Building and Environment 242 (2023) 110546

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

# ED-DQN: An event-driven deep reinforcement learning control method for multi-zone residential buildings

Qiming Fu [a,c], Zhu Li [a,c], Zhengkai Ding [a,c], Jianping Chen [b,c,d,*], Jun Luo [b], Yunzhe Wang [a,c], You Lu [a,c]

[a] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China
[b] School of Architecture and Environment, Sichuan University, Chengdu, Sichuan, 610065, China
[c] Jiangsu Province Key Laboratory of Intelligent Energy Efficiency, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China
[d] School of Architecture and Urban Planning, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China

ARTICLE INFO

*Keywords:*
Thermal comfort control
Deep reinforcement learning
Event-driven
Multi-zone residential buildings
HVAC systems

ABSTRACT

Residential Heating, Ventilation, and Air conditioning (HVAC) systems are responsible for a significant amount of energy consumption, but their management is challenging due to the complexities of building thermodynamics and human activities. Reinforcement learning (RL) has been adopted to tackle this issue, but traditional RL methods require massive training data, long learning periods, and frequent equipment adjustments. To address these issues, we construct a new event-driven Markov decision process (ED-MDP) framework, which enables adjustments of control policies triggered by events, reducing unnecessary operations. Moreover, we propose an event-driven deep Q network (ED-DQN) method, which optimizes the action selection based on the triggered events. In the HVAC control problem, the proposed ED-DQN can effectively capture dynamic non-linear features of thermal comfort, and reduce the equipment damage caused by frequent adjustments. Our experimental results show that compared to three benchmark methods and three RL methods, our ED-DQN achieved state-of-the-art performance in both energy saving and thermal comfort violations. Moreover, our method demonstrates promising performance when applied to new test thermal environments, indicating its robustness and adaptability for optimizing residential HVAC controls.

## 1. Introduction

With the current global climate change, improving building energy saving and thermal comfort has become a pressing matter. The International Energy Agency (IEA) reports that residential buildings account for the largest share of energy consumption in the building sector, which consumed 35% of the world's energy in 2020 [1]. Among building systems, HVAC systems have the highest energy consumption, which accounts for over 50% [2]. Therefore, reducing the energy consumption of HVAC systems has become one of the important research directions for improving building energy saving. However, the pursuit of energy saving in buildings must not come at the expense of thermal comfort, as compromising on the latter can result in thermal discomfort and even health issues for occupants. Especially during the COVID-19 pandemic, people are spending more time indoors [3], making it crucial to optimize the thermal comfort control of HVAC systems. However, balancing energy saving and thermal comfort is a challenging problem. As a result, researchers and industry practitioners increasingly focus on minimizing energy consumption while keeping thermal comfort for residential buildings.

Currently, the majority of HVAC systems are controlled using methods such as rule-based control (RBC), proportional integral derivative (PID), model predictive control (MPC), and their variations [4]. However, RBC is static and relies on the expertise of engineers, making it unsuitable for flexible systems. PID controllers have fixed parameters, and their performance may deteriorate when the system operates under conditions different from the tuning conditions [5]. Despite MPC having better performance, creating an accurate thermodynamic model of some building for practical applications could be challenging and time-consuming. Furthermore, the models may become inaccurate over time due to refurbishment or wear and tear of the building [6].

RL is an effective data-driven control method for solving decision problems in HVAC systems. Compared to traditional control methods, RL methods don't need accurate thermodynamic models and simplify

* Corresponding author. School of Architecture and Urban Planning, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China.
*E-mail address:* alanjpchen@aliyun.com (J. Chen).

https://doi.org/10.1016/j.buildenv.2023.110546
Received 14 April 2023; Received in revised form 13 June 2023; Accepted 16 June 2023
Available online 20 June 2023

the control of HVAC systems. However, challenges still exist when applying RL to HVAC systems. Traditional RL methods operate on fixed time steps, which can lead to data redundancy and inefficient utilization due to the similarity between successive time steps in HVAC control. Additionally, the choice of the time interval can impact control performance, with longer intervals potentially reducing control accuracy and shorter intervals leading to excessive control actions. Furthermore, slow changes in building thermodynamic processes can also slow down the RL learning rate for HVAC problems. Control problems for HVAC systems typically involve high-dimensional state spaces, which further increase the complexity of RL methods, compounded by multi-zone residential buildings. Deep reinforcement learning (DRL) methods have the potential to tackle more complex problems by leveraging the benefits of both deep learning (DL) and RL. However, they still face the challenges mentioned above [7]. Therefore, it is essential to explore new methods to improve the efficiency and performance of HVAC control.

