_i = V_{min} + i\delta z : 0 \leq i < N, \delta z = \frac{V_{max}-V_{min}}{N-1}\}$. The value output by the model corresponds to the probability of taking the current support point. The goal of the Bellman equation is achieved by projecting the return distribution $r + \gamma z_j$ onto the current distribution z_i . The overall distribution is:

$$\begin{aligned}
 (\Phi^\top Z_\theta(x, a))_i &= \sum_{j=0}^{N-1} \left[1 - \frac{|[\hat{T}z_j]_{V_{MIN}} - z_i|}{\delta z} \right]^1 \\
 &\cdot p_j(x', \pi(x')) \tag{1}
 \end{aligned}$$

QR-DQN [110]. QR-DQN differs from C51 in that it allocates a fixed uniform probability to (N) adjustable positions, rather than using (N) fixed positions to approximate its probability distribution. It employs quantile regression to randomly adjust the positions of the distribution to minimize the Wasserstein distance to the target distribution.

3.2.2. Policy-based

Policy-based reinforcement learning is a method that directly learns a policy $\pi(a|s)$, rather than indirectly deriving the policy by learning state-action values as in value-based methods. In policy-based methods, the policy function directly maps states to actions. The policy can be deterministic or stochastic (a probability distribution). Policy-based methods are better suited for handling continuous action spaces. In the context of V2G with large and complex action spaces, using policy-based methods is a better choice.

Assume the target policy π is a stochastic policy, where θ are the corresponding parameters. Given an input state, the policy outputs an action or a probability distribution over actions. The optimal policy maximizes the expected return in the environment. Therefore, we define the objective function as $J(\theta) = \mathbb{E}_{s_0}[V^{\pi_\theta}(s_0)]$, s_0 represents the initial state. We take the derivative of the objective function with respect to the policy and use gradient ascent to maximize this objective function, thereby obtaining the optimal policy.

REINFORCE [111]. REINFORCE is the most fundamental policy gradient-based reinforcement learning algorithm. Its main idea aligns with the policy-based approach, directly optimizing the policy function to maximize the cumulative reward for actions chosen in given states. The pseudocode is shown as algorithm 2.

Algorithm 2 REINFORCE Algorithm

```

1: Initialize policy parameters  $\theta$ 
2: for each episode do
3:   Initialize state  $s_0$ 
4:   for  $t = 0, 1, 2, \dots, T$  do
5:     Sample action  $a_t \sim \pi_\theta(a_t|s_t)$ 
6:     Execute action  $a_t$  and observe reward  $r_t$  and next state  $s_{t+1}$ 
7:     Store  $(s_t, a_t, r_t)$ 
8:   end for
9:   for  $t = 0, 1, 2, \dots, T$  do
10:     $G_t \leftarrow \sum_{k=t}^T \gamma^{k-t} r_k$ 
11:     $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) G_t$ 
12:   end for
13: end for
```

The REINFORCE algorithm performs poorly in terms of sample efficiency. Subsequently, many improved algorithms have been proposed to enhance its sample utilization [112].

TRPO [113]. The policy gradient algorithm primarily iterates to update the policy parameters θ in the direction of $\nabla_\theta J(\theta)$. However, when the policy network is a deep model, updating the parameters along the policy gradient can often result in the policy deteriorating significantly due to overly large step sizes, thereby affecting the training performance.

Trust Region Policy Optimization (TRPO) ensures the monotonic improvement of policy performance by guaranteeing the temporal difference residual $\mathbb{E}_{S \sim V^{\pi_\theta}} \mathbb{E}_{a \sim \pi_{\theta'}(\cdot|s)} [A^{\pi^\theta}(s, a)] \geq 0$. During the computation process, it is challenging to collect samples with target policy $\pi_{\theta'}$ as the objective function. Therefore, TRPO makes an approximation by directly using the state distribution of the old policy. In this way, TRPO uses the KL divergence to ensure that the overall policy difference is not significant.

TRPO maximizes the expected return while ensuring that the KL divergence between the new policy $\pi_{\theta_{\text{new}}}$ and the old policy $\pi_{\theta_{\text{old}}}$ remains within a specified threshold δ .

$$\max_{\theta'} L_\theta(\theta') \\ \text{s.t. } \mathbb{E}_{S \sim V^{\pi_\theta}} \left[D_{KL}(\pi_{\theta_k}(\cdot|s) \| \pi_{\theta'_k}(\cdot|s)) \right] \leq \delta \quad (2)$$

PPO [114]. TRPO has been successfully applied in many scenarios, but its computation process is very complex. Proximal Policy Optimization (PPO), based on the ideas of TRPO, implements the algorithm in a much simpler way. Numerous experimental results show that PPO can learn as well as (or even faster than) TRPO. PPO has become a very popular reinforcement learning algorithm. When trying to use a reinforcement learning algorithm in a new environment, PPO is one of the algorithms that can be tried first. PPO has two forms: one is PPO-Penalty, and the other is PPO-Clipping.

PPO-Penalty [115] uses the Lagrange multiplier method to directly incorporate the KL divergence constraint into the objective function.

$$\arg \max_{\theta} \mathbb{E}_{S \sim V^{\pi_\theta}} \mathbb{E}_{a \sim \pi_{\theta'}(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) - \beta D_{KL}[\pi_{\theta_k}(\cdot|s) \| \pi_\theta(\cdot|s)] \right] \quad (3)$$

Assume $d_k = D_{KL}(\pi_{\theta_k}, \pi_\theta)$. If $d_k < \delta$, then $\delta_{k+1} = \delta_k/2$. If $d_k > \delta$, then $\delta_{k+1} = \delta_k \times 2$. Otherwise, $\delta_{k+1} = \delta_k$. θ is a pre-set hyperparameter used to limit the difference between the learning policy and the previous round's policy.

PPO-Clip is more straightforward and commonly used. It imposes a constraint in the objective function to ensure that the difference between the new parameters and the old parameters is not too large.

$$\arg \max_{\theta} \mathbb{E}_{S \sim V^{\pi_\theta}} \mathbb{E}_{a \sim \pi_{\theta'}(\cdot|s)} \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(s, a) \right) \right] \quad (4)$$

3.2.3. Actor-Critic

Actor-Critic (AC) algorithms [116] combine the value function estimation of value-based methods and the policy optimization of policy-based methods. They use the Actor for policy updates and the Critic for value evaluation, thereby improving learning efficiency and stability. AC algorithms are essentially policy-based because their goal is to optimize a parameterized policy. They additionally learn a value function to help the policy function learn better. Considering their uniqueness, we still list them separately as another framework. The AC algorithm combines the advantages of value-based and policy-based methods, ensuring a more stable and efficient approach to reaching optimal solutions when exploring complex V2G problems.

We divide AC into two parts: the Actor (policy network) and the Critic (value network). The Actor interacts with the environment and learns a better policy using policy gradients under the guidance of the Critic's value function. The Critic learns a value function from the data collected by the Actor's interactions with the environment, which is used to judge which actions are good in the current state, thereby helping the Actor update its policy. The pseudocode is shown as algorithm 3.

Algorithm 3 Actor-Critic Algorithm

```

1: Initialize policy network parameters  $\theta$ , Initialize value network parameters  $w$ 
2: for each episode do
3:   Sample trajectories  $\{s_1, a_1, r_1, s_2, a_2, r_2, \dots\}$  using the current policy  $\pi_\theta$ 
4:   For each step in the data:  $\delta_t = r_t + \gamma V_w(s_{t+1}) - V_w(s_t)$ 
5:   Update value network parameters:  $w = w + \alpha_w \sum_t \delta_t \nabla_w V_w(s_t)$ 
6:   Update policy network parameters:  $\theta = \theta + \alpha_\theta \sum_t \delta_t \nabla_\theta \log \pi_\theta(a_t|s_t)$ 
7: end for
```

A2C/A3C [117]. A2C uses the advantage function instead of the original returns in the critic network of AC, which can serve as a measure of how good the selected action value is compared to the average value of all actions.

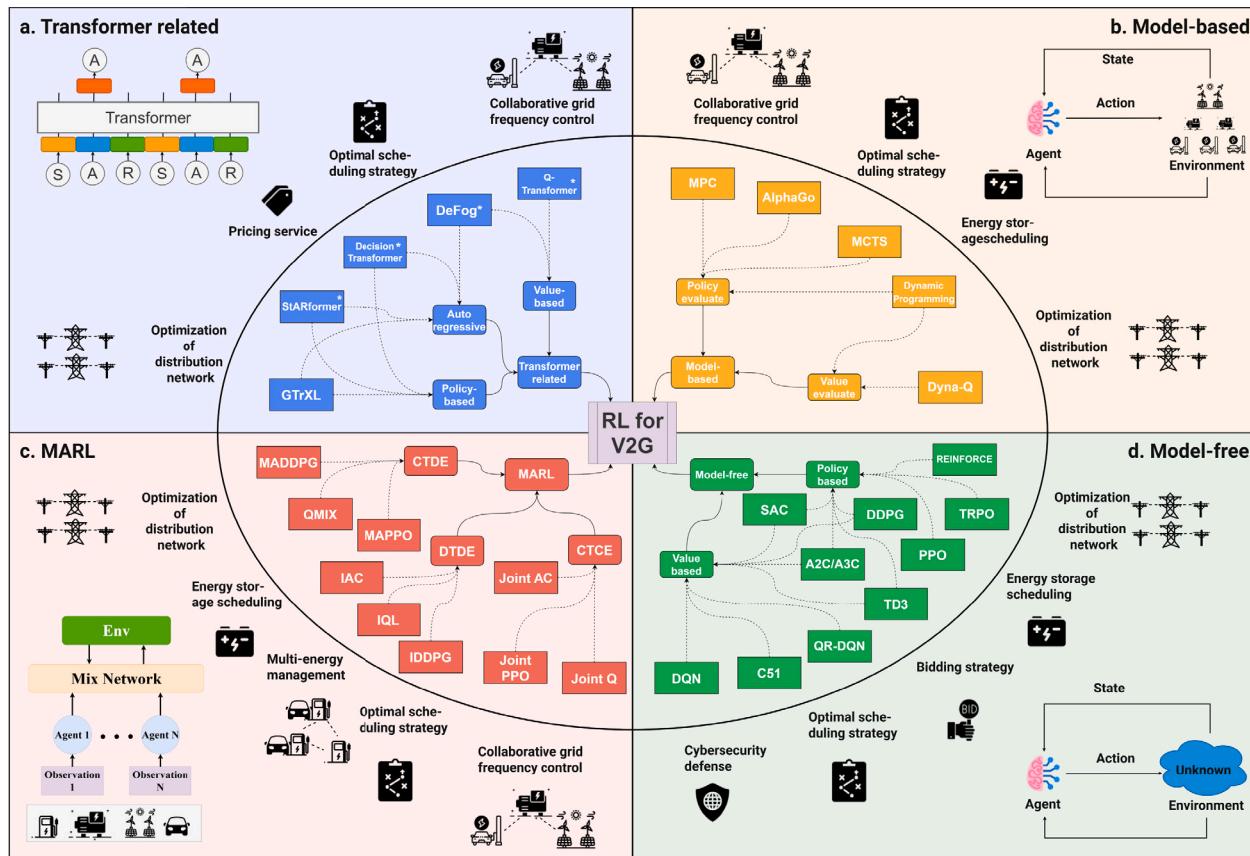


Fig. 4. Categorization of reinforcement learning for V2G. Part a describes the use of Transformer algorithms to capture time series, which are suitable for real-time and time-dependent V2G problems. Part b describes model-based algorithms, which are suitable for existing models or easily modelable V2G problems. Part c describes multi-agent algorithms, which are suitable for V2G problems involving multiple distinct agents. Part d describes model-free algorithms, which are suitable for more complex V2G problems. * indicates the method has an official code implementation available.

A3C adopts an asynchronous method to generate data. In the A3C model, each worker directly takes parameters from the Global Network and interacts with the environment to output actions. The gradients of each worker are used to update the parameters of the Global Network. Each worker can be regarded as an A2C.

DDPG [118]. The Deep Deterministic Policy Gradient algorithm (DDPG) constructs a deterministic policy and uses gradient ascent to maximize the Q-value. DDPG is an off-policy algorithm that requires maintaining four networks: the actor network, the critic network, and their corresponding target networks. In DDPG, the target networks are updated using a soft update method, which means that the target Q-network is slowly updated to gradually approach the Q-network $w^- \leftarrow +\tau w + (1 - \tau)w^-$.

TD3 [119]. Twin Delayed Deep Deterministic Policy Gradient (TD3) introduces the concept of Clipped Double-Q Learning based on DDPG, where the value estimate suffering from overestimation bias can be used as an approximate upper bound for the true value estimate. The target update for the algorithm is: $Q_{\text{target}_1} = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \pi_{\phi_i}(s'))$. Since the TD target calculation uses the same policy, we use $Q_{\text{target}_1} = Q_{\text{target}_2}$.

SAC [120]. In the Soft Actor-Critic algorithm (SAC), we model two action-value functions based on the idea of Double DQN (with parameters w_1 and w_2) and a policy function π (with parameter θ).

The goal of SAC is not only to maximize cumulative rewards but also to maximize the entropy of the policy. Entropy is a measure of the randomness of the policy, and increasing entropy can encourage more exploration, preventing the policy from prematurely converging to sub-optimal solutions. The policy objective function is $J(\pi) =$

$E_{(s,a) \sim \rho_\pi}[r(s,a) + \alpha H(\pi(.|s))]$, where α is the coefficient that balances the reward and the entropy, and $H(\pi(.|s)) = -E_{a \sim \pi}[\log \pi(a|s)]$ is the entropy of the policy.

3.2.4. Summary

This section discussed some branches and basic methods of model-free deep reinforcement learning, which are simple, lightweight, and very effective. Using these methods in the initial deployment of V2G can test the stability and effectiveness of the environment. In the actual deployment process, in addition to considering the basic functions of the model, attention should be paid to issues such as model safety, scalability, and explainability.

First, for real V2G applications, some actions may be impractical or risky. In real deployment, Safe Reinforcement Learning (Safe RL) should be used to ensure that the agent's behavior meets safety constraints and avoids unsafe actions [121].

Second, reinforcement learning is highly tied to the environment, and its scalability is greatly constrained. Theoretically, modular design and distributed processing can improve the system's scalability and ensure efficient operation in large-scale applications [122]. Another option is to train by mapping abstract actions and states, and then restore the original actions during deployment to achieve migration.

Lastly, in terms of explainability, we have a detailed discussion in the subsequent future research directions, so it will not be elaborated here.

In summary, reinforcement learning methods have the advantage of exploring the environment and solving problems. Different models are available for different scenarios, but there are still many challenges and issues to consider in the specific deployment of V2G. These will be discussed in more detail in the following sections.

3.3. Transformer related

We can see that reinforcement learning uses neural networks to make decision choices (whether policy-based or value-based). In deep learning, the invention of Transformers has greatly improved natural language processing. This implies that introducing Transformers into reinforcement learning can also significantly enhance models [123]. We will discuss Transformers separately here, mainly because these algorithms mostly abandon the traditional method of learning through temporal difference and instead use an auto-regressive approach to fit and predict the next action. Theoretically, this makes these models not belong to either policy-based or value-based categories. Introducing Transformers into reinforcement learning or multi-agent reinforcement learning can improve their representation capabilities, enabling models to better capture spatiotemporal latent features, especially in applications with complex data. However, Transformers require larger training resources and data resources, and their training complexity is also higher. In V2G systems, Transformers are more suitable for deployment in complex scenarios that generate large amounts of data or for resource scheduling problems. Besides, the Transformer model's suitability for addressing temporal issues makes it easier to capture temporal features in V2G applications. By leveraging these temporal features, better action choices can be achieved. The inclusion of these two classifications in the diagram is only to consider the final decision-making method, which is not entirely consistent with the traditional classification of reinforcement learning. The pseudocode of Self-Attention is shown as algorithm 4.

Algorithm 4 Self-Attention Mechanism

- 1: Initialize input sequence X
 - 2: Initialize weight matrices W_Q, W_K, W_V
 - 3: Compute queries: $Q = XW_Q$
 - 4: Compute keys: $K = XW_K$
 - 5: Compute values: $V = XW_V$
 - 6: Compute attention scores: $A = \frac{QK^T}{\sqrt{d_k}}$
 - 7: Apply softmax to attention scores: $A = \text{softmax}(A)$
 - 8: Compute output: $Z = AV$
 - 9: **return** Z
-

Decision Transformer [51]. Decision Transformer (DT) is a classic algorithm that introduces Transformers into reinforcement learning. Unlike traditional reinforcement learning, which updates model parameters through temporal difference learning, DT learns from trajectory data using an auto-regression approach. It does not treat data as a Markov process but rather as a sequential process. By learning from trajectory data, DT aims to find the optimal actions to optimize the overall trajectory. Therefore, the DT algorithm only uses offline data and does not have an online learning version.

DT learns offline data actions by fitting actions $L(\theta) = \frac{1}{T} \sum_{t=1}^T (a_t^{predict} - a_t^{data})^2$, and its training effectiveness is highly dependent on the quality of the offline data. Generally, DT incorporates more temporal sequence information during the learning process, allowing it to find better paths when selecting sequences. By using traditional algorithm models to generate offline data and then further training with DT, better results can typically be achieved. Additionally, DT requires a target reward value. During training, it typically does not fit the reward value directly. Instead, it sets an expected reward value for the entire sequence and calculates the target reward value accordingly. The setting of the expected reward value for the sequence can influence the overall training effectiveness of the model.

Q-Transformer [124]. Q-Transformer is a reinforcement learning algorithm that combines Transformer and Q-learning. It still employs the temporal difference learning method. Unlike traditional Q-learning, the temporal difference method focuses on the problem of multi-dimensional action spaces. By discretizing multi-dimensional continuous actions dimension by dimension (a_0, a_1, \dots, a_n), it avoids the issue

of dimensionality disaster. Additionally, to address the problem of data distribution shift in offline learning, a conservative Q-function regularization method $L_C = \mathbb{E}_{(s,a) \sim D} [\max(0, Q(s, a) - \max_{a'} Q(s, a'))]$ is introduced to ensure that the actions with the maximum Q-values remain within the data distribution.

DeFog [125]. Decision Transformer under Random Frame Dropping (DeFog) simulates random frame dropping during training, enabling the agent to better handle packet loss in real-world applications, thereby enhancing the agent's stability and performance. The Transformer's output layer is used to predict the Q-value for each action $L_Q = \mathbb{E}_{(s,a,r,s')} [(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta))^2]$. In this way, DeFog can accurately estimate Q-values even in the presence of random frame drops, allowing it to make reasonable decisions.

3.4. Multi-agent reinforcement learning

Multi-Agent reinforcement learning (MARL) differs from single-agent reinforcement learning in that it primarily addresses the interaction of multiple agents within the same environment, both with other agents and the environment itself. Compared to single-agent systems, multi-agent systems often use The Decentralized Partially Observable Markov Decision Process (Dec-POMDP). It can be composed of a tuple of five elements: (S, A, P, R, O, γ) , where O denotes the observation space. Agents need to make decisions based on limited observed information to achieve the highest reward. In the deployment of V2G, it is unrealistic to directly obtain all information. Most large-scale and complex environments contain unobservable latent variables. In such cases, MARL is more suitable for addressing these Dec-POMDP issues. [126] is used to create and observe Dec-POMDP, and solves the problem of improving joint routing and scheduling in V2G using MARL methods. MARL is suitable for environments with multiple agents. In the context of V2G, where various entities such as customers, charging stations, and the grid are involved, MARL is more apt at describing the complex environment, thus achieving better training results.

In multi-agent systems, the environment is non-stationary because each agent's strategy affects the strategies of other agents, making the learning process more complex and increasing the difficulty of strategy convergence. Communication and hardware limitations between different agents are issues that need to be addressed and resolved. In the application of V2G and similar EV systems, the environment often cannot engage in unlimited communication. Multi-agent systems can limit these issues by setting communication volumes, thereby better fitting the real deployment scenario of V2G.

In different multi-agent environments, the relationships between agents can vary, such as fully cooperative, fully competitive, or mixed relationships. Here, we will not discuss the environment but focus on learning algorithms. Based on the training and execution processes, we categorize them into three types: Centralized Training with Centralized Execution (CTCE), Decentralized Training with Decentralized Execution (DTDE), and Centralized Training with Decentralized Execution (CTDE), which shows in Fig. 5. We will introduce representative algorithms for each type.

3.4.1. CTCE

CTCE involves fully centralized training and execution phases. This means that during both training and execution, all agents share their observations and actions, treating the entire system as a single agent. This approach allows single-agent training methods to be directly applied to multi-agent problems, making it easier to find globally optimal strategies. Additionally, the behavior of all agents can be better coordinated, reducing conflicts among individuals. However, in CTCE, the joint state-action space grows exponentially with the number of agents, leading to poor scalability. Furthermore, the constant sharing of information results in high communication overhead and robustness requirements.

