# Representation Learning Enhanced Deep Reinforcement Learning for Optimal Operation of Hydrogen-based Multi-Energy Systems[*]

Zhenyu Pu[a], Yu Yang[a,*], Lun Yang[a,*], Qing-Shan Jia[b], Xiaohong Guan[a,b], Costas J. Spanos[c]

[a]*School of Automation Science and Engineering, Xi'an Jiaotong University, Shaanxi, Xi'an, 710049, China*
[b]*Center for Intelligent and Networked Systems, Department of Automation, BNRist, Tsinghua University, Beijing, 100084, China*
[c]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 94720, USA*

## Abstract

Hydrogen-based multi-energy systems (HMES) have emerged as a promising low-carbon and energy-efficient solution, as it can enable the coordinated operation of electricity, heating and cooling supply and demand to enhance operational flexibility, improve overall energy efficiency, and increase the share of renewable integration. However, the optimal operation of HMES remains challenging due to the nonlinear and multi-physics coupled dynamics of hydrogen energy storage systems (HESS) (consisting of electrolyters, fuel cells and hydrogen tanks) as well as the presence of multiple uncertainties from supply and demand. To address these challenges, this paper develops a comprehensive operational model for HMES that fully captures the nonlinear dynamics and multi-physics process of HESS. Moreover, we propose an enhanced deep reinforcement learning (DRL) framework by integrating the emerging representation learning techniques, enabling substantially accelerated and improved policy optimization for spatially and temporally coupled complex networked systems, which is not provided by conventional DRL. Experimental studies based on real-world datasets show that the comprehensive model is crucial to ensure the safe and reliable of HESS. In addition, the proposed SR-DRL approaches demonstrate superior convergence rate and performance over conventional DRL counterparts in terms of reducing the operation cost of HMES and handling the system operating constraints. Finally, we provide some insights into the role of representation learning in DRL, speculating that it can reorganize the original state space into a well-structured and cluster-aware geometric representation, thereby smoothing and facilitating the learning process of DRL.

*Keywords:* Distributed energy systems (DES), Representation learning, Deep reinforcement learning (DRL), Accelerated Learning, Hydrogen energy storage, decision-making for complex networked system

## 1. Introduction

Driven by the severe pressure of greenhouse gas effects and fossil fuel depletion, global energy sector is undergoing a profound transition toward low-carbon and sustainability. The transition had led to the proliferation of distributed energy systems that integrate on-site generation, energy conversion, storage, and management units to enable local energy supply with high share of renewables and enhanced energy efficiency. Among various emerging distributed energy systems, hydrogen-based multi-energy systems have attracted growing attention due to many unique advantages of hydrogen as an energy carrier [1]. *First*, hydrogen has a wide range of sources, including renewable-powered water electrolysis, industrial by-products and fossil-fuel based reforming processes, providing highly flexible and abundant sourcing. *Second*, hydrogen enables large-scale, long-duration storage and transport with relatively low self-discharge rate, making it particularly suitable for seasonal, cross-spatial and cross-temporal energy balancing, which is not provided by other types of energy carriers. *Third*, hydrogen exhibits prominent capability for multi-energy and cross-sector integration, as it enables flexible conversion among diverse energy carriers and coordinated operation across energy sectors. As an intermediate energy carrier, hydrogen can be converted into electricity, heat, and cooling. Moreover, it supports a wide range of applications, including transportation for fuel cell vehicles, buildings for heating and cooking, and power systems for generation and storage.

Hydrogen-based multi-energy systems typically incorporate hydrogen energy storage systems (HESS), which consist of electrolyters (ELs), fuel cells (FCs) and hydrogen tanks (HTs), enabling flexible energy conversion and storage. Specifically, ELs produce hydrogen through water electrolysis, which is subsequently stored in hydrogen tanks. When required, the stored hydrogen is converted by fuel cells into electricity and heat. HESS has experienced rapid technological progress in recent decade [1]. Several types of electrolyzers have been developed, including Alkaline Electrolyzers (AEL), Proton Exchange Membrane (PEM) Electrolyzers, Solid Oxide

Electrolyzers (SOEC) and Anion Exchange Membrane (AEM) Electrolyzer. Among them, the emerging PEM electrolyzers are particularly suitable for renewable-powered hydrogen production as they have quite fast dynamic response, relatively high energy efficiency and low operating temperatures. For hydrogen storage, high-pressure gas hydrogen storage (typically at 350–700 bar) and liquid hydrogen storage (requiring cryogenic temperatures below -253°C) are commonly adopted. For fuel cells, Proton Exchange Membrane Fuel Cells (PEMFCs) and Solid Oxide Fuel Cells (SOFCs) are two dominant types, in which the PEMFCs are well suited for distributed energy systems due to their relatively low operating temperatures (approximately 70–80 °C) and fast start-up characteristics.

Hydrogen-based multi-energy systems (HMES) have attracted extensive attention from the research community. Existing studies are spreading system architecture designing, capacity planing and optimal operation. In terms of system design, Wang et al. [2] proposed an HMES that integrates natural gas and solar energy supply with various generation, conversion, and storage technologies to achieve coordinated multi-energy provision for building communities, with its economic, environmental, and reliability advantages over conventional independent operation demonstrated by case studies. Zhang et al. [3] investigated a hydrogen-based multi-carrier energy system that incorporates renewable generation, electricity and hydrogen markets for a jointly provision of power, cooling, and heating. Liu et al. [4] studied comprehensive capacity planning for similar HMES and compared its economic benefits for different operating scenarios. The above studies generally focus on system design and planning considering the economic and environmental benefits of HMES.

The optimal operation of HMES requires to formulate and solve comprehensive optimization problems that involve multi-energy couplings, temporal dependencies, diverse energy device operating characteristics, and multiple uncertainties regarding supply and demands, which have attracted widespread attention. We have investigated and summarized existing studies in Table 1, highlighting their application scenarios, modeling approaches, uncertainty management, solution methods, as well as advantages and limitations. For application scenarios, prior works have investigated HMES for communities, microgrids, buildings and electric vehicle (EV) charging stations. Different types of models have been employed to characterize the operating behaviors of HESS. Linear models typically assume constant energy conversion efficiency across the operating ranges of ELs anf FCs. In contrast, nonlinear models explicitly account for the variation in energy conversion efficiency of ELs and FCs. Commonly-used solution methods include stochastic programming (SP), model predictive control (MPC) and deep reinforcement learning (DRL). They adopt different approaches to address uncertainties. SP methods commonly rely on scenario-based representaion to capture the daily uncertainties of renewable generation and energy demands, while MPC approaches generally depend on forecast or scenario-based chance constraints. In contrast, DRL methods learn optimal control policies under uncertainties from experience.

Overall, some existing studies (i.e., [5, 6, 7, 8]) have paid attention to the intrinsic nonlinear operating dynamics of HESS, which are essential for reliable and efficient system operation. However, these works primarily rely on SP and stochastic MPC and suffer from high computational complexity due to the need to solve constrained multi-stage nonlinear stochastic optimization problems. Typically, Dong et al. [5] studied the optimal operation of an HMES using scenario-based SP. The nonlinear behaviors of FC and absorption chiller (AC) were considered and handled by piecewise linearization. Nevertheless, this approach remains computational challenging due to the large number of mixed-integer constraints and representative scenarios for characterizing uncertainties.

Considering the system complexity and multiple uncertainties, deep reinforcement learning (DRL) has emerged as a promising solution for the optimal operation of HMES. Unlike optimization-based approaches, DRL learns optimal control strategies from the operating experience of systems, making it advantageous to accommodate the nonlinear and complex operating behaviors of HESS. In recent years, many DRL methods have been exploited for HMES. Typically, Yu et al. [9] studied the optimal operation of a building HMES based on MAADDPG. Zhu et al. [10] studied the joint operation of interconnected hydrogen-based multi-energy microgrids based on multi-agent DRL. Dolatabali et al. [11] has explored the combination of DDPG with CNN and LSTM to enable the optimal operation of an HMES. The CNN and LSTM were employed to process the high-dimensional input features associated with cloud-related weather conditions. These works have demonstrated the potential and advantage of DRL methods the optimal control of complex dynamics systems under uncertainties. However, many of the widely-used conventional DRL methods are not effective enough to accommodate the system complexity and strong spatial-temporal couplings caused by multi-energy flows and coordinated operation of diverse energy generation, conversion and storage devices. They often suffer from slow convergence and lead to deficient control policies. Specifically, conventional DRL-based control policies may exhibit a performance gap of 40%+ relative to the theoretical optimum [9]. Besides, most existing works relied on simplified linear models for HESS, without considering the intrinsic nonlinear operating characteristics and multi-physics coupled process of ELs and FCs due to computational limitations. Consequently, the resulting control policies deviate substantially from practical operating conditions and even violate device operating limits. As reported in a recent experimental study [3], both ELs and FCs exhibit pronounced nonlinear and multi-physics coupled behaviors. Specifically, the energy conversion efficiency and operational safety of HESS are jointly influenced by the coupled thermal, electrochemical, and pressure dynamics of ELs, FCs, and hydrogen tanks. Notably, increasing recent works have emphasized the importance to account for the nonlinear operating characteristics and multi-physics coupled process to ensure reliable and efficient operation of HESS. For example, [7] proposed an advanced operating model for an Electric-Hydrogen-Thermal-Gas Integrated Energy System and demonstrated a 3% annual cost reduction by accounting for the nonlinear operating characteristics of HESS. Marta et al. [6] enabled

the coordinated operation of a practical HESS and an electric vehicle charging station (EVCS) for peak demand shaving with the multi-physics coupled dynamics of HESS considered.

From the literature, we note that for the optimal operation of HMES, two critical problems remain to be addressed: (1) build a comprehensive model that considers the nonlinear and multiphysics coupled operating behaviors of HESS to ensure reliable and efficient operation of HMES. (2) handle the computational complexity caused by the system complexity and the multiple uncertainties.

### 1.1. Paper contribution

To address the existing gaps, this paper investigates the optimal operation of an HMES that integrates renewables, HESS, multiple types of energy storage and conversion devices with the goal of jointly satisfying multi-energy demands of a building community. **To enable cost-efficient, reliable and intelligent operation of the HMES, this paper makes the following main contributions:**

- We establish a comprehensive model for the optimal operation of HMES that incorporates the nonlinear operating characteristics as well as the multi-physics coupled process of HESS to ensure reliable and efficient operation.

- We propose a state-action representation learning enhanced deep reinforcement learning (SR-DRL) framework to account for the system complexity, which significantly improves learning convergence and performance over conventional DRL methods.

- We further provide some insights into the role of the state–action representation learning (SR) for DRL and conclude that it can reorganize the original state-space of HMES into a well-structured and cluster-aware presentation, thereby facilitating actor–critic learning in DRL methods.

We conduct a number of case studies in terms of the operation performance of an HMES with real-world datasets. The results show that the proposed model can enable the safe operation of the system while the existing simplified model may lead to safe constraints violation. Besides, the proposed learning framework can significantly enhance the learning rate and quality of policy, decreasing both operation cost and constraint violations under uncertainty.