To address the shortcomings of traditional RL methods when controlling multi-zone residential HVAC systems, we designed an Event-driven Deep Reinforcement Learning (ED-DRL) method. Fig. 1(a) shows the temperature variation of a traditional RL method across a range of temperature thresholds within fixed time steps, whereas Fig. 1 (b) shows the adaptability of ED-DRL optimization to environmental changes through the concept of "intermittency". The ED-DRL method employs an event-driven method that enables control actions to be performed only when there is a significant change in the HVAC systems, leading to improved data utilization. By learning the dynamic non-linear features, such as the indoor temperature in different zones, and through predefined event definition, the method can capture and exploit some states that occur infrequently. Moreover, ED-DRL can incorporate prior knowledge to assign variable weights during the event definition, thereby enhancing the control efficiency and learning speed.

In this paper, we introduce a new ED-MDP framework and a novel method called ED-DQN to tackle the control problem of multi-zone residential HVAC systems. By leveraging the powerful capabilities of ED-DQN, the agent for HVAC systems can learn from various events and optimize related control accordingly, resulting in significant improvements in energy saving and thermal comfort in dynamic and unpredictable environments. The main contributions of our work are as follows:

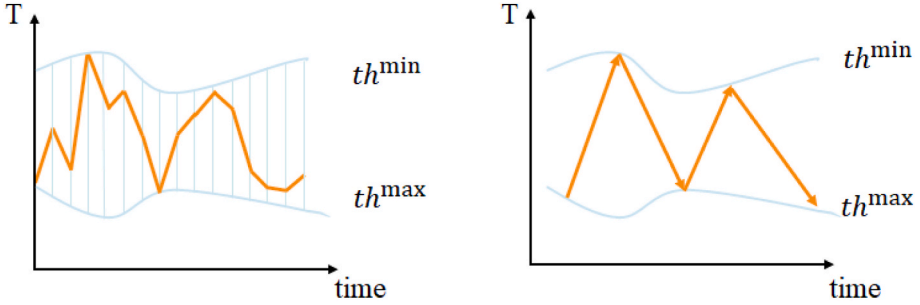
- We developed an ED-MDP framework for multi-zone residential HVAC systems that enables decision-making only when specific events occur. Compared to traditional MDP, this framework not only significantly reduces data redundancy and improves learning efficiency but also minimizes the number of unnecessary decisions.
- Our novel method, ED-DQN, optimizes control strategies for events in multi-zone residential HVAC systems. Compared to traditional RL, ED-DQN adapts control frequency to changing conditions while maintaining stability, captures dynamic non-linear features of thermal comfort, and reduces equipment damage risks from frequent switching.
- Our method is validated through simulations using real-world data, demonstrating faster learning and fewer action changes than traditional RL methods. Moreover, it achieves greater flexibility in the HVAC control, balancing the thermal comfort and the energy consumption compared to baseline methods. Additionally, its robustness is tested across multiple weather conditions. The code and data are available at https://github.com/LZLZLZlizhu/HVAC.

The remaining sections of the paper are organized as follows. Section 2 summarizes related work, section 3 presents preliminary information, section 4 introduces the multi-zone HVAC control, section 5 discusses the ED-DQN algorithm for HVAC control, section 6 presents and analyzes experimental results, and finally, section 7 concludes the paper.

## 2. Related work

There is a substantial and growing amount of literature on optimization methods for HVAC systems in buildings. Table 1 summarizes related studies for optimal control of building HVAC systems. In [8], Wemhoff presented a calibration method to optimize the set of PID coefficients for HVAC systems in order to reduce energy consumption. In [9], Baldi et al. presented an adaptive self-tuning method to dynamically regulate setpoints in large buildings. In [10], Wang et al. proposed a hierarchical control method to balance regulation capacity and thermal comfort while providing frequency regulation service in HVAC systems. In [11], Bird et al. proposed an MPC scheme to minimize HVAC overall cost and carbon usage, while ensuring the thermal comfort of occupants. In [12], Zeng et al. developed an autonomous adaptive MPC architecture to maintain the indoor temperature while reducing energy consumption. However, creating a simple but realistic building model remains challenging due to the complex and multifaceted nature of the indoor environment, which is influenced by factors such as building internal heat, layout, and exterior conditions. Despite significant progress, there is still room for improvement in developing effective optimization methods for HVAC systems in buildings.