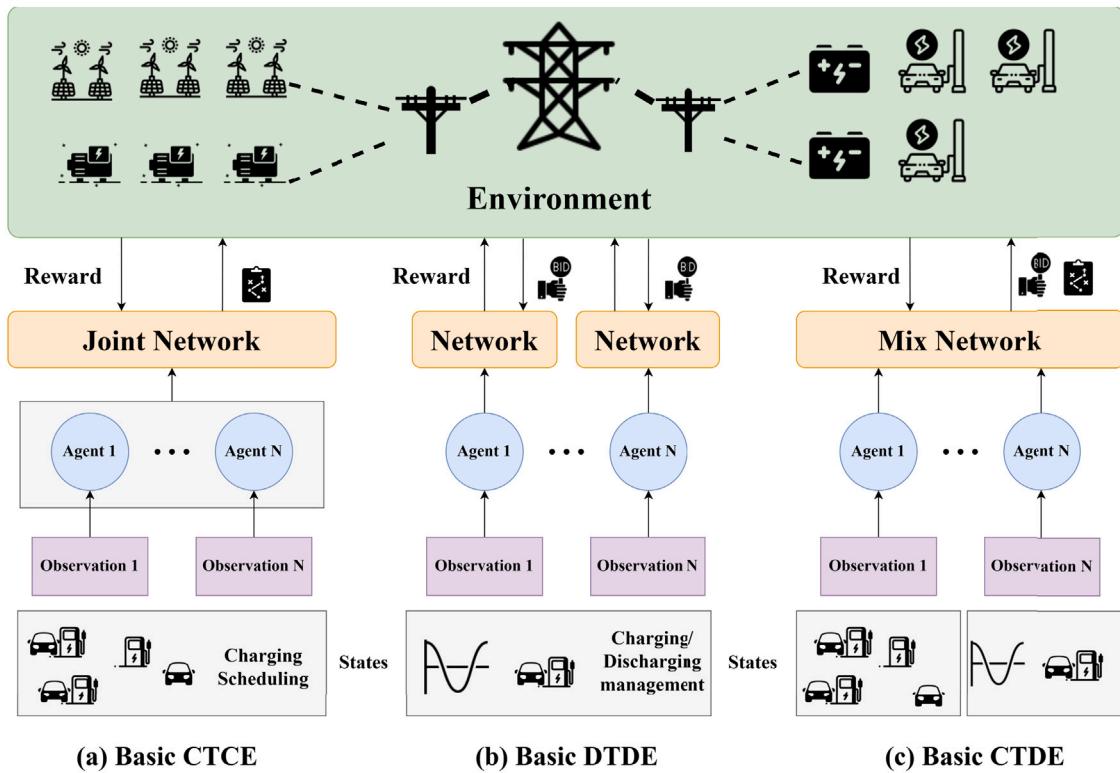


Fig. 5. Classification of multi-agent reinforcement learning in V2G.

CTCE inherently combines multiple agents for computation, making it more efficient and suitable for scenarios with fewer agents, low communication latency between agents, and a good communication environment. For V2G, edge sites are relatively small in scale, and their environments are somewhat independent. Therefore, these sites are more suitable for deploying the CTCE method when implementing MARL. For existing single-agent algorithms, there are corresponding CTCE methods, including value-based, policy-based, and actor-critic algorithms. Additionally, CTCE may be applied to low-level communication in large-scale MARL deployments through a hierarchical approach. On the other hand, in application scenarios with low communication pressure, using CTCE to accelerate training speed can be a good choice.

3.4.2. DTDE

In DTDE, both the training and execution phases are distributed, enabling efficient collaboration in large-scale systems. Due to distributed training, it offers higher scalability, less reliance on global information, and greater robustness. However, the lack of a centralized coordinator means that agents have less strategy coordination, requiring more training resources and time to achieve optimal strategies. Additionally, since global information is ignored, each agent makes decisions based only on local information, which may lead to suboptimal strategies.

DTDE is more suitable for environments with communication constraints or privacy requirements because agents do not communicate with each other. Similar to other multi-agent classifications, different types of single-agent algorithms have their developments in DTDE, such as IQN [127], IAC, and IDDPG [128]. These algorithms can achieve the effect of centralized training by aggregating all experiences for training. The choice of training method depends on the application scenario of the environment. When using these algorithms, one should adopt the appropriate training method based on their environment. DTDE has strong scalability, making it a good choice for the initial deployment of V2G when facing imperfect communication conditions. However, its limitation of relying on local information and difficulty in finding optimal solutions hinder its suitability for long-term deployment. For

long-term deployment, it is preferable to choose MARL solutions that can exchange information through communication to improve training effectiveness.

3.4.3. CTDE

In CTDE, all agents share global state information during the training phase, which effectively optimizes their strategies. During the execution phase, each agent makes decisions based only on its local observations and policies. CTDE combines the efficiency of centralized training with the flexibility of decentralized execution, enabling efficient collaboration without increasing communication overhead. However, due to the non-stationarity of the environment (as agents rely solely on their local observations during execution), the strategies may not always be effective. Additionally, the centralized training process depends on the accuracy of global information, and incorrect global information can affect the training outcomes.

QMIX [129]. In QMIX, each agent follows the DQN sampling process (such as epsilon-greedy) and stores all experience data in the replay buffer. During training, agents retrieve the current Q-value and target Q-value from the replay buffer and input these two Q-values into the mix network to obtain the overall Q-value $Q_{total} = f_{mix}(Q_1, Q_2, \dots, Q_n; \theta_s)$. In this context, θ_s denotes to the parameters generated by the hypernetwork. To ensure that the selection of locally optimal actions correctly reflects in the globally optimal actions during the optimization process, and that an increase in any local Q-value does not decrease the overall Q-value, QMIX requires the parameter weights to be non-negative, thereby satisfying the monotonicity condition $\frac{\partial Q_{total}}{\partial Q_a} \geq 0, \forall a \in A$.

MADDPG [130]. The MADDPG algorithm is an extension of the DDPG algorithm. MADDPG extends the DDPG algorithm through a centralized Q function, where all agents share a centralized Q function. This Q function can obtain observations and actions from all types of agents. In this way, MADDPG ensures that all agents evaluate state-action pairs using the same strategy, ensuring the uniformity of training

results. MADDPG updates the policy function to maximize the output of the critic function. For the critic network, we aim to minimize the difference between the output Q value and the target Q value.

$$L(\theta) = \max_{\theta} \mathbb{E}_{(o,s,a,t,o',s',d) \sim D} [Q(o, s, a, r, o', s', d; \theta) - (r + \gamma(1-d)Q_{\text{tar}}(o, s, \mu_{\theta\text{tar}}(o)))]^2 \quad (5)$$

For the actor network, we aim to maximize the final Q value $\max_{\theta} \mathbb{E}_{(o,s,a,t,o',s',d) \sim D} Q(o, s, \mu_{\theta\text{tar}}(o))$. $\mu_{\theta\text{tar}}(o)$ denotes the policy function which can be shared across agents.

MAPPO [131]. MAPPO (Multi-Agent Proximal Policy Optimization) adheres to the design principles of PPO for individual agent execution. In the overall training process, it introduces a centralized value function to compute the Generalized Advantage Estimation (GAE) and facilitate the learning process of the PPO critic. MAPPO improves the value function with each iteration $\phi_{k+1} = \text{argmin}_{\phi} \frac{1}{|D_k|T} \sum_{t \in D_k} \sum_{i=0}^T (V_{\phi}(o_t, s_t, a_t^-) - \hat{R}_t)^2$. It employs the GAE method, which estimates the advantage function by taking a weighted average of multiple time steps' TD errors, thereby reducing the high variance problem of single-step TD errors $A_t = \sum_{i=0}^{\infty} (\gamma \lambda)^i \delta_{t+i}^V$. For the policy function part, the estimated advantage can be used to update the policy function.

$$L(o, s, a, a^-, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|o)}{\pi_{\theta_k}(a|o)} A^{\theta}(o, s, a^-), \text{clip} \left(\frac{\pi_{\theta}(a|o)}{\pi_{\theta_k}(a|o)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(o, s, a^-) \right) \quad (6)$$

Here, a^- represents the actions taken by all other agents in the current state, excluding the action taken by the current agent.

CTDE combines both DTDE and CTCE, retaining shared information during training to improve training efficiency while reducing communication during execution. Specifically, in the application of V2G and similar systems, CTDE can effectively manage communication overhead, enhance strategy convergence speed, and improve overall system performance. Its efficient and flexible operations during actual execution make CTDE a promising method with broad application prospects in multi-agent systems.

3.4.4. Communication management in large-scale MARL training

Considering the execution and management of policies, V2G decisions and management are typically conducted on a city-wide basis, necessitating the consideration of large-scale MARL methods. Communication issues are inevitable in this context. To ensure the effectiveness and robustness of communication in large-scale MARL training, it is often necessary to minimize the frequency and volume of communication. Therefore, CTDE, which involves communication only during training, and DTDE, which does not continuously share information, are often more suitable choices.

Additionally, there are some viable methods to help manage and alleviate communication in large-scale MARL, such as: Reducing reliance on real-time communication through asynchronous updates. Learning to use local communication methods to reduce global communication and thus decrease communication volume [132]. Simplifying information through a hierarchical structure to reduce direct communication volume [133].

3.5. Summary

In this section, we discuss the classification of common existing reinforcement learning algorithms. These different classifications have varying effects when addressing different types of problems. When using these algorithms, different approaches should be tried based on the specific situation. In V2G systems, different problems may have different optimal solutions. In Fig. 4, we describe four different categories of reinforcement learning algorithms and their potential applications in V2G. These categories take into account different environments and problems, resulting in varying levels of performance.

4. Applications of reinforcement learning in V2G

In recent years, as the demand for the application of reinforcement learning in the V2G field has been growing, it has significantly improved the interaction efficiency between electric vehicles and the grid, enabled the intelligent and optimized scheduling of energy, and played a crucial role in reducing system uncertainty. Therefore, in this section, as shown in Fig. 6, we will outline the application of various reinforcement learning methods in several key directions of V2G from the perspectives of the three participating entities: the power grid, charging aggregators, and electric vehicle users. The aim is to comprehensively explore the mechanism of reinforcement learning in the V2G system and its role in promoting the optimized management of V2G. In order to provide a clear perspective on literature analysis, we have summarized the application path and practice methods of reinforcement learning in the V2G field in recent years in chronological order, as shown in Table 2.

4.1. Collaborative grid frequency control

In the domain of V2G, frequency control is of paramount importance [134–136]. This technology harnesses the energy storage capacity of electric vehicles, aiding not only in the stabilization of the power grid but also in the effective integration of renewable energy sources and the optimization of grid operations. By dynamically adjusting the charging and discharging processes of electric vehicles, V2G significantly bolsters the grid's frequency control capabilities, thus offering solid support for the reliability and sustainability of the power system [52,135, 137]. Recent work by the authors [138] has implemented a multi-microgrid frequency coordinated control strategy using an improved evolutionary deep reinforcement learning (EDRL) method, by considering the impact of the V2G process on the shortest full charging time of electric vehicles. This strategy effectively reduces the adjustment cost of generators and unnecessary discharge of electric vehicles. [31] realizes the dynamic optimization of V2G frequency scheduling through the application of deep reinforcement learning technology, aiming to simultaneously enhance the benefits of electric vehicle owners and aggregators, while also considering the driving needs of electric vehicle owners. To address the dynamic asymmetry issue in frequency regulation within the V2G system, researchers have proposed a Switched Integral reinforcement learning (SIRL) method [30]. This approach aims to achieve efficient regulation of cooperative grid frequency control in multi-microgrid systems. With SIRL, the system can better cope with the dynamic changes and asymmetries in the V2G environment, thereby optimizing the frequency stability and overall performance of the grid. In this paper [139], the authors construct a model predictive two-layer controller for V2G using the DDPG algorithm. This controller is capable of effectively adapting to various stochastic constraints during the control process, thereby effectively preventing the failure of the machine learning controller.

In terms of distributed control strategies, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) method can effectively implement Load Frequency Control (LFC) strategies for multiple microgrid systems in V2G [139]. This approach not only significantly reduces transmission costs but also decreases computational complexity and energy losses during transmission. In [140], researchers have employed MARL technology to meticulously design the frequency scheduling strategy for the discharge of EVs batteries. The goal is to enhance the peak load regulation capabilities of the power grid and effectively tackle the uncertainties in the electricity supply. During the implementation of this strategy, there is no need for data exchange between the EVs and a central control center, which ensures the model's autonomy and protects user privacy, achieving self-managed data control.

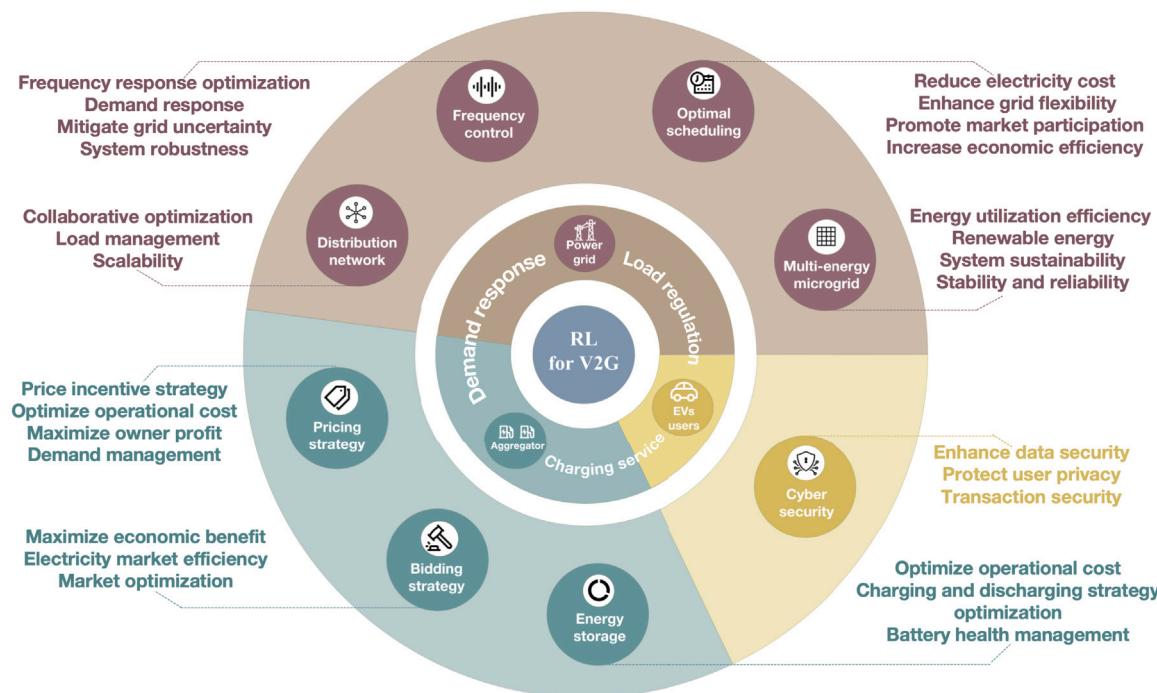


Fig. 6. Reinforcement learning in V2G applications from the perspective of the participating entity.

4.2. Optimal scheduling

With the widespread integration of electric vehicles, microgrids indeed face increasing challenges of randomness and uncertainty in operation and regulation [141]. To tackle these challenges, optimal grid scheduling becomes particularly crucial [142–144]. It not only ensures the stability of the system but also enhances the efficiency of energy utilization. To address the nonlinear challenges posed by the dynamic mobility of electric vehicles and the periodic fluctuations in user charging behavior, researchers have proposed a V2G model that considers the randomness of electric vehicle mobility and user charging behavior, which can effectively enhance the operational efficiency of the grid and the flexibility of energy management [53]. In order to facilitate the implementation of large-scale smart grids for electric vehicles, researchers have proposed a decentralized Markov Decision Process model based on multi-agents, which can effectively reduce the computational time for system scheduling and enhance the responsiveness of the grid and the accuracy of scheduling decisions [145]. In [146], the authors have employed reinforcement learning technology to construct an energy management system for V2G and Vehicle-to-Home (V2H) applications, aimed at optimizing power demand scheduling. This model not only effectively smooths the household electricity load curve and reduces the electricity costs for users but also generates additional economic benefits for them. Additionally, researchers have developed a V2G scheduling model that integrates the power and transportation dual-layer networks [147]. This model, utilizing hybrid strategy learning, is able to simultaneously address discrete routing choices and continuous scheduling decisions. This approach effectively devises reasonable driving routes for users based on traffic conditions and grid information, thereby significantly reducing carbon emissions. To ensure the effectiveness of scheduling, [148] designs a constrained SAC charging scheduling algorithm that adopts a tiered electricity pricing strategy. By incorporating a rule-based safety filtering mechanism, the algorithm maintains load balance while satisfying the charging and discharging power constraints of charging stations and electric vehicles. It not only maximizes the profits of microgrid operators but also takes into account the charging needs of electric vehicle users, and the overall scheme effectively reduces the dimensionality of the action space.

4.3. Pricing strategy

Pricing strategy are crucial for the application of V2G, as they provide real-time electricity price data to precisely control the charging of electric vehicles during low electricity demand periods and the discharge of power back into the grid during peak times, effectively optimizing the allocation of grid resources and enhancing overall grid operation efficiency [165]. Additionally, this precise pricing strategy creates economic incentives for all entities involved in V2G, promoting the in-depth development of the interaction between electric vehicles and the power grid [166]. In [149], the authors adopted a multidimensional approach, taking into account factors such as the purchase cost of aggregators, the driving behavior of electric vehicle users, and the competitive market environment. By applying the Q-learning algorithm, this research has achieved an optimal V2G pricing strategy. This strategy not only effectively coordinates the interests of power grid operators, electric vehicle aggregators, and electric vehicle users but also promotes the maximization of interests for all three parties, achieving a win-win cooperative model. In order to more effectively incentivize electric vehicle users to participate in the V2G system, researchers have analyzed user behavior patterns based on historical bidirectional charging and discharging data, and employed the DDPG method to construct a V2G dynamic real-time pricing model that updates every 15 min [152]. By increasing the frequency of pricing in the V2G system, this model not only effectively enhances user engagement but also promotes the rapid development of the V2G market. The paper [151] addresses the issue of electric vehicle charging pricing in the power system and proposes a new deep reinforcement learning method that combines Deep Deterministic Policy Gradient with Prioritized Experience Replay strategy, effectively resolving the discretization issue of charging/discharging behaviors. By optimizing in a multi-dimensional continuous space, this technique transcends the limitations of traditional discretized RL methods, enhancing the optimality and computational efficiency of the pricing scheme. The study also confirms the advantages of this method in improving the economic benefits for both aggregators and electric vehicle owners, comprehensively demonstrating the value of smart charging and V2G flexibility. To optimize the

Table 2

Summary of reinforcement learning in V2G applications.