The rest of the paper is as follows. Section 2 introduces a multi-physics coupled model of HMES. Section 3 introduces the SR-DRL method for HMES Section 4 presents the experimental results and analysis. Section 5 concludes this paper and discuss future works.

## 2. Problem formulation

### 2.1. Overview of HMES

We consider the optimal operation of a hydrogen-based multi-energy system (HMES) integrated into a building community for supplying electricity, cooling and heating energy as illustrated in Figure 1. The system integrates electrical, thermal and cooling energy storage (ESS, TES and CES) corresponding to batteries, hot and chilled water tanks, respectively. Particularly, the system involves an advanced hydrogen energy storage system (HESS) consisting of electrolyzers (EL), compressor unit, hydrogen tank, and fuel cells (FC), which enables the coordinated operation of multi-energy supply and demand. Hydrogen can be produced by renewable-powered water electrolysis and stored in the tank after cleaned and compressed by the compressor for later use. When needed, the stored hydrogen can be fueled by FC to generate electricity and heat. A heat recovery unit is coupled with the FC for capturing the residual heat, and then used to heat water in the tank or converted to cooling by the absorption chiller (AC). The hot and chilled water tanks serve as thermal energy buffers for storing heating and cooling energy to satisfy diverse forms of heating and cooling demands. A solar thermal collector is also planted for the hot water tank to gain heat from solar radiations.

This paper focuses on the optimal operation of the HMES to minimize the system's operating cost while satisfying the energy demands of the building community. We consider a discrete control framework with a decision interval of $\Delta t = 1h$ and each day equally divided into time slots $T := \{1, 2, \cdots, 24\}$.

**Notations**: This paper uses the alphabets $P, g, q, v$ to represent electricity, heating, cooling and hydrogen energy. The subscript `ces`, `tes`, `ces`, `hss` indicate the electrical, thermal, cooling and hydrogen storage, respectively.
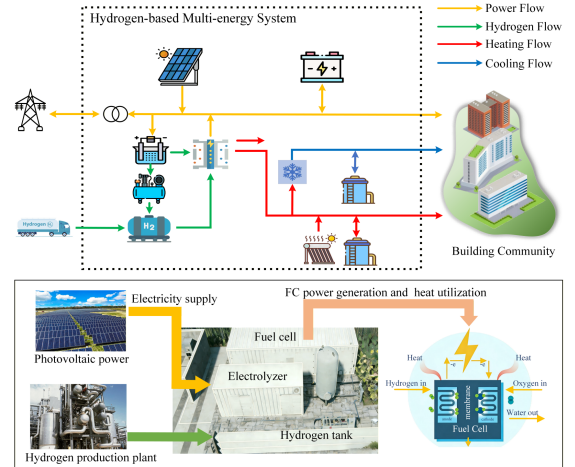


Figure 1: Hydrogen-based multi-energy system for a building community

### 2.2. Energy Storage Devices

In this section, we present mathematical models for different types of energy storage devices.

**Electrical storage system (ESS)** : The operating dynamics and physical limits of ESS can be modeled as

$$S_{t+1}^{\text{ess}} = S_t^{\text{ess}} + \Big( P_t^{\text{ess,ch}} \eta^{\text{ess,ch}} + \frac{P_t^{\text{ess,dis}}}{\eta^{\text{ess,dis}}} \Big) \Delta t, \tag{1a}$$

$$S_{\min}^{\text{ess}} \leq S_t^{\text{ess}} \leq S_{\max}^{\text{ess}}, \tag{1b}$$

$$- P_{\max}^{\text{ess,dis}} \leq P_t^{\text{ess,dis}} \leq 0, \tag{1c}$$

$$0 \leq P_t^{\text{ess,ch}} \leq P_{\max}^{\text{ess,ch}}, \tag{1d}$$

$$P_t^{\text{ess,ch}} \cdot P_t^{\text{ess,dis}} = 0, \quad t \in T. \tag{1e}$$

Table 1: Existing works of optimal operation of hydrogen-based multi-energy system (HMES)

| Scenarios | Models | Uncertainties Handing | Solution methods | Advantages | Limitations | Year | Paper |
|---|---|---|---|---|---|---|---|
| Community | Nonlinear model | Representative scenarios | SP | Capture nonlinear operating characteristics of HESS | High computational burden due to nonlinear models and scenario-based uncertainties | 2022 | [5] |
| EV charging station | Linear model | Scenario trees | SP | Mitigate forecast deviations by two-stage stochastic scheduling | High computational burden due to uncertainties; Strong dependence on forecast accuracy. | 2021 | [12] |
| Microgrid | Nonlinear model | Representative scenarios | SP | Accounting for nonlinear behaviors of HEMS | High computational complexity due to model complexity and uncertainties | 2025 | [8] |
| Microgrid | Nonlinear model | Deterministic | MILP | Consider the nonlinear operating characteristics of HESS. | Computational complexity due to model complexity; Do not account for multiple uncertainties | 2023 | [7] |
| Microgrid | Linear model | Multi-time-scale forecast of load and supply | MPC | Consider multi-time-scale characteristics of multiple energy demands | Ignores nonlinear characteristics and safety constraints of HESS | 2022 | [13] |
| EV Charging Station | Nonlinear model | Probabilistic uncertainties with chance constraints | Stochastic MPC | Account for multi-physics process of HESS and safe constraints | High computational complexity for large-scale applications | 2024 | [6] |
| Building | Linear model | Probabilistic uncertainties with chance constraints | Stochastic MPC | Jointly account for the planning and optimization of BHMES | High computational complexity; Ignores nonlinear characteristics and safety constraints of HESS | 2025 | [14] |
| Building | Linear model | Learning from experience | Multi-agent DRL | Considering building thermal dynamics for enhanced operational flexibility | Training instability and performance sensitivity to hyperparameters | 2022 | [9] |
| Microgrids | Linear model | Learning from experience | Multi-agent DRL | Considering energy sharing across microgrids for operating flexibility | Ignores nonlinear characteristics and safety constraints of HESS | 2025 | [10] |
| Microgrid | Nonlinear model | Learning from experience | DRL | Incorporate nonlinear characteristics of CHP and fuel cell | Depends on extensive data or an appropriate simulator. | 2023 | [11] |
| **Community** | **Nonlinear model** | **Learning from experience** | **SR + DRL** | **Account for multi-physics process and safe constraints of HESS; Fast convergence and high-quality policy.** | **Depends on extensive data or an appropriate simulator**. | **2026** | **This paper** |

where $S_t^{\text{ess}}$ [kWh] represents the stored electricity at time $t$, and $P_t^{\text{ess,ch}}$, $P_t^{\text{ess,dis}}$ [kW] are the charging and discharging power, with $\eta^{\text{ess,ch}}$ and $\eta^{\text{ces,dis}}$ denoting the corresponding efficiencies. Eq. (1a) models the evolution of stored electricity affected over the time. Cons. (1b) enforce the operating range of energy capacity of the storage device. Cons. (1c) and (1d) characterize the maximum charging and discharging power. Cons. (1e) are complementary constraints, enforcing non-simultaneous charging and discharging operation.

**Thermal storage system (TES)** : The thermal energy storage corresponds to hot water tank. Their operating dynamics and physical limits are modeled as

$$S_{t+1}^{\text{tes}} = S_t^{\text{tes}} + \left(g_t^{\text{tes,ch}}\eta^{\text{tes,ch}} + \frac{g_t^{\text{tes,dis}}}{\eta^{\text{tes,dis}}}\right)\Delta t, \tag{2a}$$

$$S_{\min}^{\text{tes}} \leq S_t^{\text{tes}} \leq S_{\max}^{\text{tes}}, \tag{2b}$$

$$-g_{\max}^{\text{tes,dis}} \leq g_t^{\text{tes,dis}} \leq 0, \tag{2c}$$

$$0 \leq g_t^{\text{tes,ch}} \leq g_{\max}^{\text{tes,ch}}, \tag{2d}$$

$$g_t^{\text{tes,ch}} \cdot g_t^{\text{tes,dis}} = 0, \quad t \in T. \tag{2e}$$

where $S_t^{\text{tes}}$ [kWh] represents the stored thermal energy in the hot tank at time $k$, and $g_t^{\text{tes,ch}}$, $g_t^{\text{tes,dis}}$ [kW] are the charging and discharging rate, with $\eta^{\text{tes,ch}}$ and $\eta^{\text{tes,dis}}$ denoting the corresponding efficiencies. Eq. (2a) models the dynamics of thermal energy stored in the hot water tank. Cons. (2b) enforces the permissible operating range of energy capacity of the storage device. Cons. (2c) and (2d) characterize the maximum charging and discharging rates. Cons. (2e) enforces non-simultaneous charging and discharging of thermal energy storage.

**Cooling storage system (CES)** : The cooling energy storage corresponds to the chilled water tank. Their operating dynamics and physical limits are modeled as

$$S_{t+1}^{\text{ces}} = S_t^{\text{ces}} + \left(q_t^{\text{ces,ch}}\eta^{\text{ces,ch}} + \frac{q_t^{\text{ces,dis}}}{\eta^{\text{tes,dis}}}\right)\Delta t, \tag{3a}$$

$$S_{\min}^{\text{ces}} \leq S_t^{\text{ces}} \leq S_{\max}^{\text{ces}}, \tag{3b}$$

$$-q_{\max}^{\text{ces,dis}} \leq q_t^{\text{ces,dis}} \leq 0, \tag{3c}$$

$$0 \leq q_t^{\text{ces,ch}} \leq q_{\max}^{\text{ces,ch}}, \tag{3d}$$

$$q_t^{\text{ces,ch}} \cdot q_t^{\text{ces,dis}} = 0, \quad t \in T. \tag{3e}$$

where $S_t^{\text{ces}}$ [kWh] represents the stored cooling in the chilled water tank at time $k$, and $q_t^{\text{ces,ch}}$, $P_t^{\text{ces,dis}}$ [kW] are the charging and discharging rate, with $\eta^{\text{ces,ch}}$ and $\eta^{\text{ces,dis}}$ denoting the corresponding efficiencies. Eq. (3a) models the dynamics of cooling energy stored in the tank. Cons. (3b) specifies the operating range of energy capacity of the storage. Cons. (3c) and (3d) characterize the maximum charging and discharging rates. Cons. (3e) ensures non-simultaneous charging and discharging operation of cooling energy storage.

### 2.3. Hydrogen energy storage system (HESS)

We assume the HESS consists of a Siemens Silyzer 100 electrolyzer, a Swiss Hydrogen fuel cell, a compressor, and a pressurized hydrogen storage tank, similar to [6]. The HESS suffers from pronounced nonlinear operating characteristics and admits a multi-physics coupled process. This paper adopts the formulations from [3, 6].

