Journal of Zhejiang University-SCIENCE A in press
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2500598

An optimization scheduling strategy for electric-heat-hydrogen integrated energy systems based on memory-enhanced deep reinforcement learning

Zhongli BAI[1,2], Qiang GAO[1,2,3], Hongzhi ZHANG[1,2], Junjie LIU[1,2], Yuehui JI[1,2], Yu SONG[1,2]✉, Xu CHENG[1,4]✉

[1]*Tianjin Key Laboratory of New Energy Power Conversion, Transmission, and Intelligent Control, Tianjin University of Technology, Tianjin 300384, China*
[2]*School of Electrical Engineering and Automation, Tianjin University of Technology, Tianjin 300384, China*
[3]*Maritime College, Tianjin University of Technology, Tianjin 300384, China*
[4]*School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China*

Abstract: High wind–solar penetration drives integrated energy systems (IES) to act as cross-vector buffers that absorb surplus electricity via hybrid storage, or convert it into heat and hydrogen for cross-medium energy peak shaving. However, conventional mixed-integer linear programming (MILP) solvers and static incentive schemes often struggle with the high-dimensional, strongly coupled, non-convex, and time-varying nature of electric–heat–hydrogen scheduling. Thus, we propose an end-to-end optimization framework for a renewable electric–heat–hydrogen IES, and evaluate it through offline dispatch computations and cost settlement on a 24-hour simulated case study. The proposed framework incorporates a coupled power–state of charge (SOC) dual penalty mechanism to ensure consistent storage operation over time. Additionally, it includes a bidirectional incentive-based demand response (B-IDR) mapping that is differentiable and can capture asymmetric feedback in response to price fluctuations. Furthermore, a long-term short-term memory (LSTM)-augmented maximum-entropy Soft Actor-Critic (SAC) scheduler is utilized for stable and efficient control in this continuous and high-dimensional setting. Comparisons with other methods under identical settings show that the proposed method achieves lower operating costs and carbon emissions, as well as improved training stability.

Key words: Integrated energy system; Electric-heat-hydrogen; LSTM-SAC; Reinforcement learning; Scheduling strategy

1 Introduction

Under the goal of global carbon neutrality, governments are accelerating the green and low-carbon transformation of energy systems, with renewable energy sources such as wind, solar, combined heat and power, and hydrogen storage playing prominent roles (Kurucan et al., 2024; Washizu et al., 2024) However, the intermittency and output uncertainty associated with these sources, along with the risk of Wind Turbine (WT) and Photovoltaic (PV) being curtailed, pose challenges to renewable energy consumption. Integrated Energy Systems (IES) are designed to optimize the use of electricity, heat, and gas, promoting renewable energy consumption, lowering primary energy use, and reducing carbon emissions (Koronen et al., 2020; Wu et al., 2025). IES serve as a "buffer pool" for high renewable penetration, as well as an "energy internet" that connects sources, grids, loads, and storage, thereby enhancing system flexibility and low-carbon depth.

Energy storage is critical for addressing the bottlenecks of system flexibility and consumption caused by the randomness and intermittency of high proportions of renewable energy. The integration of hydrogen storage offers a long-term and flexible “electricity-hydrogen-electricity/heat” loop, which can significantly enhance a system's ability to absorb

✉ Yu SONG, jasonsongrain@hotmail.com
Xu CHENG xu.cheng@ieee.org

Received Nov. 14, 2025; Revision accepted Jan. 28, 2026; Crosschecked

© Zhejiang University Press 2026

fluctuations in renewable energy. Recent studies have shown that hydrogen-based IES architectures can reduce operating costs and carbon emissions (Gao et al., 2023; Li et al., 2023; Liang and Pirouzi, 2024). Additionally, coupling hydrogen energy with the power system provides scalable, long-duration energy shifting, reducing peak capacity costs and decreasing the chance of WT and PV curtailment (Hu et al., 2026). Thus, hydrogen storage has enabled IES to provide long-term regulation capabilities, which is crucial for efficient low-carbon transformations of future energy systems. However, frequent charging and discharging of energy storage accelerates degradation and introduces operational risks, weakening the resilience of systems. Methods such as limiting charging and discharging amplitudes, or using external buffers, have been proposed to address these issues (Lin et al., 2024; Wang et al., 2024; Yan et al., 2024b). While these strategies improve safety, they sacrifice flexibility and renewable energy integration potential. Therefore, more refined strategies are needed to ensure flexible operation of storage units while maintaining economic efficiency.

Demand Response (DR) is widely recognized as a core economic measure for enhancing system flexibility in scenarios with high renewable energy penetration. Existing research on this topic can be broadly categorized into two types: price-driven and incentive-based compensation schemes. The former employs time-of-use tiered electricity or heat prices to guide energy consumption adjustments. Regarding this topic, Wang established an electricity-heat coupling dispatch model based on time-of-use electricity prices and time-of-use gas prices, which improved the digestibility of renewable energy output while minimizing energy procurement costs(Wang et al., 2025b). The latter focuses on directly compensating or rewarding users for load-shifting behavior. In this vein, Ampimah et al. used peak-shaving incentives to encourage residents to avoid peak periods and alleviate supply-side pressure (Ampimah et al., 2018). Also, Chen et al. combined carbon trading costs with incentive-based DR to achieve economic and emission reduction goals(Chen et al., 2025). Moreover, Luo et al. proposed a real-time hybrid price-incentive mechanism, which reduced the peak-to-average ratio in commercial parks by 18% and effectively smoothed PV fluctuations(Luo et al., 2023). Notably, existing studies have largely focused on electrical loads but overlooked dynamic compensation for heat loads or coupled electrical-heat loads, which is essential for meeting the flexibility needs of integrated energy systems.

The random fluctuations of renewable energy output, the dynamic constraints of power generation units and energy storage, and nonlinear responses on the demand side are intertwined and superimposed in the spatiotemporal dimension; this gives rise to the four characteristics of “high-dimensionality, strong coupling, non-convexity, and dynamism” in comprehensive energy dispatch problems. Aghdam et al. linearized such problems into a format for mixed integer linear programming (MILP) and solved them using GAMS-CPLEX (Hamzeh Aghdam et al., 2024). Moreover, Liu et al. converted a nonlinear IES planning problem into a two-layer MILP and solved it using the Guesthouse algorithm based on an energy sharing mechanism to minimize operating costs(Liu et al., 2022). However, these methods inevitably require radical linearization of power generation unit charging and discharging rates, multi-energy sources, and demand-side elasticity in order to achieve computational tractability. To overcome the limitations of linear solvers, recent research has begun to leverage deep reinforcement learning (DRL) to learn scheduling strategies through “state-action-reward” interactions in complex environments, thereby addressing the high-dimensional, strongly coupled, and dynamic characteristics of IES. For example, Liang et al. used a soft actor-commentator (SAC) framework to embed variables such as multi-energy power and state of charge (SOC) into a high-dimensional state-action space, and composed a reward function based on cost and carbon emissions to verify the scheduling advantages of SAC in nonlinear time-varying IES (Liang et al., 2024). However, the above approaches generally rely on feedforward neural networks, which are unable to explicitly capture the time correlations between price, load, and renewable power output, leading to strong oscillations or delays in unit output when switching between long-term and short-term scales.

Overall, while existing research has advanced IES scheduling for multi-energy coupling, storage

asset protection, demand-side response modeling, and optimization/learning solutions, several challenges remain: (i) storage usage often over-relies on rigid power limits, which protects assets but may sacrifice system flexibility under the volatility of renewables; (ii) price-based incentive schemes are frequently built upon static linear elasticity assumptions, making it difficult to capture state-dependent and asymmetric demand-side feedback; and (iii) many DRL schedulers lack temporal memory, so policies tend to oscillate when switching across multiple time scales under intertemporal constraints. In contrast to traditional DRL-based IES scheduler methodologies, in this study we leverage bidirectional incentive-based demand response modeling to address issue (ii), while using a storage-friendly power–SOC soft-penalty approach and a sequence-aware LSTM-augmented DRL scheduler to account for issues (i) and (iii), respectively. These serve as practical refinements that strengthen asset-awareness and temporal stability within the mainstream DRL dispatch framework.

This study makes three contributions:

i. SOC-consistent flexibility: a coupled power–SOC soft-penalty for stress-aware storage dispatch beyond hard limits.

ii. Bidirectional Incentive-Based Demand Response (B-IDR): a differentiable bidirectional IDR mapping that captures asymmetric price-rise/fall responses within feasibility bounds.

iii. Sequence-aware DRL scheduling: an LSTM-augmented maximum-entropy SAC scheduler that enables stabilized and continuous control across multiple time scales.

