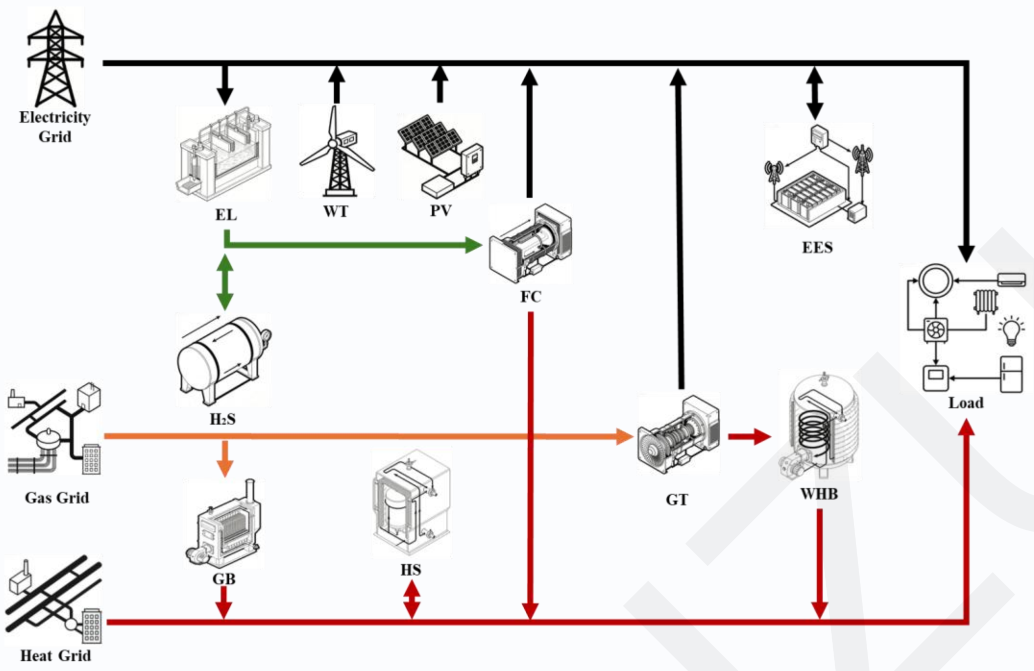
An optimization scheduling strategy for electric-heat-hydrogen integrated energy systems based on memory-enhanced deep reinforcement learning

- The paper integrates a power-SOC soft penalty, a bidirectional incentive-based demand response map, and an LSTM-SAC scheduler into one end-to-end framework.
- On a 24 h case study, the complete system reduces cost, carbon cost, curtailment penalty, and training instability at the same time.

Cite this as: Zhongli BAI, Qiang GAO, Hongzhi ZHANG, Junjie LIU, Yuehui JI, Yu SONG, Xu CHENG, 2026. An optimization scheduling strategy for electric-heat-hydrogen integrated energy systems based on memory-enhanced deep reinforcement learning. *Journal of Zhejiang University-SCIENCE A*, 27(8):809-824.
https://doi.org/10.1631/jzus.A2500598