**Electrolyzers (EL)**: The Siemens SILYZER 100 electrolyzer is a self-contained system consisting of four stacks, each of which has a nominal rated power of 25 kW and a total power of 100 kW. The stacks can be run at an overloaded power of 50 kW for a limited period of time (about 15 mins). To ensure safe and efficient operation of EL, stack temperature must be maintained below a prescribed upper limit. A cooling device with a programmed control logic is often coupled with EL for thermal management. For example, the cooling device is set to on and off when the stack temperature exceeds 50°C or drops below 40°C for EL for maintaining the target stack temperature of 60°C in [3]. The control input of EL is the injected DC power $P_t^{\text{ely}}$ in the range of $[P_{\min}^{\text{ely}}, P_{\max}^{\text{ely}}] = [15, 200]$ kW with nominal power of $P_{\text{nom}}^{\text{ely}} = 100$ kW. The energy conversion efficiency of EL depends on the ranges of DC power injections and stack temperature, exhibiting a nonlinear operating characteristic, which can be captured by piece-wise affine function

$$v_t^{\text{ely}} = \begin{cases} z_1 T_t^{\text{ely}} + z_0 + z_{\text{low}}(P_t^{\text{ely}} - P_{\text{nom}}^{\text{ely}}) & \text{if } P_{\min}^{\text{ely}} < P_t^{\text{ely}} \leq P_{\text{nom}}^{\text{ely}}, \\ z_1 T_t^{\text{ely}} + z_0 + z_{\text{high}}(P_t^{\text{ely}} - P_{\text{nom}}^{\text{ely}}) & \text{if } P_{\text{nom}}^{\text{ely}} < P_t^{\text{ely}} \leq P_{\max}^{\text{ely}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $v_t^{\text{ely}}$ [m³/h] denotes the volume rate of produced hydrogen. $P_t^{\text{ely}}$ and $T_t^{\text{ely}}$ are the injected DC power and stack temperature of EL. The model captures the power-dependent energy efficiency of the EL using parameters $z_{\text{low}}$ and $z_{\text{high}}$. Specifically, the EL exhibits higher energy conversion efficiency at lower injected power levels, with $z_{\text{low}}$ typically slightly exceeding $z_{\text{high}}$.

The stack temperature of EL are governed by internal heat generation rate $\dot{Q}_{\text{gen}}$, heat dissipation rate $\dot{Q}_{\text{loss}}$ and the cooling power $\dot{Q}_{\text{cool}}$ supplied by cooling device. This leads to the following lumped-parameter thermal balance model

$$\frac{dT_t^{\text{ely}}}{dt} = K_1 \dot{Q}_{\text{gen}} - K_2 \dot{Q}_{\text{loss}} - K_3 \dot{Q}_{\text{cool}} \tag{5}$$

We assume a similar on-off control logic of cooling device in [3] and the stack temperature dynamics can be presented in the compact form

$$T_t^{\text{ely}} = j_1 T_{t-1}^{\text{ely}} + j_2 P_{t-1}^{\text{ely}} + j_0, \quad \forall t \in T. \tag{6}$$

where the effects of heat generation, loss and cooling devices are encapsulated in model parameters $j_0, j_1, j_2$. The DC power injections $P_t^{\text{ely}}$ is involved to capture the heat generation and loss effects under different operating conditions.

To ensure safe and reliable operation of EL, stack temperature must be maintained below a upper limit $T_{\max}^{\text{ely}}$ (often 70°C), which can be modeled as

$$0 \leq T_t^{\text{ely}} \leq T_{\max}^{\text{ely}}, \quad \forall t \in T. \tag{7}$$

As introduced, though the stacks of ES have a nominal power of 25kW, they can work at power up to 50 kW for certain short

time periods (i.e., 15 minutes). This behavior can be captured by an overload counter that records the number of time slots that the stack current overloads. Specially, for a EL with four stacks, the injected DC power will be first allocated to Stack 1 and then the other three. Therefore, the first stack tends to overload first and can be used as an indicator. For the first stack, the allocated DC power $P_{t,1}^{\text{ely}}$ from the total DC power injections can ba approximated by the following PWL function

$$
P_{t,1}^{\text{ely}} = \begin{cases} P_t^{\text{ely}} & \text{if } P_{\min}^{\text{ely}} < P_t^{\text{ely}} \leq P_{\text{nom}}^{\text{ely}}/4, \\ P_{\text{nom}}^{\text{ely}}/4 & \text{if } P_{\text{nom}}^{\text{ely}}/4 < P_t^{\text{ely}} \leq P_{\text{nom}}^{\text{ely}}, \\ P_t^{\text{ely}}/4 & \text{if } P_{\text{nom}}^{\text{ely}} < P_t^{\text{ely}} \leq P_{\max}^{\text{ely}}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}
$$

The stack current depends on the allocated power and stack temperature, which can be captured by

$$
i_t^{\text{ely}} = \begin{cases} h_1 T_t^{\text{ely}} + h_0 + h_{\text{low}}(P_{t,1}^{\text{ely}} - P_{\text{nom},1}^{\text{ely}}) & \text{if } P_{\min}^{\text{ely}} < P_{t,1}^{\text{ely}} \leq P_{\text{nom},1}^{\text{ely}}, \\ h_1 T_t^{\text{ely}} + h_0 + h_{\text{high}}(P_{t,1}^{\text{ely}} - P_{\text{nom},1}^{\text{ely}}) & \text{if } P_{\text{nom},1}^{\text{ely}} < P_{t,1}^{\text{ely}} \leq P_{\max}^{\text{ely}}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}
$$

where $P_{\text{nom},1}^{\text{ely}} = 25$ kW denotes the nominal power of stack 1. The nonlinear characteristics are captured by model parameters $h_0, h_1, h_{\text{low}}, h_{\text{high}}$.

A overload counter that tacks the current of stack 1 for ensuring the safe operation of EL can be modeled as

$$
C_t^{\text{ely}} = \max\left\{0, C_{t-1}^{\text{ely}} + (i_t^{\text{ely}} - i_{\text{nom}}^{\text{ely}}) \cdot \Delta t\right\}, \tag{10a}
$$

$$
C_t^{\text{ely}} \leq C_{\max}^{\text{ely}}, \quad \forall t \in T. \tag{10b}
$$

where $i_{\text{nom}}^{\text{ely}} = 300\ A$ denotes the nominal stack current. $C^{\max}$ denotes the overload counter upper limits determined by the permissible overload time periods of EL.

**Hydrogen tank**: Hydrogen produced by EL is cleaned, purified and compressed and then storage in the hydrogen tank. The stored hydrogen is consumed by FC to generate electricity and heat. The safe and efficient operation of hydrogen storage should account for the tank capacity, temperature and pressure limits, which can be modeled as

$$
S_t^{\text{hss}} = S_{t-1}^{\text{hss}} + \left(\alpha \cdot v_t^{\text{ely}} - v_t^{\text{fc}} + v_t^{\text{buy}}\right)\Delta t, \tag{11a}
$$

$$
S_{\min}^{\text{hss}} \leq S_t^{\text{hss}} \leq S_{\max}^{\text{hss}}, \tag{11b}
$$

$$
T_t^{\text{tank}} = g_0 T_{t-1}^{\text{tank}} + g_1 T_{t-1}^{\text{amb}}, \tag{11c}
$$

$$
p_t^{\text{tank}} = (b_0 + b_1 T_t^{\text{tank}}) \cdot S_t^{\text{hss}} \cdot \rho_0/V^{\text{tank}}, \tag{11d}
$$

$$
- v_{\max}^{\text{buy}} \leq v_t^{\text{buy}} \leq v_{\max}^{\text{buy}}, \tag{11e}
$$

$$
p_{\min}^{\text{tank}} \leq p_t^{\text{tank}} \leq p_{\max}^{\text{tank}}, \quad \forall t \in T. \tag{11f}
$$

where $S_t^{\text{hss}}$ [m³] represents the volume of stored hydrogen at time $t$ (under standard atmospheric pressure). $T_t^{\text{tank}}$ and $p_t^{\text{tank}}$ characterize the tank temperate and pressure. Eq. (11a) models the dynamics of hydrogen stored in the tank influenced by the volume rate of produced hydrogen $v_t^{\text{ely}}$ by EL, consumed hydrogen $v_t^{\text{fc}}$ by FC and net purchase of hydrogen $v_t^{\text{buy}}$ in market. The relatively small non-negative constant $\alpha \in [0, 1]$ is involved to

model the loss of hydrogen through compressor. The storage capacity of hydrogen tank is implicitly determined by its pressure, which is affected by the stored hydrogen and tank temperature. Cons. (11b) capture the safe and permissible operating volume range of the hydrogen tank. Eq. (11c) capture the temperature dynamics of the tank which is affected by tank thermal inertia and heat exchange with the ambient environment captured by model parameters $g_0$ and $g_1$ with $T_t^{\text{amb}}$ denoting the ambient temperature. Eq. (11d) evaluate the tank pressure which is affected by tank temperature and the volume of stored hydrogen with $V^{\text{tank}}$ denoting the volume of tank. Cons. (11e) enforces the limits of transaction in hydrogen market. To ensure safe operation, the tank pressure must be maintained in $[0, 40]$ bar as captured by Cons. (11f).

**Full cells (FC)**: We consider a Proton Exchange Membrane Fuel Cell (PEMFC) comprising two stacks each with a maximum power of 50 kW and 100 kW in total. FC converts hydrogen $v_t^{\text{fc}}$ into stack current $i_t^{\text{fc}}$. The resulting stack current produces DC power $P_t^{\text{fc}}$. This process exhibits nonlinear behavior and can be modeled as

$$
v_t^{\text{fc}} = c \cdot i_t^{\text{fc}}, \tag{12a}
$$

$$
i_t^{\text{fc}} = \begin{cases} s_1 P_t^{\text{fc}}, & \text{if } P_{\min}^{\text{fc}} < P_t^{\text{fc}} \leq P_{\text{bp}}^{\text{fc}}, \\ s_2(P_t^{\text{fc}} - P_{\text{bp}}^{\text{fc}}) + i_{\text{bp}}^{\text{fc}}, & \text{if } P_{\text{bp}}^{\text{fc}} < P_t^{\text{fc}} \leq P_{\max}^{\text{fc}}, \\ 0, & \text{otherwise,} \end{cases} \tag{12b}
$$

$$
P_{\min}^{\text{fc}} \leq P_t^{\text{fc}} \leq P_{\max}^{\text{fc}}, \quad \forall t \in T. \tag{12c}
$$

where $v_t^{\text{fc}}$ denotes the volume rate of hydrogen injected into the FC. $c$ is an experimental constant accounting for purging and Faradaic losses. The conversion of DC power to stack current shows a nonlinear characteristics as captured by the PWL function (12b) with $(P_{\text{bp}}^{\text{fc}}, i_{\text{bp}}^{\text{fc}})$ denoting the breakpoints and $s_1, s_2$ as energy conversion efficiencies. Cons. (12c) characterize the range of the DC output of FC.

While generating electricity, the FC produces heat that are captured by the heat recover unit [15]

$$
g_t^{\text{fc}} = \eta_{\text{rec}}^{\text{fc}} \cdot (1 - \eta^{\text{fc}})/\eta^{\text{fc}} \cdot P_t^{\text{fc}}, \forall t \in T. \tag{13}
$$

where $\eta_{\text{rec}}$ denotes the heat recovery efficiency and $\eta^{\text{fc}}$ represents the heat conversion efficiency of the FC.