The remaining content is organized as follows: Section 2 introduces the mathematical model of the power generation unit; Section 3 establishes an economic dispatch model that balances energy storage characteristics using B-IDR; Section 4 presents the LSTM-SAC method and its solution process; Section 5 re-ports the experimental setup and results; Section 6 presents a discussion of the model; finally, Section 7 outlines the conclusions. The system architecture is shown in Fig. 1.

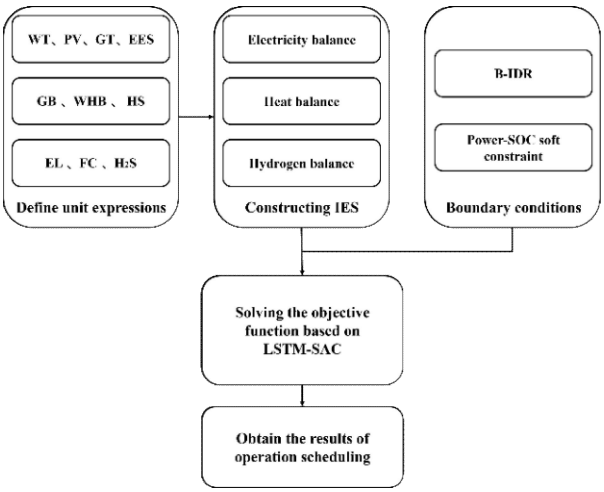

Fig. 1 Flowchart of IES with demand response

2 IES and the unit mathematical model

2.1 IES structural composition

To achieve coordinated optimization and efficient low-carbon utilization of multiple energy sources, we propose an IES that couple's electricity, heat, and hydrogen. The structure of this system is illustrated in Fig. 2. The IES has typical features such as multi-source input, multi-energy complementarity, and multi-directional output, covering three main external energy access channels: the electricity grid, gas grid, and heat grid. Within the system, various types of energy conversion and storage equipment are integrated, including electrolysis cells (EL), fuel cells (FC), gas turbines (GT), gas boilers (GB), waste heat boilers (WHB), EES, heat storage units (HS), and hydrogen storage units (H_2S). This constitutes a complete energy ecosystem loop that encompasses energy production, conversion, storage, and consumption.

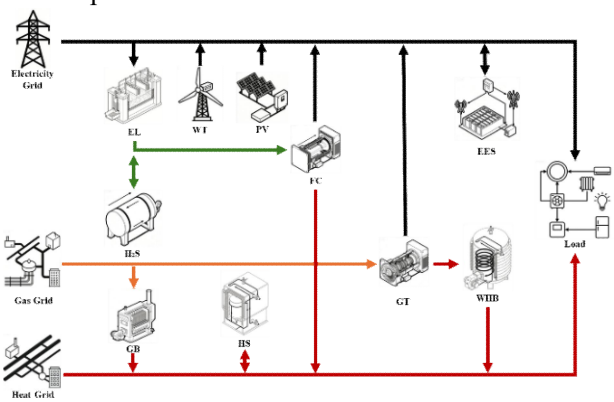

Fig. 2 The structure of the IES

2.2 Mathematical model of the unit

During system operation, the EL can utilize fluctuating renewable energy sources or external power resources during off-peak electricity price periods to produce hydrogen, thereby achieving cross-medium conversion of electrical energy to hydrogen energy (Jeon and Bae, 2025). The generated hydrogen can be stored in hydrogen storage tanks and used to provide electricity or heat across various time periods. FC generates both electricity and heat through the electrochemical reaction of hydrogen, thereby meeting end-use load requirements while reducing the need to purchase external electricity and heat(Shi et al., 2023). GT generates electricity by using gas as fuel. The associated high-temperature exhaust gases are further recovered for heat production by WHB, accordingly improving the comprehensive utilization efficiency of gas(Sadeghian et al., 2025). When heat load demand is high or renewable energy is insufficient, the system can invoke GB as peak shaving heat sources, directly converting gas into heat energy to meet heat demand(Narayanan, 2021).

The energy storage system established in this study achieves peak shaving and valley filling as well as cross-time transfer of electrical energy, heat energy, and hydrogen energy on various time scales(Nikoobakht et al., 2023). This enhances the system's load adaptability and low-carbon operational characteristics. The power function of the unit is presented in Section S1 of the Electronic Supplementary Materials.

3 Low-carbon economic dispatch model

In the dynamic operation of IES, load regulation capabilities and price response behavior on the demand side are important factors in achieving system flexibility and low-carbon goals(Jiang et al., 2024). This section therefore outlines the refinement of the energy storage system model and the integrated demand-side response (IDR) mechanism.

3.1 Energy storage operating cost modeling

In IES, energy storage systems not only perform the task of shifting energy balances in time, but also play a key role in supporting system flexibility in multi-time domain optimization(Li et al., 2021b). In order to quantify the actual operating costs and life impact of EES during frequent dispatch response processes, we construct a nonlinear operating cost model based on physical state evolution and energy coupling (Li et al., 2025). The model accounts for both the energy charge and discharge levels per unit time and the frequency of changes in the SOC, thereby reflecting the impact of its operating behavior on both economic efficiency and sustainability.

The EES operating cost function is as follows:

$$C_{\mathrm{EES}}(t) = \alpha_{\mathrm{EES}}(P_{\mathrm{EES}}^{\Sigma}(t))^{\beta_{\mathrm{EES}}} \cdot \left|\Delta E_{\mathrm{EES}}(t)\right|^{\varpi_{\mathrm{EES}}} \quad (1)$$

among these variables, $\alpha_{\mathrm{EES}} > 0$ represents the operating cost benchmark coefficient, the power index $\beta_{\mathrm{EES}} > 1$ characterizes the nonlinear response characteristics of energy storage to power intensity regulation, and $\varpi_{\mathrm{EES}} < 0$ describes the potential accelerated degradation of its lifespan and health status due to changes in SOC frequency.

The power-intensity term is evaluated in a normalized form in order to avoid unit ambiguity. Under this definition, $\beta_{\mathrm{EES}}, \varpi_{\mathrm{EES}}$ are dimensionless, while α_{EES} carries the monetary unit per time step (CNY/h) to ensure that $C_{\mathrm{EES}}(t)$ is commensurate with other cost components in the objective. The power–SOC penalty's dependence on charge/discharge power and SOC-trajectory variation is motivated by literature (Hannan et al., 2021; Lee and Kim, 2022; Yan et al., 2024a) linking cycling wear to rate/intensity and SOC excursion, as well as cycle-based degradation formulations that map SOC profiles to equivalent cycling severity. The coefficients are determined through a preliminary tuning study to achieve an acceptable balance between economic performance and cycling-intensity proxies under the present system's time discretization.

3.2 Nonlinear mapping modeling and feed-back coupling method for the incentive response mechanism

IDR has become one of the main strategies for achieving load flexibility, peak shaving, and carbon

emission reduction control. This paper proposes a bidirectional (B)-IDR with price feedback characteristics. By constructing a nonlinear response mapping based on price perturbations, the mechanism enables dynamic adaptation and adjustment of demand-side load in response to price incentive signals for electricity and heating. This mechanism is formalized as a state-dependent variational system mapping.

The behavioral response of the demand side to price incentives can be formalized as the following nonlinear mapping process:

$$\mathcal{R}:\mathbb{R}^2\times\mathbb{R}_+^2\times\mathbb{R}_+^2\to\mathbb{R}_+^2 \tag{2}$$

$$\mathcal{R}(u,p,\tau)=\begin{bmatrix}\mathcal{F}_e(u_e,p_e,\tau_e)\\ \mathcal{F}_h(u_h,p_h,\tau_h)\end{bmatrix} \tag{3}$$

where $\boldsymbol{u}=[u_e,u_h]^{\mathrm{T}}\in[-1,1]$ is the electric/heat incentive signal calculated by the agent, $\boldsymbol{p}=[p_e,p_h]^{\mathrm{T}}\in\mathbb{R}_+$ is the current actual electric/heat price vector, and $\boldsymbol{\tau}=[\tau_e,\tau_h]^{\mathrm{T}}\in[0,1]$ uses the price difference before and after the transfer as the B-IDR sharing factor. The mapping function $\mathcal{F}_x(\cdot)$ is defined as a demand-side power-law-type load reduction operation for the price difference, and has the following variational form:

$$\mathcal{F}_x(u_x,p_x,\tau_x)=\begin{cases}\tau_x\cdot\mathcal{L}_x(t)\cdot\left(\dfrac{p_x+u_x-\overline{p}_x}{\overline{p}_x}\right)^{\beta_x}, & p_x+u_x>\overline{p}_x\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

where $x\in\{e,h\}$, $\overline{p}_x$ represents the daily average reference price, and $\beta_x>1$ controls the price sensitivity of the demand side. This mapping satisfies the conditions of non-negativity, locally Lipschitz continuous, and first-order differentiable but second-order discontinuous (the proof is provided in Section S2 of the Electronic Supplementary Materials).