Year	Paper	Application	RL algorithm	State	Action	Reward	Discrete space	MARL
2019	H Ren et al. [149]		Q-Learning	Aggregator's V2G power	Variation of the V2G price	Profit of users and aggregators		
2019	X Chen et al. [71]		Q-Learning	Regulation signal, state of charge (SOC) & time, number of plug-in EVs	Charging/discharging	Revenue (cost) of selling (buying) electrical energy		
2019	X Feng et al. [150]		MARL, SARL	Average SOC of local EVs, the number of local EVs	Charging/discharging	Economy, battery degradation, and user's anxiety		
2020	D Qiu et al. [151]		PDDPG, Q-Learning, DQN, DDPG	Wholesale market prices, the demand of inflexible EVs, and the net demand of flexible EVs	Retail prices offered by the aggregator	Overall profit with penalty term		
2020	Y Zhang et al. [57]		Multi-DQN, Q-Learning	Number of available charging piles, charging state of EVs	Quantity that EVs want to charge, price willing to pay	Utility of EVs		
2021	F Alfaverh et al. [146]		Q-Learning	Electricity price, home power demand, SOC, Availability of EVs	Charging, discharging/appliances, discharging/grid and do nothing	Rating value of the action		
2021	D Liu et al. [152]		DDPG	Charging demand, EVs power load	Pricing strategy	Charging transaction revenue		
2021	Y-C Chuang et al. [153]		DDPG, PPO, Q-Learning, DQN, DRQN	Market trends, self-situation, strategies of other aggregators, and customers' behaviors	Prices offered by the aggregator to producers and customers	Aggregator's revenue minus the fee		
2021	Y Tao et al. [154]		DDPG	Current energy storage state, value of solar irradiance, wind speed, and the market cleared trading volume, observed utility	Proportion of the current energy storage, bidding price to sell that amount of energy	Direct profits		
2021	Y Lu et al. [54]		MARL, SAC, PPO, DDPG	Charging loads, historical charging prices	Charging price	Profit		
2021	S Lee et al. [79]		FRL, A2C, SAC	Price, aggregate energy demand, predicted PV generation energy	Scheduled energy charging/discharging and selling price of energy	Profit, operation cost, penalty		
2022	Y Wen et al. [53]		DQN	User electrical load demand, generation power, charging and discharging power capability	Output power, Interaction power	Overall operating cost		
2022	M Alqahtani et al. [145]		CRL, MARL	EVs' position, EVs' state of charge, solar irradiance, and power load	Mobility decisions and energy dispatch decisions	Overall cost		
2022	S Rahman et al. [87]		SAC	Average of solar generation, the moving average of day-ahead, market prices	The charge or discharge decision	The profit		
2022	A Narayanan et al. [155]		RL	The information about individual vehicle-node pairs	The (vehicle, node) pair at decision time	Total cost		
2022	A Kumari et al. [156]		DQN	Energy price and availability of EVs, EVs either consume energy or distribute energy	Charging and discharging of energy	Revenue		
2022	S R Pokhrel et al. [157]		FRL, DQN	Energy demand	Charge/discharge	Cost		
2023	P Fan et al. [139]		MA-DDPG	Frequency deviation, wind and load disturbance	Control signal	Frequency stability		
2023	F Alfaverh et al. [31]		DDPG	The battery SOC, the V2G power	Regulation-up and regulation-down	The charging demand		
2023	M B Hossain et al. [158]		GA-Based Q-Learning	The battery SOC	Charge/discharge	The cost of energy		
2023	D Liu et al. [159]		DDPG, PPO	Power demand, power output, start charging time and departure time, expected state of charge of EVs	The active and reactive power, the charging capacity, the tap position of the OLTC, and the number of SCB units	The operation cost, the voltage violations		
2024	P Fan et al. [138]		DDPG, EDRL	Frequency deviation, the charging power of EVs, cost	The control signals	Cost, penalty		
2024	P P Kumar et al. [160]		PPO	Renewable energy availability, grid load, and EVs fleet conditions	Decisions related to EVs charging and discharging	Grid stability, renewable energy utilization		
2024	W Pan et al. [161]		SAC	Different energy sources and time-varying price settings	Online V2G scheduling and time-varying prices of different energy demands	Total operation cost		
2024	L Chen et al. [162]		BLSAC, MASAC, MADDPG	Global and local observation	Charging or discharging power rate	Operational costs		
2024	M.-J Jang et al. [163]		Q-Learning	The operation time, SPG forecast error, and CS utilization	Charging or discharging quantity	SPG forecast error		
2024	P Liu et al. [164]		MJRCDD, MADDPG	Position node of the vehicle, power storage, the vehicle, attributes of the current node	Three limited optional strategies (charging, discharging and passing through the station)	Income or cost		
2024	J Sun et al. [56]		Q-learning, MPRLC	Frequency deviation	Control signals	Cost		

Frequency control

Optimal scheduling

Pricing strategy

Bidding strategy

Multi-energy microgrid

Energy storage scheduling

Optimization of distribution network

Cybersecurity defense

problem of how aggregators in the power grid adjust pricing signals to effectively coordinate charge–discharge transactions between the grid and electric vehicle users, researchers have developed an aggregator pricing strategy based on deep reinforcement learning [153]. This strategy fully considers market competition behavior, the volatility of renewable energy, and the boundaries of charging and discharging operations in a dynamic environment, aiming to maximize profits while achieving a balance between electricity supply and demand.

4.4. Bidding strategy

In the application of V2G, bidding strategies in the electricity market hold significant importance in game theory. Through carefully designed bidding strategies, electric vehicle aggregators can compete with other participants in the market, optimizing their decisions on energy buying and selling [167–169]. This strategy not only reflects the competition and game-playing in the market but also helps to enhance the stability and economy of the power grid by flexibly adjusting charging and discharging behaviors. The paper [71] addresses the issue of incomplete information games among electric vehicles in V2G systems

and proposes an algorithm called FSPEVA, which combines fictitious self-play with reinforcement learning to find the Nash equilibrium. This method effectively tackles the challenge of incomplete information by using reinforcement learning to learn the best response strategies from the interactions of electric vehicles. Simulation results show that FSPEVA can approximately converge to the Nash equilibrium in an environment with incomplete information, and by using pre-training methods with historical data, the convergence speed of the algorithm is further accelerated. To investigate the benefits of the V2G microgrid market and balance the interests of all market participants, the author has constructed a multi-agent-based residential microgrid model and an auction bidding market platform [150]. By developing an optimized Equilibrium Selection-Multi-Agent reinforcement learning (ES-MARL) algorithm, it ensures the fairness of transactions and the security of personal information while maximizing the overall benefits of the system. In [154], the authors propose a bidding strategy for electric vehicle aggregators based on deep reinforcement learning, aiming to optimize bidding decisions in the local energy market, and simulation results indicate that this method effectively increases profits and reduces risks. This paper [54] constructs a non-cooperative Stackelberg game framework, delves into the strategic pricing issues of electric vehicle charging station operators, and employs a novel multi-agent deep reinforcement learning algorithm based on the soft actor-critic to solve the game equilibrium while ensuring privacy security among operators. Based on real-world data of urban scale, a numerical case study was conducted that confirmed the significant effectiveness of this framework in optimizing the charging behavior of electric vehicle fleets and improving the overall energy efficiency of the transportation system. In [57], researchers conduct an in-depth discussion on the optimal bidding strategy for electric vehicle charging in the auction market, and innovatively proposes a reinforcement learning bidding strategy based on Multi-Deep Q-Network (Multi-DQN). This strategy effectively optimizes charging decisions by equipping each car owner with value evaluation and target networks. Experiments have proven that it significantly outperforms traditional Q-learning and random bidding methods in terms of improving economic efficiency and reducing charging time.

4.5. Multi-energy microgrid

In a V2G system that integrates diverse renewable energy sources, the complexity and difficulty of scheduling strategy significantly increase [48,170]. Reinforcement learning algorithms, with their excellent adaptive and learning capabilities, are well-suited to handle such complex dispatch issues. They provide the V2G system with precise optimization dispatch strategies, ensuring the maximization of energy utilization efficiency and the stable operation of the system. In the multi-energy system, the dispatch center based on reinforcement learning can intelligently allocate various renewable energy sources such as solar, wind, and geothermal energy according to the overall real-time charging and discharging demands of the V2G system, ensuring that power supply needs are met while promoting the sustainable development of energy. [160] has improved the utilization rate of renewable energy in the V2G system by 15.3% by employing reinforcement learning methods, which can effectively facilitate the extensive integration and profound incorporation of renewable energy, realizing long-term sustainable development. The results show that this approach effectively promotes the more comprehensive utilization of renewable energy, opening up a new path for building a resilient and efficient energy network. In [161], the authors construct a multi-energy dispatching framework for V2G based on deep reinforcement learning, which makes reasonable decisions on charging and discharging actions online using the soft actor-critic algorithm. The results demonstrate that deep reinforcement learning is significantly effective in dispatching multi-energy systems, not only improving the system's profitability but also enhancing the sustainability of energy development. To address

the profitability issues of virtual power plants in power systems with a high proportion of renewable energy, this paper [87] proposes an operation strategy based on soft actor-critic reinforcement learning. This strategy integrates solar energy systems and electric vehicle chargers for vehicle-to-grid support and maximizes profits through day-ahead and imbalance electricity market transactions. This approach achieves optimized scheduling of electric vehicle charging demands by solving a two-stage stochastic optimization problem and employing a reinforcement learning algorithm with differentiable projection layers to enforce constraints.

4.6. Energy storage scheduling

Energy storage scheduling is mainly used in V2G system to optimize the bidirectional energy flow between EVs and the grid and manage the load fluctuations of the grid [171]. This helps balance power supply and demand, and improve energy utilization efficiency. At the same time, it improves the profit of the car owners. In [162], an energy management framework based on bi-level soft actor-critic algorithm is proposed to optimize the coordinated scheduling of integrated energy and V2G systems. The bi-level agent is used to manage the scheduling of the integrated energy system and formulate the charging and discharging strategy of the EVs respectively. The bi-level reinforcement learning approach improves energy exchange between the integrated system and EVs charging stations and reduces operational costs. The rise in EVs adoption leads to increased charging loads, which can affect the reliability of the smart grid. The paper [172] proposes a fair energy scheduling approach by introducing contribution-based fairness, prioritizing EVs with higher contributions for charging energy. The scheduling problem is modeled as an infinite-horizon Markov decision process, utilizing adaptive dynamic programming to maximize long-term fairness through online training. The scheduling effectively reduces and flattens peak loads in the distribution network. Additionally, contribution-based fairness facilitates quick recovery for deeply discharged EVs. Traditionally utility providers (UPs) must monitor the state of charge of vehicle batteries (VBs). The author introduces a genetic algorithm (GA) based reinforcement learning framework to construct a demand-side energy management approach [158]. It uses rechargeable batteries (RBs) to ensure cost-effective privacy for EVs, improves scheduling efficiency, and enables accurate billing. The GA based method also accelerates convergence compared to traditional Q-Learning RL methods. As mobile energy storage capable of V2G, EVs offer a solution for power supply fluctuation. To maintain the safety and stability of multi-microgrid systems, [173] formulates a coupled system, considering the dynamic behavior of EVs users as constraints. An enhanced robust model predictive frequency control strategy for multi-microgrids incorporating EVs is proposed, which reformulates the control process into linear matrix inequalities (LMIs). The results indicate that the proposed strategy effectively mitigates frequency fluctuations and achieves a faster response time. [163] takes renewable energy into consideration, aiming to improve solar power generation (SPG) forecasts. To handle the variability in solar energy and charging station conditions, the V2G operation is formulated as a Markov decision process, with deep reinforcement learning used to adaptively correct SPG forecast errors. A significant reduction in mean squared error is achieved compared to scenarios without V2G, shows the reliability and efficiency of renewable energy integration.

4.7. Optimization of distribution network

In V2G systems, distributed networks play a crucial role, involving the coordination among multiple electric vehicles, grid infrastructure, and regional control systems [174]. In [155], the authors address the problem of routing EVs with constraints on load capacity, time windows, and V2G energy supply (CEVRPTW-D). The research introduces QuikRouteFinder which leverages reinforcement learning for EVs

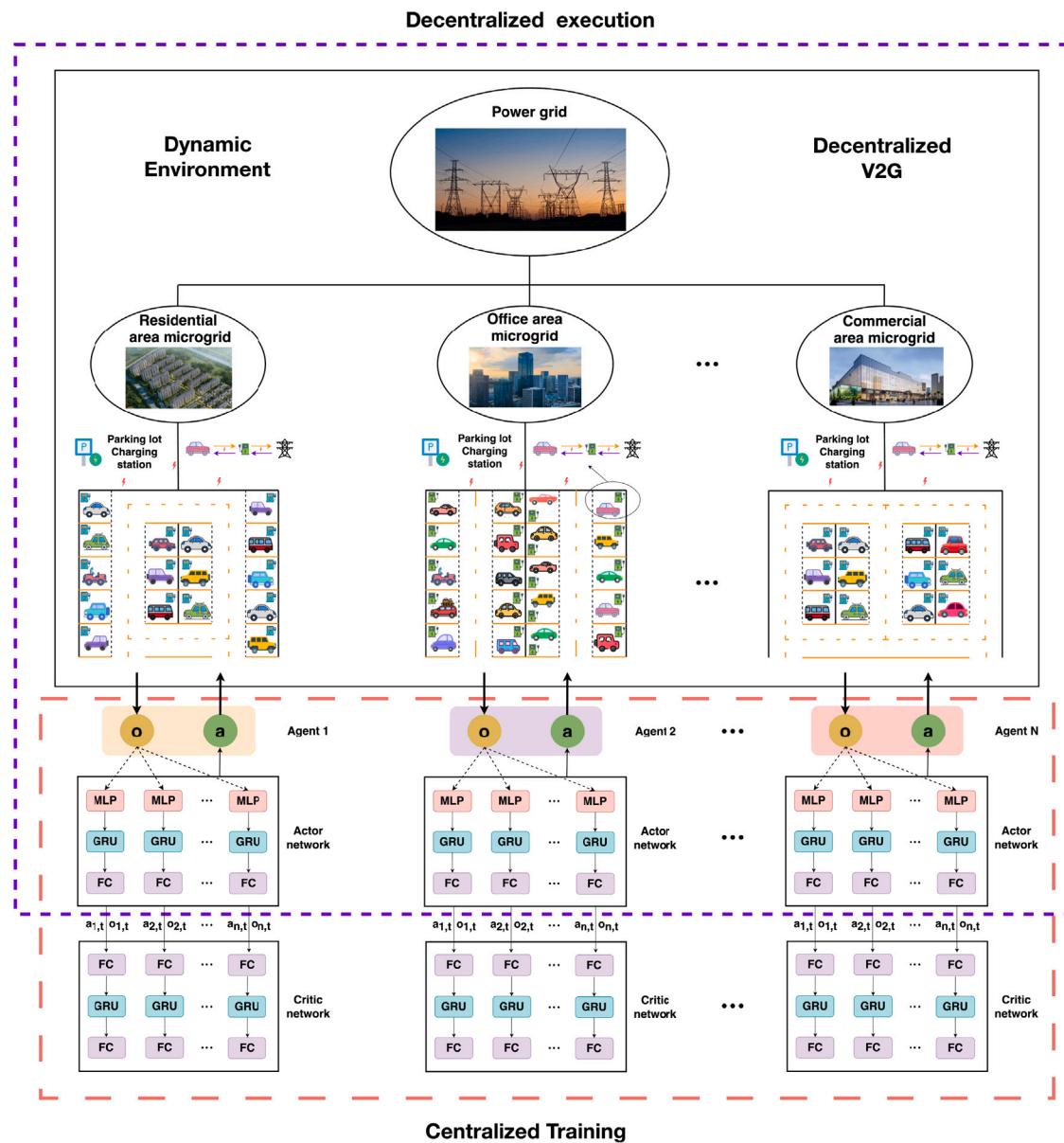


Fig. 7. Schematic diagram of multi-agent reinforcement learning for scheduling optimization in V2G distributed networks. Using the CTDE training paradigm, each agent consists of an actor network and a critic network, which can be composed of Multi-Layer Perceptrons (MLP), Gated Recurrent Units (GRU), and Fully Connected (FC) layer modules. During training, agents achieve centralized training by sharing observation state information with each other, followed by distributed execution based on the trained policies.

routing, by balancing multiple system objectives in CEVRPTW-D, a distribution network with many customers and discharge stations can be coordinated simultaneously. The paper [175] focuses on optimizing strategy of coordinated EVs charging. Firstly, it models EVs owner behaviors by Monte Carlo Simulation (MCS), including arrival and departure times, parking duration, and battery capacity. Then the objective function is optimized with the Opposition-based Competitive Swarm Optimizer (OCSO) algorithm to achieve an optimal charging schedule, aiming to minimize daily load variance and flatten the load curve. This method reduces load variance and demonstrates a flattened load curve. [159] presents a DRL approach to optimize EVs charging scheduling and voltage control in a coordinated manner. The strategy of DRL is structured in two layers, the upper layer focuses on reducing power generation costs and energy consumption costs, while the lower layer maintains voltage stability in the distribution network. The problem is formulated as a MDP, given the dynamic state space and mixed action outputs, a DDPG framework is used

to train a two-layer agent. The method of two layers effectively coordinates EVs charging and stabilizes the voltage in the distribution network. Multi-agent reinforcement learning is often considered an effective approach for solving scheduling optimization in V2G distributed networks, as shown in Fig. 7, and holds great potential for development. To overcome limitations such as EVs profitability, EVs transportation timeliness, and the high costs of central servers, a Multi-vehicle Joint Routing and Charging-Discharging Decision algorithm (MJRCDD) based on reinforcement learning is proposed [164]. In a MDP formulation, the charging and discharging behavior is integrated with route selection within the vehicle action space. Then multi-agent reinforcement learning is used to solve the joint routing together with charging decision. The MJRCDD's effectiveness is validated by PeMS data, demonstrating its capability to optimize EVs routing and energy management in V2G scenarios. In solving lack of secure and cost-effective access to real-time EVs trading data, a Secure V2G-Energy Trading (SV2G-ET) scheme is proposed [156]. Utilizing deep Q-network to schedule EVs charging and discharging, while Ethereum Blockchain

Technology (EBT) like InterPlanetary File System (IPFS) and smart contracts (SC) to securely manage and access real-time EVs trading data. The SV2G-ET scheme enhances scalability of distribution network and reduces ET data storage expenses.

4.8. Cybersecurity defense

In the V2G system, different nodes need to be comprehensively considered and deployed for cybersecurity defense to ensure the overall security and stability of the system. Firstly, endpoint devices such as electric vehicles and charging stations are one of the focal points of cybersecurity defense. These devices need to be protected from physical attacks and malware infections. Specific security measures include device authentication, data encryption, regular security updates, and intrusion detection systems. These measures ensure that only authenticated devices can access the system and that transmitted data is not intercepted or tampered with [176,177], or through predictive methods to mitigate the impact of denial-of-service (DoS) attacks [95].

Current Electric Vehicle Charging Stations (EVCS) face several challenges in maximizing profits, including insufficient data processing, inaccuracies in dynamic environment modeling, and the risk of operational data privacy breaches. Lee et al. [55] proposed an innovative privacy-preserving distributed deep reinforcement learning framework. This framework employs the soft actor-critic algorithm and integrates joint reinforcement learning strategies to effectively maximize profits under a dynamic pricing mechanism, incorporating smart EVCS, photovoltaic, and energy storage systems while strictly ensuring the privacy and security of charging data.

Secondly, at the network layer, it is equally important to protect the integrity and confidentiality of data transmission to prevent man-in-the-middle attacks and DoS attacks. To achieve this, techniques such as encrypted communication, firewall and router configuration, intrusion prevention systems (IPS), and network segmentation can be used. These measures effectively filter malicious traffic, block unauthorized access, and ensure the security of data transmission [178].

To ensure the security of the frequency control communication architecture in V2G systems, researchers have employed integral reinforcement learning technology to develop an adaptive frequency control model. This model is capable of flexibly responding to attacks of varying intensities, with its core objective being to minimize frequency deviations and mitigate the potential negative impacts of DoS attacks [32]. Through this innovative approach, researchers have provided robust security assurance for the stable operation of V2G systems.

Additionally, at the central processing nodes, such as control centers and data centers, the focus of cybersecurity defense is to protect systems that store and process data, preventing data breaches and destruction. Specific security measures include access control, data backup, log monitoring, and multi-factor authentication. By strictly controlling access to sensitive data and systems, ensuring that only authorized personnel can access them, and regularly backing up important data while monitoring system activity logs, the system can trace and analyze security incidents when they occur.

Sun et al. [56] proposed a multi-step predictive reinforcement learning V2G control (MPRLC) scheme to accurately predict multiple control steps that are blocked by DoS attacks, thereby enabling the V2G-FR controller to adapt in advance to changes in the power system.

In terms of privacy protection, researchers have addressed the data privacy issues of the interaction between electric vehicle users and the power grid in the V2G system by proposing an efficient federated reinforcement learning framework [157]. This framework uses small auxiliary batteries to generate noise to mask the real energy demands of electric vehicles and employs deep Q-learning strategies to optimize costs and privacy rewards, thereby effectively enhancing the system's privacy protection capabilities.

These measures, when implemented comprehensively across different nodes, can effectively enhance the security and stability of the V2G system. Future research should continue to explore how to achieve seamless collaboration between these nodes to further improve the system's defense capabilities. In addition, researching effective defense strategies to address the ever-changing cybersecurity challenges is also crucial.

5. Challenges for reinforcement learning in V2G

Despite the remarkable achievements of reinforcement learning in the field of V2G applications, the rapid evolution of V2G technology has led to increasingly unstable environments and a significant increase in spatial complexity [17,20]. These challenges undoubtedly bring more complex and arduous tasks for the application of reinforcement learning in the V2G field. In this section, we will delve into the core challenges faced by reinforcement learning in V2G applications, categorizing them by type such as grid fluctuation, cost optimization, and so on, from both the application and technical perspectives, and meticulously analyze the nature and impact of these challenges, as shown in Table 3.