### 2.4. Energy Conversion Devices

**Absorption Chiller (AC)**: The AC can convert heating energy to cooling energy. The produced cooling power of AC is determined by the injected heating power $g_t^{\text{ac}}$ and energy conversion efficiency $\eta^{\text{ac}}$, while respecting the operating limits. These can be modeled as

$$
\begin{aligned} q_t^{\text{ac}} &= g_t^{\text{ac}} \cdot \eta^{\text{ac}}, \\ g_t^{\text{ac}} &\leq g_{\max}^{\text{ac}}, \quad t \in T. \end{aligned} \tag{14}
$$

where $g_t^{\text{ac}}$ denotes the injected heating power to AC. $q_t^{\text{ac}}$ denotes the produced cooling energy by the AC. $g_{\max}^{\text{ac}}$ captures the working limits of the AC.

**Photovoltaic power (PV)**: Solar power generation is determined by the incident solar irradiance and the effective panel

area. The following model is commonly used to characterize PV output:

$$P_t^{\text{solar}} = \eta^{\text{pv}} \cdot A_{\text{pv}} \cdot Q_t^{\text{rad}}, \quad \forall t \in T. \qquad (15)$$

where $P_t^{\text{solar}}$ [kW] denotes the solar power output at time $t$, which is determined by the PV conversion efficiency $\eta^{\text{pv}}$, the effective solar collection area $A_{\text{pv}}$ [m²], and the incident solar radiation $Q_t^{\text{rad}}$ [kW/m²].

**Solar Thermal Connector**: The solar thermal collector absorbs solar radiation to generate thermal energy for heating water. The thermal energy injected into the hot water tank can be modeled as

$$g_t^{\text{solar}} = \eta^{\text{stc}} \cdot A_{\text{stc}} \cdot Q_t^{\text{rad}}, \quad \forall t \in T. \qquad (16)$$

where $g_t^{\text{solar}}$ [kW] denotes the absorbed heat by solar thermal collector which depends on the energy conversion efficiency $\eta^{\text{stc}}$, the collector area $A_{\text{stc}}$ [$m^2$] and incident solar radiation $Q_t^{\text{rad}}$ [kW/m²].

### 2.5. Energy Balances

The fundamental function of the HMES is to ensure the supply and demand balance at each time slot. This can be captured by the following multi-energy (i.e., electricity, heating and cooling) balance equations

$$P_t^{\text{g}} + P_t^{\text{solar}} + P_t^{\text{fc}} = P_t^{\text{ess,ch}} + P_t^{\text{ess,dis}} + P_t^{\text{ely}} + Q_t^{\text{ED}}, \qquad (17a)$$

$$g_t^{\text{solar}} + g_t^{\text{fc}} = g_t^{\text{tes,ch}} + g_t^{\text{tes,dis}} + g_t^{\text{ac}} + Q_t^{\text{HD}}, \qquad (17b)$$

$$q_t^{\text{ac}} = q_t^{\text{ces,ch}} + q_t^{\text{ces,dis}} + Q_t^{\text{CD}}, \quad \forall t \in T. \qquad (17c)$$

where $P_t^{\text{g}}$ denotes the purchase electricity from electricity market. $P_t^{\text{solar}}$ denotes the on-site PV generation. $Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}}$ denote the electricity, heating, and cooling demand of building community.

### 2.6. Objective function

Our objective is to minimize the operating cost of HMES for supplying multiple energy demands of the building community. The daily operating cost includes electricity and hydrogen trading costs

$$J_{\text{cost}} = \mathbb{E} \left\{ \sum_{t \in T} \left[ \frac{\lambda_t^{\text{b}} - \lambda_t^{\text{s}}}{2} |P_t^{\text{g}}| + \frac{\lambda_t^{\text{b}} + \lambda_t^{\text{s}}}{2} P_t^{\text{g}} + \lambda_t^{\text{h}} v_t^{\text{buy}} \right] \Delta t \right\} \qquad (18)$$

where $\lambda_t^{\text{b}}$ and $\lambda_t^{\text{s}}$ are the buying and selling prices in the electricity market at time $t$. $\lambda_t^{\text{h}}$ denotes the hydrogen trading price (i.e., both purchasing and selling). The first two terms are electricity cost, and the third is hydrogen cost.

Overall, the optimal operation of the HMES corresponds to solving the stochastic optimization problem: {min $J_{\text{cost}}$ s.t. (1) − (17)} which suffers high computational complexity from extensive decision variables, nonlinear constraints, and multiple sources of uncertainties.

## 3. Representation Learning enhanced DRL

To address the aforementioned challenges, this paper develops an advanced deep reinforcement learning (DRL) method to enable effecient and optimal operating of the HMES. Specifically, historical data of market prices, renewable output, and multiple energy demands are used to train a DRL agent, which is then deployed for real-time operation. During training, a multi-physics coupled model is employed as a simulator, enabling the DRL agent to acquire extensive operational experience. Whereas the method is model-free and does not rely on explicit mathematical formulations if a physical simulator or practical system is allowed to interact.

### 3.1. Markov Decision Process (MDP)

To develop a DRL-based approach, the problem is first reformulated as a Markov Decision Process (MDP), which is a general framework for sequential decision-making under uncertainty. An MDP is formally defined by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, including the state space $\mathcal{S}$, action space $\mathcal{A}$, state transition probability $\mathcal{P}$, reward function $r$ and discounting factor $\gamma$. In the following, we present the MDP reformulation for the optimal operation of HMES based on the mathematical model of previous section.

**State**: The state should involve all required information to make decision at each stage. For the optimal operation of the HMES, we define the system state at time $t$ as

$$s_t = \begin{pmatrix} day,\ hour,\ \lambda_t^{\text{b}},\ \lambda_t^{\text{s}},\ \lambda_t^{\text{h}},\ Q_t^{\text{rad}}, \\ S_t^{\text{ess}},\ S_t^{\text{tes}},\ S_t^{\text{ces}},\ S_t^{\text{hss}},\ T_t^{\text{ely}},\ C_t^{\text{ely}},\ T_t^{\text{tank}}, \\ Q_t^{\text{ED}},\ Q_t^{\text{HD}},\ Q_t^{\text{CD}} \end{pmatrix} \qquad (19)$$

where $day \in \{1, 2, \cdots, 7\}$ denotes the day of the weak and $hour \in \{1, 2, \cdots, 24\}$ denotes the hour. $\lambda_t^{\text{b}}, \lambda_t^{\text{s}}, \lambda_t^{\text{h}}$ are electricity and hydrogen market price. $Q_t^{\text{rad}}$ is incident solar radiation for PV cells and solar thermal collectors. $S_t^{\text{ess}}, S_t^{\text{tes}}, S_t^{\text{ces}},$ and $S_t^{\text{hss}}$ denote the energy levels of electrical, thermal, cooling, and hydrogen storage systems, respectively. $T_t^{\text{ely}}, C_t^{\text{ely}}, T_t^{\text{tank}}$ are stack temperature, overload counter of EL, and temperature of hydrogen tank of HESS. $Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}}$ are electricity, heating and cooling demands of the building community. Some studies incorporate historical information, e.g., $\lambda_{t-k:t}^{\text{b}}, \lambda_{t-k:t}^{\text{s}}, \lambda_{t-k:t}^{\text{h}}, Q_{t-k:t}^{\text{ED}}, Q_{t-k:t}^{\text{HD}}, Q_{t-k:t}^{\text{CD}}$, to enhance learning performance. This is avoided in this paper as they increase state dimensionality without performance gain.

**Action**: The action $a_t \in A$ specifies the scheduling decisions of HMES at time $t$ and is defined as

$$a_t = \left( a_t^{\text{ess}}, a_t^{\text{tes}}, a_t^{\text{css}}, a_t^{\text{hss}}, a_t^{\text{buy}} \right) \qquad (20)$$

where $a_t^{\text{ess}}, a_t^{\text{tes}}, a_t^{\text{css}}, a_t^{\text{hss}} \in [-1, 1]$ denote the normalized charging and discharging of electrical, thermal, cooling, and hydrogen storage. $a_t^{\text{buy}} \in [-1, 1]$ indicates normalized transaction in hydrogen market. A positive value indicates charging or buying, while a negative value denotes discharging or selling. Note that not all control variables of mathematical models are included as action, as the remaining variables can be derived

from system equations. While executing, the normalized action should be converted to practical device operations considering their physical limits as follows.

For electrical, thermal, cooling and hydrogen storage, the actual charging and discharging, considering physical limits, are given by

$$P_t^{\text{ess,ch}} = \text{clip}\left(a_t^{\text{ess}} P_{\max}^{\text{ess,ch}}, \quad 0, \quad \frac{S_{\max}^{\text{ess}} - S_t^{\text{ess}}}{\eta^{\text{ess,ch}}\Delta t}\right),$$

$$P_t^{\text{ess,dis}} = \text{clip}\left(a_t^{\text{ess}} P_{\max}^{\text{ess,dis}}, \frac{(S_{\min}^{\text{ess}} - S_t^{\text{ess}})\eta^{\text{ess,dis}}}{\Delta t}, 0\right), \forall t \in T. \quad (21)$$

$$g_t^{\text{tes,ch}} = \text{clip}\left(a_t^{\text{tes}} g_{\max}^{\text{tes,ch}}, \quad 0, \quad \frac{S_{\max}^{\text{tes}} - S_t^{\text{tes}}}{\eta^{\text{tes,ch}}\Delta t}\right),$$

$$g_t^{\text{tes,dis}} = \text{clip}\left(a_t^{\text{tes}} P_{\max}^{\text{tes,dis}}, \frac{(S_{\min}^{\text{tes}} - S_t^{\text{tes}})\eta^{\text{tes,dis}}}{\Delta t}, 0\right), \forall t \in T. \quad (22)$$

$$q_t^{\text{ces,ch}} = \text{clip}\left(a_t^{\text{ces}} q_{\max}^{\text{ces,ch}}, \quad 0, \quad \frac{S_{\max}^{\text{ces}} - S_t^{\text{ces}}}{\eta^{\text{ces,ch}}\Delta t}\right),$$

$$q_t^{\text{ces,dis}} = \text{clip}\left(a_t^{\text{ces}} q_{\max}^{\text{ces,dis}}, \frac{(S_{\min}^{\text{ces}} - S_t^{\text{ces}})\eta^{\text{ces,dis}}}{\Delta t}, 0\right), \forall t \in T. \quad (23)$$

$$v_t^{\text{hss,ch}} = \text{clip}\left(a_t^{\text{hss}} v_{\max}^{\text{hss,ch}}, \quad 0, \quad \frac{S_{\max}^{\text{hss}} - S_t^{\text{hss}} - v_t^{\text{buy}}}{\Delta t}\right),$$

$$v_t^{\text{hss,dis}} = \text{clip}\left(a_t^{\text{hss}} v_{\max}^{\text{hss,dis}}, \frac{S_{\min}^{\text{hss}} - S_t^{\text{hss}} - v_t^{\text{buy}}}{\Delta t}, 0\right), \forall t \in T. \quad (24)$$

where $\text{clip}(x, y, z) \triangleq \min(\max(x, y), z)$ bounds $x$ within $[y, z]$. $v_{\max}^{\text{hss,dis}}, v_{\max}^{\text{hss,ch}}$ denote the maximum charging and discharging rate of hydrogen tank, which can be determined based on experience. This ensures that electrical, thermal and cooling storage are not overcharged or over-discharged during operation.