Under the B-IDR mechanism, we introduce a backlog state variable $B_x(t+1)\in\mathbb{R}_+$ to characterize the unmet load after demand reduction, whose dynamic evolution is constrained by the replenishment mapping:

$$B_x(t+1)=B_x(t)+\delta_x(t)-\varepsilon_x(t) \tag{5}$$

$$\varepsilon_x(t)=\min\{B_x(t),\tau_x\cdot\mathcal{L}_x(t)\cdot\left(\frac{p_x+u_x-\overline{p}_x}{\overline{p}_x}\right)^{\beta_x}\cdot 1_{p_x+u_x>\overline{p}_x}\} \tag{6}$$

where $\delta_x(t),\varepsilon_x(t)$ represent the momentary reduction and replenishment of load capacity, respectively.

The adjusted load is defined as:

$$\mathcal{L}_x^{\mathrm{adj}}(t)=\mathcal{L}_x(t)-\delta_x(t)+\varepsilon_x(t),x\in\{e,h\} \tag{7}$$

Demand-side participants in this response behavior need to be given certain incentives, and so the incentive cost is given by:

$$C_x^{\mathrm{inc}}(t)=\max\{\frac{p_x+u_x-\overline{p}_x}{\overline{p}_x},0\}\cdot\overline{p}_x\cdot\delta_x(t)\cdot(1-\tau_x)$$

We also introduce a perceptual feedback coupling utility loss function as an inappropriate cost:

$$C_x^{\mathrm{unsat}}(t)=\kappa_x\cdot[\delta_x(t)]^2 \tag{9}$$

where κ_x is the inadequacy parameter.

B-IDR behavior as a feedback mapping system is controlled by the dispatching agent $\pi\in\Pi$, with continuous disturbance $\boldsymbol{u}(t)$ as the input and demand-side correction load $\mathcal{L}^{\mathrm{adj}}(t)$ as the output, satisfying:

$$\mathcal{L}^{\mathrm{adj}}(t)=\mathcal{L}(t)-\mathcal{R}(\boldsymbol{u}(t),\boldsymbol{p}(t),\boldsymbol{\tau}(t))+\mathcal{G}(\boldsymbol{u}(t),B(t)) \tag{10}$$

where $\mathcal{G}(\cdot)$ is a replenishment function that satisfies conditions of partial differentiability and boundedness.

In summary, the B-IDR mechanism can be regarded as a nonlinear compression response operator $\Gamma:\boldsymbol{u}\to\mathcal{L}^{\mathrm{adj}}$, which under the scheduling agent control strategy, constitutes a controlled feedback micro-game system. As such, it provides a continuous differentiable policy space structure for subsequent reinforcement learning optimization.

3.3 IES low-carbon economic dispatch model

To achieve IES's operational goals of meeting multi-energy load demands while balancing economic efficiency and carbon emission control, we construct a low-carbon economic dispatch model based on time-series optimization. The model comprehensively considers multiple cost factors, including energy procurement costs, carbon emission costs, equipment operating losses, energy curtailment penalties, and B-IDR costs, while also ensuring the energy supply and demand balance of the system and

equipment operating constraints. As such, this forms a complete system-level objective function.

3.3.1 Objective function

Within time period $t = 1, \cdots, T$, the scheduling objective is to minimize the weighted operating cost of the system:

$$\min F_{\mathrm{IES}} = \sum_{t=1}^{T}[C_{\mathrm{ebe}}(t) + C_{\mathrm{co_2}}(t) + C_{\mathrm{dev}}(t) + C_{\mathrm{DR}}(t) + C_{\mathrm{curt}}(t)] \tag{11}$$

where F_{IES} is the total cost, and the definitions of the other variables are provided in Section S3 of the Electronic Supplementary Materials.

3.3.2 Constraints

To ensure the feasibility and physical consistency of the IES dispatch, a series of system-level constraints that satisfy multi-energy balance between electricity, heat, and hydrogen, equipment operating boundaries, dynamic state evolution, response rules, and emissions re-strictions must be introduced based on the objective function. We formulate the IES multi-energy coordination process as a dynamic nonlinear controlled system, with its set of constraints provided in Section S4 of the Electronic Supplementary Materials (ESM).

The B-IDR behavioral boundaries and temporal balance must be explicitly modeled. First, the maximum load reduction on the demand side at any given time must be controlled within 20% of the total load, which is expressed as:

$$\delta_x^{\min}(t) \le \delta_x(t) \le \delta_x^{\max}(t) \tag{12}$$

Meanwhile, the cumulative replenishment must not exceed the maximum acceptable rebound capacity which is set to 60% of the total flexible load – in order to avoid uncontrollable rebound and excessive demand-side perception delays:

$$B_x(t) \le B_x^{\max}(t) \tag{13}$$

A sensitivity experiment for the instantaneous load reduction limit is detailed in Section S6 of the ESM.

To maintain the sustainability of the system's energy balance, we set a conservation constraint for the total amount of reduction and replenishment throughout the entire cycle, namely:

$$\sum_{t}^{T}\delta_x(t) = \sum_{t}^{T}\varepsilon_x(t) \tag{14}$$

4 LSTM-SAC construction for IES energy management

The operation of IES involves high penetration rates of random wind and solar power output, multi-energy coupling of electricity, heat, and hydrogen, and cross-temporal balancing of energy storage. Real-time dispatching issues are characterized by high dimensionality, non-linearity, and strong uncertainty(Gea-Bermúdez et al., 2023). Therefore, we propose a hybrid algorithm combining LSTM (Abbasimehr et al., 2020) and SAC (Han and Sung, 2021) to meet these challenges. First, a maximum differential entropy policy is established as the model objective function. LSTM networks are introduced at the SAC front ends to utilize gated memory mechanisms in order to extract short-term dynamic features from wind power output and load sequences, thereby obtaining more forward-looking energy management strategies.

4.1 Markov decision construction

The IES dynamic process can be regarded as a Markov decision process, which can be abstracted into three major parts: a state vector, action vector, and reward function (Wang et al., 2025a). The state of the IES determines the agent's output scheduling method for each connected component in the system. A schematic diagram of this process is shown in Fig. 3, where the state vector is as follows:

$$\begin{aligned} \mathrm{s}_t = \{&P_{\mathrm{wt}}(t), P_{\mathrm{pv}}(t), \mathcal{L}_e(t), \mathcal{L}_h(t), \\ &SOC_{\mathrm{EES}}(t), SOC_{\mathrm{HS}}(t), SOC_{\mathrm{H_2S}}(t), t\} \end{aligned} \tag{15}$$

The agent minimizes the objective function by reducing overall operating costs through energy conversion units and the charging and discharging of energy storage units. After considering load adjustment, GT, and GB power in the action space, we optimize EL and FC to determine the state of the hydrogen storage tank. Finally, the energies of the electrochemical and heat storage units are

dynamically planned. The action vector is defined as follows:

$$a_t = \{\mathcal{L}_e^{\mathrm{adj}}(t), \mathcal{L}_h^{\mathrm{adj}}(t), P_{\mathrm{gt}}(t), Q_{\mathrm{gb}}(t), P_{\mathrm{el}}(t), P_{\mathrm{fc}}(t), P_{\mathrm{EES}}^{\mathrm{ch/dis}}(t), P_{\mathrm{HS}}^{\mathrm{ch/dis}}(t)\} \tag{16}$$

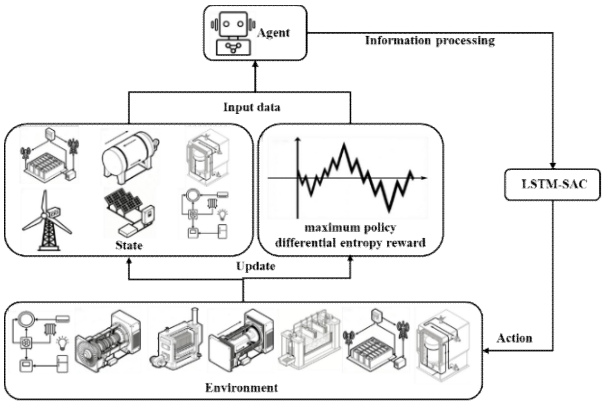

Fig. 3 Framework of the LSTM-SAC–based Markov decision process

The SAC actor outputs normalized actions via a squashed Gaussian policy. Each action dimension is linearly rescaled to the corresponding physical range, and then the feasibility layer and clipping are applied to enforce device limits prior to the environment state transition.