Research Background and System Object



System object: an electricity-heat-hydrogen coupled IES

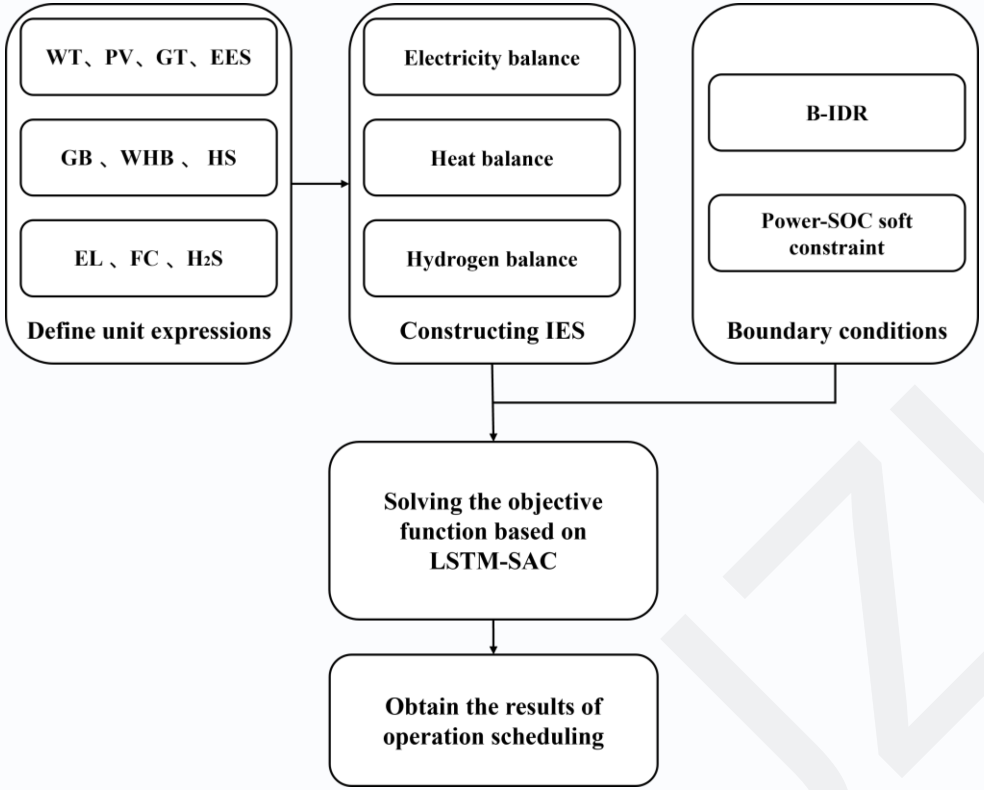
The system connects to the electricity, heat, and gas grids, and integrates EL, FC, GT, WHB, EES, HS, and H_2S.

- High renewable penetration makes the dispatch problem high-dimensional, strongly coupled, non-convex, and time-varying.
- Conventional MILP usually relies on heavy linearization, while static incentives struggle to capture asymmetric electricity/heat load responses.
- Vanilla DRL lacks temporal memory, so policy outputs can oscillate or lag when the system switches across time scales.

Target: optimize operating cost, carbon cost, and stability under multi-energy balance and device constraints.

Keywords: Integrated energy system, Electric-heat-hydrogen, LSTM-SAC, Reinforcement learning, Scheduling strategy

Overall Framework and Three Contributions



From unit modeling and multi-energy balance to boundary conditions and LSTM-SAC optimization, the paper forms one end-to-end dispatch chain.

01 SOC-consistent flexibility

A power-SOC soft penalty is introduced so storage protection is no longer handled only by rigid power limits. The penalty discourages both high charging/discharging intensity and frequent SOC swings.

02 Differentiable B-IDR mapping

A bidirectional incentive-based demand response map describes asymmetric load reduction and load rebound under price-rise and price-fall conditions.

03 Sequence-aware DRL

LSTM encoding is placed in front of SAC so the policy can exploit recent renewable and load sequences and make more stable cross-temporal decisions.

Modeling Details

Compact B-IDR formulation

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

$\boldsymbol{u}(t)$ is the agent incentive input, $\mathcal{R}$ is the load-reduction mapping, and $\mathcal{G}$ is the backlog replenishment mapping. This makes price-response behavior explicit in the dispatch environment.

Coupled EES operating cost

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

The cost penalizes high-power storage actions and adjacent-step SOC variation at the same time. The idea is not to use storage more, but to use storage more smoothly.

- B-IDR links price incentives to load-side behavior and allows reduction and rebound to be temporally separated.
- The power-SOC penalty couples dispatch economics with storage health.
- Both mechanisms provide a more realistic environment for the subsequent RL policy.

Takeaway

The paper first designs an optimization-friendly yet physically grounded environment, and then uses RL to learn within it.