The actual net purchase of hydrogen in market is

$$v_t^{\text{buy}} = a_t^{\text{buy}} v_{\max}^{\text{buy}}, \forall t \in T. \quad (25)$$

The other control variables of HMES can be derived from the system equations. Specifically, the control inputs of AC can be determined by the energy balance equations:

$$q_t^{\text{ac}} = q_t^{\text{ces,ch}} + q_t^{\text{ces,dis}} + Q_t^{\text{CD}},$$
$$g_t^{\text{ac}} = q_t^{\text{ac}}/\eta^{\text{ac}}, \forall t \in T. \quad (26)$$

Further, the working status of FC can be determined by

$$g_t^{\text{ac}} \xrightarrow{(17b)} g_t^{\text{fc}} \xrightarrow{(13)} P_t^{\text{fc}} \xrightarrow{(12b)} i_t^{\text{fc}} \xrightarrow{(12a)} v_t^{\text{fc}} \quad (27)$$

Given the purchased hydrogen $v_t^{\text{buy}}$, the hydrogen consumption of FCs $v_t^{\text{fc}}$, and the net charging and discharging volumes $v_t^{\text{hss,ch}}, v_t^{\text{hss,dis}}$, we can obtain $v_t^{\text{ely}}$ from the following hydrogen balance equation

$$v_t^{\text{hss,ch}} + v_t^{\text{hss,dis}} = \alpha \cdot v_t^{\text{ely}} - v_t^{\text{fc}} + v_t^{\text{buy}}, \forall t \in T. \quad (28)$$

We can further determine the operating status of EL as

$$v_t^{\text{ely}} \xrightarrow{(4)} P_t^{\text{ely}} \xrightarrow{(6)} T_t^{\text{ely}} \xrightarrow{(8)} P_{t,1}^{\text{ely}} \xrightarrow{(9)} i_t^{\text{ely}} \xrightarrow{(10a)} C_t^{\text{ely}} \quad (29)$$

Additionally, the amount of electricity $P_t^{\text{g}}$ purchased from the electricity market is determined by the electricity balance (17a).

**State transition**: The state transitions describe the evolution of states under given control inputs. The time components: *day* and *hour* are deterministic and follow the calendar. The external states: market prices $\lambda_t^{\text{b}}, \lambda_t^{\text{s}}, \lambda_t^{\text{h}}$, solar radiation $Q_t^{\text{rad}}$ and energy demands $Q_t^{\text{ED}}, Q_t^{\text{HD}}, Q_t^{\text{CD}}$ are not explicitly available and are captured by historical data. The storage levels $S_t^{\text{ess}}, S_t^{\text{tes}}, S_t^{\text{ces}}, S_t^{\text{hss}}$, stack temperature, overload counter and tank temperature $T_t^{\text{ely}}, C_t^{\text{ely}}, T_t^{\text{tank}}$ follow the system dynamics described in Section 2.2 and 2.3.

**Reward**: An appropriate reward should be designed to guide the learning process of DRL agent. Our objective is to minimize the operating cost for satisfying the multiple energy demands while considering physical operating limits. We therefore define the reward as the negative operating cost at each time slot. Considering the limitation of DRL in handling constraints, we also involve the hard constraints of HESS in the objective as penalty. The reward function is defined as

$$r_t = -\left[\frac{\lambda_t^{\text{b}} - \lambda_t^{\text{s}}}{2}|P_t^{\text{g}}| + \frac{\lambda_t^{\text{b}} + \lambda_t^{\text{s}}}{2}P_t^{\text{g}} + \lambda_t^{\text{h}}v_t^{\text{buy}}\right]\Delta t - \lambda \cdot penalty$$

$$\begin{aligned}penalty = \quad & \left[C_t^{\text{ely}} - C_{\max}^{\text{ely}}\right]_+ + \left[T_t^{\text{ely}} - T_{\max}^{\text{ely}}\right]_+ \\ & + \left[p_{\min}^{\text{tank}} - p_t^{\text{tank}}\right]_+ + \left[p_t^{\text{tank}} - p_{\max}^{\text{tank}}\right]_+ \\ & + \left[P_{\min}^{\text{fc}} - P_t^{\text{fc}}\right]_+ + \left[P_t^{\text{fc}} - P_{\max}^{\text{fc}}\right]_+ \\ & + \left[P_{\min}^{\text{ely}} - P_t^{\text{ely}}\right]_+ + \left[P_t^{\text{ely}} - P_{\max}^{\text{ely}}\right]_+ + \left[g_t^{\text{ac}} - g_{\max}^{\text{ac}}\right]_+\end{aligned}$$

where $\lambda \geq 0$ is a non-negative penalty factor to trading off operating cost and constraints penalties, which can be tuned by experience.
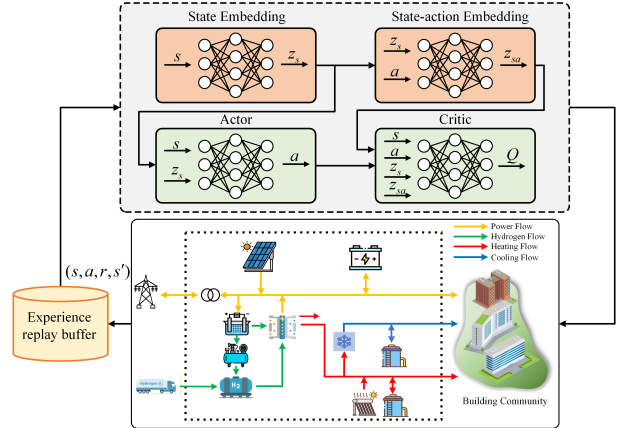


Figure 2: The architecture of representation learning enhanced DRL method

### 3.2. State-Action Reinforcement Learning for DRL

Motivated by the idea of [16], this section proposes a representation learning enhanced deep reinforcement learning (SR-DRL) framework for complex multi-energy systems involving strong temporal and spatial couplings. The key idea of the SR-DRL method is to involve the state and state-action representing learning into the DRL pipeline. Specifically, State Embedding and State-action Embedding modules are integrated to augment the inputs of the actor and critic

networks. These modules, referred to as `Representation Learning` in this paper. We emphasize that many existing works have explored representation learning for state dimension reduction or abstraction in vision-based learning tasks [17, 18, 19, 20], whereas the proposed method leverages representation learning to extract features that improve learning performance. The details of the state and state-action embedding are presented below.

**State and State-action Embedding**: We define the `state embedding` as a mapping $\Phi : \mathcal{S} \to \mathcal{Z}_s$, which transforms a raw state $s$ into a latent representation $z_s$. The `state-action embedding` is defined as $\Psi : (\mathcal{S}, \mathcal{Z}_s) \to \mathcal{Z}_{sa}$, which maps the latent state-action pair $(z_s, a)$ to an embedding $z_{sa}$. Accordingly, the embeddings are given by

$$z_s = \Phi(s)$$
$$z_{sa} = \Psi(z_s, a) \qquad (30)$$

Our objective is to learn an effective representation $(\Phi, \Psi)$ that can capture the underlying state transition dynamics in the latent space. Specifically, the latent state-action embedding $z_{sa}$ can be interpreted as the predicted next latent state when taking action $a$ at state $s$. Let $s'$ denote the corresponding next state in the original state space and $z_{s'}$ its latent representation. We therefore expect that

$$z_{sa} = z_{s'}, \forall s \in \mathcal{S}, a \in \mathcal{A}. \qquad (31)$$

Therefore, the training objective of state and state-action embedding is defined by minimizing the expected loss

$$L(\Phi, \Psi) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \|z_{sa} - z_{s'}\|$$
$$= \mathbb{E}_{s' \sim P(\cdot|s,a)} \|\Psi(\Phi(s), a) - z_{s'}\| \qquad (32)$$

**SR-enhanced DRL**: In the proposed SR-DRL, the inputs of actor and critic networks are augmented to involve the state and state-action embedding. Compared with conventional DRL, we have the following modifications

$$\begin{cases} DRL & \to & SR-DRL \\ \pi(s) & \to & \pi(s, z_s) \\ Q(s, a) & \to & Q(s, a, z_s, z_{sa}) \end{cases} \qquad (33)$$

The representation learning framework can be combined with any existing DRL methods that admit a actor-critic architecture, such as Deep Deterministic Policy Gradient (DDPG) [21] and Twin Delayed Deep Deterministic Policy Gradient (TD3) [22]. TD3 introduces three improvements: double critic, delayed actor update, and target policy smoothing. In SR-DRL, representation learning and DRL training are decoupled and alternated. The gradients for the state and state-action embeddings are

$$\nabla_\Phi L(\Phi, \Psi) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[ \nabla_\Phi \left\| \Psi(\Phi(s), a) - \left| \Phi(s') \right|_\times \right\| \right]$$
$$\nabla_\Psi L(\Phi, \Psi) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[ \nabla_\Psi \left\| \Psi(\Phi(s), a) - \left| \Phi(s') \right|_\times \right\| \right] \qquad (34)$$

where we use $|\cdot|_\times$ to denote stop-gradient operation, which can avoid the collapse of representation learning during training.

---

**Algorithm 1** Training Algorithm of the Proposed SR-DRL Framework

---

1: **Initialization:** (1) Initialize actor network $\pi^\theta$, $\pi^{\theta'}$, critic $Q^\omega$, $Q^{\omega'}$, state embedding $\Phi$, state-action embedding $\Psi$; (2) Initialize replay buffer $\mathcal{D}$.
2: **for** episode = 1 to $N_{\text{episode}}$ **do**
3:     Reset environment and obtain initial state $s_0$; initialize random noise $\mathcal{N}_t$
4:     **for** time step $t = 0$ to $T - 1$ **do**
5:         State embedding: $z_s = \Phi(s_t)$
6:         Select action: $a_t = \pi^\theta(s_t, z_s) + \mathcal{N}_t$
7:         Execute $a_t$ and get $r_t$, $s_{t+1}$
8:         Store transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}$
9:         **if** $|\mathcal{D}| > N_{\min}$ **then**
10:             Sample a mini-batch $(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}$
11:             **Update embedding** according to (34)
12:             **Update critic** according to (35) or (36)
13:             **Update actor** according to (37)
14:             **Soft update target networks** $\theta'$, $\omega'$
15:         **end if**
16:     **end for**
17: **end for**

---

**Actor and Critic Updates**: We denote the actor and target actor networks as $\pi^\theta, \pi^{\theta'}$. The critic and target critic networks are $Q^\omega, Q^{\omega'}$. For DRL, the critic is trained to minimize the Bellman error, while the actor is updated via the deterministic policy gradient. The critic updates of DDPG and TD3 are presented as follows.