In recent years, many studies have explored the potential of RL methods for HVAC control in buildings [22]. One advantage of RL is the ability to learn directly from data without requiring a model of the system. In [13], Mozer initially applied RL to learn and anticipate the lifestyle of the occupants for controlling basic residential comfort systems. In [14], Chen et al. provided a Q-learning method for HVAC and window systems to minimize both energy consumption and thermal discomfort. Despite the potential benefits of RL for HVAC control, plain RL methods such as tabular Q-learning may not be practical for problems with large state and action spaces. To address this challenge, in [15], Valladares et al. proposed a DRL-based method that leveraged the Double DQN method to maintain optimal levels of thermal comfort and air quality while minimizing energy consumption from HVAC systems.




(a) Traditional RL (b) ED-DRL

**Fig. 1.** Different methods for building under temperature threshold effect; (a) Traditional RL; (b) ED-DRL.

**Table 1**
Summary methods for optimal control of building HVAC systems.

| Classification | Reference | Optimization objective | Time step | Training time | Method |
|---|---|---|---|---|---|
| Traditional methods | AP Wemhoff (2012) [8] | Energy | 5 min | – | PID |
| | Baldi et al. (2018) [9] | Energy & Comfort | – | – | Switched self-tuning |
| | Wang et al. (2021) [10] | Energy & Comfort | 3 min | 5 days | Hierarchical optimal control |
| | Bird et al. (2022) [11] | Energy | 60 min | 3 months | MPC |
| | Zeng et al. (2021) [12] | Energy & Comfort | 5 min | 50 weeks | MPC |
| RL | Michael C.Mozer (1998) [13] | energy | – | – | Value-iteration |
| | Chen et al. (2018) [14] | Energy & Comfort | 20 min | 1 year | Q-learning |
| DRL | Valladares et al. (2019) [15] | Energy & Comfort & air quality | 20 min | 10 years | DDQN |
| | Gao et al. (2020) [16] | Energy & Comfort | 60 min | 10,000 h | DDPG |
| | Fang et al. (2022) [17] | Energy & Comfort | 10 min | 1 month | DQN |
| | Du et al. (2021) [18] | Energy & Comfort | 60 min | 1 month | DDPG |
| Event-driven | Wang et al. (2019) [19] | Energy | | | |
| | Wang et al. (2022) [20] | Energy | | | |
| | Jia et al. (2018) [21] | Energy & Comfort | | | |

In [16], Gao et al. suggested a deep deterministic policy gradient (DDPG) based HVAC control in labs by altering temperature and humidity for reducing energy consumption and maintaining the thermal comfort of occupants. Multi-zone HVAC systems have been increasingly adopted in buildings due to their numerous advantages. In [17], Fang et al. designed a DQN-based multi-objective optimal control method that effectively balances indoor air temperature with energy consumption. In [18], Du et al. proposed a DDPG method for multi-zone residential HVAC systems to minimize energy consumption costs while maintaining the thermal comfort of occupants. However, as mentioned earlier, the above-mentioned methods make decisions periodically. This can result in delayed or unnecessary actions for critical state changes, ultimately leading to a slower learning rate.

The development of event-driven methods has been prompted by the limitations of time-series data-driven methods. In [19], Wang et al. proposed an event-driven and machine-learning method for improving HVAC operation efficiency, which outperforms the traditional methods. In [20], Wang et al. proposed an event-driven method with adaptive intervals to improve optimization efficiency and overall performance. To further simplify the calculation process, in [21], Jia et al. established local and global events for HVAC control issues based on the complexity of the control strategy and demonstrated good performance experimentally. While the above-mentioned methods offer more flexible and efficient control strategies in certain scenarios, they may not be suitable for our systems since they rely heavily on events. This demands more parameter tuning and selection to ensure proper system functioning. Moreover, event-driven methods may only focus on short-term adjustments, without fully considering the long-term performance of HVAC systems.

In this paper, we propose the ED-DQN method for HVAC control in multi-zone residential buildings, which combines the advantages of event-driven methods and RL. Our goal is to minimize energy consumption while ensuring the thermal comfort of occupants. Compared to traditional RL methods, the proposed ED-DQN method can achieve better control effects. Furthermore, compared to traditional methods, ED-DQN can better balance energy consumption and thermal comfort, making it a promising method for efficient and effective HVAC control in buildings.