4.2 LSTM-SAC algorithm model

In the real-time dispatch scenarios of IES, WT, and PV, as well as demand-side loads, all exhibit significant short-cycle fluctuations (Li et al., 2021a). Embedding LSTM into the SAC policy network allows the evolution of recent random processes to be explicitly encoded in the hidden state, enabling the scheduler to have "endogenous foresight" capabilities even without an external prediction model (Abri and Abri, 2025). The proposed LSTM-SAC scheduler follows the maximum-entropy (Kangfeng et al., 2020) reinforcement learning paradigm, which augments the classical cost-minimizing objective with a differential-entropy regularization term. This entropy regularization encourages sufficient exploration in continuous action spaces and mitigates premature convergence to overly deterministic policies in model-free decision-making. For a stochastic policy $\pi_\varphi(a \mid s)$, the conditional differential entropy is defined as:

$$\mathcal{H}(\pi_\varphi(\cdot \mid s_t)) = -\int \pi_\varphi(a \mid s_t) \log \pi_\varphi(a \mid s_t) \mathrm{d}a = -\mathbb{E}_{a_t \sim \pi_\varphi(\cdot \mid s_t)}[\log \pi_\varphi(a_t \mid s_t)]. \tag{17}$$

In this study, the environment provides a cost-based reward by negating the one-step comprehensive operating cost. Let $g(s_t, a_t)$ denote the one-step cost counterpart of the dispatch objective, then:

$$\mathrm{R}_t = r(s_t, a_t) = -g(s_t, a_t) \tag{18}$$

where $g(s_t, a_t)$ represents the energy purchase, O&M costs, carbon-related costs, curtailment penalties, and B-IDR-related costs incurred at time *t*.

Maximizing the differential entropy is implemented by adding an entropy regularization term $-\Upsilon \log \pi_\varphi(a_t \mid s_t)$ to the expected return (Υ represents the differential entropy policy coefficient). In order to promote better decision making by the LSTM-SAC, the reward function results are scaled to the interval $[-10, 0]$ (Xu et al., 2023).

The entropy-regularized objective of the "maximum differential entropy policy" is written as:

$$\mathcal{J}(\pi_\varphi) = \mathbb{E}_{(s_t, a_t) \sim \rho\pi_\varphi}[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \Upsilon \mathcal{H}(\pi_\varphi(\cdot \mid s_t)))] = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (-g(s_t, a_t) + \Upsilon \cdot \log \pi_\varphi(a_t \mid s_t)],$$

where γ represents the discount factor. SAC explores strategy π_φ in a randomized manner to obtain the maximum discount utility.

The algorithm uses a double soft Q net-work to approximate the action value by minimizing the soft Bellman residual:

$$\xi_Q(\psi) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}}[(Q_\psi(s_t, a_t) - (r_t + \gamma \cdot V_{\bar{\psi}}(s_{t+1})))^2] \tag{20}$$

$$V_{\bar{\psi}}(s_t) = \mathbb{E}_{a_t \sim \pi_\varphi(\cdot \mid s_t)}[Q_{\bar{\psi}}(s_t, a_t) - \Upsilon \cdot \log \pi_\varphi(a_t \mid s_t)] \tag{21}$$

Using a Kullback–Leibler divergence drive strategy network π_φ , we converge to a soft distribution (Alexopoulos, 2021):

$$\mathcal{K}_\pi(\varphi) = \mathbb{E}_{s_t \sim \mathcal{D}}[\mathcal{D}_{KL}(\pi_\varphi(\cdot \mid s_t) \parallel Z^{-1}(s_t) \exp(Q_{\bar{\psi}}(s_t, \cdot) / \Upsilon)] \tag{22}$$

Differential entropy ensures that the strategy

remains multimodal in a continuous action space, thereby improving robustness to random fluctuations. The double- Q structure mitigates value overestimation bias and adapts to the real-time rolling optimization requirements of integrated energy systems.

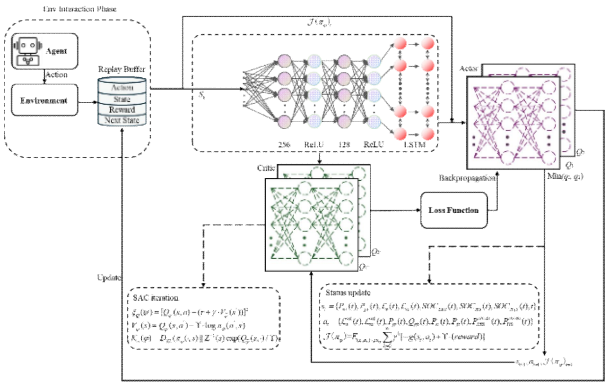

Fig. 4 Systematic process of the LSTM-SAC

In this study, a deep reinforcement learning method based on LSTM–SAC was developed for real-time scheduling of an IES with electro-thermal-hydro storage coupling. This approach follows the three main technical lines of SAC. First, the actor-critic architecture is constructed using a double-depth neural network approximating random strategy π_φ and soft action value Q_ψ. Second, the entropy maximization principle adds a differential entropy term weighted by the entropy coefficient Υ to the objective function, which enhances exploration strength and suppresses overestimation of the value function. Third, we separate the strategy update mechanism and use the experience replay buffer to repeatedly collect samples, thereby improving sample utilization and training stability. Unlike the classic SAC, LSTM encoders are pre-positioned at the strategy and value ends. We then map historical WT and PV power output and multi-energy load sequences to latent vectors, endowing the intelligent agent with endogenous predictive capabilities for short-term trends, and pre-adjusting the unit status to adapt to load changes. The overall structure of the system is shown in Fig. 4. The feature sequence obtained from the two-layer feedforward MLP is fed into the LSTM, and the hidden state is projected linearly to generate global features. After receiving the characteristics, the actor samples actions based on the current state. The critic network calculates the loss function for actor s_{t+1}, a_{t+1} and $\mathcal{J}(\pi_\varphi)_{t+1}$ outputs soft Q network updates to the actor with backpropagation, and continuously stores quadruplets $(s_t, a_{t+1}, \mathcal{J}(\pi_\varphi)_{t+1}, s_{t+1})$ in the replay buffer. After each fixed number of steps, the algorithm extracts a batch of samples from the replay pool and alternately updates ψ, φ, Υ according to Eqs. (20)-(22). The expected utility is iterated through via backpropagation until convergence, completing one off-policy optimization cycle.

5 Case study and results

5.1 Scenario description and simulation

This paper proposes a method with a system environment constructed using Python 3.11, PyTorch, and OpenAI Gym. Experiments were conducted on a workstation with an Intel Core i5 processor, 32GB of memory, and an NVIDIA RTX 4070 SUPER GPU. The parameter settings were: learning rate=1e-3, batch size=256, gamma=0.99 (discount factor), tau=0.005 (soft target update rate), with approximately 1.48 MiB of trainable parameters. The IES model represents a mixed industrial and commercial park in northern China, where energy coupling equipment operates stably across various load types. The scheduling analysis uses a 24-hour cycle with a 1-hour time unit, and the experimental data and parameters are provided in Section S3 of the ESM.

Five scenarios are constructed with the goal of increasing system economic benefit and reducing carbon emissions through the use of energy storage and B-IDR. Scenario 1 quantifies the costs under the baseline for the region, providing a basis for analysis in other scenarios. Scenario 2 includes B-IDR, and is used to analyze its impact on the system. Scenario 3 includes energy storage units, and scenario 4 removes hydrogen conversion and hydrogen storage units from scenario 3. We accordingly explore the impact of energy storage units and the introduction of hydrogen energy on the economic efficiency and carbon emissions of the system. Scenarios 3 and 4 are designed as a controlled ablation pair to isolate the contribution of the hydrogen pathway: Scenario 3 enables the EL–H_2S–FC loop (electricity – hydrogen – electricity/heat) while Scenario 4 removes hydrogen

conversion and storage, keeping the remaining settings identical. Scenario 5 includes a complete energy storage system and B-IDR mechanism, and all scenarios include carbon trading.