5.1. Grid fluctuation

The fluctuations and instability of the power grid pose significant challenges to the promotion of V2G technology [196–198], mainly in two aspects: Firstly, the large-scale discharge of electric vehicles in a short period of time can easily cause huge fluctuations in the grid, leading to an increase in the peak-valley difference of grid load, thereby affecting the stability and power quality of the grid. Secondly, the disordered charging behavior of electric vehicles may exacerbate the burden on the grid during peak periods, causing an imbalance between power supply and demand, and increasing the complexity and uncertainty of grid operation. For reinforcement learning, the challenge brought by grid fluctuations lies in the need to identify and respond to rapidly changing grid conditions in real-time, which places higher demands on the adaptability and learning capabilities of the algorithm. At the same time, when dealing with grid fluctuations, reinforcement learning must also overcome the incompleteness of data and the uncertainty of predictions, factors that collectively increase the difficulty of optimization for reinforcement learning.

Addressing the fluctuation issues in electric vehicle charging management within V2G technology, researchers [28] propose a reinforcement learning method based on Deep Q-Networks, which learns the optimal charging strategies to adapt to the uncertainty of electricity prices and the heterogeneity of vehicle usage patterns, significantly reducing power costs and improving charging efficiency. To achieve collaborative optimization of power grid frequency deviation, this paper [30] proposes a Switching Integral reinforcement learning (SIRL) scheme, which effectively addresses the dynamically asymmetric frequency regulation capacity issue by integrating V2G control with power plant frequency control. In [179], researchers propose a charging control strategy based on deep reinforcement learning, utilizing its perception and learning abilities to address the uncertainties of wind power fluctuations and user demand, thereby optimizing the charging process of electric vehicles. Experimental results show that the strategy can effectively converge in uncertain environments and meet charging requirements, while also allowing users to flexibly adjust charging times. To cope with the randomness and uncertainty caused by the integration of distributed energy sources and electric vehicles in V2G microgrids, researchers [53] have developed an optimized scheduling strategy for electric vehicle microgrids driven by deep Q-learning. By utilizing the real-time learning and experience replay of reinforcement learning, this strategy effectively optimizes the charging and discharging behaviors of electric vehicles in the microgrid. This article [139]

Table 3

Summary of the challenges faced by reinforcement learning in V2G applications.

Challenges	Level	Participants	Year	Paper	Specific issues		
Grid fluctuation	Power grid		2021	A Yang, et al. [179]	① Optimization of charging control strategy. ② Uncertainties of wind power output and user demand. ③ Complex dynamic environmental interference.		
			2022	Y Wen, et al. [53]	① Randomness and uncertainty of distributed energy sources in V2G microgrid. ② Nonlinear influence caused by the mobility of EVs and the periodicity of user behavior.		
			2023	X Hao, et al. [28]	① Uncertainty of electricity prices in V2G systems. ② Heterogeneity of vehicle usage patterns. ③ Sensitivity.		
			2024	X Song, et al. [30]	① Collaborative optimization of power grid frequency deviation. ② Dynamically asymmetric frequency regulation capacity issue.		
	Application		2020	NBG Brinkel, et al. [60]	① Minimizing the cost and carbon dioxide emissions of electric vehicle charging. ② Reduction in V2G charging costs. ③ Multi-objective optimization.		
			2023	P Zhang, et al. [91]	① Optimize energy distribution decisions in V2G. ② Improve long-term average returns and cost-effectiveness ratio. ③ Minimize the electricity cost.		
			2024	M Yavuz, et al. [28]	① Reduction in overall energy costs. ② Enhance the energy self-sufficiency rate of charging stations.		
	EVs users		2023	S Wen, et al. [180]	① Uncertainty of battery degradation. ② Estimation of battery health status in V2G. ③ Interpretability.		
			2023	MM Shibli, et al. [181]	① Electric vehicle charging management strategy. ② Efficient and sustainable charging modes. ③ Effectiveness and robustness.		
			2024	J Xie, et al. [182]	① Health status of the battery. ② Rewards based on battery health information. ③ Battery long-term performance.		
			2021	L Yan, et al. [183]	① Uncertainty of the EVs charging demand Analysis of Dynamic. ② User Charging and Discharging Behavior. ③ The optimal sequential charging decision.		
			2023	T Zhu, et al. [184]	① The impact of user charging behavior on V2G system. ② Balance power control and incentive consumption. ③ User service experience.		
			2023	J Maeng, et al. [185]	① User Commute Behavior Analysis. ② Charging preference. ③ Reducing charging cost and maximizing the use of EVs battery.		
Sample efficiency	Technology		2020	F Zhang, et al. [186]	① Learning efficiency of electric vehicle charging control strategies. ② Sparse rewards in charging and discharging phases. ③ Optimal exploration of charging control strategies.		
			2020	L Hou, et al. [187]	① Design of the reward function. ② Exploration of dynamic electricity pricing optimization strategies.		
			2022	Z Ye, et al. [188]	① The computational efficiency of V2G optimization problems. ② The learning efficiency of charger control strategies. ③ Scalability in different charging scenarios.		
			2024	N Kumar, et al. [189]	① Computational efficiency of large-scale V2G infrastructure data. ② Learning efficiency in high-dimensional state spaces.		
			2024	M Yavuz, et al. [190]	① Efficient learning from small-scale sampled data. ② Scalability.		
			2021	A Omara, et al. [191]	① Data integrity attacks in V2G. ② Charging Service Stability. ③ V2G System vulnerabilities.		
			2023	A Novak, et al. [192]	① Security of V2G systems and electric vehicle owners. ② Physical tampering. ③ Replay attacks.		
			2024	J Sun, et al. [95]	① DoS attacks in V2G. ② Effectively compensate for control signals interfered by attacks.		
Cyber attack			2022	I Beil, et al. [193]	① Increasing user engagement. ② Design of attractive reward mechanism.		
			2023	P Prakash, et al. [194]	① Complexity and variability of the real V2G environment. ② High-risk trial-and-error cost.		
			2024	M Javed, et al. [195]	① Privacy-preserving protocol for V2G. ② Stability and reliability over long-term operation.		

effectively addresses the uncertainty and fluctuations in V2G by designing a reinforcement learning-based model predictive control (MPC) strategy, which combines DDPG with MPC in a dual-layer controller to enhance the robustness and responsiveness of the frequency control process in multi-microgrid systems.

Although reinforcement learning methods have shown excellent performance in Grid fluctuation applications, they are limited by the need for a large amount of high-quality data for effective training. Reinforcement learning models, especially deep reinforcement learning, need to be thoroughly trained in various scenarios to learn robust and generalizable strategies. High-quality data ensures that the model is exposed to a complete state and action space during training, enabling it to handle various uncertainties and fluctuations in the real-world environment. However, in practical applications, collecting and labeling sufficient high-quality data is both costly and time-consuming. This data requirement becomes a key challenge limiting the application of reinforcement learning methods in Grid fluctuation scenarios.

5.2. Cost optimization

The cost optimization of V2G technology is a key factor in its widespread application, as it helps reduce energy consumption and

operating costs while enhancing the flexibility and economic benefits of grid dispatch [199–201]. However, in the process of cost optimization for V2G systems, factors such as electric vehicle battery technology, the layout and capacity limitations of charging infrastructure, and the constant changes in grid regulations and policies are all key complex factors affecting cost optimization. In addition, the uncertainty of market electricity prices and the volatility of renewable energy also pose additional challenges for the cost optimization of V2G systems. In this context, reinforcement learning algorithms face severe challenges, as they must continuously adapt and update strategies to cope with these dynamic changes and uncertainties, thereby ensuring the sustainability and effectiveness of cost optimization. In order to minimize electricity costs, researchers [202] have proposed an agent model based on deep reinforcement learning by combining Virtual Network Embedding (VNE) and DRL algorithms. This model can adaptively perceive environmental characteristics and optimize energy distribution decisions, significantly improving the long-term average revenue and benefit-cost ratio of V2G applications, effectively addressing the V2G cost optimization issue. In [190], researchers utilize deep reinforcement learning with multi-double deep Q-network (DDQN) agents to optimize energy management in V2G systems for peer-to-peer (P2P) energy trading, resulting in a significant reduction in overall energy costs and

an increase in self-sufficiency rate. By formulating the environment as a Markov decision process, the method effectively eliminates the complexities and uncertainties in energy management, achieving notable cost savings. Addressing the trade-off between minimizing the cost and carbon dioxide emissions of electric vehicle charging, this article [60] proposes a method using reinforcement learning for multi-objective optimization. By considering V2G, battery degradation, and transformer capacity, it achieves a significant reduction in V2G charging costs while maintaining low computational costs. Additionally, the study shows that under increased transformer capacity limits, the additional costs or emission benefits of electric vehicle charging do not exceed the expenses and emissions associated with enhancing the power grid itself.

As V2G technology and the market continue to rapidly develop, the environment will become increasingly complex, posing even greater challenges for reinforcement learning in the optimization of V2G costs, necessitating the continuous evolution and enhancement of algorithms to adapt to this growing complexity.

Although reinforcement learning performs excellently in V2G cost optimization, its limitations cannot be ignored. RL models face stability issues in dynamic environments and uncertain electricity prices, which can lead to slow strategy updates. These limitations need to be continuously addressed and improved in future research.

5.3. Degradation of the battery

As the number of charging and discharging cycles for electric vehicles increases, the degradation of battery has become a significant barrier to inhibiting electric vehicle users from participating in V2G interactions [61,203–206]. How to optimize the management of battery charging and discharging to enhance battery efficiency while maximizing user profits is a key challenge currently faced in the development of V2G technology. In order to mitigate the impact of battery degradation, researchers [181] propose an innovative electric vehicle charging management strategy based on deep reinforcement learning technology. This strategy simulates charging facilities as a learning environment, with users playing the role of learning agents. The strategy comprehensively considers fast charging, regular charging, and V2G applications affected by battery aging. Extensive testing with actual electric vehicle charging data has confirmed the system's effectiveness and robustness in ensuring the stability of the power distribution network and meeting user charging needs. This solution ingeniously incorporates considerations for battery degradation within the reinforcement learning framework, aiming to coordinate the needs of electric power companies and electric vehicle users for efficient and sustainable charging modes. The deep reinforcement learning framework proposed by researchers [182] for EVs prosumer scheduling effectively balances the autonomy of electric vehicles with the implementation of grid policies, while considering the health of the battery. By dynamically adjusting the feasible operating range of voltage and current to optimize scheduling, the framework can improve charging efficiency without sacrificing battery life. The training reward mechanism takes into account the operational costs of electric vehicles, incentives from the power grid, and rewards based on battery health information to ensure long-term battery performance.

In the deep development of V2G technology, facing the continuous progress of battery technology, one of the key challenges will be how to adjust and optimize reinforcement learning models so that they can better adapt to the unique characteristics of new types of batteries, thereby accurately predicting and effectively addressing battery degradation issues. Meanwhile, the diversity and complexity of electric vehicle charging scenarios require that reinforcement learning algorithms must be continuously improved to more effectively handle various uncertainties and nonlinear challenges in the battery degradation process. Furthermore, as the mode of electric vehicle charging continues to innovate, reinforcement learning strategies need to be updated in sync to adapt to the impact of new charging behaviors on battery health

status [180].

Reinforcement learning for optimizing EV battery charging and discharging management may involve users' private data, which needs careful handling. Sensitive data must be dealt with cautiously to ensure privacy. Additionally, the training process may be susceptible to malicious data contamination, potentially leading to ineffective training results. The computational requirements for handling large numbers of users must also be considered. These topics will be discussed in detail in the subsequent sections.

5.4. User behavior analysis

The importance of user charging and discharging behavior analysis in the V2G field lies in its ability to uncover the deep-seated patterns and habits of electric vehicle usage [207,208], providing the power grid with valuable foresight and regulatory capabilities, thereby achieving a balance in grid load and optimizing the allocation of electrical resources. Additionally, by incorporating other human behaviors, such as human mobility, as an auxiliary validation method for user behavior, human mobility analysis already has excellent predictive models. V2G is highly correlated with such information, and by integrating this information, the model's expressive power can be enhanced [209–211]. At the same time, the application of reinforcement learning in this field faces challenges such as accurately capturing the dynamics and diversity of user behavior, designing effective reward mechanisms to guide the learning process, and ensuring the robustness and safety of the algorithms in practical applications.

In order to finely analyze the specific impact of electric vehicles on V2G microgrid loads during charging and discharging processes, researchers [184] have developed a novel model-free learning algorithm. The algorithm indirectly guides user behavior by displaying price-linked incentives at charging stations. To enhance the flexibility of the strategy's implementation, it employs a two-layer optimization framework combined with primal-dual theory, using reinforcement learning to balance power control and incentive consumption, thereby maximizing the improvement of user service experience. Rigorous theoretical analysis has confirmed that the algorithm possesses bounded sub-optimality, and simulation experiments have verified the significant effectiveness of the two-layer optimization framework in optimizing V2G systems. In addressing the analysis of dynamic user charging and discharging behavior, researchers [183] have developed a model-free deep reinforcement learning approach that learns the optimal charging control strategy through interaction with a dynamic environment to solve the electric vehicle charging scheduling problem considering dynamic user behavior and electricity prices. By adopting the continuous soft actor-critic framework and combining stages of supervised learning and reinforcement learning, this method has demonstrated its significant effectiveness in dealing with dynamic user behavior at various charging locations. In [185], the authors employ a model-free reinforcement learning approach to learn the optimal charging/discharging decisions of electric vehicle users, by analyzing their dynamic behaviors, charging preferences, and energy demands to optimize charging strategies and maximize battery utilization. Experimental results indicate that this method effectively considers the uncertainty of user behavior, thereby reducing charging costs and improving the efficiency of electric vehicle batteries.

Although the aforementioned research has made significant progress in the analysis of user behavior in V2G, reinforcement learning still faces challenges in handling large-scale heterogeneous user data and improving the real-time response capabilities of algorithms. Moreover, ensuring the long-term stability and adaptability of the algorithms in the face of variable market conditions and changes in user behavior is also a serious issue.

5.5. Sample efficiency

Sampling efficiency is a key challenge in the application of reinforcement learning to V2G, as it requires a large amount of interaction data to learn effective strategies, which is time-consuming and costly in practice. Furthermore, the dynamic and complex nature of the V2G environment exacerbates the issue of sampling efficiency [212,213], making it difficult for algorithms to quickly adapt to the ever-changing market and grid conditions. To enhance the learning efficiency and reduce the cost of trial and error for the algorithm in the optimization of charging scheduling problems, researchers have proposed a Centralized Allocation and Decentralized Execution (CADE) reinforcement learning framework [188]. By utilizing shared experience replay memory, this enables chargers to learn in parallel and optimize charging strategies, significantly improving sampling efficiency and learning efficiency, while also enhancing the scalability of the algorithm. To overcome the constraints of environmental uncertainty on the performance of multi-target charging control, in [186], the authors model the charging control as an MDP and apply the CDDPG algorithm, significantly improving the efficiency of policy learning. By using LSTM networks to analyze historical data, the identification of key patterns is accelerated, thereby speeding up the learning process. Moreover, by introducing Gaussian noise and a dual-buffer mechanism, the issue of sparse rewards is effectively addressed, further enhancing the efficiency of sampling.

Despite the extensive efforts made in previous research to improve sampling efficiency, reinforcement learning still faces numerous challenges in enhancing sampling efficiency within the diverse application scenarios of V2G. Particularly, the unpredictability brought by dynamic environmental changes forces the algorithm to continuously adapt, which increases the complexity of the learning process. Additionally, existing technologies are limited in efficiency when dealing with real-time data from large-scale V2G systems, urgently necessitating further optimization of algorithms to alleviate the computational burden of high-dimensional state spaces [189]. Moreover, how to utilize limited interactive data more effectively while maintaining learning efficacy remains a key issue to be addressed in the field of reinforcement learning for V2G [190]. Ultimately, precisely designing reward functions is crucial to the performance of the algorithm, which typically requires a deep understanding of the V2G domain and is accomplished through continuous experimentation and adjustment [87,187].

5.6. Cyber attack

Cyber attacks pose a significant threat to V2G systems, potentially leading to power supply interruptions, data leaks, and system paralysis [214–216]. In [191], the researchers point out that as electric vehicles are integrated as mobile energy storage units in smart grids, data integrity attacks have become a serious issue in V2G applications, which could significantly impact the operation of the system. In addition, Denial-of-Service (DoS) attacks pose a serious threat to the performance of grid frequency regulation based on V2G technology [56]. The challenges in defending against such attacks mainly include: the scale and speed of attack traffic often exceed the capabilities of traditional defense systems, and attackers may exploit system vulnerabilities to launch more covert attacks, making it difficult for defensive measures to be updated in a timely manner to counter new threats. Additionally, researchers [192] have pointed out that attack threats such as Eavesdropping, Man-in-the-Middle Attacks, Physical Tampering, and Replay Attacks severely impact the security of V2G systems and electric vehicle owners, necessitating urgent development of robust security measures to safeguard against these vulnerabilities.

In addressing the aforementioned cyber threats in V2G systems, reinforcement learning faces challenges such as how to quickly identify abnormal behaviors, design effective defense strategies, and maintain

learning efficiency in unstable environments. In addition, reinforcement learning models also need to overcome the difficulty of maintaining adaptability and robustness in the face of constantly evolving attack patterns. At the same time, reinforcement learning must enhance the interpretability of the model to ensure that the logic of defense decisions can be clearly articulated. Furthermore, reinforcement learning needs to integrate interdisciplinary knowledge during the implementation process to enhance comprehensive defense capabilities against cyber attacks.

5.7. Real-world deployment

Although reinforcement learning has achieved remarkable results in scheduling decisions in the V2G field and shown great potential for application, its deployment in real-world environments still faces numerous challenges [217–219]. Firstly, the complexity and variability of the real V2G environment lead to deviations between simulated training and practical application, and any missteps in model decision-making could result in unstable power supply or damage to electric vehicle batteries [194], a high-risk trial-and-error cost that is not tolerable. Secondly, the demand for real-time decision-making sets stringent requirements for the computational capabilities of reinforcement learning models. In addition, the application of reinforcement learning in V2G systems must adhere to the constraints of policies and regulations, such as the rules of the electricity market and battery safety standards. Meanwhile, increasing user engagement is also a significant challenge [193,200], requiring the design of more attractive reward mechanisms while protecting privacy [195]. Ultimately, the model needs to maintain stability and reliability over long-term operation to ensure the sustained functioning of the V2G system.

In response to these challenges, the application of reinforcement learning in real V2G environments needs to adopt more efficient training strategies [212], such as combining offline pre-training with online fine-tuning, to enhance the model's generalization ability. At the same time, the development of real-time response algorithms is required to meet the V2G system's demand for immediate decision-making. In addition, strict data privacy and security protection measures must be implemented, and continuous monitoring and evaluation of the model's performance are necessary to ensure its stability and reliability in long-term operation.

When deploying, it is essential to consider real-world latency issues and hardware limitations. For the previously mentioned applications, such as distribution network optimization and bidding strategies, which have large-scale and spatiotemporal characteristics, overall efficiency is crucial. These problems usually do not change frequently after initial computation, so during usage, reinforcement learning models with larger parameters can be chosen to improve training accuracy without worrying too much about latency and hardware limitations. For optimal scheduling and pricing strategies involving individual car owners' decisions, due to hardware limitations, information security, computational resource allocation, and efficiency become particularly critical. Therefore, it is not suitable to use complex reinforcement learning algorithms locally. Instead, sharing training parameters through the cloud and then deploying them to users' personal hardware devices can improve efficiency and reduce the burden on local hardware. For tasks requiring real-time resource allocation, such as collaborative grid frequency control, multi-energy microgrids, and energy storage scheduling, which involve complex real-world frameworks, it is necessary to consider both the representation capabilities of reinforcement learning models and the real-time nature of communication. To ensure efficiency while reducing communication, suitable reinforcement learning frameworks (such as transformer models to capture temporal changes in data and MARL frameworks to simulate real-world framework distribution) can be used to solve these problems. Additionally, in model training, soft constraints on communication data transmission (such as adding communication volume indicators as part of the loss function to reduce the model's demand for communication volume) can be implemented to achieve communication efficiency optimization while ensuring accuracy.