## 3. Preliminaries

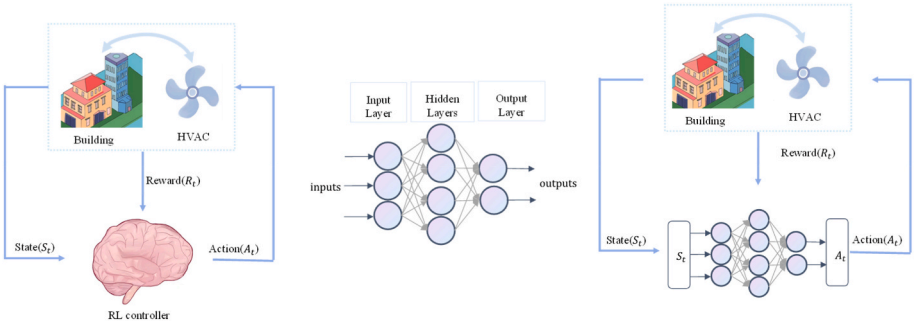
### 3.1. Deep reinforcement learning

#### 3.1.1. MDP

RL is a kind of machine-learning method that allows agents to learn how to make decisions through trial-and-error interactions with the environment [23]. Usually, the RL problem can be formulated as an MDP, which can be represented as a quintuple $(S,A,\mathrm{r},p,\gamma)$, as illustrated in Fig. 2(a).

(1) $S$ is the state space, $s(t) \in S$ indicates the state at time $t$.
(2) $A$ is the action space, $a(t) \in A$ represents the action taken by the agent at time $t$.
(3) $r : S \times A \rightarrow R$ is the reward function, $r(t) \sim r\{s(t), a(t)\}$ indicates the immediate reward obtained by the agent executing the action $a_t$ in the state $s_t$.
(4) $P{:}S\times S\times A\rightarrow[0,1]$ is state-transition probability, satisfying the Markov property $p\{s(t+1)|s(1),a(1),\ldots s(t),a(t)\}=p\{s(t+1)|s(t),a(t)\}$ for any trajectory $s(1),a(1),\ldots s(t),a(t)$ in state-action space.
(5) $\gamma \in [0,1]$, a discount factor, a smaller value indicates less concern for long-term payback.

During the learning process, at each time step $t$, the agent initially receives an state $s(t) \in S$ and selects an action $a(t) \in A$ according to




(a) RL (b) DNNs (c)DRL

**Fig. 2.** Illustration of RL and DRL [22].

policy $\pi(a|s)$, where denotes the probability distribution of the mapping from $s$ to $a$. Subsequently, the agent receives a new state $s(t+1)$ and an immediate reward $r(t+1)$. These interactions with the environment give rise to a trajectory $\partial$ that extends until the agent reaches a terminal state. The agent aims to maximize the expected return $E[G(t)]$ from each state, where the return is defined as a sum of discounted rewards, as follows:

$$R(\partial) = \sum_{t=\Delta t}^{T-1} \gamma^{t-1} r(t) \tag{1}$$

The value function is typically defined to measure the optimality of some policy $\pi$, including the state-value function (V function) and the action-value function (Q function). The former function represents the expected cumulative reward at a given state $s$, as defined in Equation (2):

$$V_\pi(s) = \mathbb{E}_\pi[G(t)|S(t) = s] \tag{2}$$

Meanwhile, the latter function represents the expected cumulative reward when taking action $a$ in state $s$, as defined in Equation (3):

$$Q_\pi(s,a) = \mathbb{E}_\pi[G(t)|S(t) = s, A(t) = a] \tag{3}$$

The optimal policy is determined by maximizing either the state-value function $V_*$ or the action-value function $Q_*$, respectively. The following definitions apply to these two functions:

$$V_*(\mathrm{s}) = \max_\pi V_\pi(s) \tag{4}$$

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = \mathbb{E}\left[R\left(t+1\right) + \gamma \max_{a'} Q_*\left(s\left(t+1\right), a'\right) \middle| S\left(t\right) = s, A\left(t\right) = a\right] \tag{5}$$

#### 3.1.2. Deep Q network algorithm (DQN)

DeepMind's [24] pioneering work in DRL has greatly impacted the field of building energy saving, spurring its rapid development. Advancements in deep neural network (DNN) methods and computing power have enabled the emergence of DRL as a powerful method, as shown in Fig. 2(c). By combining the strengths of DL and RL, DRL has become an effective tool for tackling complex tasks and high-dimensional problems [25].

Notably, the DQN method is a representative DRL method that boasts three significant improvements over Q-learning. Firstly, DQN utilizes DNN to approximate Q values because of its ability to extract complex features and superior generalization capabilities. Secondly, during training, DQN utilizes a replay buffer to break correlations between data, avoiding convergence problems caused by high correlations between training samples. Finally, DQN utilizes a target network to calculate the target Q value, which avoids updates to the current state Q value and improves the stability of the training process. These characteristics make DQN widely applicable in various fields, such as games [26], robot control [27], and cloud broadcasting [28]. DQN is known to outperform traditional control methods in complex control problems, making it an attractive option for HVAC control in multi-zone residential buildings. Moreover, DQN is particularly effective in handling large numbers of discrete actions and is easier to implement and tune than other DRL methods, especially for problems with high-dimensional state spaces.