5.2 Experimental results

5.2.1 Optimized scheduling results and analysis

In Scenario 5, the energy storage system and B-IDR demand response are jointly optimized (Figs. 5–8). B-IDR shifts the morning peak by accounting for cumulative transfer constraints and uses the 18:00 price signal to pre-release a small portion of demand, alleviating the 19:00 peak and reducing procurement costs (Fig. 5). For the electricity balance (Fig. 6), 01:00–07:00 is off-peak (25–35 MW), and mainly supplied by WT and FC with minor GT support. Low prices drive the power-to-hydrogen subsystem into "production mode" to generate hydrogen/heat (Figs. 7–8), while remaining renewables charge EES; however, a small amount of renewable curtailment still occurs. From 08:00–11:00, rising demand and insufficient renewables increase GT output to ~20 MW, with EES and H_2S discharging for coordinated peak shaving, and EL reducing consumption. During 12:00–16:00, PV/WT peak output reduces GT. From 17:00–21:00, PV drops to zero and the evening peak requires EES support, GT ramp-up, and partial external procurement. From 22:00–24:00, WT and GT re-balance the system for the next cycle.

The heat trajectory shows a nighttime plateau, midday trough, and evening rise (Fig. 7). In the early hours, low gas prices enable GB to reach 30 MW while HS discharges to meet demand. Between 06:00–10:00, as gas prices rise and heat demand falls, GB output decreases; FC begins and WHB recovers GT waste heat, covering >80% of demand at low marginal cost. A brief GT power peak around 10:00 raises the recoverable heat to 16.8 MW, minimizing GB output and avoiding forced HS discharge. During the midday trough, FC+WHB with HS discharge supplies >70% of the heat, keeping GB at low duty. From 15:00–21:00, increasing electric load raises GT to 20 MW (20:00), yielding up to 17 MW waste heat (~57% of heat load), while EL and FC co-operate; the resulting CHP net cost is at least 40% lower than external procurement. At 24:00, HS releases 1.5 MWh to cap GB ramping at 6 $MW \cdot h^{-1}$ and maintain safe operation.

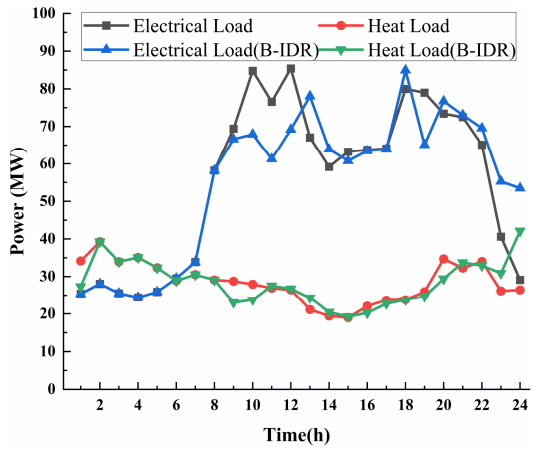

Fig. 5 The load adjustment through B-IDR

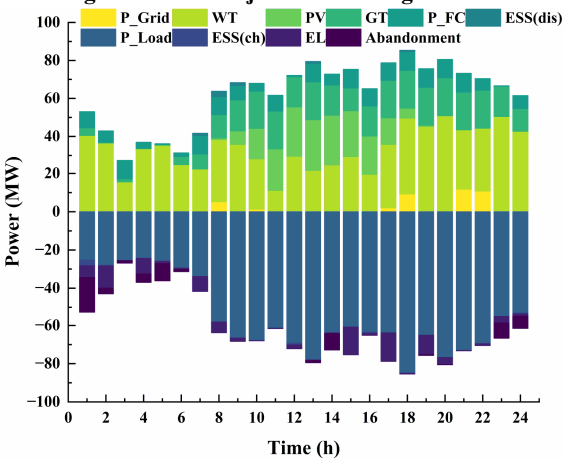

Fig. 6 The electricity balance of IES

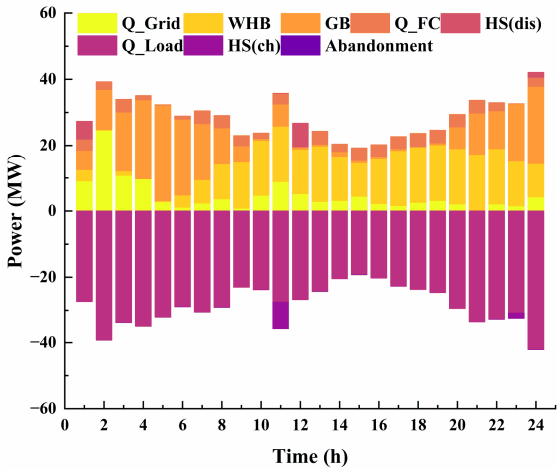

Fig. 7 The heat balance of IES

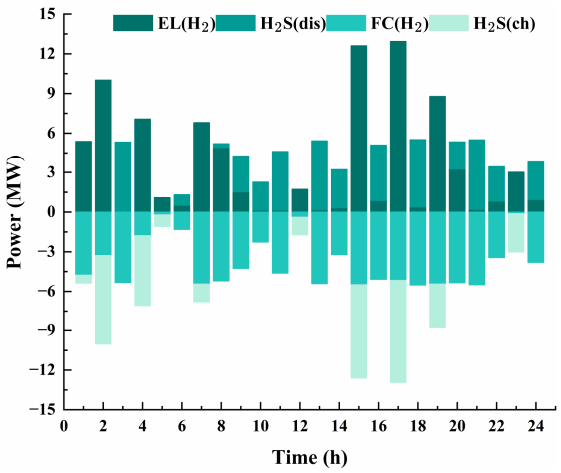

Fig. 8 The hydrogen balance of IES

5.2.2 Results comparison between different scenarios

Fig. 9 compares the 24-h economic costs of the five scenarios. In Scenario 1 (baseline), no flexibility is available, yielding a total cost of 3.80 × 10^5 CNY (~55,000 USD); energy procurement contributes above 70% and leads to an energy penalty of 9.03 × 10^3 CNY (~1,306 USD). With high renewable penetration, dispatch is dominated by the trade-off between high-carbon grid purchases and curtailment penalties. Introducing B-IDR (Scenario 2) transfers marginal costs to the demand side via real-time prices, suppressing peaks and reducing procurement costs by 18.8%. The total cost and carbon cost drop by 13.95% and 13.17%, respectively. Although incentives of 1.22 × 10^4 CNY (~1,765 USD) are paid, the demand-side "excess cost" is only equivalent to 1.22 × 10^3 CNY (~177 USD), implying that price signals can mobilize participation without materially degrading user experience.

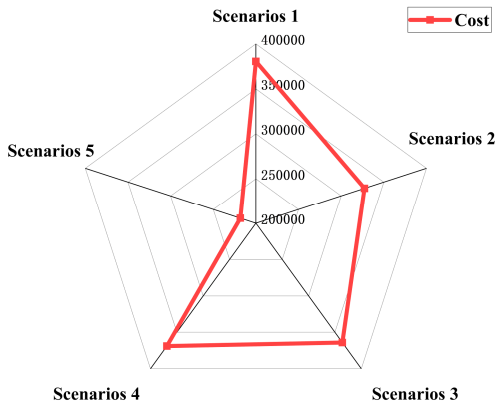

Fig. 9 Total cost comparison between different scenarios

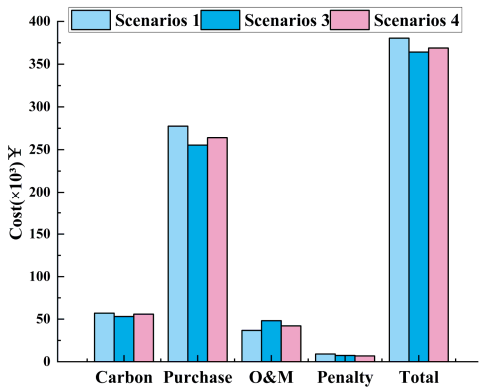

Fig. 10 IES cost comparison between different energy storage systems

In storage-only cases (Fig. 10), Scenario 3 (EES+HS+H_2S) uses the EL–H_2S–FC pathway to convert surplus/low-price renewable power into hydrogen and discharge via FC at peak hours, reducing procurement costs from 2.78 × 10^5 CNY (~40,237 USD) to 2.55 × 10^5 CNY (~36,908 USD, -8.27%) and lowering carbon-related costs as well (-6.94%), though at the cost of higher EL/FC O&M. Scenario 4 (EES+HS) mitigates short-/mid-term fluctuations with lower O&M, achieving a slightly lower total cost but weaker reductions in procurement and carbon costs (4.84% and 1.97%, respectively). This shows hydrogen's value in longer-horizon electricity–hydrogen–electricity/heat shifting and improved low-carbon substitution, with its competitiveness depending on EL–FC O&M. When B-IDR is coordinated with three storage systems (Scenario 5, Fig. 11), the total and carbon costs decrease by 42.65% and 44.49%, respectively, surpassing the additive effects of Scenarios 2 and 3. Demand shifting increases the FC discharge window, storage buffers IDR, and multi-energy coupling reduces peak-shifting costs, with curtailment costs dropping to 4.78× 10^3 CNY (~691 USD). Overall, EES/HS handle short-to-mid-term deficits, H_2S addresses intraday energy mismatches, and B-IDR unlocks demand flexibility for cross-scale peak shaving and valley filling.