LSTM-SAC Scheduler

MDP design

- State st: wind, PV, electric/heat loads, three storage SOCs, and time index.
- Action at: load adjustment, GT/GB output, EL/FC output, and EES/HS charge-discharge levels.
- Reward Rt = -g(st, at): energy purchase cost, O&M cost, carbon-related cost, curtailment penalty, and B-IDR cost.

Why LSTM?

It encodes recent renewable and load sequences into the hidden state, so the policy gains endogenous short-term foresight even without an external predictor.

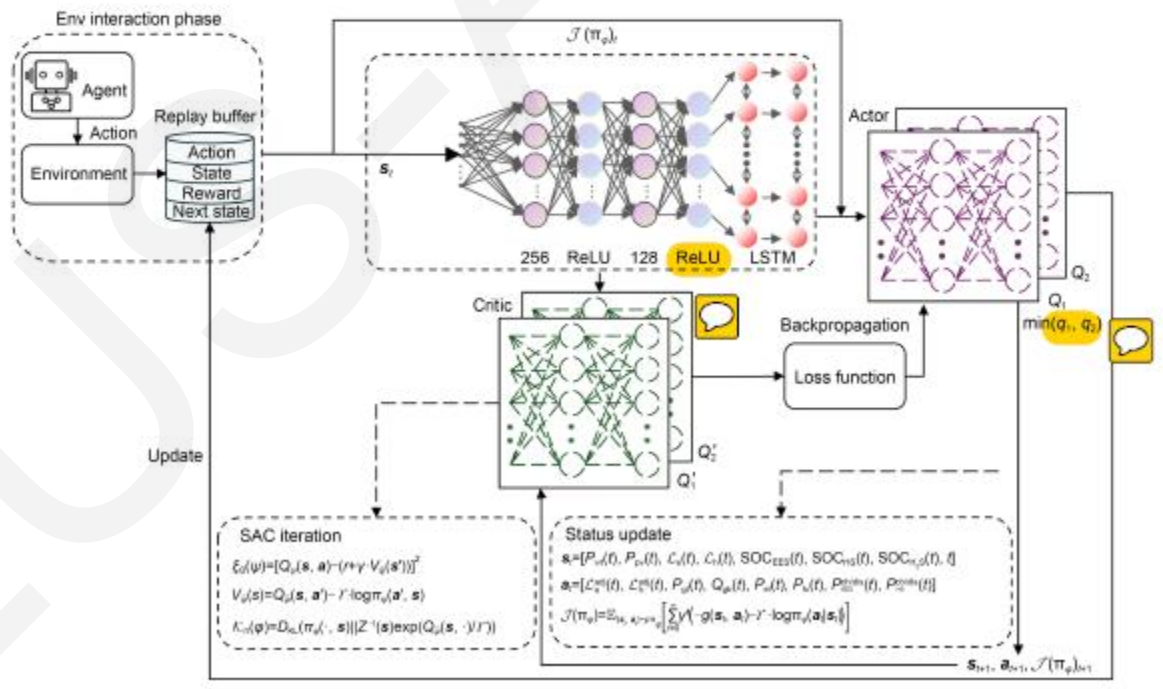


Fig. 4 Systematic process of the LSTM-SAC

Fig. 4 shows the full pipeline from environment interaction and replay buffer to actor, critic, and loss updates.

Case Study and Five Scenarios

System and training environment

- Case: a mixed industrial-commercial park in northern China, with 24 h dispatch and a 1 h time step.
- Environment: Python 3.11, PyTorch, and OpenAI Gym.
- Key hyperparameters: batch size 256, γ = 0.99, τ = 0.005.