## 4. MULTI-ZONE HVAC control

In this section, we provide an overview of the optimal control framework for HVAC systems and then present a case study of HVAC control.

### 4.1. Overall framework

In this section, we outline the three main steps of our paper as shown in Fig. 3. Firstly, we develop a simulation testbed using Python to assess the performance of the ED-DQN method. Secondly, we model the HVAC control problem as an ED-MDP and design a set of event trigger rules to determine when to update the value function. Lastly, we evaluate the performance of the ED-DQN method and present the experimental results.

### 4.2. Case study

In this section, we will develop a case study of a residential building and model the HVAC control problem as an ED-MDP for evaluating the ED-DQN method.

#### 4.2.1. Simulation environment

For the case study, we select a conventional three-person occupancy apartment model with five rooms, as previously used by Deng et al. [29]. The apartment consists of two bedrooms (Room 1 and Room 3) and a living room (Room 5), which are the functioning rooms for training and testing HVAC conditioning. The toilet and kitchen, which are occupied only under specific circumstances, are not considered. The room layout is shown in Fig. 4.

The occupancy of the apartment varies depending on the day of the week and the time of the day. Specifically, there are two occupants when Room 1 is occupied and one occupant when Room 3 is occupied. The schedule of occupancy is provided in Table 2, where people are dispersed differently based on their jobs on weekdays and weekends.

The weather data is obtained from the Bureau [30], where the weather in Changsha is used for training while the weather in Shanghai and Chongqing are used for testing. The study focuses on cooling and is limited to the cooling season from May to September. Additionally, a simulated electricity pricing sequence is created to provide price signals, where electricity prices alternate between high and low values every 4 h. This simulated pricing sequence is designed to test whether the DRL agent could detect the impact of price signals on the reward function and adjust its control strategies accordingly.

#### 4.2.2. A brief introduction to the multi-zone HVAC systems control problem

In this paper, we focus on optimizing energy consumption and thermal comfort control in a residential building with 3 thermal zones using HVAC systems. The simulation environment relies on a state-space method, which enables the formulation of state equations to accurately describe the heat balance [31,32]. We adopt a discretized time model where time is represented by t = 0, 1, 2, … with each time step lasting for half an hour. At each time step, monitor the room temperature and determine if adjustments are necessary. If the temperature exceeds or falls below the target range, adjustments are required. To achieve indoor temperature control, calculate the cooling load of the space, which represents the amount of heat that needs to be removed from the room. By calculating the cooling load, the HVAC system controls the room temperature and keeps it in line with the target range in the corresponding time steps. For more details about the calculation methods and model, please refer to Appendix A in the description.