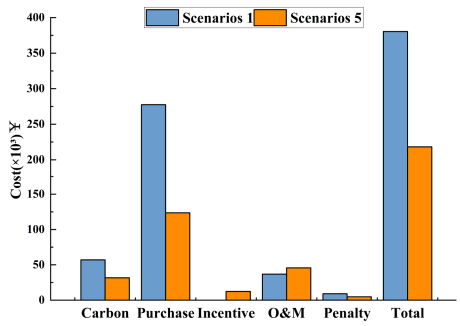

Fig. 11 Cost comparison of scenario 5 and scenario 1 at each stage

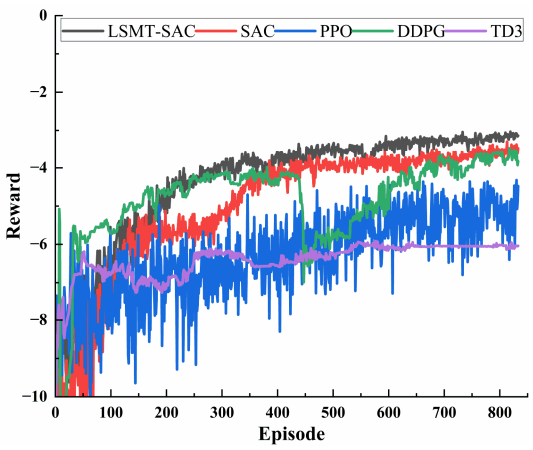

Fig. 12 Reward function curve

6 Discussion

6.1 Model performance and mechanism analysis

We implemented SAC, PPO, DDPG, and TD3 in the same environment as LSTM-SAC, comparing all algorithms on Scenario 5 (DR + storage). To ensure fairness, the training budget, uncertainty setting (seed = 2025), state/action definitions, and constraint-handling procedures were identical across all algorithms, with hyperparameters tuned within narrow windows around the default values. The final configurations are summarized in the ESM. As shown in Fig. 12, LSTM-SAC and DDPG enter a sustained improvement phase around 100 episodes, while PPO and TD3 show oscillations until 200 episodes, indicating less sample-efficient updates. The SAC baseline – aided by entropy regularization – accelerates early exploration and surpasses PPO around 150 episodes. After convergence, LSTM-SAC stabilizes around −3.0∼−3.2 after 600 episodes, outperforming SAC (−3.9), DDPG (−4.9), PPO (−5.2), and TD3 (−6.1). This corresponds to a 7% reduction in daily operating costs compared to SAC, and nearly 31% savings compared to TD3. The stability advantage is confirmed by the results in Table 1, where LSTM-SAC's plateau-stage standard deviation (0.3) is smaller than SAC's (0.45) and DDPG's (0.6). Overall, LSTM-SAC offers faster convergence, lower cost, and higher stability, benefiting from entropy-regularized exploration and sequence-aware temporal encoding.

The EL–H_2S–FC loop provides an additional intertemporal flexibility channel that complements EES/HS by decoupling electricity production and consumption over longer intra-day intervals. As illustrated by the hydrogen balance (Fig. 8), hydrogen is typically produced during low - price/renewable - surplus hours and converted back during peak demand, which reduces high-carbon grid purchases and improves carbon-related outcomes. The corresponding economic gain, however, competes with the O&M cost of the hydrogen conversion chain; therefore, the hydrogen subsystem becomes increasingly advantageous under larger price spreads, higher renewable surpluses, or longer scheduling horizons where long duration shifting is more valuable.

Table 1 Performance comparison between continuous-control DRL algorithms

Algorithm	Mean episodic reward	Std. dev.	Converged level (approx.)
LSTM-SAC	-3.18	0.30	~ -3.2
SAC	-3.92	0.45	~ -3.9
DDPG	-4.90	0.60	~ -4.9
PPO	-5.23	0.85	~ -5.2
TD3	-6.10	0.72	~ -6.1

The proposed LSTM-SAC is lightweight (≈1.48 MiB parameters) and requires 20,000 timesteps for training, equating to approximately 833 episodes with 1-hour resolution (T=24 steps). Early stopping is used to halt training if the reward standard deviation over 20 episodes falls below 0.5; this typically results in convergence under 1 hour on an RTX 4070 SUPER GPU. After training, offline dispatch is generated in seconds with each step involving policy evaluation, feasibility mapping, and state transition/cost updates. Memory is mainly used by the replay buffer, requiring ~0.3 GB for a full buffer of 1e6 transitions. For larger systems, simulation and constraint-handling costs scale linearly with the number of components, while

learning complexity increases with state–action dimensions and sample demand. The scalability is enhanced via action factorization, parameter sharing, and modular state encoding. The learning and feasibility layers also extend to hierarchical day-ahead and real-time operations, where the day-ahead layer generates schedules and the real-time layer updates schedules with receding-horizon corrections.

6.2 Comparison of energy storage cost functions

The physical state evolution and power coupling cost model introduced in this paper links instantaneous charging and discharging power with SOC gradients in adjacent time periods as a binary term. By simultaneously depicting deep charge/discharge cycles and high-frequency action penalties, it more accurately reflects the marginal cost of batteries under high-frequency large-scale scheduling. To test the impact of this nonlinear model on scheduling strategies and system economics, we employ a traditional linear cost function in the comparative experiment, namely.

By comparing the performance of the "coupling cost" and "linear cost" models under the same optimization framework, we can intuitively reveal the impact of cost function morphology on EES performance. The results of this test are shown in Table 2. Relative to traditional line-based energy pricing, the proposed coupled cost function exhibits completely different "self-constraining" characteristics at the dispatch level. The daily charge-discharge capacity of the EES converged from 21.69 MWh to 12.68 MWh, with high-power charge-discharge events sharply decreasing from 3 and 4 times to 1 time each. The total number of operations was also reduced from 14 to 12.

Table 2 Experimental results for different EES functions

	Coupling function	Linear function
Daily capacity	12.6824 MWh	21.69412 MWh
High-power charges	1 time	3 times
High-power discharges	1 time	4 times
Operations	12 times	14 times

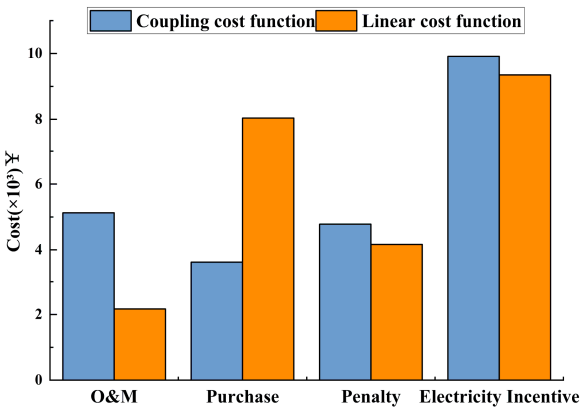

Fig. 13 Cost comparison between different EES cost functions

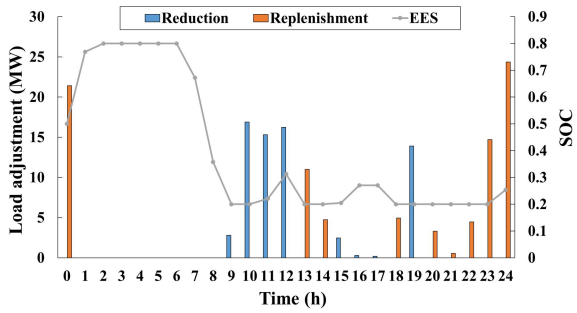

Fig. 14 Load adjustments and EES dynamics

Although the coupling cost increases the EES's own operating and maintenance costs, it triggers multiple chain benefits at the system level; the comparison results of different cost functions are shown in Fig. 13. Due to the suppression of deep cycling, conventional units can maintain stable power operation, reducing frequency changes and avoiding energy gaps caused by ramping restrictions during sudden peak loads. While unit operating costs experience a slight increase, the strategy effectively reduces the need for costly external power purchases and decreases the incentives required during peak periods. At the same time, the moderate retention of reserve energy improves the flexibility of renewable energy absorption. The slight uptick in penalty costs is outweighed by the larger reductions in electricity purchases and incentive payments. Ultimately, the total daily cost decreased from 2.69 × 10^5 CNY (~38,935 USD) in the linear model to 2.18× 10^5 CNY (~31,553 USD, a reduction of 18.9%), and the number of cycles also decreased, indicating that the EES life and operational safety margin improved simultaneously. It is evident that simple linear energy pricing cannot reveal the hidden costs of charging and

discharging rates and SOC fluctuations, while the power-SOC coupling function achieves a higher-order dynamic balance between economy, reliability, and asset health by penalizing peaks and encouraging smoothness.

