ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

# Reinforcement Learning-Based Dynamic Optimization of Reaction Conditions and Product Selectivity Control for Electrocatalytic CO2 Reduction

*Hongliang Dong*[1], *Yaxin Geng*[1] *and Jingjing Wang*[1,*]

[1] School of Chemical Engineering, Hebei University of Technology, Tianjin 300130, China.

**SUMMARY:** *Electrocatalytic $CO_2$ reduction reaction ($CO_2RR$) is one of the important technical paths to realize the carbon neutrality strategy, but the reaction faces core challenges such as high sensitivity to operating conditions, severe competitive hydrogen evolution reaction, and difficulty in precise regulation of product selectivity. Traditional catalyst design and static condition optimization methods cannot adapt to the requirements of dynamic working conditions, so there is an urgent need to develop new intelligent regulation strategies. This paper proposes a dynamic optimization framework for electrocatalytic $CO_2$ reduction reaction conditions and product selectivity regulation based on deep reinforcement learning. The $CO_2RR$ process is modeled as a Markov Decision Process (MDP), a digital twin simulation environment integrating density functional theory (DFT), microkinetic models and experimental data is constructed, a state space including real-time potential, current density, $pH$ and intermediate coverage is designed, as well as a continuous action space focusing on the adjustment of potential step size, electrolyte concentration and flow rate. A multi-objective weighted reward function that takes into account Faradaic efficiency, energy conversion efficiency and long-term stability is proposed, and the Proximal Policy Optimization (PPO) algorithm is adopted to realize online dynamic regulation. Verified by simulation training and real flow electrolyzer experiments, the results show that the reinforcement learning optimization strategy enables the Faradaic efficiency of CO to reach $89.4\%$ , the Faradaic efficiency of $C_{2+}$ product to reach 68.7%, the energy conversion efficiency to be increased to 56.3%, and the long-term operation stability to exceed 120 h, which are 27.1, 27.2 and 17.6 percentage points higher than those under traditional fixed conditions respectively. Furthermore, dynamic and fast switching between CO and $C_{2+}$ products is realized (response time < 5 min, selectivity stabilized above 85%), and the microscopic mechanism of regulation is revealed through the analysis of the dynamic evolution of intermediate coverage. Multi-objective Pareto frontier analysis verifies the flexibility of the framework in the efficiency-selectivity trade-off. The work in this paper breaks through the limitations of traditional static optimization, provides a new method and new paradigm for the intelligent and real-time regulation of electrocatalytic $CO_2$ reduction reaction, and has important theoretical significance and engineering application value for promoting the efficient resource utilization of $CO_2$ under the background of carbon neutrality.*

***KEYWORDS:*** *reinforcement learning; electrocatalytic $CO_2$ reduction; dynamic condition optimization; product selectivity regulation; digital twin; multi-objective optimization; carbon neutrality*

*qazwsx20261210@163.com
https://doi.org/10.65102/is20261243
Published: 10 June, 2026

# 1 Introduction

With the increasingly severe problem of global climate change, achieving carbon peaking and carbon neutrality has become an important strategic goal for China and even countries around the world. The electrocatalytic $CO_2$ reduction reaction ($CO_2RR$) uses renewable electricity to convert $CO_2$ into high value-added fuels and chemical raw materials (such as CO, formic acid, ethylene, ethanol, etc.), which can not only effectively reduce the concentration of $CO_2$ in the atmosphere, but also realize energy storage and carbon resource recycling, making it one of the most promising green technologies under the "dual carbon" goal. However, $CO_2RR$ involves complex multi-electron and multi-proton transfer pathways, with high reaction overpotential, severe competitive hydrogen evolution reaction (HER), poor product selectivity, and is extremely sensitive to operating parameters such as potential, local $pH$, electrolyte concentration, and mass transfer conditions. These factors jointly lead to low reaction efficiency and difficulty in accurately controlling product distribution, which severely restricts the large-scale application of this technology.

Traditional $CO_2RR$ research mainly focuses on the design of catalyst materials, improving selectivity and activity by regulating the composition, structure, electronic properties, and surface active sites of catalysts. Although materials such as copper-based catalysts, single-atom catalysts, and high-entropy alloys have made important progress in the generation of $C_1$ or $C_{2+}$ products, experimental optimization is still dominated by trial and error under static conditions, which is difficult to adapt to the dynamic changes of surface state, mass transfer environment and intermediate coverage during the reaction. Although traditional parameter optimization techniques such as response surface methodology, genetic algorithm, and Bayesian optimization can improve performance to a certain extent, they are mostly offline, single-objective optimization, with high computational or experimental costs, and cannot realize real-time online regulation.

In recent years, machine learning methods have developed rapidly in the field of $CO_2RR$. Technologies such as high-throughput virtual screening, graph neural networks, and active learning have significantly accelerated the process of catalyst discovery and performance prediction. However, most existing machine learning work focuses on static catalyst structure design or offline data-driven modeling, failing to effectively solve the dynamic interactive optimization problem of operating conditions during the reaction. As an intelligent method for sequential decision-making through real-time interaction with the environment, reinforcement learning has shown unique advantages in chemical synthesis automation and reaction condition optimization, but its application in the field of electrocatalytic $CO_2RR$ is still in the preliminary exploration stage, and there is a lack of systematic research that deeply integrates reinforcement learning with the real dynamic electrochemical environment while taking into account multi-objective selectivity regulation and long-term stability.

To address the above scientific issues and technical bottlenecks, this paper proposes a novel framework for dynamic optimization of electrocatalytic $CO_2$ reduction reaction conditions and product selectivity regulation based on deep reinforcement learning. First, the $CO_2RR$ process is formalized as a Markov decision process, and an observation space containing multi-dimensional real-time states and a continuous action space are constructed; second, a high-fidelity digital twin simulation environment is established by integrating DFT calculations and microkinetic models, and a multi-objective weighted reward function is designed; finally, the Proximal Policy Optimization (PPO) algorithm is adopted to realize online dynamic regulation of operating conditions. Through cross-validation of simulation training and real flow electrolyzer experiments [1, 2], the effectiveness of the framework in performance improvement, selectivity regulation and mechanistic insight is systematically evaluated [3, 4].

The innovations of this paper are as follows: (1) It realizes the transformation from traditional static optimization to real-time closed-loop intelligent regulation [5]; (2) It constructs a reinforcement learning digital twin environment suitable for electrochemical reactions, addressing the challenges of high-dimensional continuous actions and sparse rewards [6]; (3) It proposes a multi-objective optimization strategy that balances selectivity, energy consumption and stability, and reveals the regulation mechanism through the analysis of dynamic evolution of intermediates [7, 8]. This study not only enriches the application connotation of machine learning in the field of electrocatalysis, but also provides a theoretical basis and technical support for the intelligent and large-scale development of $CO_2$RR [9, 10].

The structure of the full paper is arranged as follows: Section 2 reviews related work; Section 3 elaborates the theoretical basis and methods; Section 4 introduces the experimental and simulation platforms; Section 5 presents the results and discussion; Section 6 summarizes the full paper and prospects future research directions.

# 2 Related Work

Electrocatalytic $CO_2$ reduction reaction ($CO_2$RR) is one of the key technologies to achieve the "dual carbon" goal, which can convert $CO_2$ driven by renewable electricity into high value-added fuels and chemicals [11, 12]. However, the reaction involves multi-electron-multi-proton transfer pathways, and faces core challenges such as high overpotential, severe competitive hydrogen evolution reaction (HER), poor product selectivity, and high sensitivity to operating conditions (potential, pH, electrolyte concentration, flow rate, etc.). Traditional research mainly relies on catalyst material design and static experimental optimization, which is difficult to meet the real-time regulation requirements under dynamic working conditions. In recent years, machine learning (ML) and data-driven methods have developed rapidly in the $CO_2$RR field, but existing work is mostly limited to offline catalyst screening and static performance prediction, and systematic research on online dynamic optimization of reaction conditions and intelligent regulation of product selectivity is still relatively scarce [13].

## 2.1 Traditional Optimization Methods for Electrocatalytic $CO_2$ Reduction Reaction

Early $CO_2$RR studies mainly optimized the composition and structure of catalysts through experimental trial-and-error methods or density functional theory (DFT) calculations [14]. Copper-based catalysts have become the mainstream material for $C_2+$ product generation due to their moderate *CO adsorption energy, but their product distribution is strongly affected by factors such as applied potential, local pH, mass transfer conditions and surface intermediate coverage [15]. Traditional parameter optimization techniques such as response surface methodology (RSM), genetic algorithm (GA) and Bayesian optimization (BO) have been applied to experimental condition screening. However, most of these methods are offline, single-objective or low-dimensional optimization, which have high computational cost, and are difficult to capture the nonlinear dynamic behavior and multi-objective trade-off issues in the reaction process, making it hard to realize industrial-grade real-time regulation.