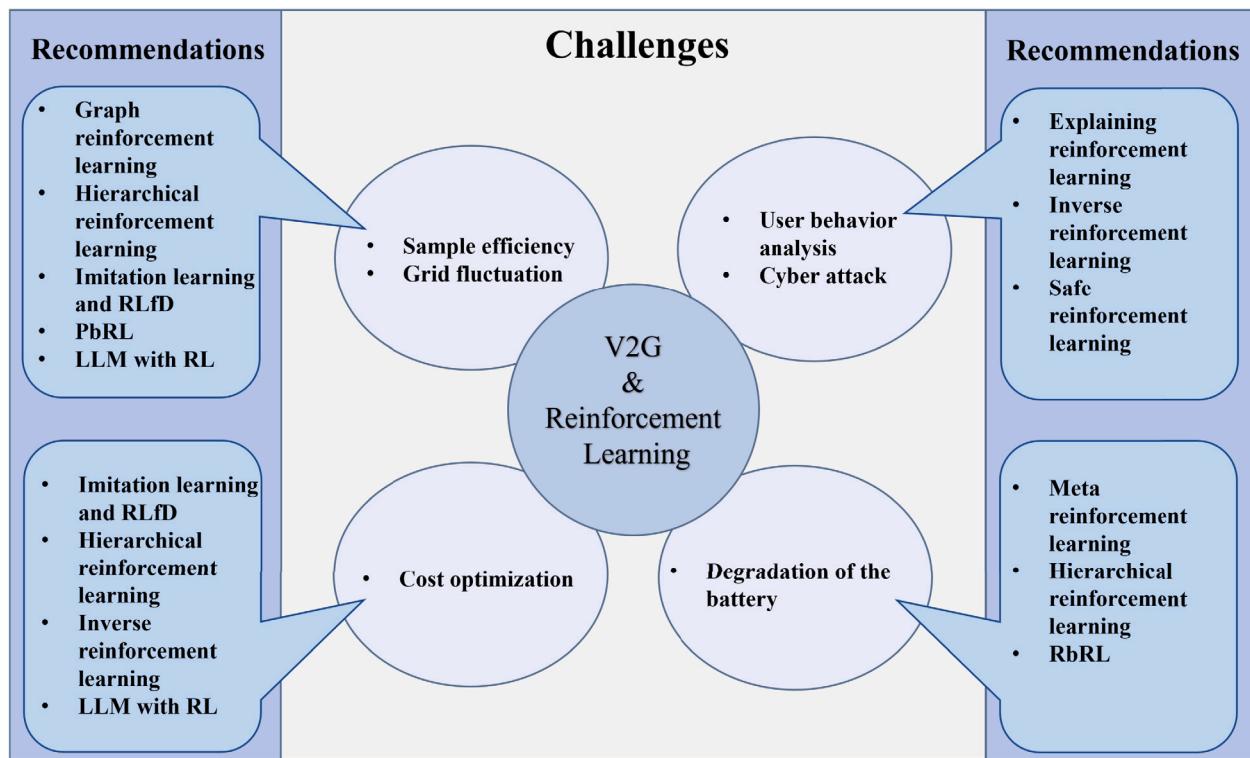


Fig. 8. Future research directions to overcome challenges in reinforcement learning for V2G.

6. Future research directions

As reinforcement learning continues to develop, its powerful compatibility is demonstrated through its combination with different methods. It can serve as the main component of training and assist in enhancing other methods. In the V2G context, reinforcement learning has already shown feasible applications in various scenarios. These applications can be further improved if combined with more novel methods.

In this section, we consider the existing issues in V2G and, based on these issues, summarize some of the current cutting-edge reinforcement learning algorithms. According to the characteristics of these methods, we explore their potential applications in V2G scenarios. We hope that this paper will not only advance the development of reinforcement learning itself but also promote its practical application, especially in the V2G domain. Fig. 8 describes the potential challenges V2G may face and the cutting-edge reinforcement learning algorithms that can be applied to address them.

The subsequent sections will discuss various directions of reinforcement learning, including explaining reinforcement learning, inverse reinforcement learning, and many other reinforcement learning methods. These sections will showcase the advantages and disadvantages of each direction and their potential applications in the V2G field.

It is important to note that the classifications in this paper do not rigorously cover all the cutting-edge aspects of reinforcement learning. However, we discuss some of the forefront technologies in these areas and the problems they are suited to solve. When addressing V2G, we will analyze the specific problems that different reinforcement learning approaches are best suited to tackle. We hope this will be beneficial for future research.

6.1. Explaining reinforcement learning

In the process of promoting and implementing V2G systems, explainability is crucial for enhancing investor confidence and user trust.

In reinforcement learning, causal explainability aims to elucidate the decision-making process and behavior of agents through causal relationships. This approach not only focuses on the actions taken by the agent in specific states but also on how these actions influence environmental variables through causal chains, ultimately leading to rewards or outcomes. We hope that this explainability can help people understand why an agent makes certain decisions, thereby improving reinforcement learning algorithms through these explanations. This can enhance the overall safety and controllability of the system. Additionally, such explainable methods can increase user trust [220], which is extremely important both from practical application and societal perspectives.

For V2G systems, we aim to ensure that while businesses and individuals benefit, the stability of the power grid remains unaffected. [221] demonstrates the feasibility of explainable reinforcement learning in large-scale applications. We speculate that explainable reinforcement learning can also be applied in the deployment of V2G systems to assist in decision-making. By employing explainable reinforcement learning, we can enhance the acceptance and trust in V2G. Businesses can confidently invest more funds based on the insights provided by intelligent agents, and the public can safely use the V2G network to increase their earnings, thereby facilitating the adoption of V2G. Regarding grid stability, grid operators can use explainability to understand the behavior of intelligent agents in maintaining the grid and improve their decision-making processes to reduce grid fluctuation. We believe that explainable reinforcement learning can ensure stable operation and increase overall profitability in the construction and utilization of V2G systems. COViz [222] analyzes user behavior to generate multiple behavior chains, broadening explainability and helping people understand the reasons behind different decisions during the charging and discharging process. SVERL [62] improves upon traditional Shapley values and proposes a new algorithm to explain agent performance. It anticipates that meaningful explanations can be generated across various domains, aligning with and complementing

Table 4

Summary of future research directions in reinforcement learning for V2G.

Algorithm	Year	Paper	Potential applications					
			Sample efficiency	Degradation of the battery	Cost optimization	User behavior analysis	Cyber attack	Grid fluctuation
Explaining reinforcement learning	2021	J. Druce et al. [220]	✓	✗	✗	✓	✓	✓
	2022	Kravaris et al. [221]	✓	✗	✗	✗	✓	✗
	2023	Y. Amitai et al. [222]	✓	✗	✓	✗	✓	✓
	2023	Beechey et al. [62]	✓	✗	✗	✓	✓	✓
Meta reinforcement learning	2016	Y. Dan et al. [223]	✓	✓	✓	✓	✗	✗
	2020	Y. Tianhe et al. [224]	✓	✓	✗	✗	✗	✗
	2023	H. HongCai et al. [63]	✓	✓	✗	✓	✗	✓
Inverse reinforcement learning	2000	Andrew Y. et al. [225]	✓	✗	✓	✓	✗	✓
	2021	S. Arora et al. [226]	✓	✗	✓	✓	✗	✓
	2023	R. Serra et al. [64]	✓	✗	✓	✓	✗	✓
	2023	Metelli et al. [227]	✓	✗	✓	✓	✗	✓
	2023	Swamy et al. [228]	✓	✗	✓	✓	✗	✓
Hierarchical reinforcement learning	1999	Thomas G. et al. [229]	✓	✗	✗	✓	✗	✓
	2022	H. Buysse et al. [230]	✓	✗	✗	✓	✗	✓
	2023	Z. Hailong et al. [65]	✓	✓	✓	✓	✗	✓
	2023	I. Jendoubi et al. [231]	✓	✗	✓	✓	✗	✓
	2024	L. Kun et al. [232]	✓	✗	✓	✗	✗	✗
Imitation learning and RLfD	2017	Chelsea Finn et al. [233]	✓	✗	✓	✓	✗	✓
	2019	J. Mingxuan et al. [234]	✓	✗	✓	✗	✗	✗
	2024	Zare et al. [235]	✓	✗	✓	✗	✗	✗
Graph reinforcement learning	2017	Barret Zoph et al. [236]	✓	✗	✗	✗	✗	✓
	2018	V. Zambaldi [237]	✓	✗	✗	✗	✗	✓
	2023	J. Fan et al. [238]	✓	✗	✗	✗	✓	✓
	2023	Munikoti et al. [239]	✓	✗	✗	✓	✗	✓
LLM with reinforcement learning	2024	J. Wen et al. [240]	✓	✗	✓	✓	✗	✓
	2024	L. Wen et al. [241]	✓	✗	✓	✓	✗	✓
	2024	Caldwell et al. [242]	✓	✗	✓	✓	✗	✓
Other reinforcement learning	2017	C. Finn et al. [243]	✓	✓	✓	✗	✗	✗
	2023	J. Dai et al. [244]	✓	✗	✗	✓	✓	✗
	2024	Metcalf et al. [245]	✓	✗	✗	✗	✗	✗
	2024	D. White et al. [246]	✓	✗	✗	✓	✗	✗
	2024	Z. Zhu et al. [247]	✓	✓	✓	✗	✗	✓

human intuition. These allow companies to offer various V2G solutions based on user preferences, thereby increasing public trust in V2G.

Currently, although explainable reinforcement learning has improved interpretability and made it easier for humans to understand, its performance is often affected. When using this method, it is important to consider the balance between interpretability and performance.

6.2. Meta reinforcement learning

In the deployment and usage of V2G systems, strategies often encounter different environments with similar distributions. Therefore, scalability and generalizability of these strategies are crucial. In Deep Reinforcement Learning (DRL), learning algorithms rely heavily on extensive interactions between the agent and the environment, leading to high training costs. When the environment changes, previously

learned optimal strategies become obsolete, necessitating retraining for the new environment. Meta-reinforcement learning [223] introduces meta-learning, allowing the agent to quickly adapt to new tasks using historical experience from past tasks with only a few samples from the new environment. Additionally, as the model evolves, its generalization capability continues to increase [224].

For enterprises, using reinforcement learning alone to deploy V2G systems in different regions is highly inefficient. Not only do they need to undergo repetitive training processes, but data acquisition also becomes increasingly difficult and costly. Meta-reinforcement learning can train a more general model that can be adapted to different regions with minimal samples and training resources, thereby improving sample efficiency and reducing actual research and operational costs. By proposing a framework to enhance the generalization of prior experience, He et al. [63] use Gaussian quantization variational autoencoders for task context and skill clustering, making it more suitable for V2G deployment in previously unseen cities.

However, the high generalization capability of meta-reinforcement learning implies that its training costs are high, the training process is more complex and unstable, and it requires the precondition that generalized tasks have the same distribution. Although we generally assume that within V2G, similar tasks under different conditions still have similar distributions, if the distributions differ significantly, the model's performance may degrade considerably. Alternatively, predictive methods can be employed to mitigate the impact of DoS attacks, as proposed by Sun et al. [56], who developed a multi-step predictive reinforcement learning V2G control (MPRLC) scheme to accurately predict and adapt to changes in the power system caused by DoS attacks.

This enhanced discussion aims to provide a more comprehensive understanding of the potential challenges and solutions for deploying V2G systems using meta-reinforcement learning and other advanced reinforcement learning methods.

6.3. Inverse reinforcement learning

In V2G systems, directly constructing reward functions often requires significant human and material resources, and the data utilization efficiency is relatively low. Inverse reinforcement learning [64, 225] (IRL) aims to infer the reward function of an agent from observed behavior. Unlike traditional reinforcement learning, IRL deduces the underlying reward function by observing the agent's actions. For example, in autonomous driving tasks, an agent can infer the reward function from human driver behavior and then learn strategies based on this reward function. By learning from expert behavior in different environments, IRL offers higher interpretability and can be applied to various settings, providing better generalization. This allows the agent to adapt more effectively to environmental changes and make optimal decisions.

Currently, the high generalization and interpretability of IRL make it suitable for addressing the issue of low sample efficiency in V2G systems. It also shows potential in optimizing deployment costs and analyzing user behavior.

However, IRL requires high-quality input data to avoid ambiguity in the solutions [226]. It necessitates multiple runs of reinforcement learning algorithms to find the optimal strategy, with computational costs increasing as the scale grows. These factors make inverse reinforcement learning unsuitable for addressing V2G issues that require high real-time responsiveness, such as user explanations and grid fluctuation, especially in environments with significant and frequent changes. By framing IRL as the problem of estimating the feasible reward set, the reward region compatible with expert behavior [227] avoids the ambiguity in reward selection. By leveraging the state distribution of experts, Swamy et al. [228] can alleviate the global exploration component of RL subroutines, theoretically providing exponential acceleration. These advancements suggest that IRL may be capable of

addressing these V2G issues in the future.

In summary, while IRL currently demonstrates potential in enhancing the generalizability and interpretability of models within V2G systems, it is also associated with high computational costs and requires a precondition of high-quality data. Future research could focus on addressing these challenges to further optimize the effectiveness and applicability of IRL in real-time V2G scenarios.

6.4. Hierarchical reinforcement learning

V2G systems often involve large-scale models, and the issues of dimensionality explosion and sparse rewards in reinforcement learning urgently need to be addressed. Hierarchical reinforcement learning (HRL) decomposes complex reinforcement learning tasks into multiple sub-tasks, solving each sub-task individually to complete the overall task [229]. This approach is more effective in handling tasks with long time spans and complex decision-making. HRL addresses the sparse reward problem by dividing strategies into different hierarchical sub-strategies, where each sub-strategy receives rewards passed down from the higher level during the learning process. This hierarchical structure not only improves sample efficiency but also reduces computational complexity, enabling the model to adapt and solve complex tasks more quickly.

HRL is suitable for solving complex, long-duration tasks [65, 232]. In V2G systems, pre-training high-level strategies in different grid environments can accelerate the optimization and updating of charging and discharging strategies in new environments [231]. Additionally, using HRL models for load prediction and management can enhance grid stability and efficiency. By predicting user charging behavior through a hierarchical structure, HRL can optimize grid resource allocation.

Despite these theoretical advantages, designing a reasonable hierarchical structure and balancing exploration and exploitation across different levels remain critical issues. Currently, HRL frameworks are often tailored to specific applications [230]. For V2G systems, the development of a well-designed HRL framework is still an ongoing research topic.

The application of HRL in V2G systems presents the potential to enhance model efficiency and effectiveness, but there are significant challenges to overcome. Future research should focus on refining HRL frameworks to address these challenges, ensuring they can be applied effectively to V2G scenarios. Additionally, exploring how HRL can be integrated with other advanced techniques to further improve its performance and adaptability in V2G systems could provide valuable insights.

6.5. Imitation learning and RLfD

The complexity of V2G tasks makes direct exploration through reinforcement learning very challenging. Additionally, the high cost and risk associated with direct exploration can lead to severe consequences. Both imitation learning and Reinforcement Learning from Demonstrations (RLfD) are techniques that learn from demonstration data. The former focuses on imitating expert behavior, while the latter also considers the rewards from the environment. By introducing a demonstration guidance term into the reward, the agent is guided to explore like an expert [234, 235].

By using imitation learning and RLfD, we can enhance the efficiency of data learning in models. A common approach to improve learning accuracy and speed is to initially employ imitation learning followed by reinforcement learning algorithms. This is particularly effective in scenarios like V2G, such as grid fluctuation management, where rapid and efficient learning is required, leveraging the advantage of large data volumes. For user data filled with impurities (e.g., data on charging and discharging strategy selection), RLfD [233] can be effectively utilized.

However, it is important to note that imitation learning and RLfD have different application scenarios. Imitation learning heavily relies on the quality and quantity of expert demonstration data. If the demonstration data is imperfect, it may lead to suboptimal strategies. On the other hand, RLfD can explore and compensate for the shortcomings of demonstration data through reinforcement learning [248]. For V2G, a large amount of data is generated daily, but if imitation learning is to be used, some filtering and cleaning are required. Conversely, RLfD is recommended for data processing.

6.6. Graph reinforcement learning

In V2G systems, both the grid distribution and the operation distribution of electric vehicles naturally conform to a graph structure. Therefore, we can consider introducing graph neural networks (GNNs) to handle these graph-structured data, thereby improving the generalization ability of reinforcement learning models. Graph reinforcement learning (GRL) mainly has two combination methods: GNN-enhanced RL and RL-enhanced GNN. GNN-enhanced RL [237,249] utilizes GNNs to model relationships between agents in multi-agent reinforcement learning, thereby improving cooperation and competition capabilities. RL-enhanced GNN [236] employs reinforcement learning for neural architecture search (NAS).

By modeling these graph-structured data, GRL can better capture and utilize the relationships and interactions between EVs, thereby optimizing the overall system performance. Additionally, GRL has strong real-time optimization capabilities, making it suitable for network architectures to address network attacks [238] in V2G systems, thereby enhancing system robustness.

However, the GRL model itself is quite complex, leading to high computational complexity and training difficulty, which may pose challenges in development and deployment. Although V2G inherently has a large amount of data, if the data is not cleaned, it may contain noise and be incomplete, which will affect the final performance of the GRL model. GRL models usually lack interpretability [239], which may not be ideal for user-facing decisions.

To address these challenges, future research should focus on simplifying GRL model structures to reduce computational complexity and enhance interpretability. Moreover, exploring effective data pre-processing techniques to clean and enrich the data can significantly improve GRL model performance. Additionally, integrating GRL with other advanced methodologies could further enhance its applicability and robustness in V2G systems.

6.7. LLM with reinforcement learning

LLMs are machine learning models trained on vast amounts of text data for natural language processing tasks. They can understand and generate human language. In V2G systems, they can be applied to analyze human behavior and address complex, large-scale issues due to their superior performance in capturing information over long time sequences and large data volumes [240].

However, using LLMs requires a substantial amount of labeled data to achieve good results, which incurs high manual costs for labeling in V2G. Additionally, deploying large models demands significant resources, making it impractical to generate and deploy a large model specifically for V2G. There are also concerns about whether the computational power can meet the real-time processing needs for large-scale problems. Reinforcement Learning (RL), through interaction with the environment, can generate large amounts of data with reward values. Enhancing RL with LLMs can leverage LLMs' ability to capture spatiotemporal information while using existing models for training, thus reducing deployment costs [241]. Fine-tuning LLMs with reinforcement learning is also a common training method [242]. Combining both methods can improve each other's representation capabilities, and the choice of method should be based on specific needs and resources in

practical applications.

LLMs have high resource demands and are more difficult and costly to maintain in the face of attacks. When using commercial LLMs for training, sensitive user data may face privacy and security challenges [250]. Therefore, using LLMs for security protection is not recommended.

The high resource demands and potential privacy concerns associated with LLMs make them challenging for use in V2G systems. Future research could focus on developing more efficient and secure methods to leverage the capabilities of LLMs in V2G applications. This could involve exploring techniques to reduce the amount of labeled data required, enhancing the computational efficiency of LLMs, and addressing privacy and security concerns.

6.8. Other reinforcement learning

Apart from the aforementioned reinforcement learning methods, there are still many other sub-methods that we cannot cover one by one. Here, we briefly introduce a few reinforcement learning algorithms that might be applicable in the V2G field. Safe reinforcement learning [244] introduces constrained Markov decision processes (CMDPs) and employs safety constraints, such as intrusion detection in cybersecurity, to prevent the selection of unexpected actions during the exploration process. Preference-based Reinforcement Learning (PbRL) [245] and Rating-based Reinforcement Learning (RbRL) [246] analyze human preference behavior samples to optimize strategies. PbRL compares different samples to select better strategies, while RbRL learns and optimizes based on sample ratings. This allows users to train agents according to their preferences, better planning their charging and discharging schemes. Additionally, [247] mentions using transfer reinforcement learning (RL) to transfer knowledge across different tasks or environments, leveraging previously learned experiences to accelerate the learning process of new tasks. We hope that in the future, V2G can utilize these methods to improve its usability and service experience, thereby promoting the overall development of the V2G industry.

At the same time, the algorithms mentioned above are not mutually exclusive. Combining them may improve the performance of the algorithms to some extent [243]. However, considering the real-time requirements and model complexity constraints in V2G deployment, we do not recommend arbitrarily combining these algorithms. It is essential to select the method that most suitably aligns with specific requirements. Here, we have provided a brief summary in [Table 4](#) for easy reference.

7. Conclusion

In this paper, we have conducted an in-depth survey on the application of reinforcement learning in the V2G field, systematically reviewing and analyzing the development progress of reinforcement learning in addressing the complex scheduling issues in V2G. Based on the analysis of the development trajectory of reinforcement learning in the V2G field, a classification framework for the application of reinforcement learning methods is designed from a structured perspective. Moreover, from the perspective of different stakeholders, we have deeply analyzed eight key application scenarios and their effectiveness of reinforcement learning in the V2G field. This paper also discusses the main risks and challenges that reinforcement learning faces in V2G applications from both application and technical dimensions, and offers targeted solutions and future development directions.

The core contribution of this review lies in filling the gap in the literature analysis of reinforcement learning in the V2G field. Furthermore, through in-depth analysis, we have revealed the significant effectiveness of reinforcement learning in mitigating the fluctuations and disturbances caused by the random integration of electric vehicles into the grid. This fully demonstrates the efficiency and advantages of

reinforcement learning in addressing the dynamic uncertainty scheduling optimization challenges in V2G systems, thereby providing valuable references for enhancing the stability and reliability of V2G systems during their future large-scale development.