6.3 The coupling effect of demand-side response on energy storage unit output

In this study we found that coordinating B-IDR with energy storage systems can form a "economic signal-physical flexibility" closed loop (Fig. 9), which yields greater benefits than the isolated application of either approach. Fig. 14 presents the dynamic changes in B-IDR and EES. From 0:00 to 6:00, the EES is charged (SOC initial state 0.5). In the morning, as the load increases but electricity prices remain at peak levels, there is no B-IDR signal, and the EES discharges to alleviate load pressure. From 10:00-12:00, the B-IDR generates peak-shaving signals, and the EES pauses discharging to store energy. Later, from 13:00-18:00, the B-IDR and EES dynamically adjust so as to optimize economic efficiency. At 23:00, the EES has made up for the capacity gap.

We further employ a $\pm$ 6-hour window for correlation analysis: when the load reduction reached a peak of 16.9 MW (10:00), the B-IDR adjustment amount showed a positive correlation (0.44) with EES charging, indicating that the system tended to use the reduced load to charge the battery. At the same time, there was a negative correlation (-0.30) with EES discharge, indicating that the discharge was suppressed and the battery remained in a storage state for later use. When the load reaches a maximum replenishment of 24.35 MW (24:00), B-IDR shows a significant negative correlation with EES charging (-0.70), meaning that battery charging stops immediately during the increasing load phase. However, this effect is essentially unrelated to EES discharge (0.02), because the replenishment occurs at the end of the day, and battery discharge does not significantly intervene at this point. The results demonstrate that B-IDR has a coupling characteristic of "peak shaving and energy storage, replenishment, and charging suspension" on the EES unit. Under the given conditions, the introduction of B-IDR can effectively reduce the charging and discharging frequency and thus extend battery life.

6.4 Comparative experiments with representative SAC variants

To evaluate the economic differences between various policy learning paradigms under identical data and constraints within the same domain, we conduct comparative experiments with a feasible unified domain, unified training budget, and unified network scale. These experiments encompass four representative SAC variants (SAC, PER-SAC, CVaR-SAC, and DSAC) alongside the proposed approach. All methods share identical data and parameter configurations, with only the minimum necessary adjustments being made within their respective hyperparameter ranges. The results are presented in Table 3.

Compared to the baseline SAC (2.348 $\times 10^5$ CNY, ~33,984 USD), PER-SAC reduces the total cost to 2.310 $\times 10^5$ CNY (~33,435 USD) by prioritizing sampling of key experiences (a reduction of approximately 1.62%). CVaR-SAC utilizes risk-sensitive targets to suppress periods of adverse tail events, reducing the total cost to 2.269 $\times 10^5$ (a decrease of about 3.37%). On this dataset, DSAC's distributed estimation exhibits a conservative bias under the price tail distribution, resulting in a total cost of 2.480 $\times 10^5$ (representing a 5.62% increase over the baseline). The proposed method further reduces the total cost to 2.182 $\times 10^5$ (a 7.07% decrease from the baseline), delivering the most favorable performance. Thus, the proposed framework is readily extendable to market-oriented operation because its incentive-based DR decisions can be directly linked to settlement, and the electricity–hydrogen coupling provides additional value under time-varying prices. These market-facing benefits are consistent with recent studies on electricity–hydrogen market mechanisms and energy–carbon price coordination (Mochi et al., 2025; Zuo et al., 2025).

Table 3 Experimental results for different variants

Method	Ref.	Cost total ($\times 10^5$ CNY)
SAC	(Haarnoja et al., 2018)	2.348
PER-SAC	(Saglam et al., 2023)	2.310
CVaR-SAC	(Ning et al., 2024)	2.269
DSAC	(Duan et al., 2021)	2.480
LSTM-SAC	Ours	2.182

7 Conclusions

This paper proposes an integrated energy system (IES) framework combining B-IDR, multi-energy storage, and time-encoded LSTM-SAC optimization for high-renewable energy parks. A power-SOC soft penalty significantly reduces EES usage, extending service life and increasing safety margins. Compared to the baseline, the integrated scheme (scenario 5) reduces total system costs by 42.6%, carbon costs by 44.5%, and the renewable curtailment penalty by 47.0%. These improvements highlight the synergistic effect of B-IDR and energy storage. The temporal sequence encoding and adaptive exploration improve the RL scheduling efficacy in high-dimensional action spaces with coupled price signals and multi-energy constraints. Our analysis shows that the learned strategy suppresses EES discharging, reducing cyclic loading and thermal stress.

In future work, we will extend the framework to multi-site coordinated operations, and explore blockchain-enabled incentive/settlement mechanisms to enhance transparency and unlock the value of distributed renewables and extended storage. Long-term evaluations using multi-day/seasonal datasets and calibrated degradation models will also be conducted, in order to quantify storage aging impacts beyond a single typical day. In addition, by building on the findings regarding price-coupled B-IDR, we will investigate deployable real-time pricing schemes with multiple-time-scale consistency to support market-oriented operations and electricity spot markets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. T2422015, the Marie Skłodowska-Curie Actions (MSCA) under Project No. 101111188, the Key Technologies R&D Program of Tianjin under Project No. 24YFZCSN00030, and the Research and Reform Fund for Postgraduate Education and Teaching of Tianjin University of Technology with Grant No. ZDXM2502.

Author contributions

Zhongli Bai was responsible for the conceptualization, methodology, investigation, and writing of the original draft. Hongzhi Zhang contributed to data curation and validation. Qiang Gao and Junjie Liu provided supervision and reviewed the manuscript. Yuehui Ji contributed to writing, reviewing, and editing the manuscript. Yu Song played a key role in the conceptualization, methodology, and project administration. Xu Cheng provided resources, project administration, and funding acquisition. All authors contributed to the final manuscript.

Conflict of interest

All authors declare that they have no conflict of interest.

References

Abbasimehr H, Shabani M, Yousefi M, 2020. An optimized model using lstm network for demand forecasting. *Computers & industrial engineering*, 143:106435.

Abri S, Abri R, 2025. Deep learning methods for lstm-based personalized search: A comparative analysis. *International Journal of Machine Learning and Cybernetics*, 16(4):2747-2759.

Alexopoulos A, 2021. The fractional kullback–leibler divergence. *Journal of Physics A: Mathematical and Theoretical*, 54(7):075001.

Ampimah BC, Sun M, Han D, et al., 2018. Optimizing sheddable and shiftable residential electricity consumption by incentivized peak and off-peak credit function approach. *Applied Energy*, 210:1299-1309.

Chen H, Wu H, Li H, et al., 2025. Bi-level optimal scheduling of integrated energy systems considering incentive-based demand response and green certificate-carbon trading mechanisms. *Energy Reports*, 13:330-344.

Duan J, Guan Y, Li SE, et al., 2021. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. *IEEE transactions on neural networks and learning systems*, 33(11):6584-6598.

Gao MF, Han ZH, Zhao B, et al., 2023. Optimal planning method of multi-energy storage systems based on the power response analysis in the integrated energy system. *Journal of Energy Storage*, 73 https://doi.org/ARTN10901510.1016/j.est.2023.109015

Gea-Bermúdez J, Bramstoft R, Koivisto M, et al., 2023. Going offshore or not: Where to generate hydrogen in future integrated energy systems? *Energy Policy*, 174:113382.

Haarnoja T, Zhou A, Hartikainen K, et al., 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*,

Hamzeh Aghdam F, Mohammadi-Ivatloo B, Abapour M, et al., 2024. Enhancing the risk-oriented participation of wind power plants in day-ahead, balancing, and hydrogen markets with shared multi-energy storage systems.

Han S, Sung Y, 2021. A max-min entropy framework for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:25732-25745.

Hannan MA, Wali S, Ker PJ, et al., 2021. Battery energy-storage system: A review of technologies, optimization objectives, constraints, approaches, and outstanding issues. *Journal of Energy Storage*, 42:103023.

Hu H, Zhao X, Shang G, et al., 2026. Low-carbon optimization scheduling of hybrid energy storage in integrated energy system considering bidirectional interaction between green certificate and carbon trading. *Energies*, 19(1):70.