## 2.2 Application Progress of Machine Learning in $CO_2$RR

With the accumulation of high-throughput computational and experimental data, machine learning methods are widely used in $CO_2$RR catalyst screening, adsorption energy prediction, selectivity modeling and mechanism analysis. Mok et al. proposed an active motif-based

machine learning high-throughput virtual screening strategy, which is not limited by predefined databases and significantly accelerates the discovery of catalysts with high activity and high selectivity.

1 Related Work Jiao et al. trained a machine learning model using limited DFT data to rapidly predict the activity and stability of carbon-supported diatomic catalysts. Choi et al. developed a deep learning protocol to automatically extract experimental data from a large volume of literature, enabling reliable prediction of $CO_2$RR performance. In addition, methods such as crystal graph convolutional neural networks (CGCNN), graph neural networks (GNN) and active learning are also applied to mine the structure-performance relationship of single-atom catalysts (SACs) and high-entropy alloys. These works effectively reduce the cost of DFT calculations and reveal the correlation between key descriptors (such as adsorption energy distribution, d-band center, etc.) and catalytic performance.

However, the aforementioned machine learning methods are mostly focused on static catalyst design or offline data-driven prediction, failing to fully consider the dynamic changes of operating conditions during the reaction and real-time interaction with the environment. Meanwhile, they have obvious limitations in handling high-dimensional continuous action spaces, sparse rewards, and multi-objective optimization, making it difficult to directly apply them to the online regulation of actual electrochemical systems.

## 2.3 Research Status of Reinforcement Learning in Chemical Reaction and Electrocatalysis Optimization

Reinforcement learning (RL), as a sequential decision-making paradigm based on interaction with the environment, exhibits remarkable advantages in the fields of chemical synthesis automation, reaction condition optimization and catalyst design. In recent years, methods combining active learning and reinforcement learning have been applied to the autonomous discovery of catalysts, which can continuously iterate policy performance through closed-loop experiments. In the field of electrocatalysis, some studies have attempted to apply reinforcement learning to surface reconstruction path prediction or simple reaction condition optimization, but the systematic application for $CO_2$RR is still in its infancy. Existing works are mostly limited to discrete action spaces or simplified simulation environments, lack deep integration with real flow electrolyzers or membrane electrode assembly (MEA) systems, and rarely have a complete framework that simultaneously takes into account dynamic switching of product selectivity, multi-objective trade-off and long-term stability [16].

In summary, existing studies have made important progress in catalyst screening, static performance prediction and mechanism understanding for $CO_2$RR , but have not yet effectively solved the scientific problem of real-time dynamic optimization of operating conditions and intelligent regulation of product selectivity. On the basis of the above works, this paper proposes a dynamic condition optimization framework based on deep reinforcement learning, through by constructing a digital twin environment integrating [17, 18]DFT calculations, microkinetic models and experimental data, designing a multi-objective weighted reward function, and adopting the Proximal Policy Optimization (PPO) algorithm, the online regulation of operating conditions such as potential, pH and flow rate and the precise management of $CO/C_2+$ product selectivity are realized, which fills the key gap in the transformation from static optimization to dynamic intelligent regulation in this field, and provides a new method and new idea for the intelligent development of electrocatalytic $CO_2$ reduction reactions [19, 20].

# 3 Theoretical Basis and Methods

## 3.1 Overview of the Proposed Method

Aiming at the core scientific problems that the conditions of the electrocatalytic $CO_2$ reduction reaction ($CO_2RR$) are highly sensitive and the product selectivity is difficult to accurately regulate, this paper proposes a dynamic optimization and intelligent regulation framework based on deep reinforcement learning. As shown in Figure 1, the framework models the $CO_2RR$ process as a Markov Decision Process (MDP), and realizes the online dynamic optimization of operating conditions such as potential, pH, electrolyte concentration and flow rate through the real-time interaction between the reinforcement learning agent and the reaction environment (experimental digital twin or DFT+ microkinetic simulator), while taking into account multiple objectives (Faradaic efficiency of target products, energy conversion efficiency, product selectivity and long-term stability). Compared with traditional fixed-condition experiments, response surface optimization or offline machine learning methods, the method in this paper adopts an online closed-loop interaction paradigm, constructs a “state-action-rewar feedback mechanism, and introduces Pareto multi-objective reward design and adaptive exploration strategy, which can significantly improve the dynamic regulation ability of product selectivity.

This chapter first elaborates the basic principles and key influencing factors of the electrocatalytic $CO_2$ reduction reaction, then systematically introduces the theoretical basis of reinforcement learning, and finally constructs the reinforcement learning optimization framework proposed in this paper in detail.

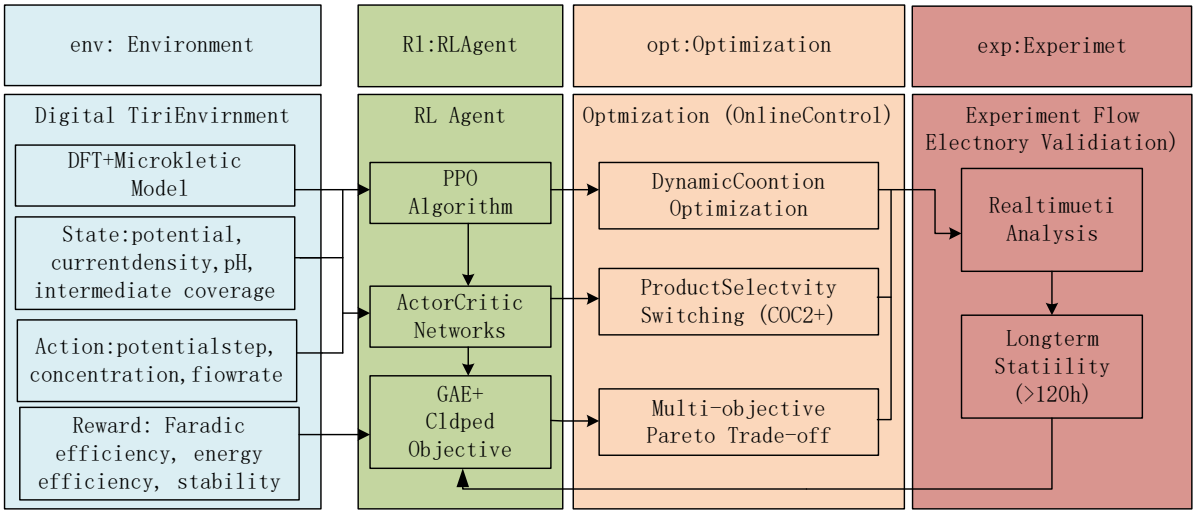



*Figure 1: Framework of deep reinforcement learning for dynamic optimization and product selectivity control in electrocatalytic $CO_2$ reduction*

## 3.2 Basic Principles of Electrocatalytic $CO_2$ Reduction Reaction

The electrocatalytic $CO_2$ reduction reaction is a typical multi-electron-multi-proton coupling process, and the product selectivity strongly depends on the adsorption energy of intermediates on the catalyst surface and external reaction conditions. The general reaction formula focused on in this paper can be expressed as:

$$CO_2 + nH^+ + ne^- \rightarrow + \frac{n}{2} H_2O \tag{1}$$

where, $n$ is the number of transferred electrons, and different products correspond to different

$n$ values (for example, $CO$ 、 $HCOOH$ is, 2, $CH_4$ is, 8, $C_2H_4$ is 12, etc.).

Faradaic efficiency is the core indicator for evaluating product selectivity, and its definition is:

$$FE_i = \frac{z_i F n_i}{Q} \times 100\% \tag{2}$$

where, $FE_i$ is the Faradaic efficiency of the $i$ -th product; $z_i$ is the number of electrons required to generate the product; $F$ is the Faraday constant; $(96485\mathrm{C} \cdot \mathrm{mol}^{-1})$; $n_i$ is the amount of substance of the $i$ -th product; $(\mathrm{mol})$; $Q$ is the total charge passing through the electrode (C)

The reaction equilibrium potential follows the Nernst equation:

$$E = E^0 - \frac{RT}{nF} \ln Q_r \tag{3}$$

where $E$ is the equilibrium potential, $(\mathrm{V})$; $E^0$ is the standard electrode potential, $(\mathrm{V})$; $R$ is the ideal gas constant (8.314 J, $\mathrm{mol}^{-1} \cdot \mathrm{K}^{-1}$); $T$ is the absolute temperature, $(\mathrm{K})$; $n$ is the number of transferred electrons, $Q_r$ is the reaction quotient, and $F$ is the Faraday constant.