This review provides an important reference for the systematic analysis of the application of reinforcement learning in the V2G field, offering strong guidance to researchers and engineers both in theory and practice. It aims to further attract more attention from readers to the development of reinforcement learning in V2G scheduling optimization, and to collectively promote and enhance the digitalization and intelligence of V2G systems.

CRediT authorship contribution statement

Hongbin Xie: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Ge Song:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **Zhuoran Shi:** Visualization, Investigation. **Jingyuan Zhang:** Writing – original draft, Investigation. **Zhenjia Lin:** Supervision, Methodology. **Qing Yu:** Supervision, Funding acquisition. **Hongdi Fu:** Visualization. **Xuan Song:** Supervision, Project administration, Funding acquisition, Conceptualization. **Haoran Zhang:** Supervision, Project administration, Methodology, Conceptualization.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant numbers 52472316, 52341203, and 52461160297. Additionally, it received funding from the China Postdoctoral Science Foundation, China (Certificate Number: 2024M760083). Furthermore, this research was supported by the National Key Research and Development Project of China (2021YFB1714400). Support was also provided by the High-performance Computing Platform of Peking University.

Data availability

No data was used for the research described in the article.

References

- [1] Ukoba Kingsley, Medupin Rasaq Olawale, Yoro Kelvin O, Eterigho-Ikelegbe Oraoghene, Jen Tien-Chien. Role of the fourth industrial revolution in attaining universal energy access and net-zero objectives. *Energy* 360 2024;1:100002.
- [2] Nna Théodore Patrice Nna, Sapnken Flavian Emmanuel, Tamba Jean Gaston. Economic determinants and forecasting of electricity demand in Cameroon: A policy-driven approach using multilinear regression. *Energy* 360 2024;100013.
- [3] Yeung Chakhung, Wang Jianguo, Du Yaping, Cao Jinxin, Zhou Quan, Du Zhen-tao, et al. Clearness index cluster analysis for photovoltaic weather classification based on solar irradiation measurement data: Theory and application. *Energy* 360 2024;2:100010.
- [4] Zhang Haoran, Zhao Pengjun, Zhang Wenqian, Zeng Zhenzhong, Wu Yegang, Li Peiran, et al. Promoting sustainable solar-energy development in harmony with global threatened bird ranges. *Nexus* 2024;1(2).
- [5] Azim Mohseni Naser, Bayati Navid, Ebel Thomas. Energy management strategies of hybrid electric vehicles: A comparative review. *IET Smart Grid* 2024;7(3):191–220.
- [6] Sui Yi, Zhang Haoran, Song Xuan, Shao Fengjing, Yu Xiang, Shibusaki Ryosuke, et al. GPS data in urban online ride-hailing: A comparative analysis on fuel consumption and emissions. *J Clean Prod* 2019;227:495–505.
- [7] Wang Feiran, Zhuang Lu, Cheng Shasha, Zhang Yue, Cheng Shulei. Spatiotemporal variation and convergence analysis of China's regional energy security. *Renew Sustain Energy Rev* 2024;189:113923.
- [8] Yao Yuhao, Zhang Haoran, Lin Lifeng, Lin Guixu, Shibusaki Ryosuke, Song Xuan, et al. Internet of things positioning technology based intelligent delivery system. *IEEE Trans Intell Transp Syst* 2022;24(11):12862–76.
- [9] Dik Abdullah, Omer Siddig, Boukhanouf Rabah. Electric vehicles: V2G for rapid, safe, and green EV penetration. *Energies* 2022;15(3):803.
- [10] Wohlschläger Daniela, Kigle Stephan, Schindler Vanessa, Neitz-Regett Anika, Fröhling Magnus. Environmental effects of vehicle-to-grid charging in future energy systems—A prospective life cycle assessment. *Appl Energy* 2024;370:123618.
- [11] Mojumder Md Rayid Hasan, Ahmed Antara Fahmida, Hasanuzzaman Md, Alamri Basem, Alsharef Mohammad. Electric vehicle-to-grid (V2G) technologies: Impact on the power grid and battery. *Sustainability* 2022;14(21):13856.
- [12] Sufyan Muhammad, Rahim Nasrudin Abd, Muhammad Munir Azam, Tan Chia Kwang, Raihan Siti Rohani Sheikh, Bakar Ab Halim Abu. Charge coordination and battery lifecycle analysis of electric vehicles with V2G implementation. *Electr Power Syst Res* 2020;184:106307.
- [13] Abubakr Hussein, Lashab Abderezak, Vasquez Juan C, Mohamed Tarek Hassan, Guerrero Josep M. Novel V2G regulation scheme using dual-PSS for PV island microgrid. *Appl Energy* 2023;340:121012.
- [14] Viswanath Belagavi, Khated Dheeraj Kumar, Padhy Narayana Prasad. Vehicle-to-grid (V2G) technology for optimal grid demand profile and sustainable energy management: A state-of-the-art survey. In: 2023 IEEE 3rd international conference on sustainable energy and future electric transportation. IEEE; 2023, p. 1–6.
- [15] Oad Ammar, Ahmad Hafiz Gulfi, Talpur Mir Sajjad Hussain, Zhao Chenglin, Pervez Amjad. Green smart grid predictive analysis to integrate sustainable energy of emerging V2G in smart city technologies. *Optik* 2023;272:170146.
- [16] Yu Qing, Wang Zhen, Song Yancun, Shen Xinwei, Zhang Haoran. Potential and flexibility analysis of electric taxi fleets V2G system based on trajectory data and agent-based modeling. *Appl Energy* 2024;355:122323.
- [17] Wan Muchun, Yu Heyang, Huo Yingning, Yu Kan, Jiang Quanyuan, Geng Guangchao. Feasibility and challenges for vehicle-to-grid in electricity market: A review. *Energies* 2024;17(3):679.
- [18] Hossain Sagar, Rokonuzzaman Md, Rahman Kazi Sajedur, Habib AKM Ahasan, Tan Wen-Shan, Mahmud Md, et al. Grid-vehicle-grid (g2v2g) efficient power transmission: an overview of concept, operations, benefits, concerns, and future challenges. *Sustainability* 2023;15(7):5782.
- [19] Zhang Haoran, Song Xuan, Xia Tianqi, Yuan Meng, Fan Zipei, Shibusaki Ryosuke, et al. Battery electric vehicles in Japan: Human mobile behavior based adoption potential analysis and policy target response. *Appl Energy* 2018;220:527–35.
- [20] Escoto Marc, Guerrero Antoni, Ghorbani Elnaz, Juan Angel A. Optimization challenges in vehicle-to-grid (V2G) systems and artificial intelligence solving methods. *Appl Sci* 2024;14(12):5211.
- [21] Hussein Mahmoud M, Mohamed Tarek Hassan, Mahmoud Mohamed Metwally, Aljohania Mansour, Mosaad Mohamed I, Hassan Ammar M. Regulation of multi-area power load frequency in presence of V2G scheme. *PLoS One* 2023;18(9):e0291463.
- [22] Li Xinzhou, Tan Yitong, Liu Xinxin, Liao Qiangqiang, Sun Bo, Cao Guangyu, et al. A cost-benefit analysis of V2G electric vehicles supporting peak shaving in shanghai. *Electr Power Syst Res* 2020;179:106058.
- [23] Wu Shengcheng, Pang Aiping. Optimal scheduling strategy for orderly charging and discharging of electric vehicles based on spatio-temporal characteristics. *J Clean Prod* 2023;392:136318.
- [24] Khezri Rahmat, Steen David, Wikner Evelina, et al. Optimal V2G scheduling of an EV with calendar and cycle aging of battery: An MILP approach. *IEEE Trans Transp Electrification* 2024.
- [25] Kaelbling Leslie Pack, Littman Michael L, Moore Andrew W. Reinforcement learning: A survey. *J Artificial Intelligence Res* 1996;4:237–85.
- [26] Arulkumaran Kai, Deisenroth Marc Peter, Brundage Miles, Bharath Anil Anthony. Deep reinforcement learning: A brief survey. *IEEE Signal Process Mag* 2017;34(6):26–38.
- [27] Wang Xu, Wang Sen, Liang Xingxing, Zhao Dawei, Huang Jincai, Xu Xin, et al. Deep reinforcement learning: A survey. *IEEE Trans Neural Networks Learn Syst* 2022;35(4):5064–78.
- [28] Hao Xu, Chen Yue, Wang Hewu, Wang Han, Meng Yu, Gu Qing. A V2G-oriented reinforcement learning framework and empirical study for heterogeneous electric vehicle charging management. *Sustain Cities Soc* 2023;89:104345.
- [29] Qiu Dawei, Wang Yi, Hua Weiqi, Strbac Goran. Reinforcement learning for electric vehicle applications in power systems: A critical review. *Renew Sustain Energy Rev* 2023;173:113052.
- [30] Song Xin, Sun Jian, Tan Shanwen, Ling Rui, Chai Yi, Guerrero Josep M. Cooperative grid frequency control under asymmetric V2G capacity via switched integral reinforcement learning. *Int J Electr Power Energy Syst* 2024;155:109679.
- [31] Alfaverh Fayiz, Denai Mouloud, Sun Yichuang. Optimal vehicle-to-grid control for supplementary frequency regulation using deep reinforcement learning. *Electr Power Syst Res* 2023;214:108949.
- [32] Sun Jian, Qi Guanqiu, Chai Yi, Zhu Zhiqin, Guerrero Josep M. An adaptive V2G capacity-based frequency regulation scheme with integral reinforcement learning against dos attacks. *IEEE Trans Smart Grid* 2023;15(1):834–47.

- [33] Sovacool Benjamin K, Kester Johannes, Noel Lance, de Rubens Gerardo Zarazua. Actors, business models, and innovation activity systems for vehicle-to-grid (V2G) technology: A comprehensive review. *Renew Sustain Energy Rev* 2020;131:109963.
- [34] Bibak Bijan, Tekiner-Moğulkoç Hatice. A comprehensive analysis of vehicle to grid (V2G) systems and scholarly literature on the application of such systems. *Renew Energy Focus* 2021;36:1–20.
- [35] İnci Mustafa, Savrun Murat Mustafa, Çelik Özgür. Integrating electric vehicles as virtual power plants: A comprehensive review on vehicle-to-grid (V2G) concepts, interface topologies, marketing and future prospects. *J Energy Storage* 2022;55:105579.
- [36] Viswanath Belagavi, Khatod Dheeraj Kumar, Padhy Narayana Prasad. Vehicle-to-grid (V2G) optimization for grid peak demand reduction and decarbonization: A state-of-the-art review. In: 2022 IEEE 10th power India international conference. IEEE; 2022, p. 1–6.
- [37] Panchanathan Suresh, Vishnuram Pradeep, Rajamanickam Narayananmoorthi, Bajaj Mohit, Blazek Vojtech, Prokop Lukas, et al. A comprehensive review of the bidirectional converter topologies for the vehicle-to-grid system. *Energies* 2023;16(5):2503.
- [38] Vishnu Gayathry, Kaliyaperumal Deepa, Jayaprakash Ramprabhakar, Karthick Alagar, Kumar Chinnaian V, Ghosh Aritra. Review of challenges and opportunities in the integration of electric vehicles to the grid. *World Electr Veh J* 2023;14(9):259.
- [39] Uribe-Pérez Noelia, Gonzalez-Garrido Amaia, Gallarreta Alexander, Justel Daniel, González-Pérez Mikel, González-Ramos Jon, et al. Communications and data science for the success of vehicle-to-grid technologies: Current state and future trends. *Electronics* 2024;13(10):1940.
- [40] Goncaruc Andrei, De Cauwer Cedric, Sapountzoglou Nikolaos, Van Kriekinge Gilles, Huber Dominik, Messagie Maarten, et al. The barriers to widespread adoption of vehicle-to-grid: A comprehensive review. *Energy Rep* 2024;12:27–41.
- [41] Chen Guangjie, Zhang Zhaoyun. Control strategies, economic benefits, and challenges of vehicle-to-grid applications: Recent trends research. *World Electr Veh J* 2024;15(5):190.
- [42] Ganesh Akhil Hannegudda, Xu Bin. A review of reinforcement learning based energy management systems for electrified powertrains: Progress, challenge, and potential solution. *Renew Sustain Energy Rev* 2022;154:111833.
- [43] Yang Ting, Zhao Liyuan, Li Wei, Zomaya Albert Y. Reinforcement learning in sustainable energy and electric systems: A survey. *Annu Rev Control* 2020;49:145–63.
- [44] Zhang Dongxia, Han Xiaoqing, Deng Chunyu. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE J Power Energy Syst* 2018;4(3):362–70.
- [45] Li Yuanzheng, Yu Chaofan, Shahidehpour Mohammad, Yang Tao, Zeng Zhi-gang, Chai Tianyou. Deep reinforcement learning for smart grid operations: algorithms, applications, and prospects. *Proc IEEE* 2023.
- [46] Hassouna Mohamed, Holzhüter Clara, Lytaev Pawel, Thomas Josephine, Sick Bernhard, Scholz Christoph. Graph reinforcement learning in power grids: A survey. 2024, arXiv preprint [arXiv:2407.04522](https://arxiv.org/abs/2407.04522).
- [47] Vamvakas Dimitrios, Michailidis Panagiotis, Korkas Christos, Kosmatopoulos Elias. Review and evaluation of reinforcement learning frameworks on smart grid applications. *Energies* 2023;16(14):5326.
- [48] Zhang Yubao, Chen Xin, Zhang Yuchen. Transfer deep reinforcement learning-based large-scale V2G continuous charging coordination with renewable energy sources. 2022, arXiv preprint [arXiv:2210.07013](https://arxiv.org/abs/2210.07013).
- [49] Bellemare Marc G, Dabney Will, Munos Rémi. A distributional perspective on reinforcement learning. 2017, arXiv preprint [arXiv:1707.06887](https://arxiv.org/abs/1707.06887).
- [50] Silver David, Huang Aja, Maddison Chris J, Guez Arthur, Sifre Laurent, van den Driessche George, et al. Mastering the game of go with deep neural networks and tree search. *Nature* 2016;529(7587):484–9.
- [51] Chen Lili, Lu Kevin, Rajeswaran Aravind, Lee Kimin, Grover Aditya, Laskin Michael, et al. Decision transformer: Reinforcement learning via sequence modeling. 2021, arXiv preprint [arXiv:2106.01345](https://arxiv.org/abs/2106.01345).
- [52] Rao Yingqing, Yang Jun, Xiao Jinxing, Xu Bingyan, Liu Wenjing, Li Yonghui. A frequency control strategy for multimicrogrids with V2G based on the improved robust model predictive control. *Energy* 2021;222:119963.
- [53] Wen Yuxin, Fan Peixiao, Hu Jia, Ke Song, Wu Fuzhang, Zhu Xu. An optimal scheduling strategy of a microgrid with V2G based on deep Q-learning. *Sustainability* 2022;14(16):10351.
- [54] Lu Ying, Liang Yanchang, Ding Zhao-hao, Wu Qiuwei, Ding Tao, Lee Wei-Jen. Deep reinforcement learning-based charging pricing for autonomous mobility-on-demand system. *IEEE Trans Smart Grid* 2021;13(2):1412–26.
- [55] Lee Sangyoon, Choi Dae-Hyun. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach. *Appl Energy* 2021;304:117754.
- [56] Sun Jian, Wang Xin, Qi Guanqiu, Li Huaqing, Wang Huiwei, Vasquez Juan C, et al. Resilient frequency regulation for DoS attack intensity adaptation via predictive reinforcement V2G control learning. *IEEE Trans Smart Grid* 2024.
- [57] Zhang Yang, Zhang Zhengfeng, Yang Qingyu, An Dou, Li Donghe, Li Ce. EV charging bidding by multi-DQN reinforcement learning in electricity auction market. *Neurocomputing* 2020;397:404–14.
- [58] Jeong Seong Il, Choi Dae-Hyun. Electric vehicle user data-induced cyber attack on electric vehicle charging station. *IEEE Access* 2022;10:55856–67.
- [59] Akbarian Amirhossein, Bahrami Mahdi, Ahmadi Mehdi, Vakilian Mehdi, Lehtonen Matti. Detection of cyber attacks to mitigate their impacts on the manipulated EV charging prices. *IEEE Trans Transp Electrification* 2024.
- [60] Brinkel NBG, Schram WL, AlSaif TA, Lamproupolous I, Van Sark WGJHM. Should we reinforce the grid? Cost and emission optimization of electric vehicle charging under different transformer limits. *Appl Energy* 2020;276:115285.
- [61] Sagaria Shemin, van der Kam Mart, Boström Tobias. Vehicle-to-grid impact on battery degradation and estimation of V2G economic compensation. *Appl Energy* 2024;377:124546.
- [62] Beechey Daniel, Smith Thomas MS, Şimşek Özgür. Explaining reinforcement learning with Shapley values. In: Krause Andreas, Brunskill Emma, Cho Kyunghyun, Engelhardt Barbara, Sabato Sivan, Scarlett Jonathan, editors. Proceedings of the 40th international conference on machine learning. Proceedings of machine learning research, vol. 202, PMLR; 2023, p. 2003–14.
- [63] He Hongcai, Zhu Anjie, Liang Shuang, Chen Feiyu, Shao Jie. Decoupling meta-reinforcement learning with Gaussian task contexts and skills. 2023, arXiv preprint [arXiv:2312.06518](https://arxiv.org/abs/2312.06518).
- [64] Ruiz-Serra Jaime, Harré Michael S. Inverse reinforcement learning as the algorithmic basis for theory of mind: current methods and open problems. *Algorithms* 2023;16(2):68.
- [65] Zhang Hailong, Peng Jiankun, Dong Hanxuan, Tan Huachun, Ding Fan. Hierarchical reinforcement learning based energy management strategy of plug-in hybrid electric vehicle for ecological car-following process. *Appl Energy* 2023;333:120599.
- [66] Chae Jongseong, Han Seungyul, Jung Whiyoung, Cho Myungsik, Choi Sungho, Sung Youngchul. Robust imitation learning against variations in environment dynamics. In: International conference on machine learning. PMLR; 2022, p. 2828–52.
- [67] Qiu Dawei, Wang Yi, Ding Zhaohao, Strbac Goran. Graph reinforcement learning for carbon-aware electric vehicles in power-transport networks. *IEEE Trans Smart Grid* 2024.
- [68] Zhang Qianzhi, Yan Jinyue, Gao H Oliver, You Fengqi. A systematic review on power systems planning and operations management with grid integration of transportation electrification at scale. *Adv Appl Energy* 2023;11:100147.
- [69] Dong Zihang, Zhang Xi, Zhang Ning, Kang Chongqing, Strbac Goran. A distributed robust control strategy for electric vehicles to enhance resilience in urban energy systems. *Adv Appl Energy* 2023;9:100115.
- [70] Wan Zhiqiang, Li Hepeng, He Haibo, Prokhorov Danil. Model-free real-time EV charging scheduling based on deep reinforcement learning. *IEEE Trans Smart Grid* 2019;10(5):5246–57.
- [71] Chen Xiangyu, Leung Ka-Cheong. Fictitious self-play for vehicle-to-grid game with imperfect information. In: ICC 2019-2019 IEEE international conference on communications. IEEE; 2019, p. 1–6.
- [72] Dorokhova Marina, Martinson Yann, Ballif Christophe, Wyrtsch Nicolas. Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. *Appl Energy* 2021;301:117504.
- [73] Monfaredi Farzam, Shayeghi Hossein, Siano Pierluigi. Multi-agent deep reinforcement learning-based optimal energy management for grid-connected multiple energy carrier microgrids. *Int J Electr Power Energy Syst* 2023;153:109292.
- [74] Wang Fan, Gao Jie, Li Mushu, Zhao Lian. Autonomous PEV charging scheduling using dyna-Q reinforcement learning. *IEEE Trans Veh Technol* 2020;69(11):12609–20.
- [75] Kiaee Farkhondeh. Integration of electric vehicles in smart grid using deep reinforcement learning. In: 2020 11th international conference on information and knowledge technology. 2020, p. 40–4.
- [76] Qiu Dawei, Ye Yujian, Papadaskalopoulos Dimitrios, Strbac Goran. A deep reinforcement learning method for pricing electric vehicles with discrete charging levels. *IEEE Trans Ind Appl* 2020;56(5):5901–12.
- [77] Li Hepeng, Wan Zhiqiang, He Haibo. Constrained EV charging scheduling based on safe deep reinforcement learning. *IEEE Trans Smart Grid* 2020;11(3):2427–39.
- [78] Yan Linfang, Chen Xia, Zhou Jianyu, Chen Yin, Wen Jinyu. Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. *IEEE Trans Smart Grid* 2021;12(6):5124–34.
- [79] Lee Sangyoon, Choi Dae-Hyun. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach. *Appl Energy* 2021;304:117754.
- [80] Wang Yi, Qiu Dawei, Strbac Goran, Gao Zhiwei. Coordinated electric vehicle active and reactive power control for active distribution networks. *IEEE Trans Ind Informatics* 2023;19(2):1611–22.
- [81] Tao Yuechuan, Qiu Jing, Lai Shuying, Zhang Xian, Wang Yunqi, Wang Guibin. A human-machine reinforcement learning method for cooperative energy management. *IEEE Trans Ind Inform* 2022;18(5):2974–85.