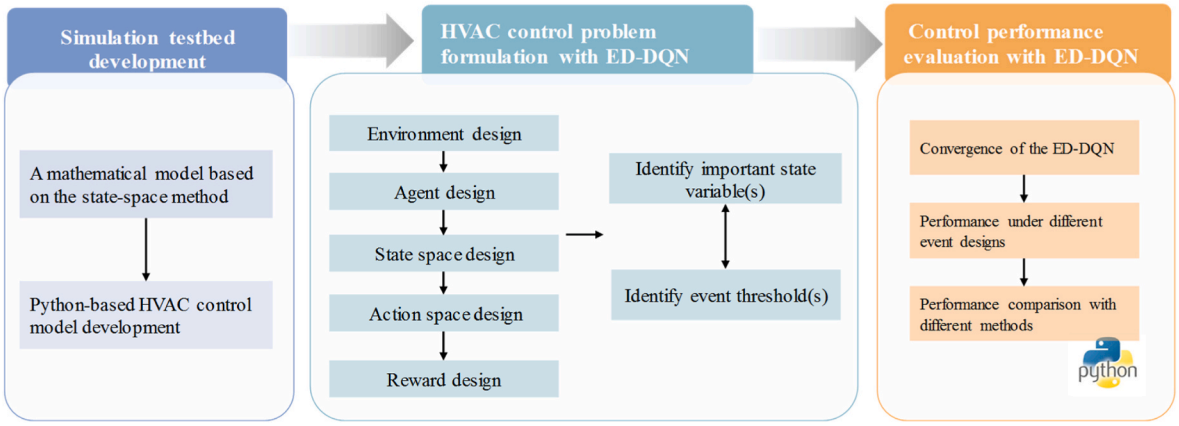



**Fig. 3.** The overall framework.

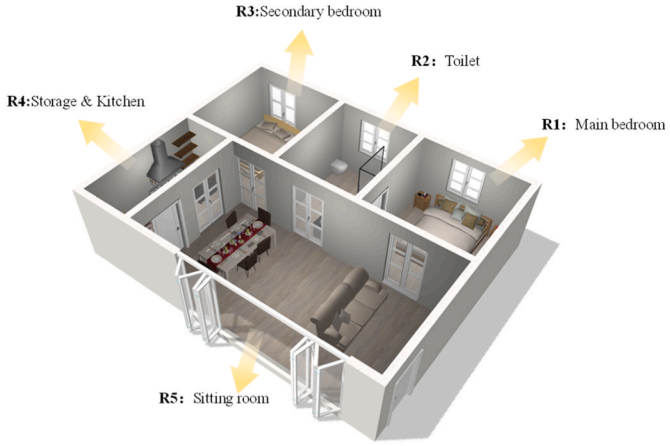



**Fig. 4.** The layout of the 3-occupant apartment.

However, for the traditional RL method involving a periodic and discrete learning process, the learning process may be computationally inefficient especially when the learning environment is stable. To address this problem, we model the HVAC control as an ED-MDP and propose an event-driven method to improve learning performance.

*4.2.3. Modeling the HVAC control problem into an ED-MDP*

In this section, we formulate the multi-zone HVAC control problem as an ED-MDP, represented by a tuple $(S, A, \mathrm{r}, p, \gamma, e)$. As the indoor temperature at each time step is only influenced by the previous interval, the problem can be viewed as a finite Markov process and is independent of the temperature in other time steps. Specifically, for the multi-zone residential HVAC control problem, the details are shown as follows:

**State:** Current outdoor temperature $T_{out}(t)$; current indoor temperature $T_{in,z}(t)$ for the zone $z$; electricity price $\lambda^{electricity}(t)$; current people number $K_z(t)$ in zone $z$, where $t$ is the current time step. Therefore, the state can be represented as $[T_{out}(t), T_{in,1}(t), T_{in,2}(t), T_{in,3}(t), \lambda^{electricity}(t), K_1(t), K_2(t), K_3(t)]$.

**Action:** The setpoint $Setpt_z(t)$ for the zone $z$. Therefore, the action can be represented as $[Setpt_1(t), Setpt_2(t), Setpt_3(t)]$.

The HVAC unit in each zone has a binary on/off state, which is controlled by the following logic based on the current indoor temperatures:

$$HVAC\ state = \begin{cases} 0, if\ T_{in}(t) < setpt_z(t) \\ 1, if\ T_{in}(t) > setpt_z(t) \\ remain\ at\ current\ status, else \end{cases} \tag{6}$$

In Equation (6), it is specified that HVAC systems will stop operating if the indoor temperature falls below the setpoint. Conversely, if the indoor temperature exceeds the setpoint, the HVAC systems will adjust the temperature towards the setpoint to ensure thermal comfort. If the indoor temperature is already at the setpoint, the HVAC systems will maintain their current state to sustain the desired thermal comfort for occupants.

**Reward:** We establish a reward and penalty formula to optimize indoor thermal comfort and energy saving in the HVAC system:

$$r(t) = \begin{cases} -\alpha \sum_{t'=t-\Delta t}^{t} E_{HVAC}(t') + \beta \sum_{t'=t-\Delta t}^{t} R^{comfort}(t') - \sum_{t'=t-\Delta t}^{t} SW^{penalty}(t'), if \mathrm{T}_{in}(t) > th^{\min} \\ \sum_{t'=t-\Delta t}^{t} R^{off}(t') - \sum_{t'=t-\Delta t}^{t} SW^{penalty}(t'), if \mathrm{T}_{in}(t) < th^{\min} \end{cases} \tag{7}$$

In Equation (7), the first line consists of three terms, where the first term represents energy consumption of HVAC systems in each zone. The second term denotes the reward for thermal comfort in each zone, and the third term $SW^{\mathrm{penalty}}(t')$ represents the penalty for frequent switching of HVAC systems. To be more specific, the reward for thermal comfort is defined as follows:

**Table 2**
Schedule of people.

| Weekday/Weekend | Schedule | Room 1 | Room 3 | Room 5 |
|---|---|---|---|---|
| Weekday | 0:00–7: 00, 20:00–24:00 | 2 | 1 | 0 |
| | 7:00–8:00, 18:00–20:00 | 0 | 0 | 3 |
| weekend | 0:00–7: 00, 20:00–24:00 | 2 | 1 | 0 |
| | 7:00–20:00 | 0 | 0 | 3 |

$$R^{comfort}(t) = \frac{1}{1 + e^{\left(T_{in}(t) - th^{best}\right)^2}} - R^{\text{penalty}}(t), \tag{8}$$