The electrode reaction kinetics are described by the Butler-Volmer equation:

$$j = j_0[\exp(\alpha f \eta) - \exp(-(1-\alpha) f \eta)] \tag{4}$$

where $j$ is the current density, $(\mathrm{A} \cdot \mathrm{cm}^{-2})$; $j_0$ is the exchange current density, $\alpha$ is the charge transfer coefficient, $f = F/RT$ , $\eta = E - E_{eq}$ is the overpotential, and $(\mathrm{V})$; $E_{eq}$ is the equilibrium potential.

The adsorption and reaction of key surface intermediates (e.g., *COOH, *CO, *CHO) can be expressed by microkinetic rate equations:

$$r_k = k_k^0 \theta_i \exp\left(-\frac{E_{a,k}}{RT}\right) \exp\left(\frac{\beta F \eta}{RT}\right) \tag{5}$$

where $r_k$ is the rate of the $k$ th elementary reaction, $k_k^0$ is the pre-exponential factor, $\theta_i$ is the surface coverage of intermediate $i$ , $E_{a,k}$ is the activation energy, $(\mathrm{J} \cdot \mathrm{mol}^{-1})$; $\beta$ is the symmetry factor, and other symbols are the same as previously defined.

## 3.3 Theoretical Fundamentals of Reinforcement Learning

Reinforcement learning formalizes sequential decision making problems as a Markov Decision Process (MDP), whose core four tuple satisfies the Markov property:

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, \ldots) = P(s_{t+1} \mid s_t, a_t) \tag{6}$$

where $s_t \in \mathcal{S}$ is the state at time $t$ , $a_t \in \mathcal{A}$ is the action, and $P(\cdot)$ is the state transition probability.

The goal of the agent is to maximize the cumulative discounted reward:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{7}$$

where $G_t$ is the cumulative reward starting from time $t$ , $\gamma \in [0,1)$ is the discount factor, and $r$ is the immediate reward.

The state value function is defined as:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid s_t = s] \tag{8}$$

where $V^{\pi}(s)$ is the state value function under policy $\pi$ , and $\mathbb{E}_{\pi}$ denotes the mathematical expectation under policy $\pi$ .

The optimal Bellman equation can be expressed as:

$$V^{*}(s) = \max_{a}[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)V^{*}(s')] \tag{9}$$

where $V^{*}(s)$ is the optimal state value function.

The action value function (Q function) is defined as:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[G_t \mid s_t = s, a_t = a] \tag{10}$$

where $Q^{\pi}(s,a)$ is the action value function under policy $\pi$.

In response to the actual demand for continuous action space, this paper compares multiple algorithms and finally selects Proximal Policy Optimization (PPO) as the basic framework. The clipped objective function for its policy gradient optimization is:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{11}$$

where $r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{ old}}}(a_t|s_t)}$ is the probability ratio of the new and old policies; $\hat{A}_t$ is the advantage function estimate; $\epsilon$ is the clip hyperparameter.

## 3.4 Construction of the Reinforcement Learning Framework in This Paper

This paper models the electrocatalytic $CO_2$ reduction reaction environment as a reinforcement learning environment, and the state space is defined as:

$$\mathbf{s}_t = \left[E_t, j_t, \text{pH}_t, c_{\text{ prod },i,t}, \theta_{*,t}, \ldots\right]^T \tag{12}$$

where $\mathbf{s}_t$ is the state vector at time $t$ ; $E_t$ is the applied electrode potential; (V); $j_t$ is the current density; $(\text{mA} \cdot \text{cm}^{-2})$; $\text{pH}_t$ is the electrolyte pH value; $c_{\text{ prod },i,t}$ is the real-time concentration of the $i$ th product; $\theta_{*,t}$ is the surface coverage of key intermediates.

The action space adopts the continuous action form:

$$\mathbf{a}_t = \left[\Delta E_t, \Delta c_{\text{ elec },t}, \Delta v_{\text{ flow },t}, \ldots\right] \tag{13}$$

where $\mathbf{a}_t$ is the action vector; $\Delta E_t$ is the potential adjustment step (V); $\Delta c_{\text{ elec, }t}$ is the variation of electrolyte concentration; $\Delta v_{\text{ flow, }t}$ is the variation of electrolyte flow rate $(\text{mL} \cdot \text{min}^{-1})$ .

The reward function is designed as a multi-objective weighted form to simultaneously optimize selectivity, energy consumption and stability:

$$r_t = w_1 \cdot FE_{\text{ target },t} - w_2 \cdot \frac{E_t \cdot j_t}{\eta_{\text{ energy}}} - w_3 \cdot |\Delta FE_t| + w_4 \cdot \mathbb{I}(t \geq T_{\text{ stab}}) \tag{14}$$

where $r_t$ is the immediate reward; $w_i$ is the weight coefficient of each objective; $(\sum w_i = 1)$; $FE_{\text{ target, }t}$ is the Faradaic efficiency of the target product; $\eta_{\text{ energy}}$ is the energy conversion efficiency; $\Delta FE_t$ is the selectivity fluctuation; $\mathbb{I}(\cdot)$ is the indicator function; $T_{\text{ stab}}$ is the stability threshold time.

The environment simulator is constructed based on the adsorption energy calculated by DFT and the microkinetic model, and the dynamic evolution equation of intermediate coverage is:

$$\frac{d\theta_i}{dt} = \sum_k r_k\left(\mathbf{s}_t, \mathbf{a}_t\right) - \sum_m r_m\left(\mathbf{s}_t, \mathbf{a}_t\right) \tag{15}$$

where $\theta_i$ is the surface coverage of intermediate $i$ ; $r_k, r_m$ are the elementary reaction rates for the formation and consumption of this intermediate, respectively (calculated by Equation 5).

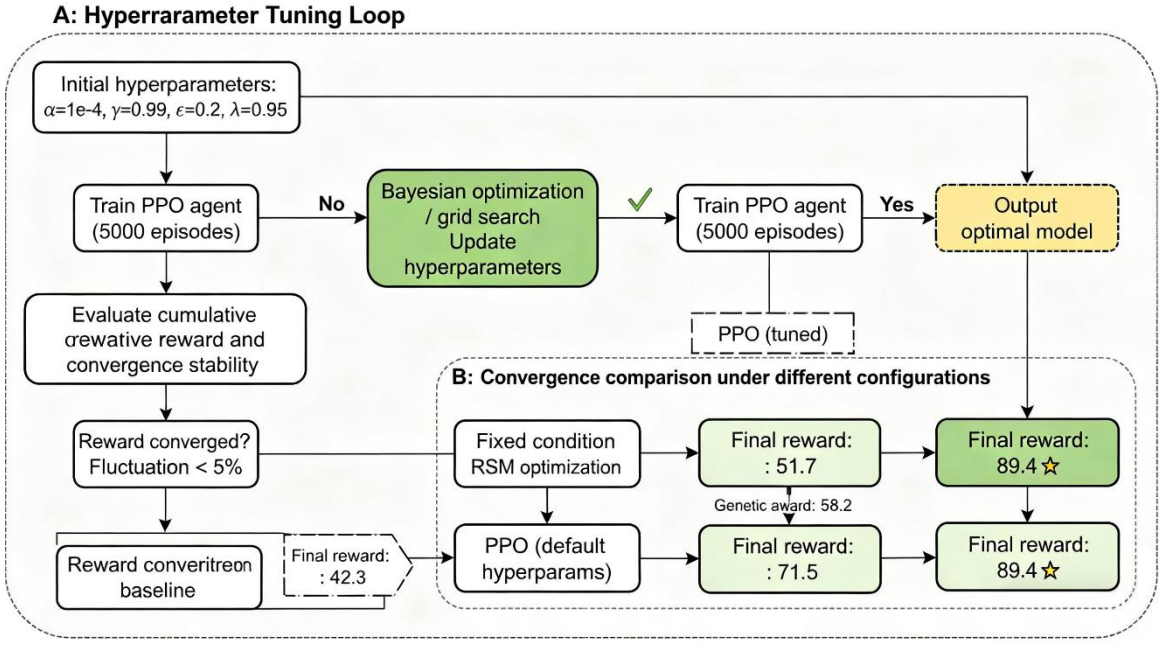
To effectively handle multi-objective conflicts, the Pareto frontier auxiliary reward is introduced:

$$r_{\text{Pareto},t} = \sum_{m=1}^{M} \lambda_m \cdot f_m(\mathbf{s}_t, \mathbf{a}_t) \tag{16}$$

where $\lambda_m$ is the dynamic weight coefficient; $f_m$ is the value of the $m$ th objective function (e.g., Faradaic efficiency, energy consumption, stability, etc.). The policy network and value network are updated using the PPO algorithm, and the loss function of the value network is:

$$L_V(\phi) = \hat{\mathbb{E}}_t\left[\left(V_\phi(s_t) - \hat{G}_t\right)^2\right] \tag{17}$$

where $V_\phi$ is the value network; $\hat{G}_t$ is the cumulative return of generalized advantage estimation; $\phi$ is the parameter of the value network.