Scenario design

S1	Baseline without flexibility resources
S2	B-IDR only
S3	EES + HS + H_2S, with hydrogen pathway enabled
S4	Hydrogen conversion and H_2S removed; EES + HS retained
S5	Complete system: storage + B-IDR + carbon trading

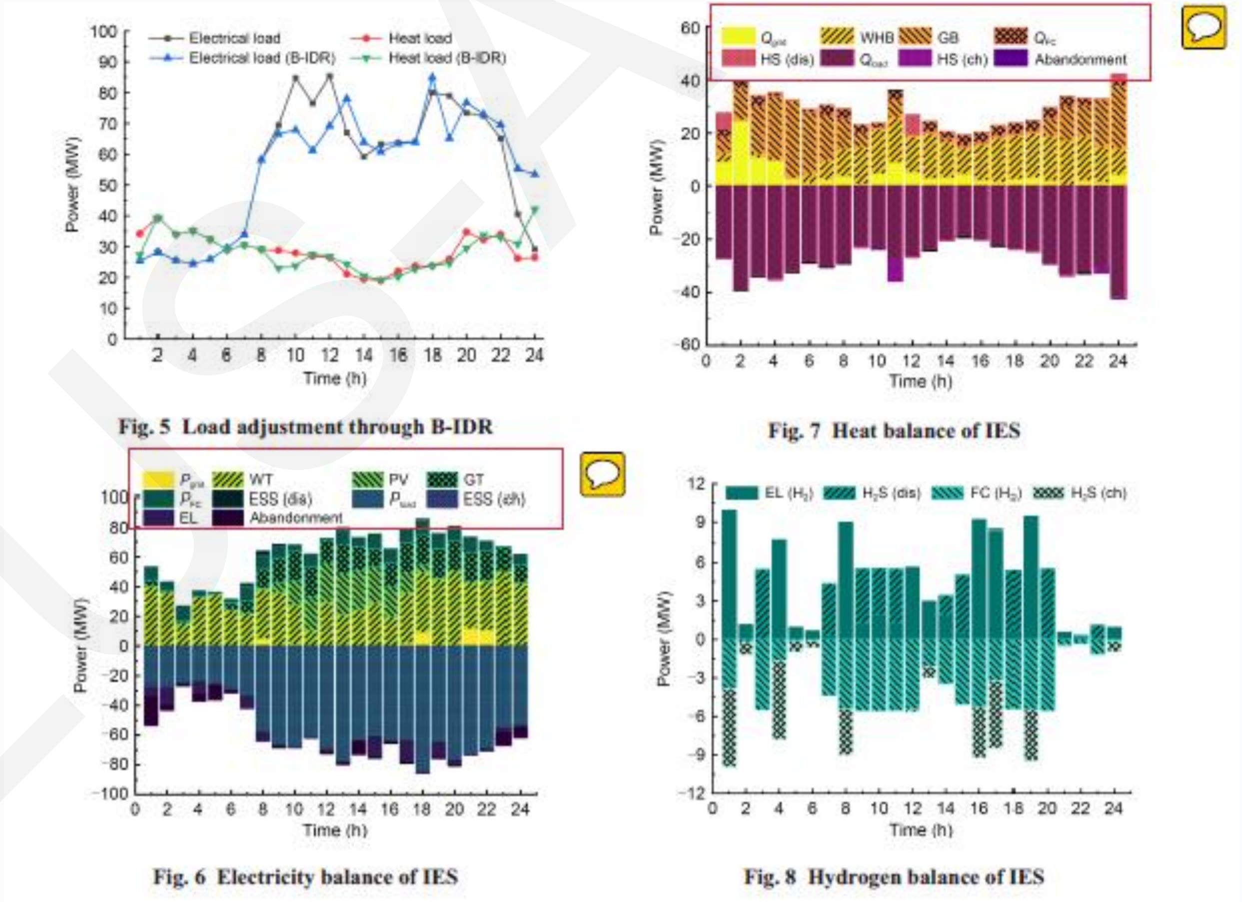


Fig. 5 Load adjustment through B-IDR

Fig. 7 Heat balance of IES

Fig. 6 Electricity balance of IES

Fig. 8 Hydrogen balance of IES

Scenario 5 jointly optimizes B-IDR and storage, while Figs. 5-8 visualize load adjustment and the electricity, heat, and hydrogen balances.