Equation (8) employs a Gaussian equation to define the current indoor thermal comfort level. In each zone, a thermal comfort zone $[th^{\min}, th^{\max}]$ is defined, where $th^{\min}$ and $th^{\max}$ represent the lowest and highest temperatures that occupants feel comfortable, respectively. The most comfortable temperature is represented by $th^{best}$. $R^{penalty}(t)$ denotes the penalty term for deviations from the comfort zone as shown in Equation (9):

$$R^{\text{penalty}}(t) = \begin{cases} if\, \mathrm{T}_{in}(t) < th^{\min}, th^{\min} - \mathrm{T}_{in}(t) \\ if\, \mathrm{T}_{in}(t) > th^{\max}, \mathrm{T}_{in}(t) - th^{\max} \\ otherwise, 0 \end{cases} \tag{9}$$

Equation (7) uses weights $\alpha$ and $\beta$ to balance energy consumption and thermal comfort in each zone, based on factors such as room occupancy and electricity prices. To further reduce energy consumption, it is important to avoid overcooling by shutting down HVAC systems. The second line of Equation (7) includes a reward term for shutting down HVAC systems, denoted as $R^{off}(t)$. However, to prevent frequent switching that may incur additional costs, the model also considers the cost of switching the HVAC system $SW^{\text{penalty}}(t)$.

When the trigger function $e$ exceeds a certain threshold, the agent is triggered and the state transition occurring. The state transition function $P$ is modified as $P\{s(t+1)|s(t), a(t), e\}$ which denotes the probability that action $a_t$ at event $e$ will cause the system transition. However, defining the state transition function for ED-MDP in multi-zone residential buildings is a challenge due to the complexity of the environment. Even if a model is established, defining the transition function may not guarantee that the agent will obtain the optimal strategy. Therefore, a model-free ED-DRL method can avoid these problems while improving adaptability and robustness.

#### 4.2.4. Definition of HVAC control events

ED-MDP leverages the event-driven method to optimize system performance, which enhances system responsiveness and real-time performance by responding to events only when it is necessary, rather than performing operations periodically. The crucial aspect of triggering the event-driven method is to identify events, which are changes in the environment that prompt the agent to take action. In practice, events are identified based on predefined triggering rules that involve identifying important state variables and event thresholds [19]. These rules specify that when the values of these variables change or exceed specific thresholds, the corresponding events are triggered. For instance, in an HVAC system, a trigger rule can be set when the room temperature surpasses a certain threshold, the system will automatically adjust the temperature to maintain comfort. By predetermining events, event-driven methods can accurately respond to environmental changes and perform the required control actions, thereby improving the system's efficiency and adaptability.

Different state changes can have varying impacts on HVAC optimal control. In the previous modeling of ED-MDP, the state variables included outdoor temperature, indoor temperature, electricity price, and the number of people in the room. With the optimization objectives being thermal comfort and energy consumption, we consider that the changes in the latter three variables can be regarded as potential events. The trigger rule used in this paper is detailed in Table 3. The trigger rule is based on the price signal ($\lambda$), where 0 indicates low prices and 1 for high prices. K denotes the occupancy status of the room, with 1 indicating the presence of occupants and 0 indicating no occupants.

**Table 3**
Event trigger rule.

| Events 1-6 | |
|---|---|
| Event 1 | $\lambda$ : 0→1 *or* 1→0 |
| Event 2 | $K$ : 0→1 *or* 1→0 |
| Event 3 | $K = 1, \lambda = 0$ and $TH^{comfort}(t) < a$ |
| Event 4 | $K = 1, \lambda = 1$ and $TH^{comfort}(t) < b$ |
| Event 5 | $K = 0, \lambda = 0$ and $TH^{comfort}(t) < c$ |
| Event 6 | $K = 0, \lambda = 1$ and $TH^{comfort}(t) < d$ |

*4.2.4.1. State transition events.* Event 1 is associated with price changes [33], as they have a direct impact on the energy consumption. Prices are categorized into two states: high prices and low prices. When prices are low, the emphasis on energy savings can be moderately relaxed, whereas when prices are high, the need for energy savings becomes more critical. Therefore, triggering event 1 enables an accurate assessment of energy consumption by accounting for the uncertainty in prices. We define event 1 as triggered when the current price differs from the price in the previous moment.

Similarly, event 2 is linked to the room occupancy [34], which captures the variability in the number of people. The room occupancy is categorized into two states: occupied and unoccupied. When a room is unoccupied, the demand for thermal comfort can be relaxed. We define event 2 as triggered when the current number of people in the room differs from the number of people in the room at the previous moment.