Jeon S, Bae S, 2025. Integrated optimization for sizing, placement, and energy management of hybrid energy storage systems in renewable power systems. *Journal of Energy Storage*, 106:114793.

Jiang M, Xu Z, Zhu H, et al., 2024. Integrated demand response modeling and optimization technologies supporting energy internet. *Renewable and Sustainable Energy Reviews*, 203:114757.

Kangfeng Z, Wang X, Wu B, et al., 2020. Feature subset selection combining maximal information entropy and maximal information coefficient. *Applied intelligence*, 50(2):487-501.

Koronen C, Åhman M, Nilsson LJ, 2020. Data centres in future european energy systems—energy efficiency, integration and policy. *Energy efficiency*, 13(1):129-144.

Kurucan M, Özbaltan M, Yetgin Z, et al., 2024. Applications of artificial neural network based battery management systems: A literature review. *Renewable and Sustainable Energy Reviews*, 192:114262.

Lee J-O, Kim Y-S, 2022. Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems. *International Journal of Electrical Power & Energy Systems*, 137:107795.

Li F, Shi Z, Zhu Z, et al., 2025. Energy management strategy for direct current microgrids with consideration of photovoltaic power tracking optimization. *Energies*, 18(2):252.

Li P, Wang Z, Wang N, et al., 2021a. Stochastic robust optimal operation of community integrated energy system based on integrated demand response. *International Journal of Electrical Power & Energy Systems*, 128:106735.

Li Q, Xiao X, Pu Y, et al., 2023. Hierarchical optimal scheduling method for regional integrated energy systems considering electricity-hydrogen shared energy. *Applied Energy*, 349:121670.

Li Y, Wang B, Yang Z, et al., 2021b. Optimal scheduling of integrated demand response-enabled community-integrated energy systems in uncertain environments. *IEEE Transactions on Industry Applications*, 58(2):2640-2651.

Liang H, Pirouzi S, 2024. Energy management system based on economic flexi-reliable operation for the smart distribution network including integrated energy system of hydrogen storage and renewable sources. *Energy*, 293:130745.

Liang T, Zhang X, Tan J, et al., 2024. Deep reinforcement learning-based optimal scheduling of integrated energy systems for electricity, heat, and hydrogen storage. *Electric Power Systems Research*, 233:110480.

Lin J, Gu Y, Wang Z, et al., 2024. Operational characteristics of an integrated island energy system based on multi-energy complementarity. *Renewable Energy*, 230:120890.

Liu J, Samson SY, Hu H, et al., 2022. Demand-side regulation provision of virtual power plants consisting of interconnected microgrids through double-stage double-layer optimization. *IEEE Transactions on Smart Grid*, 14(3):1946-1957.

Luo Y, Gao Y, Fan D, 2023. Real-time demand response strategy base on price and incentive considering multi-energy in smart grid: A bi-level optimization method. *International Journal of Electrical Power & Energy Systems*, 153:109354.

Mochi P, Espegren KA, Korpås M, 2025. Local electricity-hydrogen market. *International Journal of Hydrogen Energy*, 116:17-22.

Narayanan M, 2021. Annual evaluation of a model predictive controller in an integrated thermal-electrical renewable energy system using clustering technique. *Journal of Energy Resources Technology*, 143(5):051302.

Nikoobakht A, Mokarram MJ, Moghaddam EM, 2023. Enhancing integrated energy systems' resilience against windstorms through a decentralized cooperation model. *Electric Power Systems Research*, 225:109801.

Sadeghian O, Shotorbani AM, Ghassemzadeh S, et al., 2025. Energy management of hybrid fuel cell and renewable energy based systems-a review. *International Journal of Hydrogen Energy*, 107:135-163.

Saglam B, Mutlu FB, Cicek DC, et al., 2023. Actor prioritized experience replay. *Journal of Artificial Intelligence Research*, 78:639-672.

Shi M, Vasquez JC, Guerrero JM, et al., 2023. Smart communities-design of integrated energy packages considering incentive integrated demand response and optimization of coupled electricity-gas-cooling-heat and hydrogen systems. *International Journal of Hydrogen Energy*, 48(80):31063-31077.

Wang X, Yang C, Zhao J, et al., 2025a. Energy management of solar spectral beam-splitting integrated energy systems using soft actor-critic method. *Applied Thermal Engineering*, 269:125966.

Wang Y, Dong H, Ma K, et al., 2024. Multi frequency stability optimization of integrated energy systems considering virtual energy storage characteristics of heating networks. *Applied Thermal Engineering*, 257:124254.

Wang Y, Han L, Deng X, et al., 2025b. Optimization scheduling of household integrated energy systems for improving thermal comfort with low cost. *Energy and Buildings*, 329:115229.

Washizu A, Ju Y, Yoshida A, et al., 2024. Modeling the distributed energy resource aggregator services in a macroeconomic framework: The application to japan. *Energy*, 312:133561.

Wu L, Zhang W, Chen W, et al., 2025. A multi-time scale optimal scheduling strategy for integrated energy systems considering the power randomness of wind and photovoltaic. *Electrical Engineering*, 107(7):9109-9123.

Xu Y, Wei Y, Jiang K, et al., 2023. Action decoupled sac reinforcement learning with discrete-continuous hybrid

action spaces. *Neurocomputing*, 537:141-151.
Yan N, Zhao Z, Li X, et al., 2024a. Multi-time scales low-carbon economic dispatch of integrated energy system considering hydrogen and electricity complementary energy storage. *Journal of Energy Storage*, 104:114514.
Yan Z, Zhang Y, Yu J, 2024b. Allocative approach to multiple energy storage capacity for integrated energy systems based on security region in buildings. *Journal of Energy Storage*, 84:110951.
Zuo L, Xi Y, Zhang J, 2025. Leveraging electrochemical co2 reduction for optimizing comprehensive benefits of multi-energy systems: A collaborative optimization approach driven by energy-carbon integrated pricing. *Energy*, 322:135413.

Electronic supplementary materials

Section S1. Unit Mathematical Model.
Section S2. Proof of the mathematical properties of the B-IDR nonlinear mapping.
Section S3. The cost of the sub-items.
Section S4. The cost of the sub-items.
Section S5. Experimental Data and Parameters.
Section S6. Sensitivity to instantaneous load reduction cap.

中文概要

题　目：基于记忆增强深度强化学习的电-热-氢集成能源系统优化调度策略

作　者：白忠立[1,2]，高强[1,2,3]，张宏志[1,2]，刘俊杰[1,2]，吉月辉[1,2]，宋雨[1,2]，程徐[1,4]

机　构：[1]天津理工大学，天津市新能源电力变换传输与智能控制重点实验室，中国天津，300384；[2]天津理工大学，电气工程与自动化学院，中国天津，300384；[3]天津理工大学，海运学院，中国天津，300384；[4]天津理工大学，计算机学院，中国天津，300384；

目　的：本研究专注于高可再生能源渗透的能源园区，利用 B-IDR 机制实现需求侧响应，通过电-SOC 软惩罚降低储能使用频率，延长储能设备使用寿命。

创新点：1. 提出了一种电-SOC 软惩罚机制，以保持储能设备的灵活性和延长使用寿命。2. 提出了基于双向激励需求响应（B-IDR）的机制，通过反馈价格波动捕捉不对称需求响应。3. 结合 LSTM 的时序感知能力，提高了强化学习调度的稳定性和效率。

方　法：本研究提出了一种基于 LSTM-SAC 调度器的优化框架，用于解决高渗透可再生能源下的电-热-氢综合能源系统（IES）调度问题。该方法结合了多能储存与双向激励需求响应（B-IDR）机制，通过 LSTM 增强的 SAC 调度器捕捉风电、光伏输出及负荷变化等短期动态特征，优化能源管理策略。在低碳经济调度模型中，综合考虑了能源采购、碳排放、设备运行损耗等多个成本因素，并通过多场景模拟验证了该方法在降低系统总成本、碳排放和可再生能源削减成本方面的有效性。实验结果表明，结合 B-IDR 和多能储能系统的调度方法显著提高了系统的经济性与低碳性能。

结　论：通过结合 B-IDR 和储能系统的调度优化，本研究提出的方法显著降低了储能系统的循环负荷，提高了系统的经济性和碳减排效果。随着多能耦合技术的应用，尤其是在氢气储能路径的引入下，系统的低碳调度效果得到了显著改善。本研究的成果为高渗透可再生能源的低碳优化调度提供了新思路，并为未来能源市场的优化调度策略提供了实践依据。

关键词：综合能源系统，电-热-氢，LSTM-SAC，强化学习，调度策略