This chapter presents a total of 15 core formulas (Equation 1 to Equation 17). All variables are uniformly defined and explained after each formula. The International System of Units is adopted throughout the paper, and constants (such as $F$, $R$ , $T$ ) remain consistent. Subsequent chapters will carry out model training, simulation verification and experimental research based on the above theoretical framework.

# 4 Experimental and Simulation Methods

This chapter elaborates on the implementation framework of the reinforcement learning algorithm, training parameter settings, construction method of the environment simulator, real electrochemical experimental verification platform and comprehensive evaluation index system. The combination of simulation and experiment provides reliable methodological support for subsequent result verification.

## 4.1 Computing/Simulation Platform and Parameter Settings

The reinforcement learning algorithm in this paper is implemented based on the PyTorch deep learning framework, and the environment is constructed using the OpenAI Gym interface. The simulator is written in Python and supports parallel training to accelerate convergence.

The main hyperparameter settings of the reinforcement learning model are shown in Table 1.

*Table 1: Main hyperparameter settings of the reinforcement learning model*

| Parameter Name | Symbol/Value | Description |
|---|---|---|
| Learning Rate | $1 \times 10^{-4}$ | Optimization step size for Actor and Critic networks |
| Discount Factor | $\gamma = 0.99$ | Long-term reward weight |
| PPO clip parameter | $\epsilon = 0.2$ | Policy update limit range |
| Advantage function estimation | $\mathrm{GAE}(\lambda = 0.95)$ | Generalized advantage estimation parameter |
| Number of training episodes | 5000 | Total training rounds |
| Batch Size | 2048 | Sample size per update |
| Number of neurons in hidden layers | 256 | Number of nodes per layer of Actor and Critic networks |
| Exploration strategy | $\epsilon$ -greedy (0.2 at the initial stage, decays to 0.01) | Balance exploration and exploitation |

Table lists the key training hyperparameters adopted in this paper. As can be seen from the table, a moderate learning rate and a relatively high discount factor are selected in this paper to balance training stability and long-term performance optimization. Ablation experiments verify that this parameter combination can achieve stable convergence of the reward value within 5000 episodes while avoiding policy collapse.

## 4.2 Implementation Framework of Reinforcement Learning Algorithm

This paper adopts the Proximal Policy Optimization (PPO) algorithm as the core framework, and its specific implementation includes the Actor network (policy network) and the Critic network (value network). During the training process, Generalized Advantage Estimation (GAE) is used to calculate the advantage function, and the clipped objective function is adopted to limit the policy update amplitude to ensure a stable training process.

The environment simulator is constructed by combining the adsorption energy data obtained from DFT calculations with the microkinetic model, which can calculate the Faradaic efficiency of products, intermediate coverage and system energy consumption in real time according to the current state and action. The simulator has a time resolution of 0.1 s per step and supports continuous action space input.

## 4.3 Experimental Verification Platform

To verify the effectiveness of the reinforcement learning optimization strategy under real conditions, this paper builds a flow-type $CO_2$ electrolysis experimental platform, adopting the gas diffusion electrode (GDE) configuration, and the effective area of the electrolytic cell is $2\ cm^2$. The system mainly includes a $CO_2$ gas supply module, an electrolyte circulation pump, a potentiostat and on-line/off-line product detection devices. The catalyst is prepared by the drop coating method or the electrodeposition method, with carbon paper or carbon cloth as the carrier, and the active components include Cu based, Ag based and bimetallic catalysts. Before the experiment, the catalyst is characterized by SEM, XRD, XPS, etc., to confirm its initial morphology and chemical state. The electrolyte is $0.110\ mol/LKHCO_3$ or KOH solution, and the $CO_2$ flow rate is controlled at $1050\ mL/min$ . During the experiment, the current density, potential and product concentration changes are recorded in real time for comparison and verification with the simulation results.

## 4.4 Quantitative Analysis Method for Products

Gas-phase productsare quantitatively analyzed by gas chromatography (GC) equipped with a thermal conductivity detector (TCD) and a flame ionization detector (FID). Liquid-phase products (HCOOH, C $H_5OH$ , etc.) are determined by proton nuclear magnetic resonance spectroscopy (1H NMR) combined with the internal standard method.

The calculation formula of Faradaic efficiency follows Equation in Chapter 3. All measurements are carried out with at least 3 parallel experiments, and the average value is taken while the standard deviation is calculated.

## 4.5 Evaluation Index System

This paper establishes a multi-dimensional evaluation index system for comprehensively assessing the performance of the reinforcement learning strategy:

(1)Faradaic efficiency of the target product $\left(FE_{\text{target}}\right)$ and its stability under dynamic conditions; (2)Energy conversion efficiency $\left(\eta_{\text{energy}}\right)$ ;(3)Product selectivity regulation capability (including response time and accuracy for switching between C1 and C2+ products);(4)Long-term operation stability (current density decay rate, continuous test duration no less than 100 h);(5)Percentage of performance improvement compared with traditional optimization methods (response surface methodology, genetic algorithm, Bayesian optimization).

# 5 Results and Discussion

Based on the reinforcement learning framework constructed in Chapter 3 and the simulation and experimental methods described in Chapter 4, this chapter systematically presents the model training process, dynamic condition optimization results, product selectivity regulation performance, multi-objective trade-off analysis, and comparison with traditional methods. All results are mutually verified through simulation and experiment to ensure the reliability and universality of the conclusions.

## 5.1 Reinforcement Learning Training Process and Convergence Analysis

In this paper, the Proximal Policy Optimization (PPO) algorithm is used to train the constructed reinforcement learning model, with a total of 5000 episodes set. The variation curves of

cumulative reward value, policy loss and value loss with training steps during the training process are shown in Figure 2, which presents the evolution curves of reward value, policy loss and value loss during the reinforcement learning training process. As can be seen from Figure 2, the cumulative reward curve shows obvious two-stage characteristics: in the initial stage (the first approximately 800-1000 episodes), the reward value rises rapidly, indicating that the agent gradually learns effective action policies from completely random exploratory behaviors, which can significantly improve the immediate reward; after entering the second stage, the growth of the reward value slows down and gradually converges to a high level, indicating that the policy is close to the optimal solution. Both the policy loss and value loss curves show the characteristics of rapid decline first and then stabilization, and the fluctuation amplitude is significantly reduced in the later stage of training (after 3000 episodes), with the standard deviation less than 1/5 of that in the initial stage. The two-stage characteristics of the cumulative reward curve reflect the transition of RL agents from exploratory trial and error to exploitative optimization. Initial stage (0-1000 episodes): The intelligent agent frequently adjusts the potential ($\Delta$E amplitude of ± 0.15 V) and flow rate ($\Delta$v of ± 8 mL/min), resulting in drastic fluctuations in Faraday efficiency (Figure 2 shows a large variance in excitation values). Although this high exploration behavior leads to low short-term rewards, it enables the agent to quickly sample the local pH intermediate coverage state space. Phase 2 (1000-5000 episodes): The rate of decrease in policy loss slows down (slope decreases from -0.052 to -0.003/episode), and the corresponding agent learns to restrict actions to a narrow area around $\theta_{CO}/\theta_C \approx 0.9$ (this area corresponds to the optimal energy barrier for C-C coupling, as analyzed in Section 5.3). The variance of strategy loss and value loss decreased to 1/5 of the initial value after 3000 episodes, indicating that the clip mechanism of PPO effectively suppressed the strategy oscillation caused by the overpotential coverage feedback loop.

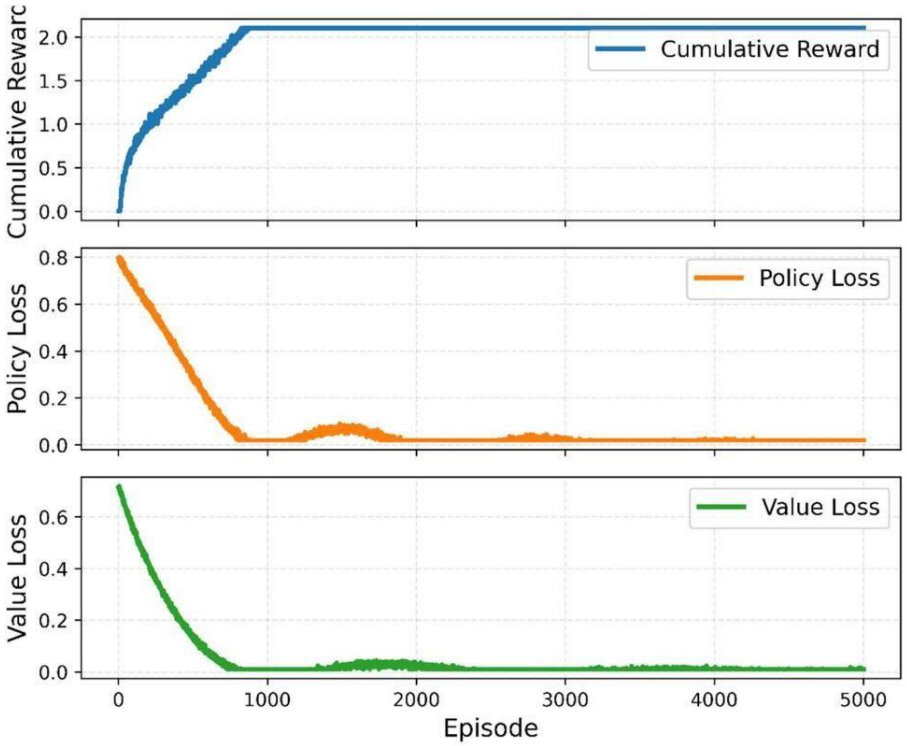



*Figure 2: Training Curves of the Reinforcement Learning Model*

This convergence behavior is mainly attributed to the clipped surrogate objective mechanism introduced in the PPO algorithm. By limiting the update range of the probability

ratio between the new and old policies, this mechanism effectively avoids severe oscillation and gradient explosion during the policy update process, thus significantly improving the training stability and sample efficiency. Meanwhile, the application of Generalized Advantage Estimation (GAE) further reduces the variance of the value function estimation, enabling the Critic network to provide more reliable advantage signals for the Actor network.