*DDPG Critic Updates:*.

$$\begin{cases} L_{\text{critic}}(\omega) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1}) \sim D} \left[ (Q^\omega(s_t, a_t, z_s, z_{sa}) - y_t)^2 \right], \\ y_t = r_t + \gamma \, Q^{\omega'}(s_{t+1}, \pi^{\theta'}(s_{t+1}, z_{s'}), z_{s'}, z_{s'a'}). \\ \omega \leftarrow \omega - \eta_Q \nabla_\omega L_{\text{critic}}(\omega), \end{cases} \qquad (35)$$

*TD3 Critic Updates:*.

$$\begin{cases} L_{\text{critic}}(\omega_i) = \mathbb{E}\left[ (Q^{\omega_i}(s_t, a_t, z_s, z_{sa}) - y_t)^2 \right], i = 1, 2 \\ y_t = r_t + \gamma \, \min_{i=1,2} Q^{\omega'_i}\left( s_{t+1}, \pi^{\theta'}(s_{t+1}, z_{s'}) + \epsilon, z_{s'}, z_{s'a'} \right), \\ \omega_i \leftarrow \omega_i - \eta_Q \nabla_{\omega_i} L_{\text{critic}}(\omega_i), \quad i = 1, 2 \end{cases} \qquad (36)$$

where $\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$ is the clipped target policy noise, and $Q^{\omega'_1}, Q^{\omega'_2}$ are the two target critic networks.

*DDPG/TD3 Actor Updates:*.

$$\begin{cases} \nabla_\theta J \approx \mathbb{E}_{s \sim D}\left[ \nabla_a Q^\omega(s_t, a_t, z_s, z_{sa}) \big|_{a=\pi^\theta(s, z_s)} \nabla_\theta \pi^\theta(s, z_s) \right], \\ \theta \leftarrow \theta + \eta_\pi \nabla_\theta J. \end{cases}$$
$$(37)$$

For target actor and critic networks, updates are often performed at a slower pace for stable training. The widely used soft update scheme is adopted in our framework. The overall algorithm for the proposed SR-DRL is summarized in Algorithm 1. Note that the above framework applies to all DRL methods adopting the same actor-critic architecture.

9

## 4. Experiments

This section evaluates the proposed model and enhanced DRL-based method for the optimal operation of an HMES based on real-world datasets. The experimental configurations are summarized as follows. The model parameters of the involved HESS (including a Siemens Silyzer 100 electrolyzer, a Swiss Hydrogen fuel cell, a compressor, and a pressurized hydrogen storage tank) are adopted from the experimental studies [3, 23]. The configurations of other energy generation, conversion and storage devices follow the specifications of [5, 24]. The time-of-use (ToU) electricity price and multi-energy demand profiles (electricity, heating, and cooling) of a building community are taken from the CityLearn dataset [25]. We assume symmetric purchasing and selling prices of electricity market. The hydrogen market is not explicitly considered in the studies but can be readily incorporated into the framework. The penalty factor for constraints are determined by experience and set as 0.1. For ease of reference, all system parameters are summarized in Table A.3. The following methods for the optimal operation of HMES are compared in the experiments:

- `Perfect MPC`: A deterministic model predictive control (MPC) approach is implemented by solving the optimization problem {min $J_{cost}$ s.t. (1) − (17)} with deterministic renewable supply and multi-energy demand profiles, representing the theoretical optima.

- `DDPG` and `TD3`: Standard DDPG and TD3 agents are trained based on the formulated MDP to enable learning-based optimal operation of HMES.

- `SR-DDPG` and `SR-TD3`: The proposed state-action representation learning (SR) module is integrated with DDPG and TD3, and trained for the optimal operation of HMES.

For model and method evaluation, three representative cases of different time periods (i.e., `Case 1`, `Case 2`, and `Case 3`) are considered. For each case, a total of 30 days of data are used for training and evaluation. For the DRL (`DDPG` and `TD3`) and SR-DRL (`SR-DDPG` and `SR-TD3`), 20 days are used for training and 10 days for testing. The model architecture and training hyperparameters are summarized in Table A.4. For `Perfect MPC`, the optimal control policy for each day is obtained using the Gurobi solver.

### 4.1. Training performance

This section investigates the training performance of the DRL (`DDPG` and `TD3`) and SR-DRL (`SR-DDPG` and `SR-TD3`) methods. For fair comparisons, all methods share identical model and training parameters except for the state-action representation modules used by the SR-DRL methods. For all cases, the evolution of episode return during training are recorded and visualized in Figure 3. The theoretical optima for each case provided by `Perfect MPC` by averaging the daily cost are indicated in the figure titles. The results show that the SR-DRL methods consistently outperform conventional DRL in terms of both convergence rate and final episode return across

all cases. Moreover, SR-DRL approaches exhibit a markedly more stable learning process compared with their conventional counterparts. Specifically, the SR-DRL methods (`SR-TD3` and `SR-DDPG`) improve final episode returns by 16.7% and 13.4% (`Case 1`), 15.5% and 17.2% (`Case 2`), 13.5% and 11.5% (`Case 3`) compared with their conventional counterparts (`TD3` and `DDPG`). Notably, these results highlight the superior capability of SR-DRL approaches in handling the complexity of engineering systems. For conventional `TD3`, the performance gaps relative to the theoretical optima are 45.1% (`Case 1`), 43.3% (`Case 2`), and 38.7% (`Case 3`). In contrast, the proposed `SR-TD3` significantly reduces these gaps to 20.9%, 21.1%, and 20.1%, respectively. Similar performance improvements are observed with the DDPG methods. While conventional `DDPG` exhibits performance gaps of 37.1%, 42.3%, and 31.0% across the three cases, the proposed `SR-DDPG` reduces these gaps to 18.7%, 17.8%, and 15.9%, respectively. These results demonstrate the superior capability of SR-DRL methods in approaching high-quality control policies for complex networked energy systems.

### 4.2. Operating performance

We test the performance of trained DRL and SR-DRL agents for the optimal operation of HMES using the 10-day testing datasets. The average daily episode return, operation cost, and penalty incurred by constraint violations are used as performance metrics and indicated by `Cost + λ· Penalty`, `Cost`, and `Penalty`, respectively. The results for all three cases with the considered performance metrics under the different methods are reported in Table 2 with the results with our proposed methods highlighted in bold. The results show that the SR-DRL methods (`SR-TD3` and `SR-DDPG`) apparently outperform conventional DRL (`TD3` and `DDPG`) considering all performance metrics. Specifically, for `Case 1`, the overall performance gaps of DRL, measured by `Cost + λ· Penalty`, are 42.8% (`TD3`) and 36.5% (`DDPG`), which are reduced to 25.4% and 21.0% by the SR-DRL methods (`SR-TD3` and `SR-DDPG`). Similarly, the operation cost gaps are reduced from 39.1% (`TD3`) and 32.1% (`DDPG`) to 24.0% (`SR-TD3`) and 20.3% (`SR-DDPG`), respective. Similar improvements can be observed for `Case 2` and `Case 3`. Notably, the SR-DRL methods exhibit superior advantages in handling system constraints as substantial reduction of penalties incurred by constraint violations are observed with the SR-DRL methods compared with the DRL counterparts across all cases.

The superior performance of the SR-DRL methods over conventional DRL is further evidenced by the distributions of operational cost and constraint violation penalties shown in Figure 4. Since the `Perfect MPC` assumes complete system information and explicitly enforces all constraints, no constraint violation penalties are observed across all cases. In contrast, both DRL and SR-DRL methods incur certain penalties due to system uncertainties. Notably, the SR-DRL methods (`SR-TD3` and `SR-DDPG`) consistently achieve lower daily operating costs and reduced constraint violation penalties compared with their DRL counterparts (`TD3` and `DDPG`).
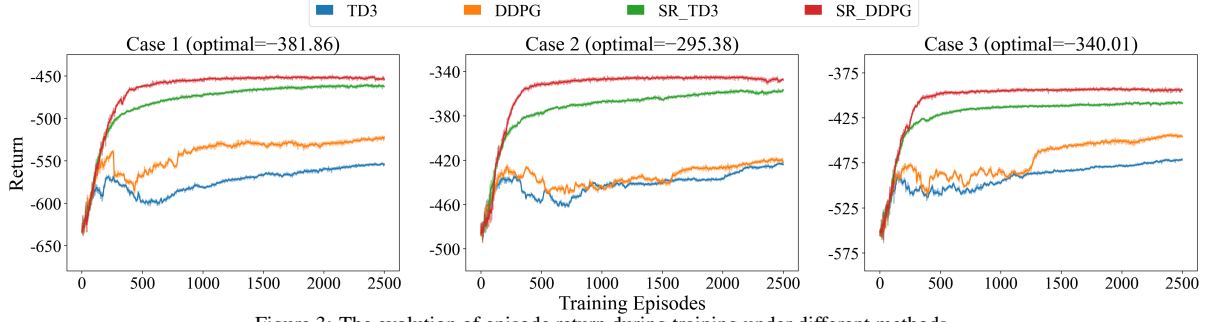
Figure 3: The evolution of episode return during training under different methods.

Table 2: Performance of DRL and SR-DRL methods compared with Perfect MPC for different cases

| Cases | Methods | Cost + λ· Penalty | | Cost | | Penalty |
|-------|---------|-------|--------|------|--------|---------|
| | | [$] | Gap[%] | [$] | Gap[%] | |
| Case 1 | Perfect MPC | 426.44 | – | 426.44 | – | 0 |
| | TD3 | 608.83 | 42.8 | 593.15 | 39.1 | 156.77 |
| | **SR-TD3** | **534.92** | **25.4** | **528.62** | **24.0** | **63.01** |
| | DDPG | 582.11 | 36.5 | 563.30 | 32.1 | 188.14 |
| | **SR-DDPG** | **516.17** | **21.0** | **512.97** | **20.3** | **31.96** |
| Case 2 | Perfect MPC | 373.34 | – | 373.34 | – | 0 |
| | TD3 | 572.04 | 53.2 | 552.35 | 47.9 | 196.88 |
| | **SR-TD3** | **505.48** | **35.4** | **493.59** | **32.2** | **118.90** |
| | DDPG | 541.31 | 45.0 | 525.48 | 40.8 | 158.26 |
| | **SR-DDPG** | **464.88** | **24.5** | **458.49** | **22.8** | **63.90** |
| Case 3 | Perfect MPC | 250.57 | – | 250.57 | – | 0 |
| | TD3 | 443.36 | 76.9 | 421.56 | 68.2 | 217.97 |
| | **SR-TD3** | **395.97** | **58.0** | **380.04** | **51.7** | **159.27** |
| | DDPG | 410.61 | 63.9 | 389.23 | 55.3 | 213.78 |
| | **SR-DDPG** | **353.75** | **41.2** | **346.68** | **38.4** | **70.70** |

To further investigate the operational performance of different methods, we examine the responses of all energy storage devices (i.e., ESS, TES, CES, and HSS) to the time-varying electricity prices. We use `Case 1` and the TD3 methods as an example and present the evolution of state-of-charge (SOC) of all energy storage devices in Figure 5. Overall, all storage devices exhibit clear price-sensitive behaviors: they generally charge during periods of low electricity prices and discharge when prices are high, which is consistent with our operating cost reduction objective. Whereas, for the DRL methods (`TD3`), the hydrogen storage does not respond well compared with the proposed SR-DRL (`SR-TD3`). Specifically, the hydrogen storage remains close to full over the testing days and is not effectively discharged during high-price periods, when storage backup is expected to function for operating cost reduction. This actually demonstrates the limitations of conventional DRL methods for handling complex engineering systems with strong spatial-temporal couplings.