- [82] Tao Yuechuan, Qiu Jing, Lai Shuying. Deep reinforcement learning based bidding strategy for EVAs in local energy market considering information asymmetry. *IEEE Trans Ind Inform* 2022;18(6):3831–42.
- [83] Li Sichen, Hu Weihao, Cao Di, Dragičević Tomislav, Huang Qi, Chen Zhe, et al. Electric vehicle charging management based on deep reinforcement learning. *J Mod Power Syst Clean Energy* 2022;10(3):719–30.
- [84] Qiu Dawei, Wang Yi, Zhang Tingqi, Sun Mingyang, Strbac Goran. Hybrid multiagent reinforcement learning for electric vehicle resilience control towards a low-carbon transition. *IEEE Trans Ind Inform* 2022;18(11):8258–69.
- [85] Qiu Dawei, Wang Yi, Sun Mingyang, Strbac Goran. Multi-service provision for electric vehicles in power-transportation networks towards a low-carbon transition: A hierarchical and hybrid multi-agent reinforcement learning approach. *Appl Energy* 2022;313:118790.
- [86] Wang Jianing, Guo Chunlin, Yu Changshu, Liang Yanchang. Virtual power plant containing electric vehicles scheduling strategies based on deep reinforcement learning. *Electr Power Syst Res* 2022;205:107714.
- [87] Rahman Saidur, Punt Linda, Ardakanian Omid, Ghiasi Yashar, Tan Xiaoqin. On efficient operation of a V2G-enabled virtual power plant: when solar power meets bidirectional electric vehicle charging. In: Proceedings of the 9th ACM international conference on systems for energy-efficient buildings, cities, and transportation. 2022, p. 119–28.
- [88] Fu Liyue, Wang Tong, Song Min, Zhou Yuhu, Gao Shan. Electric vehicle charging scheduling control strategy for the large-scale scenario with non-cooperative game-based multi-agent reinforcement learning. *Int J Electr Power Energy Syst* 2023;153:109348.
- [89] Pokhrel Shiva Raj, Hossain Mohammad Belayet, Walid Anwar. Modeling practically private wireless vehicle to grid system with federated reinforcement learning. *IEEE Trans Serv Comput* 2024;17(3):1044–55.
- [90] Fan Peixiao, Ke Song, Yang Jun, Li Rui, Li Yonghui, Yang Shaobo, Liang Jifeng, Fan Hui, Li Tiecheng. A load frequency coordinated control strategy for multimicrogrids with V2G based on improved MA-DDPG. *Int J Electr Power Energy Syst* 2023;146:108765.
- [91] Zhang Peiyang, Chen Ning, Kumar Neeraj, Abualigah Laith, Guizani Mohsen, Duan Youxiang, Wang Jian, Wu Sheng. Energy allocation for vehicle-to-grid settings: A low-cost proposal combining DRL and VNE. *IEEE Trans Sustain Comput* 2024;9(1):75–87.
- [92] Zhang Borui, Li Chaojie, Hu Boyang, Li Xiangyu, Wang Rui, Dong Zhaoyang. Graph reinforcement learning for securing critical loads by E-mobility. In: Luo Biao, Cheng Long, Wu Zheng-Guang, Li Hongyi, Li Chaojie, editors. *Neural information processing*. Singapore: Springer Nature Singapore; 2024, p. 303–14.
- [93] Dong Jiawei, Yassine Abdulsalam, Armitage Andy, Hossain M Shamim. Multi-agent reinforcement learning for intelligent V2G integration in future transportation systems. *IEEE Trans Intell Transp Syst* 2023;24(12):15974–83.
- [94] Fonseca Tiago, Ferreira Luis, Cabral Bernardo, Severino Ricardo, Praca Isabel. EnergAlze: Multi agent deep deterministic policy gradient for vehicle to grid energy management. 2024, arXiv preprint [arXiv:2404.02361](#).
- [95] Sun Jian, Wang Xin, Qi Guanqiu, Li Huaqing, Wang Huiwei, Vasquez Juan C, et al. Resilient frequency regulation for DoS attack intensity adaptation via predictive reinforcement V2G control learning. *IEEE Trans Smart Grid* 2024;1.
- [96] Song Xin, Sun Jian, Tan Shanwen, Ling Rui, Chai Yi, Guerrero Josep M. Cooperative grid frequency control under asymmetric V2G capacity via switched integral reinforcement learning. *Int J Electr Power Energy Syst* 2024;155:109679.
- [97] Sun Jian, Qi Guanqiu, Chai Yi, Zhu Zhiqin, Guerrero Josep M. An adaptive V2G capacity-based frequency regulation scheme with integral reinforcement learning against DoS attacks. *IEEE Trans Smart Grid* 2024;15(1):834–47.
- [98] Otterlo Martijn, Wiering Marco. Reinforcement learning and Markov decision processes. *Reinf Learning: State the Art* 2012;3–42.
- [99] Sutton Richard S, Barto Andrew G. Reinforcement learning: An introduction. 2nd ed.. The MIT Press; 2018.
- [100] Geramifard Alborz, Walsh Thomas J, Tellex Stefanie. A tutorial on linear function approximators for dynamic programming and reinforcement learning. Hanover, MA, USA: Now Publishers Inc.; 2013.
- [101] Szepesvári Csaba, Jozsef Attila, Littman Michael. Generalized Markov decision processes: Dynamic-programming and reinforcement-learning algorithms. Technical report, USA: Brown University; 1996.
- [102] Sutton Richard S. Learning to predict by the methods of temporal differences. *Mach Learn* 1988;3(1):9–44.
- [103] Ajagekar Akshay, Mattson Neil S, You Fengqi. Energy-efficient AI-based control of semi-closed greenhouses leveraging robust optimization in deep reinforcement learning. *Adv Appl Energy* 2023;9:100119.
- [104] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Graves Alex, Antonoglou Ioannis, Wierstra Daan, et al. Playing atari with deep reinforcement learning. 2013, arXiv preprint [arXiv:1312.5602](#).
- [105] Hessel Matteo, Modayil Joseph, van Hasselt Hado, Schaul Tom, Ostrovski Georg, Dabney Will, et al. Rainbow: Combining improvements in deep reinforcement learning. 2017, arXiv preprint [arXiv:1710.02298](#).
- [106] van Hasselt Hado, Guez Arthur, Silver David. Deep reinforcement learning with double Q-learning. 2015, arXiv preprint [arXiv:1509.06461](#).
- [107] Schaul Tom, Quan John, Antonoglou Ioannis, Silver David. Prioritized experience replay. 2016, arXiv preprint [arXiv:1511.05952](#).
- [108] Wang Ziyu, Schaul Tom, Hessel Matteo, van Hasselt Hado, Lanctot Marc, de Freitas Nando. Dueling network architectures for deep reinforcement learning. 2016, arXiv preprint [arXiv:1511.06581](#).
- [109] Dabney Will, Ostrovski Georg, Silver David, Munos Rémi. Implicit quantile networks for distributional reinforcement learning. 2018, arXiv preprint [arXiv:1806.06923](#).
- [110] Dabney Will, Rowland Mark, Bellemare Marc G, Munos Rémi. Distributional reinforcement learning with quantile regression. 2017, arXiv preprint [arXiv:1710.10044](#).
- [111] Sutton Richard S, McAllester David, Singh Satinder, Mansour Yishay. Policy gradient methods for reinforcement learning with function approximation. In: Solla S, Leen T, Müller K, editors. In: *Advances in neural information processing systems*, vol. 12. MIT Press; 1999.
- [112] Zhang Junzi, Kim Jongho, O'Donoghue Brendan, Boyd Stephen. Sample efficient reinforcement learning with REINFORCE. 2020, arXiv preprint [arXiv:2010.11364](#).
- [113] Schulman John, Levine Sergey, Moritz Philipp, Jordan Michael I, Abbeel Pieter. Trust region policy optimization. 2017, arXiv preprint [arXiv:1502.05477](#).
- [114] Schulman John, Wolski Filip, Dhariwal Prafulla, Radford Alec, Klimov Oleg. Proximal policy optimization algorithms. 2017, arXiv preprint [arXiv:1707.06347](#).
- [115] Heess Nicolas, Dhruba TB, Sriram Srinivasan, Lemmon Jay, Merel Josh, Wayne Greg, et al. Emergence of locomotion behaviours in rich environments. 2017, arXiv preprint [arXiv:1707.02286](#).
- [116] Konda Vijay, Tsitsiklis John. Actor-critic algorithms. In: Solla S, Leen T, Müller K, editors. In: *Advances in neural information processing systems*, vol. 12. MIT Press; 1999.
- [117] Mnih Volodymyr, Badia Adrià Puigdomènec, Mirza Mehdi, Graves Alex, Lillicrap Timothy P, Harley Tim, et al. Asynchronous methods for deep reinforcement learning. 2016, arXiv preprint [arXiv:1602.01783](#).
- [118] Lillicrap Timothy P, Hunt Jonathan J, Pritzel Alexander, Heess Nicolas, Erez Tom, Tassa Yuval, et al. Continuous control with deep reinforcement learning. 2019, arXiv preprint [arXiv:1509.02971](#).
- [119] Fujimoto Scott, van Hoof Herke, Meger David. Addressing function approximation error in actor-critic methods. 2018, arXiv preprint [arXiv:1802.09477](#).
- [120] Haarnoja Tuomas, Zhou Aurick, Abbeel Pieter, Levine Sergey. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. 2018, arXiv preprint [arXiv:1801.01290](#).
- [121] Gu Shangling, Yang Long, Du Yali, Chen Guang, Walter Florian, Wang Jun, et al. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Trans Pattern Anal Mach Intell* 2024.
- [122] Zhan Yusen, Ammar Haitham Bou, Taylor Matthew E. Scalable lifelong reinforcement learning. *Pattern Recognit* 2017;72:407–18.
- [123] Li Wenzhe, Luo Hao, Lin Zichuan, Zhang Chongjie, Lu Zongqing, Ye Deheng. A survey on transformers in reinforcement learning. 2023, arXiv preprint [arXiv:2301.03044](#).
- [124] Chebotar Yevgen, Vuong Quan, Irpan Alex, Hausman Karol, Xia Fei, Lu Yao, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive Q-functions. 2023, arXiv preprint [arXiv:2309.10150](#).
- [125] He Kaiming, Sun Jian, Tang Xiaou. Single image haze removal using dark channel prior. In: 2009 IEEE conference on computer vision and pattern recognition. 2009, p. 1956–63.
- [126] Wang Yi, Qiu Dawei, Strbac Goran. Multi-agent reinforcement learning for electric vehicles joint routing and scheduling strategies. In: 2022 IEEE 25th international conference on intelligent transportation systems. 2022, p. 3044–9.
- [127] Kostrikov Ilya, Nair Ashvin, Levine Sergey. Offline reinforcement learning with implicit Q-learning. 2021, arXiv preprint [arXiv:2110.06169](#).
- [128] Zhou Shiyang, Ren Weiya, Ren Xiaoguang, Wang Yanzhen, Yi Xiaodong. Independent deep deterministic policy gradient reinforcement learning in cooperative multiagent pursuit games. In: Farkaš Igor, Masulli Paolo, Otte Sebastian, Wermter Stefan, editors. *Artificial neural networks and machine learning – ICANN 2021*. Cham: Springer International Publishing; 2021, p. 625–37.
- [129] Rashid Tabish, Samvelyan Mikayel, de Witt Christian Schroeder, Farquhar Gregory, Foerster Jakob, Whiteson Shimon. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. 2018, arXiv preprint [arXiv:1803.11485](#).
- [130] Lowe Ryan, Wu Yi, Tamar Aviv, Harb Jean, Abbeel Pieter, Mordatch Igor. Multi-agent actor-critic for mixed cooperative-competitive environments. 2020, arXiv preprint [arXiv:1706.02275](#).
- [131] Yu Chao, Velu Akash, Vinitsky Eugene, Gao Jiaxuan, Wang Yu, Bayen Alexandre, et al. The surprising effectiveness of PPO in cooperative, multi-agent games. 2022, arXiv preprint [arXiv:2103.01955](#).
- [132] Müller Robert, Turalic Hasan, Phan Thomy, Kölle Michael, Nüßlein Jonas, Linnhoff-Popien Claudia. ClusterComm: Discrete communication in decentralized MARL using internal representation clustering. 2024, arXiv preprint [arXiv:2401.03504](#).

- [133] Wang Yi, Qiu Dawei, Teng Fei, Strbac Goran. Two-stage TSO-DSO services provision framework for electric vehicle coordination. *IEEE Trans Power Syst* 2024;1:13.
- [134] Iqbal Sheeraz, Xin Ai, Jan Mishkat Ullah, Salman Salman, Zaki Atta Ul Munim, Rehman Haseeb Ur, et al. V2G strategy for primary frequency control of an industrial microgrid considering the charging station operator. *Electronics* 2020;9(4):549.
- [135] Irfan Muhammad, Deilami Sara, Huang Shujuan, Tahir Tayyab, Veetil Binesh Putthen. Optimizing load frequency control in microgrid with vehicle-to-grid integration in Australia: Based on an enhanced control approach. *Appl Energy* 2024;366:123317.
- [136] Wang Jingxiang, Wang Zhaojian, Yang Bo, Liu Feng, Wei Wei, Guan Xinping. V2G for frequency regulation service: a stackelberg game approach considering endogenous uncertainties. *IEEE Trans Transp Electrification* 2024.
- [137] Yoo Yeong, Al-Shawesh Yousef, Tchagang Alain. Coordinated control strategy and validation of vehicle-to-grid for frequency control. *Energies* 2021;14(9):2530.
- [138] Fan Peixiao, Ke Song, Yang Jun, Wen Yuxin, Xie Lilong, Li Yonghui, et al. A frequency cooperative control strategy for multimicrogrids with EVs based on improved evolutionary-deep reinforcement learning. *Int J Electr Power Energy Syst* 2024;159:109991.
- [139] Fan Peixiao, Yang Jun, Ke Song, Wen Yuxin, Li Yonghui, Xie Lilong. Load frequency control strategy for islanded multimicrogrids with V2G dependent on learning-based model predictive control. *IET Gener Transm Distrib* 2023;17(21):4763–80.
- [140] Dong Jiawei, Yassine Abdulsalam, Armitage Andy, Hossain M Shamim. Multi-agent reinforcement learning for intelligent V2G integration in future transportation systems. *IEEE Trans Intell Transp Syst* 2023.
- [141] Pan Chaofeng, Li Yuan, Wang Jian, Liang Jun, Jinyama Ho. Research on multi-lane energy-saving driving strategy of connected electric vehicle based on vehicle speed prediction. *Green Energy and Intelligent Transportation* 2023;2(6):100127.
- [142] Yin Wanjun, Jia Leilei, Ji Jianbo. Energy optimal scheduling strategy considering V2G characteristics of electric vehicle. *Energy* 2024;294:130967.
- [143] Pan Weiqi, Zou Bokang, Li Fengtao, Luo Yifu, Chen Qirui, Zhang Yuansi, et al. Collaborative operation optimization scheduling strategy of electric vehicle and steel plant considering V2G. *Energies* 2024;17(11):2448.
- [144] Shibli Mostafa, Ismail Loay, Massoud Ahmed. Electric vehicles charging management using machine learning considering fast charging and vehicle-to-grid operation. *Energies* 2021;14(19):6199.
- [145] Alqahtani Mohammed, Scott Michael J, Hu Mengqi. Dynamic energy scheduling and routing of a large fleet of electric vehicles using multi-agent reinforcement learning. *Comput Ind Eng* 2022;169:108180.
- [146] Alfaverh Fayiz, Denai Mouloud, Sun Yichuang. Electrical vehicle grid integration for demand response in distribution networks using reinforcement learning. *IET Electr Syst Transp* 2021;11(4):348–61.
- [147] Wang Yi, Qiu Dawei, He Yinglong, Zhou Quan, Strbac Goran. Multi-agent reinforcement learning for electric vehicle decarbonized routing and scheduling. *Energy* 2023;284:129335.
- [148] Zhang Shulei, Jia Runda, Pan Hengxin, Cao Yankai. A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid. *Appl Energy* 2023;348:121490.
- [149] Ren Hui, Zhang Aiwei, Li Wei. Study on optimal V2G pricing strategy under multi-aggregator competition based on game theory. In: 2019 IEEE sustainable power and energy conference. IEEE; 2019, p. 1027–32.
- [150] Fang Xiaohan, Wang Jinkuan, Song Guanru, Han Yinghua, Zhao Qiang, Cao Zhiao. Multi-agent reinforcement learning approach for residential microgrid energy scheduling. *Energies* 2019;13(1):123.
- [151] Qiu Dawei, Ye Yujian, Papadaskalopoulos Dimitrios, Strbac Goran. A deep reinforcement learning method for pricing electric vehicles with discrete charging levels. *IEEE Trans Ind Appl* 2020;56(5):5901–12.
- [152] Liu Dunnan, Wang Weiye, Wang Lingxiao, Jia Heping, Shi Mengshu. Dynamic pricing strategy of electric vehicle aggregators based on DDPG reinforcement learning algorithm. *IEEE Access* 2021;9:21556–66.
- [153] Chuang Yu-Chieh, Chiu Wei-Yu. Deep reinforcement learning based pricing strategy of aggregators considering renewable energy. *IEEE Trans Emerg Top Comput Intell* 2021;6(3):499–508.
- [154] Tao Yuechuan, Qiu Jing, Lai Shuying. Deep reinforcement learning based bidding strategy for EVAs in local energy market considering information asymmetry. *IEEE Trans Ind Inform* 2021;18(6):3831–42.
- [155] Narayanan Ajay, Misra Prasant, Ojha Ankush, Bandhu Vivek, Ghosh Supratim, Vasan Arunchandar. A reinforcement learning approach for electric vehicle routing problem with vehicle-to-grid supply. 2022, arXiv preprint arXiv:2204.05545.
- [156] Kumari Aparna, Trivedi Mihir, Tanwar Sudeep, Sharma Gulshan, Sharma Ravi. SV2G-ET: A secure vehicle-to-grid energy trading scheme using deep reinforcement learning. *Int Trans Electr Energy Syst* 2022;2022(1):9761157.
- [157] Pokhrel Shiva Raj, Hossain Mohammad Belayet, Walid Anwar. Modeling practically private wireless vehicle to grid system with federated reinforcement learning. *IEEE Trans Serv Comput* 2023.
- [158] Hossain Mohammad Belayet, Pokhrel Shiva Raj, Vu Hai L. Efficient and private scheduling of wireless electric vehicles charging using reinforcement learning. *IEEE Trans Intell Transp Syst* 2023;24(4):4089–102.
- [159] Liu Ding, Zeng Peng, Cui Shijie, Song Chunhe. Deep reinforcement learning for charging scheduling of electric vehicles considering distribution network voltage stability. *Sensors* 2023;23(3):1618.
- [160] Kumar Polamarasetty P, Nuvvula Ramakrishna SS, Tan Chai Ching, Al-Salman Ghafar Ahmed, Guntreddi Venkataramana, Raj V Arun, et al. Energy-aware vehicle-to-grid (V2G) scheduling with reinforcement learning for renewable energy integration. In: 2024 12th international conference on smart grid. IEEE; 2024, p. 345–9.
- [161] Pan Weiqi, Yu Xiaorong, Guo Zishan, Qian Tao, Li Yang. Online EVs vehicle-to-grid scheduling coordinated with multi-energy microgrids: A deep reinforcement learning-based approach. *Energies* 2024;17(11):2491.
- [162] Chen Longxiang, He Huan, Jing Rui, Xie Meina, Ye Kai. Energy management in integrated energy system with electric vehicles as mobile energy storage: An approach using bi-level deep reinforcement learning. *Energy* 2024;307:132757.
- [163] Jang Moon-Jong, Oh Eunsung. Deep-reinforcement-learning-based vehicle-to-grid operation strategies for managing solar power generation forecast errors. *Sustainability* 2024;16(9).
- [164] Liu Peng, Liu Zhe, Fu Tingting, Garg Sahil, Kaddoum Georges, Hassan Mohammad Mehedi. Optimization of multi-vehicle charging and discharging efficiency under time constraints based on reinforcement learning. *Alex Eng J* 2024;105:724–35.
- [165] Chen Ping, Han Lu, Xin Guoyu, Zhang Aiwei, Ren Hui, Wang Fei. Game theory based optimal pricing strategy for V2G participating in demand response. *IEEE Trans Ind Appl* 2023;59(4):4673–83.
- [166] Saha Manoj, Thakur Sidhartha Sankar, Bhattacharya Aniruddha. Optimal scheduling of electric vehicles: Leveraging grid-to-vehicle (G2V) and vehicle-to-grid (V2G) prices: A game-theoretic approach. In: 2023 IEEE 3rd international conference on sustainable energy and future electric transportation. IEEE; 2023, p. 1–6.
- [167] Lotfi Siamak, Sedighizadeh Mostafa, Abbasi Rezvan, Hosseiniyan Seyed Hossein. Vehicle-to-grid bidding for regulation and spinning reserve markets: A robust optimal coordinated charging approach. *Energy Rep* 2024;11:925–36.
- [168] Lyu Ruike, Guo Hongye, Zheng Kedi, Sun Mingyang, Chen Qixin. Co-optimizing bidding and power allocation of an EV aggregator providing real-time frequency regulation service. *IEEE Trans Smart Grid* 2023;14(6):4594–606.
- [169] Lei Xiang, Yu Hang, Shao Ziyun, Jian Linni. Optimal bidding and coordinating strategy for maximal marginal revenue due to V2G operation: Distribution system operator as a key player in China's uncertain electricity markets. *Energy* 2023;283:128354.
- [170] Li Chuan, Carta Daniele, Benigni Andrea. EV charging station placement considering V2G and human factors in multi-energy systems. *IEEE Trans Smart Grid* 2024.
- [171] Li Shuangqi, Zhao Pengfei, Gu Chenghong, Bu Siqi, Chung Edward, Tian Zhongbei, et al. Energy storage capacity estimation and charging management for electric vehicle grid integration. *CSEE J Power Energy Syst* 2024;1–10.
- [172] Xie Shengli, Zhong Weifeng, Xie Kan, Yu Rong, Zhang Yan. Fair energy scheduling for vehicle-to-grid networks using adaptive dynamic programming. *IEEE Trans Neural Networks Learn Syst* 2016;27(8):1697–707.
- [173] Nogay H Selcuk. Estimating the aggregated available capacity for vehicle to grid services using deep learning and nonlinear autoregressive neural network. *Sustain Energy, Grids Networks* 2022;29:100590.
- [174] Li Mince, Wang Yujie, Peng Pei, Chen Zonghai. Toward efficient smart management: a review of modeling and optimization approaches in electric vehicle-transportation network-grid integration. *Green Energy and Intelligent Transportation* 2024;100181.
- [175] Das Soumyabrata, Thakur Padmanabh, Singh Asheesh K, Singh SN. Optimal management of vehicle-to-grid and grid-to-vehicle strategies for load profile improvement in distribution system. *J Energy Storage* 2022;49:104068.
- [176] Ahmed Shafiq, Anisi Mohammad Hosseini. Optimizing V2G dynamics: An AI-enhanced secure protocol for energy management in industrial cyber-physical systems. *IEEE Trans Ind Cyber- Phys Syst* 2024.
- [177] Mekkaoui Kheireddine. Enhancing V2G network security: A novel cockroach behavior-based machine learning classifier to mitigate MitM and DoS attacks. *Adv Electr Comput Eng* 2024;24(2).
- [178] Sepehrzad Reza, Faraji Mohammad Javad, Al-Durra Ahmed, Sadabadi Mahdieh S. Enhancing cyber-resilience in electric vehicle charging stations: A multi-agent deep reinforcement learning approach. *IEEE Trans Intell Transp Syst* 2024;25(11):18049–62.
- [179] Yang Anyun, Sun Hongbin, Zhang Xiao. Deep reinforcement learning strategy for electric vehicle charging considering wind power fluctuation.. *J Eng Sci Technol Rev* 2021;14(3).
- [180] Wen Shuang, Lin Ni, Huang Shengxu, Wang Zhenpo, Zhang Zhaosheng. Lithium battery health state assessment based on vehicle-to-grid (V2G) real-world data and natural gradient boosting model. *Energy* 2023;284:129246.
- [181] Shibli Mostafa M, Ismail Loay S, Massoud Ahmed M. Electric vehicles charging management using deep reinforcement learning considering vehicle-to-grid operation and battery degradation. *Energy Rep* 2023;10:494–509.