Key Results: System-Level Gains

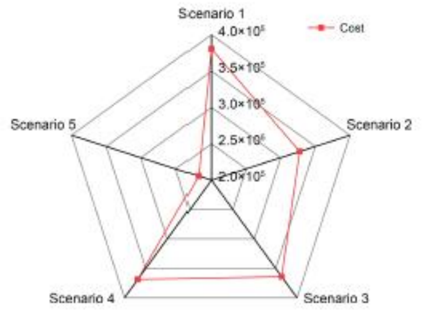


Fig. 9 Total cost comparison between different scenarios

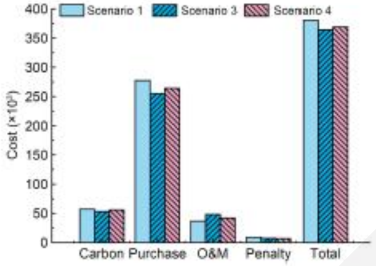


Fig. 10 IES cost comparison between different energy storage systems

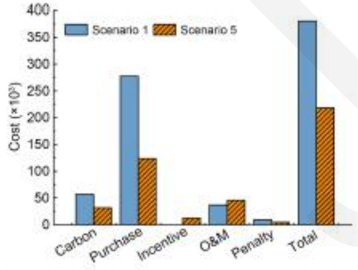


Fig. 11 Cost comparison of scenario 5 and scenario 1 at each stage

Total cost

-42.65%

Carbon cost

-44.49%

Curtailment penalty

$9.03\times10^3 \rightarrow 4.78\times10^3$ CNY

- B-IDR shifts part of the morning and evening demand away from high-price periods.
- EES and HS handle short-to-mid-term fluctuations, while H_2S provides an intraday long-duration shifting channel.
- Demand-side transfer and storage release work synergistically rather than additively.

Further Results: Algorithm Comparison and EES Cost Function

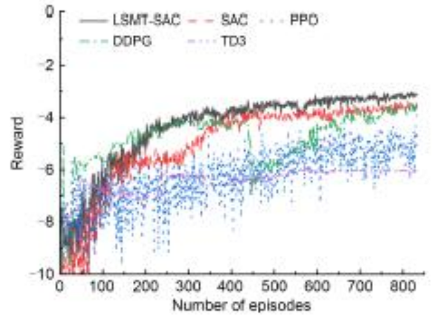


Fig. 12 Reward function curve

Reward curve (Fig. 12)

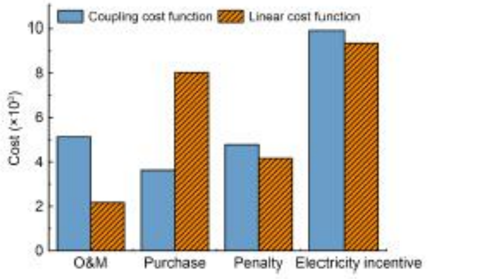


Fig. 13 Cost comparison between different EES cost functions

EES cost-function comparison (Fig. 13)

Table 1 Performance comparison between continuous control DRL algorithms

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

Table 1: continuous-control DRL comparison

- LSTM-SAC converges around -3.2, outperforming SAC (-3.9), DDPG (-4.9), PPO (-5.2), and TD3 (-6.1).
- Relative to SAC, daily operating cost is further reduced by about 7%, and the plateau-stage standard deviation drops to 0.30.
- The coupled EES cost cuts daily charge-discharge energy from 21.69 MWh to 12.68 MWh, while high-power charge/discharge events fall from 3/4 to 1/1.
- Compared with the linear cost, total daily cost decreases by about 19.0%, showing system-level benefits from smoother storage trajectories.

Conclusion and Suggested Talking Points

Main conclusions

- The paper unifies B-IDR, a storage power-SOC penalty, and LSTM-SAC into one low-carbon economic dispatch framework.
- The complete scenario achieves substantial improvements in total cost, carbon cost, and curtailment penalty.
- The real contribution is not just replacing the solver with RL, but jointly designing the environment and the learning policy.

Limitations / future work

Current validation is mainly based on a single 24 h simulation. Future work should add multi-day and seasonal data, calibrated degradation models, and stronger links to market settlement and real-time pricing.