### 4.3. Model evaluation

This section evaluates the importance of explicitly considering the multi-physics coupled processes of the HESS for ensuring the safe and reliable operation of the HMES. For the
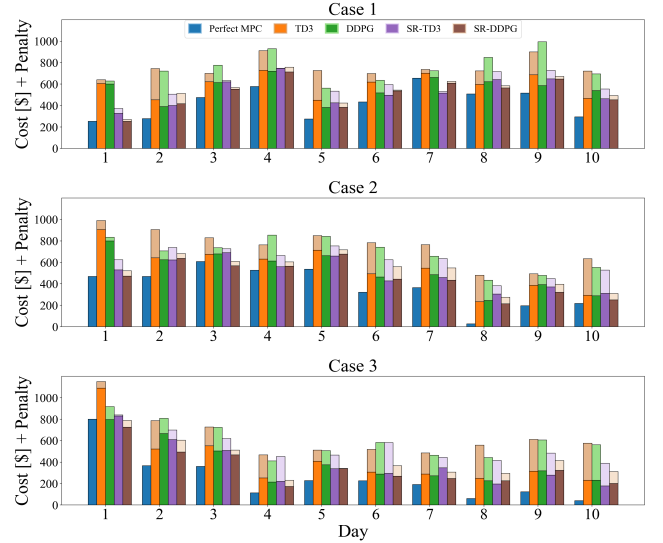


Figure 4: Distribution of stacked cost and penalty of constraint violations under different methods.

proposed model, the safety-critical variables of the HESS, including stack current, stack temperature, and hydrogen tank pressure are explicitly modeled. We first examine the proposed model with `Perfect MPC` by comparing two scenarios: with and without the safety constraints on stack temperature, stack current, and tank pressure, as defined in (6), (10b), and (11f) with the 10-day testing datasets of `Case 2`. Omitting these constraints is equivalent to neglecting the multi-physics interactions of current, temperature, and pressure with the HESS. For the considered case, we note that though the stack temperature and current mostly remain within their feasible ranges with both scenarios, the hydrogen tank pressure exhibits severe violations when the safety constraints are removed as shown in Figure 6(a). Whereas the tank pressure has been strictly maintained within the allowable limits when the safety constraints are explicitly enforced. This demonstrates the importance of explicitly considering the multi-physics process of HESS to ensure reliable operation of the HMES.

We further evaluate the effectiveness of the proposed SR-DRL methods in ensuring the safe operation of HESS under uncertainties. Specifically, we examine the evolution of hydrogen tank pressure over a 10-day testing period using `DDPG` and `SR-DDPG`. As shown in Figure 6(b), conventional `DDPG` frequently violates the pressure limits, whereas `SR-DDPG` consistently maintains safe operation throughout the period. These
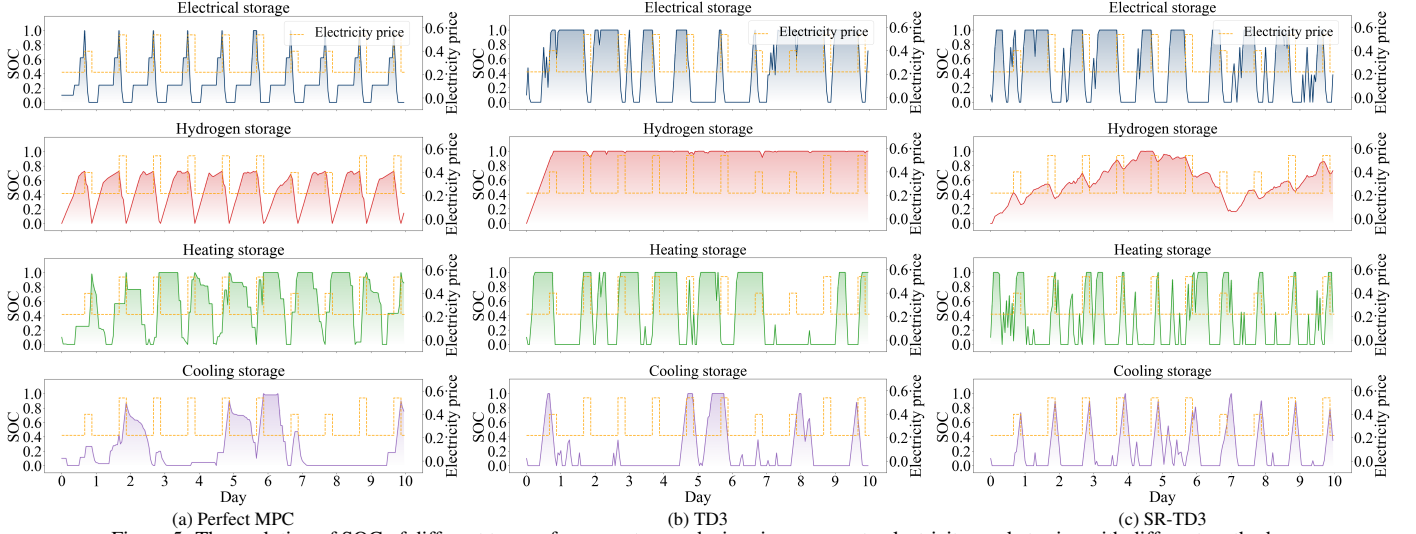
11

Figure 5: The evolution of SOC of different types of energy storage devices in response to electricity market price with different methods.

results highlight the advantage of SR-DRL in enforcing physical constraints, which is essential for the practical deployment of learning-based control strategies in HMES.
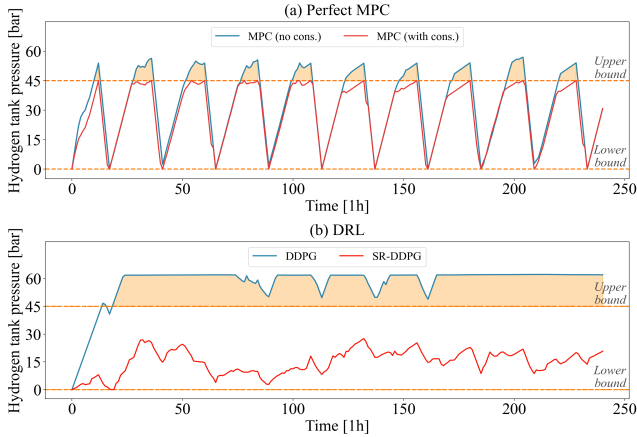


Figure 6: Evolution of hydrogen tank pressure over the 10-day testing periods.

### 4.4. Function of Representation Learning

In this section, we provide some insights on the role of state-action representation learning (SR) for DRL. we rely on principal component analysis (PCA) and clustering analysis for intuitive illustrations. The functions of SR can be interpreted from two perspectives.

*First, representation learning can reshape the geometric structure of state space, making it more amenable to policy and value function learning in DRL.* To illustrate this effect, we analyze the state-space distributions of DRL (`DDPG`) and SR-DRL (`SR-DDPG`) using PCA. We compare the outputs of the first layer of the actor networks of DRL and SR-DRL, which respectively convert the raw states $s$ and augmented state $[s, z_s]$ into the same dimensionality and enable a fair comparison. A total of 10,000 state samples are collected and projected onto the first two principal components, with the resulting 3D and 2D visualizations shown in Figure 7(a)-(d). For conventional DDPG, we observe that the state space exhibits a

multi-scattered and irregular geometric structure, characterized by multiple high-density regions. Such a poorly structured state space is expected to pose challenges for the actor-critic learning of DRL, as it hinders discriminative feature extraction and increases the difficulty of accurately approximating the policy and value functions, ultimately slowing convergence. In contrast, when the state-action representation learning modules are incorporated with SR-DDPG, the resulting state space exhibits a compact and well-structured geometry, which likely underlies the improved convergence and overall performance observed with SR-DRL approaches.
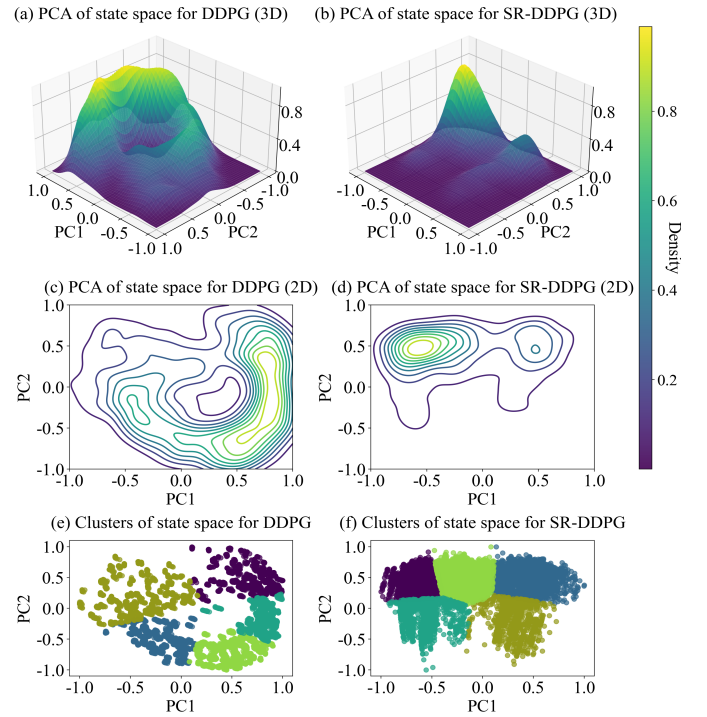


Figure 7: 3D, 2D, and clustering analysis of PCA-projected state space with DDPG and SR-DDPG.

*Second, representation learning seems capable of reorganizing the state space from a decision-oriented perspective.* To

12

examine this effect, we further conduct a clustering analysis on the PCA-projected state space with the DRL and SR-DRL methods. For a fair comparison, the state spaces are normalized and the resulting clusters are displayed in Figure 7(e) and (f). We note that, for conventional DRL, the clusters of state space are quite loose and diffuse, indicating the presence of substantial task-irrelevant noise. In contrast, the SR-DRL method reorganizes the states into compact and well-separated clusters with clear boundaries. Each cluster corresponds to a coherent decision-relevant mode, within which states share similar physical interpretations and control implications. From a learning perspective, such a cluster-aware state space is expected to significantly reduces the complexity of policy and value function learning, improves learning stability and convergence rate.

## 5. Conclusion and future work

This paper investigates the optimal operation of a hydrogen-based multi-energy system (HMES) that integrates diverse energy generation, conversion, and storage devices to provide coordinated electricity, heating, and cooling services for a building community. We propose a comprehensive operational model that explicitly captures the nonlinear operating characteristics and multi-physics processes of hydrogen energy storage system (HESS), which are crucial for safe and reliable HMES operation but have been mostly overlooked.

To address computational challenges posed by complex system dynamics, strong spatial-temporal couplings, and multiple uncertainties, a representation-learning enhanced deep reinforcement learning (SR-DRL) approach is proposed, significantly accelerating convergence rates and improving performance over conventional DRL. We further give some illustrative insights on the role of representation learning for DRL based on principal component analysis and clustering analysis of a state space, concluding that it can reshape the state space of HMES into a well-structured and clustered geometry, facilitating the actor-critic learning with DRL methods.