- [182] Xie Jiahang, Vorobev Petr, Yang Rufan, Nguyen Hung Dinh. Battery health-informed and policy-aware deep reinforcement learning for EV-facilitated distribution grid optimal policy. *IEEE Trans Smart Grid* 2024.
- [183] Yan Linfang, Chen Xia, Zhou Jianyu, Chen Yin, Wen Jinyu. Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. *IEEE Trans Smart Grid* 2021;12(6):5124–34.
- [184] Zhu Tianxiang, Zhang Xiaoxi, Duan Jingpu, Zhou Zhi, Chen Xu. A budget-aware incentive mechanism for vehicle-to-grid via reinforcement learning. In: 2023 IEEE/ACM 31st international symposium on quality of service. IEEE; 2023, p. 1–10.
- [185] Maeng Julie, Min Daiki, Kang Yuncheol. Intelligent charging and discharging of electric vehicles in a vehicle-to-grid system using a reinforcement learning-based approach. *Sustain Energy, Grids Networks* 2023;36:101224.
- [186] Zhang Feiye, Yang Qingyu, An Dou. CDDPG: A deep-reinforcement-learning-based approach for electric vehicle charging control. *IEEE Internet Things J* 2020;8(5):3075–87.
- [187] Hou Luyang, Ma Shuai, Yan Jun, Wang Chun, Yu Jia Yuan. Reinforcement mechanism design for electric vehicle demand response in microgrid charging stations. In: 2020 international joint conference on neural networks. IEEE; 2020, p. 1–8.
- [188] Ye Zuzhao, Gao Yuanqi, Yu Nanpeng. Learning to operate an electric vehicle charging station considering vehicle-grid integration. *IEEE Trans Smart Grid* 2022;13(4):3038–48.
- [189] Kumar Navin, Sood Sandeep Kumar, Saini Munish. Internet of vehicles (IoV) based framework for electricity demand forecasting in V2G. *Energy* 2024;297:131199.
- [190] Yavuz Mete, Kivanç Ömer Cihan. Optimization of a cluster-based energy management system using deep reinforcement learning without affecting prosumer comfort: V2X technologies and peer-to-peer energy trading. *IEEE Access* 2024.
- [191] Omara Ahmed, Kantarci Burak. On the impact of data integrity attacks on vehicle-to-microgrid services. In: 2021 IEEE 26th international workshop on computer aided modeling and design of communication links and networks. IEEE; 2021, p. 1–7.
- [192] Novak Andrej, Ivanov Alexei. Network security vulnerabilities in smart vehicle-to-grid systems identifying threats and proposing robust countermeasures. *J Artif Intell Mach Learn Manag* 2023;7(1):48–80.
- [193] Beil Ian, Whittemore Luke, Shrestha Anik. Utility experience with vehicle-to-grid regulatory and technology challenges, and the final hurdles to large-scale V2G deployment. In: 2022 IEEE power & energy society general meeting. IEEE; 2022, p. 1–5.
- [194] Prakash P. Deployment strategies of EV school buses with vehicle to grid (V2G) in the US school system. *Int J Supply Chain Manag* 2023;12(5):22–33.
- [195] Javed Mubeen, Arslan Akram Muhammad, Noor Mian Adnan, Kumari Saru. On the security of a novel privacy-preserving authentication scheme for V2G networks. *Secur Priv* 2024;7(2):e357.
- [196] Jie Bo, Baba Jumpei, Kumada Akiko. Contribution to V2G system frequency regulation by charging/discharging control of aggregated EV group. *IEEE Trans Ind Appl* 2023.
- [197] Bibak Bijan, Tekiner-Mogulkoc Hatice. Influences of vehicle to grid (V2G) on power grid: An analysis by considering associated stochastic parameters explicitly. *Sustain Energy, Grids Networks* 2021;26:100429.
- [198] Amir Mohammad, Haque Ahteshamul, et al. Integration of EVs aggregator with microgrid and impact of V2G power on peak regulation. In: 2021 IEEE 4th international conference on computing, power and communication technologies. IEEE; 2021, p. 1–6.
- [199] Cheikh-Mohamed Saleh, Celik Berk, Sechilaru Manuela, Loemont Fabrice. PV-powered charging station with energy cost optimization via V2G services. *Appl Sci* 2023;13(9):5627.
- [200] Al-obaidi Abdullah Azhar, Farag Hany EZ. Optimal design of V2G incentives and V2G-capable electric vehicles parking lots considering cost-benefit financial analysis and user participation. *IEEE Trans Sustain Energy* 2023;15(1):454–65.
- [201] Daramola Alex S, Ahmadi Seyed Ehsan, Marzband Mousa, Ikpehai Augustine. A cost-effective and ecological stochastic optimization for integration of distributed energy resources in energy networks considering vehicle-to-grid and combined heat and power technologies. *J Energy Storage* 2023;57:106203.
- [202] Zhang Peiyi, Chen Ning, Kumar Neeraj, Abualigah Laith, Guizani Mohsen, Duan Youxiang, et al. Energy allocation for vehicle-to-grid settings: A low-cost proposal combining DRL and VNE. *IEEE Trans Sustain Comput* 2023.
- [203] Harnischmacher Christine, Markefke Lukas, Brendel Alfred Benedikt, Kolbe Lutz. Two-sided sustainability: Simulating battery degradation in vehicle to grid applications within autonomous electric port transportation. *J Clean Prod* 2023;384:135598.
- [204] Preis Valentin, Biedenbach Florian. Assessing the incorporation of battery degradation in vehicle-to-grid optimization models. *Energy Inform* 2023;6(Suppl 1):33.
- [205] Lee Hyeon-Gyu, Choi Jong-Won, Ryu Seong-Taek, Lee Kyu-Jin. Life degradation of lithium-ion batteries under vehicle-to-grid operations based on a multi-physics model. *Int J Automot Technol* 2024;1–12.
- [206] Leippi Andre, Fleschutz Markus, Davis Kevin, Klingler Anna-Lena, Murphy Michael D. Optimizing electric vehicle fleet integration in industrial demand response: Maximizing vehicle-to-grid benefits while compensating vehicle owners for battery degradation. *Appl Energy* 2024;374:123995.
- [207] Shin Gwang-Su, Kim Ho-Young, Mahseredjian Jean, Kim Chul-Hwan. Smart vehicle-to-grid operation of power system based on EV user behavior. *J Electr Eng Technol* 2024;1–12.
- [208] Chen Jinhui, Li Wenyi. Analysis of V2G stakeholder interaction behavior based on evolutionary game theory. In: 2024 3rd international conference on energy, power and electrical technology. IEEE; 2024, p. 1238–43.
- [209] Yao Yuhan, Zhang Haoran, Shi Xiaodan, Chen Jinyu, Li Wenjing, Song Xuan, et al. LTP-net: Life-travel pattern based human mobility signature identification. *IEEE Trans Intell Transp Syst* 2023;24(12):14306–19.
- [210] Liu Yapan. Urban scale vehicle-to-building-to-grid integration leveraging human mobility modeling [Ph.D. thesis]. Syracuse University; 2024.
- [211] Yao Yuhan, Zhang Haoran, Chen Jinyu, Li Wenjing, Shibasaki Ryosuke, Song Xuan. Mobility tableau: Human mobility similarity measurement for city dynamics. *IEEE Trans Intell Transp Syst* 2023;24(7):7108–21.
- [212] Jiao Zihao, Ran Lun, Zhang Yanzi, Ren Yaping. Robust vehicle-to-grid power dispatching operations amid sociotechnical complexities. *Appl Energy* 2021;281:115912.
- [213] Sharma S, Jain Prerna. Risk-averse integrated DR and dynamic V2G scheduling of parking lot operator for enhanced market efficiency. *Energy* 2023;275:127428.
- [214] An Haopeng, Yi Jianbo, Zhang Guangdou, Bamisile Olusola, Li Jian, Huang Qi, et al. A robust V2G voltage control scheme for distribution networks against cyber attacks and customer interruptions. *IEEE Trans Smart Grid* 2024.
- [215] Rajasekaran Arun Sekar, Azees Maria, Al-Turjman Fadi. A comprehensive survey on security issues in vehicle-to-grid networks. *J Control Decis* 2023;10(2):150–9.
- [216] Warraich ZS, Morsi WG. Early detection of cyber-physical attacks on fast charging stations using machine learning considering vehicle-to-grid operation in microgrids. *Sustain Energy, Grids Networks* 2023;34:101027.
- [217] Zhao Zhonghao, Lee Carman KM, Huo Jiage. EV charging station deployment on coupled transportation and power distribution networks via reinforcement learning. *Energy* 2023;267:126555.
- [218] Nairi Nivedita, Vijayanthi C. Optimization of vehicle-to-grid (V2G) services for development of smart electric grid: A review. In: 2021 international conference on smart generation computing, communication and networking. IEEE; 2021, p. 1–6.
- [219] Li Shuangqi, Zhao Alexis Pengfei, Gu Chenghong, Bu Siqi, Chung Edward, Tian Zhongbei, et al. Interpretable deep reinforcement learning with imitative expert experience for smart charging of electric vehicles. *IEEE Trans Power Syst* 2024.
- [220] Druce Jeff, Harradon Michael, Tittle James. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. 2021, arXiv preprint [arXiv:2106.03775](https://arxiv.org/abs/2106.03775).
- [221] Kravaris Theocharis, Lentzos Konstantinos, Santipantakis Georgios, Vouros George A, Andrienko Gennady, Andrienko Natalia, et al. Explaining deep reinforcement learning decisions in complex multiagent settings: towards enabling automation in air traffic flow management. *Appl Intell* 2022;53(4):4063–98.
- [222] Amitai Yotam, Septon Yael, Amir Ofra. Explaining reinforcement learning agents through counterfactual action outcomes. 2023, arXiv preprint [arXiv:2312.11118](https://arxiv.org/abs/2312.11118).
- [223] Duan Yan, Schulman John, Chen Xi, Bartlett Peter L, Sutskever Ilya, Abbeel Pieter. RL²: Fast reinforcement learning via slow reinforcement learning. 2016, arXiv preprint [arXiv:1611.02779](https://arxiv.org/abs/1611.02779).
- [224] Yu Tianhe, Quillen Deirdre, He Zhanpeng, Julian Ryan, Hausman Karol, Finn Chelsea, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Kaelbling Leslie Pack, Kragic Danica, Sugiria Komei, editors. Proceedings of the conference on robot learning. Proceedings of machine learning research, vol. 100, PMLR; 2020, p. 1094–100.
- [225] Ng Andrew Y, Russell Stuart J. Algorithms for inverse reinforcement learning. In: Proceedings of the seventeenth international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2000, p. 663–70.
- [226] Arora Saurabh, Doshi Prashant. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 2021;297:103500.
- [227] Metelli Alberto Maria, Lazzati Filippo, Restelli Marcello. Towards theoretical understanding of inverse reinforcement learning. In: Krause Andreas, Brunskill Emma, Cho Kyunghyun, Engelhardt Barbara, Sabato Sivan, Scarlett Jonathan, editors. Proceedings of the 40th international conference on machine learning. Proceedings of machine learning research, vol. 202, PMLR; 2023, p. 24555–91.
- [228] Swamy Goluk, Wu David, Choudhury Sanjiban, Bagnell Drew, Wu Steven. Inverse reinforcement learning without reinforcement learning. In: Krause Andreas, Brunskill Emma, Cho Kyunghyun, Engelhardt Barbara, Sabato Sivan, Scarlett Jonathan, editors. Proceedings of the 40th international conference on machine learning. Proceedings of machine learning research, vol. 202, PMLR; 2023, p. 33299–318.

- [229] Dietterich Thomas G. Hierarchical reinforcement learning with the MAXQ value function decomposition. 1999, ArXiv Preprint ArXiv:Cs/9905014.
- [230] Hutsebaut-Buyse Matthias, Mets Kevin, Latré Steven. Hierarchical reinforcement learning: A survey and open research challenges. *Mach Learn Knowl Extr* 2022;4(1):172–221.
- [231] Jendoubi Imen, Bouffard François. Multi-agent hierarchical reinforcement learning for energy management. *Appl Energy* 2023;332:120500.
- [232] Lei Kun, Guo Peng, Wang Yi, Zhang Jian, Meng Xiangyin, Qian Linmao. Large-scale dynamic scheduling for flexible job-shop with random arrivals of new jobs by hierarchical reinforcement learning. *IEEE Trans Ind Inform* 2024;20(1):1007–18.
- [233] Zha Yantian, Guan Lin, Kambhampati Subbarao. Learning from ambiguous demonstrations with self-explanation guided reinforcement learning. 2024, arXiv preprint [arXiv:2110.05286](#).
- [234] Jing Mingxuan, Ma Xiaojian, Huang Wenbing, Sun Fuchun, Yang Chao, Fang Bin, et al. Reinforcement learning from imperfect demonstrations under soft expert guidance. 2019, arXiv preprint [arXiv:1911.07109](#).
- [235] Zare Maryam, Kebria Parham M, Khosravi Abbas, Nahavandi Saeid. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Trans Cybern* 2024.
- [236] Zoph Barret, Le Quoc V. Neural architecture search with reinforcement learning. 2017, arXiv preprint [arXiv:1611.01578](#).
- [237] Zambaldi Vinicius, Raposo David, Santoro Adam, Bapst Victor, Li Yujia, Babuschkin Igor, et al. Relational deep reinforcement learning. 2018, arXiv preprint [arXiv:1806.01830](#).
- [238] Ju Mingxuan, Fan Yujie, Zhang Chuxu, Ye Yanfang. Let graph be the go board: Gradient-free node injection attack for graph neural networks via reinforcement learning. *Proc AAAI Conf Artif Intell* 2023;37(4):4383–90.
- [239] Munikoti Sai, Agarwal Deepesh, Das Laya, Halappanavar Mahantesh, Nataraajan Balasubramaniam. Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications. *IEEE Trans Neural Networks Learn Syst* 2023;1–21.
- [240] Wen Jinbo, Zhang Ruichen, Niyato Dusit, Kang Jiawen, Du Hongyang, Zhang Yang, et al. Generative AI for low-carbon artificial intelligence of things with large language models. 2024, arXiv preprint [arXiv:2404.18077](#).
- [241] Wen Licheng, Fu Daocheng, Li Xin, Cai Xinyu, Ma Tao, Cai Pinlong, et al. DiLu: A knowledge-driven approach to autonomous driving with large language models. 2024, arXiv preprint [arXiv:2309.16292](#).
- [242] Luong Trung Quoc, Zhang Xinbo, Jie Zhanming, Sun Peng, Jin Xiaoran, Li Hang. ReFT: Reasoning with reinforced fine-tuning. 2024, arXiv preprint [arXiv:2401.08967](#).
- [243] Finn Chelsea, Yu Tianhe, Zhang Tianhao, Abbeel Pieter, Levine Sergey. One-shot visual imitation learning via meta-learning. 2017, arXiv preprint [arXiv:1709.04905](#).
- [244] Dai Josef, Pan Xuehai, Sun Ruiyang, Ji Jiaming, Xu Xinbo, Liu Mickel, et al. Safe RLHF: Safe reinforcement learning from human feedback. 2023, arXiv preprint [arXiv:2310.12773](#).
- [245] Metcalf Katherine, Sarabia Miguel, Fedzechkina Masha, Theobald Barry-John. Can you rely on synthetic labellers in preference-based reinforcement learning? It's complicated. *Proc AAAI Conf Artif Intell* 2024;38(9):10128–36.
- [246] White Devin, Wu Mingkang, Novoseller Ellen, Lawhern Vernon J, Waytowich Nicholas, Cao Yongcan. Rating-based reinforcement learning. 2024, arXiv preprint [arXiv:2307.16348](#).
- [247] Zhu Zhuangdi, Lin Kaixiang, Jain Anil K, Zhou Jiayu. Transfer learning in deep reinforcement learning: A survey. *IEEE Trans Pattern Anal Mach Intell* 2023;45(11):13344–62.
- [248] Hester Todd, Vecerik Matej, Pietquin Olivier, Lanctot Marc, Schaul Tom, Piot Bilal, et al. Deep Q-learning from demonstrations. 2017, arXiv preprint [arXiv:1704.03732](#).
- [249] Xing Qiang, Xu Yan, Chen Zhong. A bilevel graph reinforcement learning method for electric vehicle fleet charging guidance. *IEEE Trans Smart Grid* 2023;14(4):3309–12.
- [250] Caldwell Thomas. The challenge of securing electric vehicle charger infrastructure. *Cyber Security: A Peer- Rev J* 2024;7(4):371–82.