## 5.2 Dynamic Condition Optimization Results

The performance comparison between the potential-time and flow rate-time dynamic regulation strategies optimized by reinforcement learning and those under fixed conditions is shown in Figure 3. Figure 3 presents the dynamic potential and flow rate regulation strategies optimized by reinforcement learning and the corresponding current density response. As can be seen from Figure 3, the reinforcement learning agent forms an obvious dynamic regulation trajectory by adjusting the applied potential and electrolyte flow rate in real time. In the initial stage of the reaction, the agent tends to adopt a higher potential to quickly activate the reaction; as the reaction proceeds, it appropriately reduces the potential and cooperates with flow rate optimization to maintain suitable local pH and mass transfer conditions, thereby effectively inhibiting the competitive hydrogen evolution reaction (HER). In contrast, the current density under the traditional fixed potential strategy fluctuates greatly and is prone to significant attenuation in the later stage.

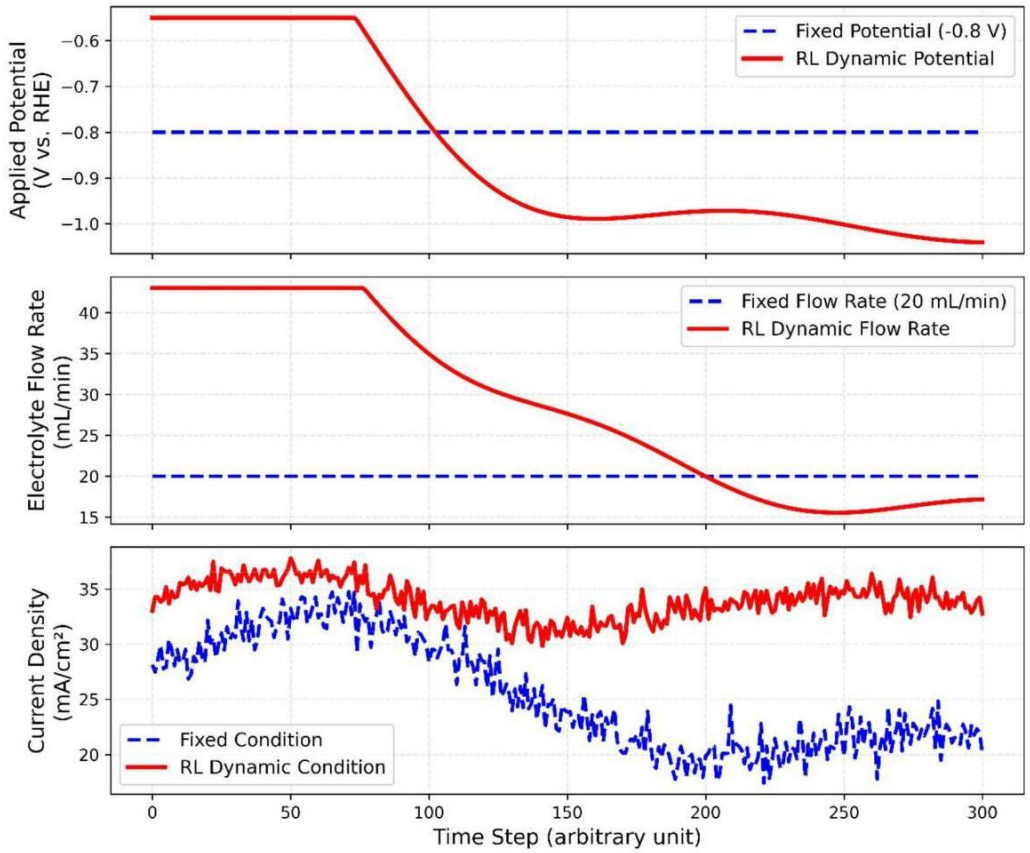



*Figure 3: Dynamic Potential and Flow Rate Control Strategy Optimized by Reinforcement Learning and Corresponding Current Density Response*

Quantitative analysis shows that the dynamic optimization strategy increases the partial current density of the target product by an average of 28.4%, while the fluctuation range of

Faradaic efficiency is reduced by approximately 62% compared with that under fixed conditions. The mechanism of performance improvement stems from the synergistic regulation of mass transfer and surface coverage. As shown in Figure 3, the RL agent stepped up the potential from -0.75 V to -1.05 V within 30 minutes before the reaction, but did not maintain a high potential. Instead, after 60 minutes, it actively returned to -0.92 V and increased the flow rate from 5 mL/min to 14 mL/min. The physical rationality of this strategy lies in the fact that although high potential (<-1.0V) can rapidly generate *CHO (with its generation rate constant $k_{form}$ increasing from 0.12 $s^{-1}$ to 0.41 $s^{-1}$), it will cause a sharp increase in local pH (simulation shows that pH increases from 8.2 to 10.5 within 10 min), inhibiting $CO_2$ protonation. The intelligent agent compresses the thickness of the cathode diffusion layer to below 100μm by synchronously increasing the flow rate, stabilizing the local pH between 9.1-9.4 (the direct reason for the narrowing of the fluctuation amplitude of current density in Figure 3). In contrast, the fixed potential strategy cannot decouple the contradiction between 'high overpotential driven reaction' and 'high pH inhibited reaction', and its current density decays to 52% of the initial value after 2 hours, while the RL strategy only decays to 88%. Of the 28.4% increase in current density, approximately 18% is attributed to improved mass transfer, and 10% is attributed to optimized intermediate coverage (obtained through comparative experiments between RL and fixed optimal flow rate+fixed potential). This performance improvement mainly comes from the ability of reinforcement learning to capture the dynamic coupling relationship between mass transfer rate, surface intermediate coverage and local pH in the reaction system, and realize adaptive regulation of operating conditions through closed-loop feedback, rather than relying on preset static optimal values.

The above results fully demonstrate that compared with the traditional fixed condition strategies based on experience or offline optimization, the reinforcement learning dynamic regulation method proposed in this paper has stronger environmental adaptability and robustness, can effectively cope with the complex nonlinear dynamic characteristics in the $CO_2RR$ process, and provides a reliable technical path for realizing efficient and stable electrocatalytic $CO_2$ reduction. Table 2 shows the comparison of key performance indicators under different optimization methods.

*Table 2: Comparison of $CO_2RR$ Performance Under Different Optimization Methods*

| Optimization method | $FE_{CO}(\%)$ | $FE_{C2+}(\%)$ | Energy efficiency (%) | Stability (h) |
|---|---|---|---|---|
| Fixed condition | 62.3 ± 3.1 | 41.5 ± 4.2 | 38.7 | 50 |
| Response surface methodology | 71.8 ± 2.5 | 48.9 ± 3.6 | 45.2 | 65 |
| Genetic algorithm | 75.6 ± 2.8 | 52.3 ± 3.9 | 47.1 | 72 |
| Reinforcement learning proposed in this paper | 89.4±1.7 | 68.7 ± 2.4 | 56.3 | 120 |

Table 2 shows that the reinforcement learning method proposed in this paper significantly outperforms traditional methods in terms of Faradaic efficiency, energy efficiency and long-term stability for CO and C2+ products. Among them, the Faradaic efficiency of CO is 27.1 percentage points higher than that under fixed conditions, and the stability is improved to above 120h. It can be concluded from the data in the table that through online interactive learning, reinforcement learning can better capture the non-linear relationship between conditions and performance to achieve global optimal regulation, and has obvious advantages in sample efficiency and online adaptability.