Future promising directions include the extension of the representation learning approach to enhance multi-agent deep reinforcement learning with its application to the coordinated operation of interconnected energy systems.

## References

[1] R. Last Name, S. Coauthor, I. Coauthor, S. Coauthor, A. Coauthor, A critical review of green hydrogen production by electrolysis: From technology and modeling to performance and cost, Energies 19 (1) (2026) 59. doi:10.3390/en19010059. URL https://www.mdpi.com/1996-1073/19/1/59

[2] J. Wang, K. Xue, Y. Guo, J. Ma, X. Zhou, M. Liu, J. Yan, Multi-objective capacity programming and operation optimization of an integrated energy system considering hydrogen energy storage for collective energy communities, Energy Conversion and Management 268 (2022) 116057.

[3] B. Flamm, C. Peter, F. N. Büchi, J. Lygeros, Electrolyzer modeling and real-time control for optimized production of hydrogen gas, Applied Energy 281 (2021) 116031.

[4] J. Liu, Z. Xu, J. Wu, K. Liu, X. Guan, Optimal planning of distributed hydrogen-based multi-energy systems, Applied Energy 281 (2021) 116107.

[5] X. Dong, J. Wu, Z. Xu, K. Liu, X. Guan, Optimal coordination of hydrogen-based integrated energy systems with combination of hydrogen and water storage, Applied energy 308 (2022) 118274.

[6] M. Fochesato, C. Peter, L. Morandi, J. Lygeros, Peak shaving with hydrogen energy storage: From stochastic control to experiments on a 4 MWh facility, Applied Energy 376 (2024) 123965.

[7] H. Dong, Z. Shan, J. Zhou, C. Xu, W. Chen, Refined modeling and co-optimization of electric-hydrogen-thermal-gas integrated energy system with hybrid energy storage, Applied Energy 351 (2023) 121834.

[8] N. Qi, K. Huang, Z. Fan, B. Xu, Long-term energy management for microgrid with hybrid hydrogen-battery energy storage: A prediction-free coordinated optimization framework, Applied Energy 377 (2025) 124485.

[9] L. Yu, Z. Xu, X. Guan, Q. Zhao, C. Dou, D. Yue, Joint optimization and learning approach for smart operation of hydrogen-based building energy systems, IEEE Transactions on Smart Grid 14 (1) (2022) 199–216.

[10] B. Zhang, W. Hu, A. M. Ghias, X. Xu, Z. Chen, Multi-agent deep reinforcement learning based distributed control architecture for interconnected multi-energy microgrid energy management and optimization, Energy Conversion and Management 277 (2023) 116647.

[11] A. Dolatabadi, H. Abdeltawab, Y. A.-R. I. Mohamed, A novel model-free deep reinforcement learning framework for energy management of a pv integrated energy hub, IEEE Transactions on Power Systems 38 (5) (2022) 4840–4852.

[12] J. Mei, Y. Zuo, C. H. Lee, X. Wang, J. L. Kirtley, Stochastic optimization of multi-energy system operation considering hydrogen-based vehicle applications, Advances in Applied Energy 2 (2021) 100031.

[13] X. Fang, W. Dong, Y. Wang, Q. Yang, Multiple time-scale energy management strategy for a hydrogen-based multi-energy microgrid, Applied Energy 328 (2022) 120195.

[14] J. Liu, Y. Zhao, Z. Xu, J. Wu, K. Liu, X. Guan, Coordinated integration of hydrogen-enabled multi-energy systems to achieve zero-carbon buildings, Building and Environment (2025) 113433.

[15] J. Assaf, B. Shabani, Transient simulation modelling and energy performance of a standalone solar-hydrogen combined heat and power system integrated with solar-thermal collectors, Applied Energy 178 (2016) 66–77.

[16] S. Fujimoto, W.-D. Chang, E. Smith, S. S. Gu, D. Precup, D. Meger, For sale: State-action representation learning for deep reinforcement learning, Advances in neural information processing systems 36 (2023) 61573–61624.

[17] A. Zhang, R. McAllister, R. Calandra, Y. Gal, S. Levine, Learning invariant representations for reinforcement learning without reconstruction, arXiv preprint arXiv:2006.10742 (2020).

[18] A. X. Lee, A. Nagabandi, P. Abbeel, S. Levine, Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model, Advances in Neural Information Processing Systems 33 (2020) 741–752.

[19] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, P. Abbeel, Deep spatial autoencoders for visuomotor learning, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 512–519.

[20] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, A. Farhadi, Target-driven visual navigation in indoor scenes using deep reinforcement learning, in: 2017 IEEE international conference on robotics and automation (ICRA), IEEE, 2017, pp. 3357–3364.

[21] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971 (2015).

[22] S. Fujimoto, H. van Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 1587–1596.

[23] M. Fochesato, C. Peter, L. Morandi, J. Lygeros, Peak shaving with hydrogen energy storage: From stochastic control to experiments on a 4 mwh facility, Applied Energy 376 (2024) 123965.

[24] J. Liu, Z. Xu, J. Wu, K. Liu, X. Guan, Optimal planning of distributed hydrogen-based multi-energy systems, Applied Energy 281 (2021) 116107.

[25] K. Nweye, K. Kaspar, G. Buscemi, T. Fonseca, G. Pinto, D. Ghose, S. Duddukuru, P. Pratapa, H. Li, J. Mohammadi, et al., Citylearn v2: energy-flexible, resilient, occupant-centric, and carbon-aware management of grid-interactive communities, Journal of Building Performance Simulation 18 (1) (2025) 17–38.

# Appendix A. Parameters

Table A.3: Energy system configuration and device parameters

| Device | Param. | Value | Units | Param. | Value | Units |
|---|---|---|---|---|---|---|
| ESS | $S^{\text{ess,min}}$ | 0 | kWh | $S^{\text{ess,max}}$ | 50 | kWh |
| | $P^{\text{ess,ch}}_{\max}$ | 20 | kW | $P^{\text{ess,dis}}_{\max}$ | 20 | kW |
| | $S^{\text{ess}}_{\text{init}}$ | 5 | kWh | $\eta^{\text{ess,ch/dis}}$ | 0.95 | – |
| TES | $S^{\text{tes,min}}$ | 0 | kWh | $S^{\text{tes,max}}$ | 100 | kWh |
| | $S^{\text{tes}}_{\text{init}}$ | 10 | kWh | $g^{\text{tes,ch}}_{\max}$ | 50 | kW |
| | $g^{\text{tes,dis}}_{\max}$ | 50 | kW | $\eta^{\text{tes,ch/dis}}$ | 0.9 | – |
| CES | $S^{\text{ces,min}}$ | 20 | kWh | $S^{\text{ces,max}}$ | 200 | kWh |
| | $S^{\text{ces}}_{\text{init}}$ | 20 | kWh | $q^{\text{ces,ch}}_{\max}$ | 40 | kW |
| | $q^{\text{ces,dis}}_{\max}$ | 40 | kW | $\eta^{\text{ces,ch/dis}}$ | 0.9 | – |
| EL | $P^{\text{ely}}_{\min}$ | 15 | kW | $P^{\text{ely}}_{\max}$ | 200 | kW |
| | $P^{\text{ely}}_{\text{nom}}$ | 100 | kW | $P^{\text{ely}}_{\text{nom},1}$ | 25 | kW |
| | $z_1$ | $1.618 \cdot 10^{-5}$ | m$^3$/[°C $\cdot$ s] | $z_0$ | $1.490 \cdot 10^{-2}$ | m$^3$/s |
| | $z_{\text{low}}$ | $1.530 \cdot 10^{-4}$ | m$^3$/[s $\cdot$ kW] | $z_{\text{high}}$ | $1.195 \cdot 10^{-4}$ | m$^3$/[s $\cdot$ kW] |
| | $T^{\text{ely}}_{\max}$ | 70 | °C | $T^{\text{ely}}_{\text{init}}$ | 20 | °C |
| | $j_0$ | 3.958 | °C | $j_1$ | 0.551 | – |
| | $j_2$ | 0.430 | °C/kWh | $h_0$ | 235.254 | °C |
| | $h_1$ | 0.673 | – | $h_{\text{low}}$ | 0.987 | A/kW |
| | $h_{\text{high}}$ | 9.0 | A/kW | $i^{\text{ely}}_{\text{nom}}$ | 300 | A |
| | $C^{\text{ely}}_{\max}$ | 75 | Ah | $\alpha$ | 0.697 | – |
| HT | $S^{\text{hss,min}}$ | 0 | kg | $S^{\text{hss,max}}$ | 50 | kg |
| | $S^{\text{hss}}_{\text{init}}$ | 0 | kg | $V^{\text{tank}}$ | 10 | m$^3$ |
| | $p^{\text{tank}}_{\min}$ | 0 | bar | $p^{\text{tank}}_{\max}$ | 45 | bar |
| | $m^{\text{hss,ch}}_{\max}$ | 30 | m$^3$/h | $m^{\text{hss,dis}}_{\max}$ | 30 | m$^3$/h |
| | $\eta^{\text{hss,ch/dis}}$ | 1.0 | – | $\rho_0$ | $8.99 \cdot 10^{-2}$ | kg/m$^3$ |
| | $b_0$ | $11.5 \cdot 10^5$ | m$^2$/s$^2$ | $b_1$ | $4.16 \cdot 10^3$ | m$^2$/[°C $\cdot$ s$^2$] |
| | $g_0$ | 0.94 | – | $g_1$ | $5.91 \cdot 10^{-2}$ | – |
| FC | $P^{\text{fc}}_{\max}$ | 160 | kW | $P^{\text{fc}}_{\min}$ | 0 | kW |
| | $P^{\text{fc}}_{\text{bp}}$ | 47.97 | kW | $\eta_{\text{fc}}$ | 0.3 | – |
| | $\eta^{\text{fc}}_{\text{rec}}$ | 0.8 | – | $i^{\text{fc}}_{\text{bp}}$ | 122.80 | A |
| | $s_1$ | 2.56 | 1/kV | $s_2$ | 3.31 | 1/kV |
| | $c$ | 0.21 | Nm$^3$/C | – | – | – |
| AC | $g^{\text{ac}}_{\max}$ | 200 | kW | $\eta_{\text{ac}}$ | 0.94 | – |
| PV | $A_{\text{PV}}$ | 1500 | m$^2$ | $A_{\text{stc}}$ | 400 | m$^2$ |
| | $\eta^{\text{stc}}$ | 0.762 | – | $\eta^{\text{PV}}$ | 0.2 | – |

Table A.4: DRL and SR-DRL model and training parameters

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Optimizer | Adam | Epochs | 2500 |
| Batch size | 1024 | Buffer size | $10^6$ |
| Discount rate | 1.0 | Learning rate | $3 \times 10^{-4}$ |
| Soft update rate | 0.01 | | |
| State embedding | [256, 256, 256, 512] | State-action embedding | [256, 256, 256, 512] |
| Actor | [512, action_dim] | Critic | [512, 512, 1] |