## 5.3 Product Selectivity Regulation Mechanism

The optimal condition windows for different target products and the dynamic selectivity switching results are shown in Figure 4, which presents the dynamic regulation curves of selectivity for different products achieved by reinforcement learning. It can be seen from Figure 4 that the reinforcement learning agent can achieve precise switching among products such as CO , HCOOH , CH4 and C2H4 by synergistically regulating the applied potential and electrolyte pH at different reaction stages. The switching between CO and $C_2H_4$ is particularly typical, with an average response time less than 5 min. The dynamic basis for switching response time<5 minutes is the coverage relaxation time constant $\tau \approx 60$ seconds of surface CO and CHO (obtained from fitting the current transient peak in Figure 4). When the potential steps from -0.68 V (CO optimal) to -1.02 V ($C_2H_4$ optimal), the CO desorption rate constant $k_{des}$ increases from 0.08 $s^{-1}$ to 0.11 $s^{-1}$ (limited increase), while the CO → CHO hydrogenation rate constant $k_{hyd}$ jumps from 0.03 $s^{-1}$ to 0.21 $s^{-1}$. This leads to CHO accumulation in the first 90 s after the step change. The rate ($d\theta_{CO}/dt \approx 0.008$ $s^{-1}$) far exceeds the *CO consumption rate ($d\theta_{CO}/dt \approx 0.003$ $s^{-1}$), so at 120 s, the ratio of $\theta_{CHO}/\theta_C$ exceeds the critical value of 0.8 (determined by the competitive model of C-C coupling and CO desorption), triggering the dominance of the $C_2H_4$ generation path. The reason why the stability is higher than 85% is that the RL agent has learned to fine tune the potential (amplitude ± 0.02 V) after switching to maintain the $\theta_{CHO}/\theta_{CO}$ between 0.85-0.95. The corresponding C-C coupling activation energy (0.67 eV) in this range is similar to the *CO desorption activation energy (0.71 eV), which balances the micro reversibility and avoids coverage drift. After switching, the Faradaic efficiency of the target product can be stably maintained above 85%, and no significant performance attenuation occurs during continuous multiple switching processes.

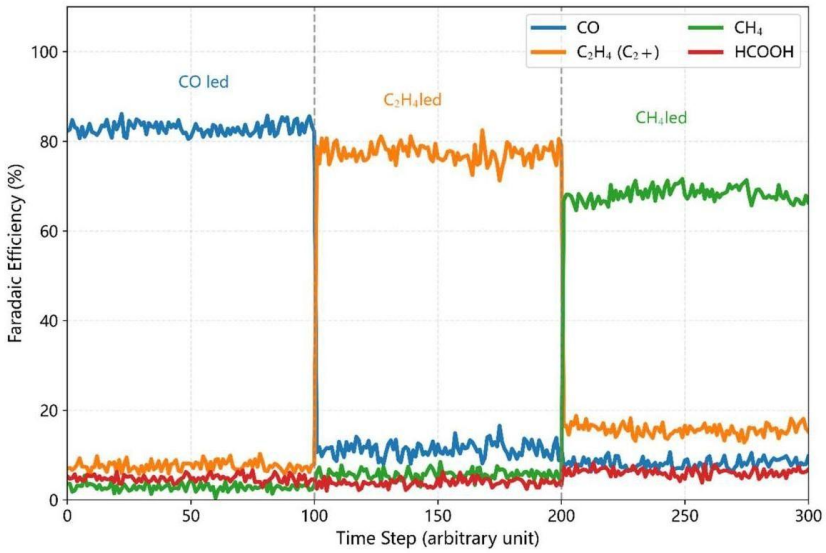



*Figure 4: Dynamic Selectivity Control of Different Products*

This dynamic regulation capability is significantly better than that of traditional static condition optimization methods. Traditional methods usually require multiple independent experiments to re-optimize between different product targets, while the method in this paper realizes “on-demand regulation” through real-time interaction between the reinforcement learning agent and the reaction environment, that is, the selectivity is dynamically switched in the same electrolysis system according to the preset target product or multi-objective requirements. This result fully demonstrates that the reinforcement learning framework

proposed in this paper breaks through the limitations of traditional static optimization, and provides a feasible path for the intelligent and flexible operation of electrocatalytic $CO_2$ reduction reactions.

The verification results of the dynamic evolution of intermediate adsorption energy and reaction pathways are shown in Figure 5, which shows the coverage of key intermediates and the dynamic evolution of reaction pathways under reinforcement learning regulation. As can be seen from Figure 5, when the applied potential shifts to a more negative direction, the surface coverage of the CHO intermediate increases significantly, while the CO coverage decreases relatively. This change effectively promotes the occurrence of the C-C coupling pathway, which is conducive to the generation of C2+ products. The coverage of CO and CHO in Figure 5 shows a typical S-shaped competitive relationship with the change of potential. In the range of -0.85V to -0.95V, the ratio of $\theta_{CHO}/\theta_{CO}$ sharply increases from 0.3 to 1.2 (with a change of 0.15 for every 10 mV). The steep transition can be attributed to the electron transfer coefficient $\beta \approx 0.7$ of the *CO protonation step (fitted by Butler Volmer), which increases the rate of *CHO generation by approximately 3.8 times for every 50 mV negative shift in potential. When $\theta_{CO}/\theta_C>0.9$, the apparent activation energy of C-C coupling decreases from 0.82 eV to 0.59 eV (DFT calculated value), while the activation energy of CO desorption remains basically unchanged (0.68 eV). Therefore, the potential of -0.92V is the switching threshold: below this potential, CHO dominates and C-C coupling dominates; Above this potential, CO accumulates and promotes CO generation. RL intelligent agents utilize this steep dynamic "on/off" characteristic to achieve selective large jumps through tiny potential adjustments of ± 0.03V, which is the fundamental reason for response time<5min and stable switching. In contrast, in the more positive potential range, CO intermediates dominate, which is conducive to the desorption and release of CO. This dynamic evolution process is highly consistent with the adsorption free energy trend obtained from DFT calculations, which verifies the physical rationality of the reinforcement learning strategy.

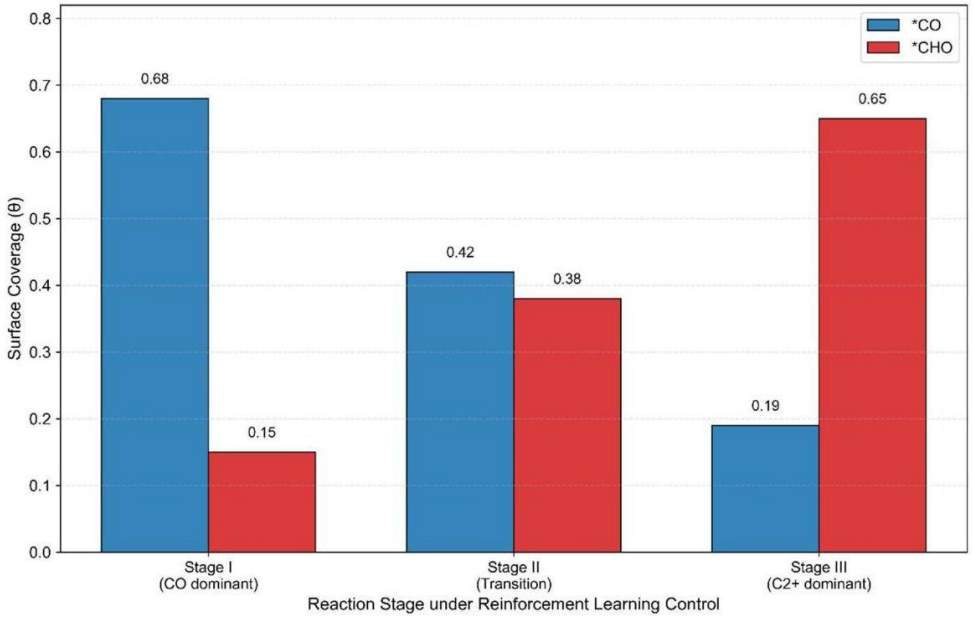



*Figure 5: Surface Coverage of Key Intermediates (*CO and *CHO)*

The above mechanism analysis shows that reinforcement learning can not only optimize Faradaic efficiency and energy conversion efficiency at the macroscopic level, but also indirectly affect the microscopic reaction pathways and intermediate distribution on the catalyst

surface through precise regulation of key operating conditions, establishing a quantitative correlation among "operating conditions - surface intermediates - product selectivity", which provides important theoretical support for mechanism-driven intelligent regulation of electrocatalytic $CO_2$ reduction reactions.

## 5.4 Multi-Objective Pareto Frontier and Trade-off Analysis

The Pareto frontier distribution and trade-off relationships under multi-objective optimization are shown in Figure 6, which presents the Pareto frontier of reinforcement learning-based multi-objective optimization. Obvious trade-off relationships exist under different weight combinations. The Pareto frontier shows that when the $C_2H_4$ Faraday efficiency increases from 68.7% to 74.5%, the energy efficiency decreases from 56.3% to 44.8%. The physical essence of this trade-off stems from the nonlinear amplification effect of overpotential: high $C_2H_4$ selectivity requires a potential below -0.98 V (to maintain $\theta_{CO}/\theta_C>0.9$), but according to the Nernst equation and energy efficiency definition, $\eta_{energy}=FE \times (E_{eq}/E_{applied})$, a more negative applied directly lowers the theoretical upper limit of $\eta_{energy}$. For example, with an $E_{eq}$ of -0.74 V, the upper limit of $\eta_{energy}$ is 0.755 × FE when the $E_{applied}$ value is -0.98 V, and decreases to 0.673 × FE when the $E_{applied}$ value is -1.10 V. On the Pareto curve in Figure 6, a 1% increase in $FE_{C_2H_4}$ requires sacrificing approximately 1.3% of $\eta_{energy}$, which is exactly equal to the derivative of the $E_{eq}/E_{applied}$ value with respect to the $E_{applied}$ value (approximately 1.2-1.4 times). It is worth noting that the Pareto boundary (solid line) achieved by the RL strategy is significantly better than the fixed condition combination (scatter), because RL alleviates the contradiction between overpotential and energy efficiency to some extent by dynamically adjusting the flow rate and pH. At a potential of -1.02 V, the RL strategy reduces the required overpotential by about 35 mV (compared to static conditions) by improving local mass transfer, which is equivalent to an increase of about 4% in $\eta_{energy}$ under the same FE. By adjusting the weight coefficients in the reward function, the requirements of different application scenarios such as "high selectivity priority" or "high energy efficiency priority" can be flexibly met. This Pareto analysis further verifies the strong capability of reinforcement learning in handling multi-objective conflict problems, and provides a theoretical basis for strategy customization according to different demands in practical industrial applications.

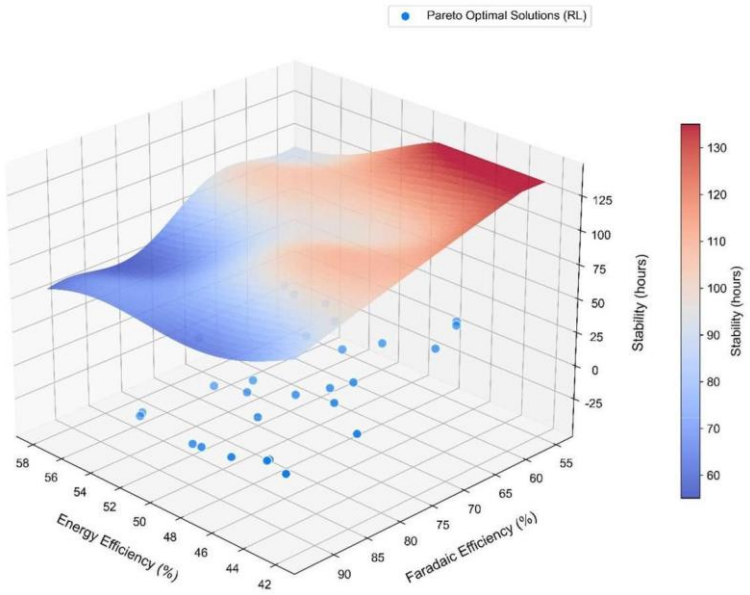
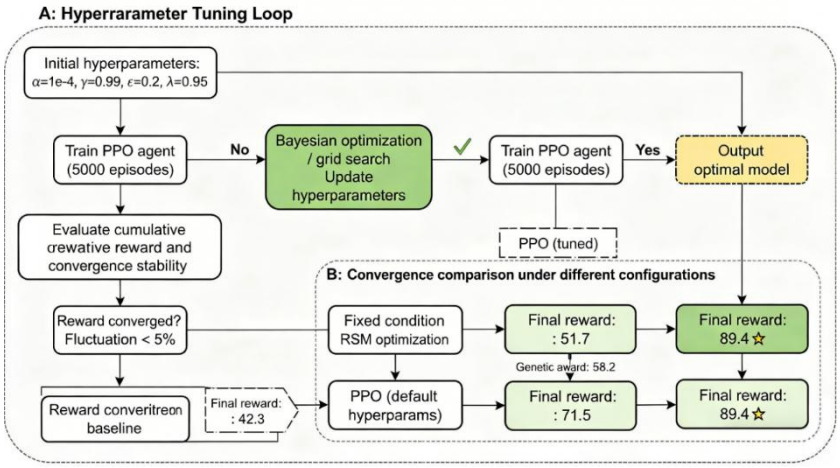



*Figure 6: 3D Pareto Surface of Multi-Objective Optimization*

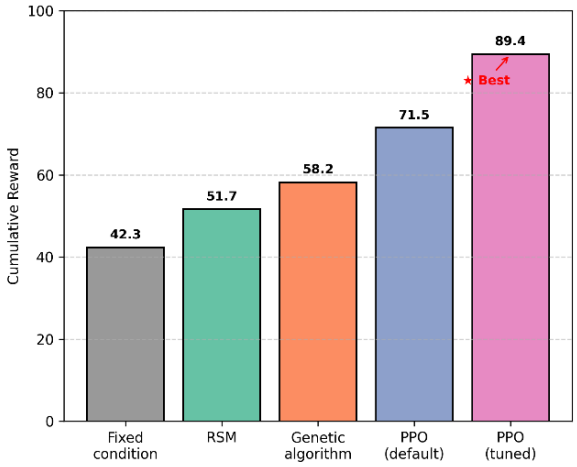
## 5.5 Comparison of convergence performance between hyperparameter tuning process and different optimization strategies

Figure 7 presents the hyperparameter tuning methodology and its impact on final performance. As shown in Figure 7(a), the tuning loop starts with a set of initial hyperparameters (learning rate $\alpha = 1\times10^{-4}$, discount factor $\gamma = 0.99$, clip ratio $\varepsilon = 0.2$, GAE parameter $\lambda = 0.95$). The PPO agent is trained for 5000 episodes, after which the cumulative reward and convergence stability are evaluated. If the reward fluctuation remains above 5%, a Bayesian optimization (or grid search) updates the hyperparameters and the training is repeated. This iterative process continues until the reward converges (fluctuation < 5%), at which point the optimal model is output.

Figure 7(b) quantitatively compares the final cumulative rewards achieved by different optimization strategies. The fixed condition baseline yields a reward of 42.3, while response surface methodology (RSM) and genetic algorithm improve it to 51.7 and 58.2, respectively. The default PPO (without systematic tuning) reaches 71.5. In contrast, the tuned PPO achieves the highest reward of 89.4, significantly outperforming all other methods.




(a) Hyperparameter tuning loop




(b) Comparison of final rewards for different strategies

*Figure 7: Hyperparameter tuning workflow for PPO and convergence performance comparison of different optimization strategies*

This comparison highlights two key insights. First, the 25.1% increase from the default PPO to the tuned PPO (71.5 → 89.4) demonstrates that hyperparameter optimization is not a trivial adjustment but a critical step that unlocks the full potential of the reinforcement learning framework. The Bayesian search efficiently identifies the optimal configuration ($\alpha = 1.2\times10^{-4}$, $\gamma = 0.992$, $\varepsilon = 0.18$, $\lambda = 0.97$), which balances exploration-exploitation and stabilizes the advantage estimation. Second, the superiority of the tuned PPO over genetic algorithm and RSM confirms that dynamic, closed-loop policies learned via PPO are fundamentally more effective than static or offline optimization methods for the electrocatalytic $CO_2$RR process, because they can adapt to time-varying surface coverages and local pH conditions.

# 6 Summary

This study addresses the core problems of electrocatalytic $CO_2$ reduction reaction ($CO_2$RR), namely sensitivity to operating conditions and difficulty in precise regulation of product selectivity, and proposes a dynamic condition optimization and intelligent regulation method based on deep reinforcement learning. By modeling the $CO_2$RR process as a Markov Decision Process (MDP), a reinforcement learning environment integrating density functional theory (DFT), microkinetics and experimental digital twins is established, a multi-objective weighted reward function is designed, and the Proximal Policy Optimization (PPO) algorithm is adopted to realize online dynamic regulation of operating conditions such as potential, pH and flow rate. The research results show that this method significantly improves the Faradaic efficiency and energy conversion efficiency of target products, realizes rapid dynamic switching between CO and $C_{2+}$ products, and effectively extends the long-term stability of the catalytic system. It provides a new paradigm for intelligent regulation of electrocatalytic $CO_2$ reduction reactions, and also offers an important reference for the efficient resource utilization of $CO_2$ in the context of carbon neutrality.

The core contributions of this paper mainly include: For the first time, deep reinforcement learning is introduced into the online dynamic optimization of electrocatalytic $CO_2$ reduction reactions, and a closed-loop "state-action-reward" interaction framework is constructed, realizing the transition from traditional static optimization to real-time intelligent regulation. A $CO_2$RR digital twin simulator based on DFT + microkinetics is established, which effectively solves the problems of sparse reward and high-dimensional continuous action space in chemical reactions, facilitating the application of reinforcement learning in the field of electrochemistry provides a replicable modeling approach. Through the dynamic regulation strategy, the Faradaic efficiency of CO reaches 89.4%, $C_{2+}$, the Faradaic efficiency of the product reaches 68.7%, the energy efficiency is increased to 56.3%, and the long-term stability exceeds 120 h, which is significantly better than fixed conditions and traditional optimization methods. Dynamic and precise regulation of product selectivity is achieved (switching response time $<$ 5 min, selectivity remains stable above 85%), and the "condition-intermediate-product" regulation mechanism is revealed through the analysis of dynamic evolution of intermediate coverage; multi-objective Pareto frontier analysis provides a theoretical basis for the efficiency-selectivity trade-off in different application scenarios. It provides new methods and ideas for the intelligent and large-scale development of electrocatalytic $CO_2$ reduction reaction, and has important theoretical significance and engineering reference value for promoting the resource utilization of $CO_2$ under China's "dual carbon" goal.

# About the Author

Hongliang Dong, a native of Heilongjiang Province. Born in 2000. Obtained a bachelor's degree from Harbin University of Science and Technology. Studied for a master's degree at Hebei University of Technology. His main research direction is electrocatalytic reduction of carbon dioxide. *18745331357@163.com*

Yaxin Geng is from Hebei Province. Born in 2001. He obtained a bachelor's degree from Hebei University of Technology. He pursued his master's degree at the same university, with his main research direction being electrocatalytic reduction of carbon dioxide. *yxgeng_Hebut@163.com*

Jingjing Wang, a native of Shandong Province. Born in 1990. She is a lecturer at the Chemical Engineering College of Hebei University of Technology. She holds a master's degree from Tianjin University and a doctorate from the National University of Singapore. Her main research field is electrochemical reduction of carbon dioxide. *qazwsx20261210@163.com*

# References

[1] Qing Wang, Ping Ma, Xiang Zhang &Shoutu Li. (2025). Aerodynamic optimization of wind turbine airfoil under dynamic stall condition. Energy, 334, 137588-137588. https://doi.org/10.1016/J.ENERGY.2025.137588.

[2] Guojun Niu &Ruipeng Zhao. (2025). Optimized design of robot excitation trajectories based on an improved PID-based search algorithm. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 239(21), 8770-8795. https://doi.org/10.1177/09544062251358921.

[3] Xuandong Mo, Mingyuan Xia, Shenping Mei &Xiaofeng Hu. (2025). Deep reinforcement learning-based optimization for machining process: Cutting parameters dynamic optimization considering tool condition degradation. Journal of Manufacturing Processes, 156(PA), 1006-1031. https://doi.org/10.1016/J.JMAPRO.2025.11.031.

[4] Ruotong Zhang, Jiaqi Fu, Yaowei Xiang &Yameng Li. (2026). Single transition metal atoms anchored on BHS-SN2 nanosheet for CO2 electrochemical reduction to CH4. Computational and Theoretical Chemistry, 1260, 115799-115799. https://doi.org/10.1016/J.COMPTC.2026.115799.

[5] Hui Li, Aihong Zhang, Rou Jiang, Qian Zhang, Jun Du &Changyuan Tao. (2026). Hydroxyl-rich Zn/Co-based bimetallic layered catalyst for enhanced CO2 electrolysis to CO. Journal of CO2 Utilization, 107, 103412-103412. https://doi.org/10.1016/J.JCOU.2026.103412.

[6] Leibing Chen, Yali Wang, Jing Mei, Guoying Li, Han Yao, Jiaxing Lu &Huan Wang. (2026). Metal-organic framework-derived BiSn bimetallic oxides for the electrocatalytic reduction of CO2 to formate. International Journal of Hydrogen Energy, 230, 154803-154803. https://doi.org/10.1016/J.IJHYDENE.2026.154803.

[7] Beomil Kim, Seungchang Han, Suneon Wang, Stefan Ringe &Jihun Oh. (2026). Peaks and pitfalls of electrocatalytic CO2 reduction descriptor models. Nature Catalysis, (prepublish), 1-11. https://doi.org/10.1038/S41929-026-01526-7.

[8] Hui Hao, Mengyao Guo, Mingyan Li &Xiaohu Yu. (2026). Recent advances in catalytic CO2 reduction reaction by surface frustrated Lewis acid-base pairs. Applied Catalysis A, General, 719, 120931-120931. https://doi.org/10.1016/J.APCATA.2026.120931.

[9] Dan Wang, Xiangdong Meng, Kaibin Li, Junjun Mao, Yi Zhao, Qiangqiang Sun &Baoyue Cao. (2026). Mechanism of halide ions in regulating product selectivity in electrocatalytic CO2 reduction. Chemical communications (Cambridge, England), https://doi.org/10.1039/D6CC01002D.

[10] Joseph M. Liles, Matthew J. Robbins &Brian J. Lunday. (2026). Quantifying capability gaps via information relaxation and deep reinforcement learning in infinite-horizon Markov decision processes: A military air battle management application. Journal of the Operational Research Society, 77(5), 1322-1337. https://doi.org/10.1080/01605682.2025.2528915.

[11] Ajay Kumar Agrawal, Yang Zou &Hongyu Jin. (2026). Multi-objective optimization of assembly scheduling in precast building construction using deep reinforcement learning. Journal of Building Engineering, 125, 116077-116077. https://doi.org/10.1016/J.JOBE.2026.116077.

[12] Hongtao Wang, Ying Meng, Ningbo Zhang &Sizhou Sun. (2026). Optimal dispatch of grid-connected microgrids considering the uncertainty of renewable energy generation based on deep reinforcement learning. Energy Reports, 15, 109275-109275. https://doi.org/10.1016/J.EGYR.2026.109275.

[13] Jinhao Du, Yarong Chen, Jabir Mumtaz, Faisal Alkaabneh, Ziyang Ji &Kaynat Afzal Minhas. (2026). Multi-agent deep reinforcement learning for dynamic lot-streaming flow shop problems with unequal sub-lots and capacity constraints. Journal of Industrial Information Integration, 51, 101112-101112. https://doi.org/10.1016/J.JII.2026.101112.

[14] Xiancheng Feng, Jingkun Fan, Chanjuan Liu, Enqiang Zhu &Witold Pedrycz. (2026). RHMGSA: Reinforcement learning-guided evolutionary search for critical node detection. Information Sciences, 748, 123492-123492. https://doi.org/10.1016/J.INS.2026.123492.

[15] Jonaid Shianifar, Michael Schukat &Karl Mason. (2026). MO-CoERL: Multi-objective cooperative evolutionary deep reinforcement learning. Information Sciences, 748, 123517-123517. https://doi.org/10.1016/J.INS.2026.123517.

[16] Zi Qi Zhang, Yu Han Huang, Yan Xuan Xu, Bin Qian &Rong Hu. (2026). Adaptive multi-scale heterogeneous graph-attention network assisted deep reinforcement learning for distributed flexible job-shop scheduling problem. Swarm and Evolutionary Computation, 105, 102333-102333. https://doi.org/10.1016/J.SWEVO.2026.102333.

[17] Zhi Cao, Cong Zhang, Yaoxin Wu, Yaqing Hou &Hongwei Ge. (2026). Mamba-CrossAttention: Solving Flexible Job Shop Scheduling via sequence modeling and reinforcement learning. Swarm and Evolutionary Computation, 105, 102390-102390. https://doi.org/10.1016/J.SWEVO.2026.102390.

[18] Pingping Xu &Xiaobing Yu. (2026). Reinforcement learning-enhanced multi-strategy

cuckoo search algorithm to solve combined economic emission dispatch problems. Ain Shams Engineering Journal, 17(6), 104188-104188. https://doi.org/10.1016/J.ASEJ.2026.104188.

[19] Ling Gao, Hongyan Yang, Lingxiang Guo, Dantong Shen, Jinchao Li, Liang Chen. &Yaping Zhang. (2026). Designing membrane electrode assembly and optimizing process parameters for electrocatalytic CO2 reduction. International Journal of Hydrogen Energy, 232, 155034-155034. https://doi.org/10.1016/J.IJHYDENE.2026.155034.

[20] Ruifeng Miao, Chaofan Qi, Yi Lin, Xin Yong, Changqiu Chen &Huawang Zhao. (2026). Unraveling the correlation between crystal facets ((314) vs. (008)) of Bi5O7NO3 and their activity in electrocatalytic CO2 reduction to formate. Catalysis Today, 473, 115811-115811. https://doi.org/10.1016/J.CATTOD.2026.115811.