*4.2.4.2. Combination events.* Thermal comfort is influenced by a combination of indoor temperature, price, and room occupancy, with room temperature, in particular, having a direct impact on thermal comfort [35]. Therefore, events 3–6 are defined based on different thermal comfort conditions. The value of thermal comfort is represented by the first half of Equation (8), which corresponds to a positive reward denoted as $TH^{comfort}$. This design enables the agent to effectively respond to thermal comfort. To define events 3–6, we have established distinct thresholds: $a$, $b$, $c$, and $d$, with the order $a > b > c > d$. When the room is occupied, event 3 is triggered if the price is low and the thermal comfort value is below threshold $a$, while event 4 is triggered if the price is high and the thermal comfort value is below threshold $b$. On the other hand, when the room is unoccupied, event 5 is triggered if the price is low and the thermal comfort value is below threshold $c$, and event 6 is triggered if the price is high and the thermal comfort value is below threshold $d$. Thus, the triggering of events 3–6 ensures that the occupants' thermal comfort needs are met with minimal energy consumption.

By defining these events and thresholds, we can allow the agent to learn useful information faster and make better decisions.

## 5. Algorithm for HVAC control

In this section, we present the ED-DQN method for controlling HVAC systems. The ED-DQN framework is illustrated in Fig. 5.

The details of training ED-DQN is summarized in Algorithm 1. At the beginning of the algorithm, the Q-network and target Q-network are initialized with the same parameters. During each iteration, the DQN acquires state information about residential buildings and decides whether to trigger the events predefined in Section 4.4. These events take into account price changes, room occupancy, and thermal comfort. In lines 4–8, if an event is triggered, the agent takes an action based on the current state. Otherwise, it continues to perform the current action until the event is triggered again. Equation (10) is used to select the action after the trigger event, following an ε-greedy policy with ε denoting the "greedy factor". Random actions are selected with probability ε, while the action with the largest value function is selected with probability 1-ε. Notably, ε decreases progressively during the training process until it reaches its minimum value.

$$\mathrm{a}(n) = \begin{cases} Random\ \mathrm{a} \in \mathrm{A} | \varepsilon \\ \underset{a \in A}{\operatorname{argmax}}\, Q(s, a) \Big| 1 - \varepsilon \end{cases} \tag{10}$$

Following that, the selected action is carried out, and the obtained

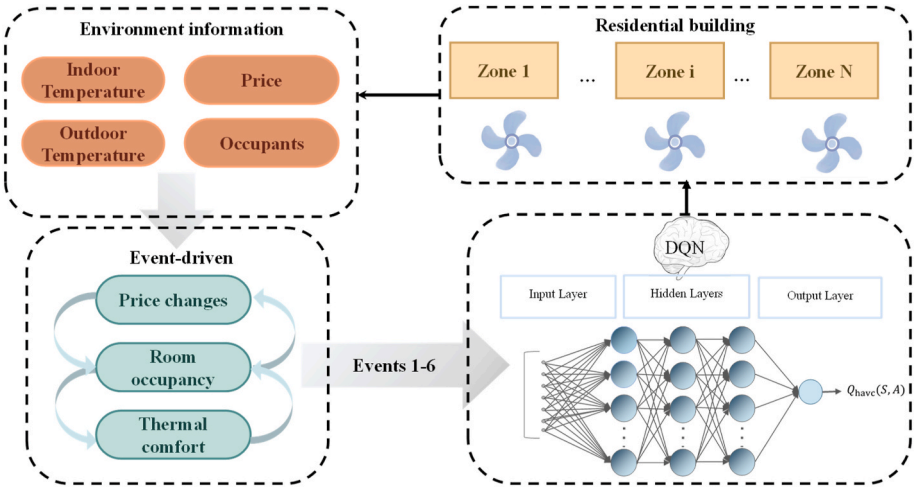



**Fig. 5.** The proposed ED-DQN framework for building HVAC systems control.

**Algorithm 1** ED-DQN for multi-zone HVAC systems control

**Require**: Initialize action-value function $Q$ with random weight $\omega$

**Require**: Initialize target action-value function $Q^{'}$ with random weight $\omega^{'}$

**Require**: Initialize minibatch size B, replay buffer $R$, and the total number of iterations $N$

1. **for** episode=1 to $N$ do
2. Initialize system state $s(T_{out}(0), T_{in,z}(0), \lambda^{electricity}(0), K_z(0))$
3. **for** t=1, T do
4. **if an event occurs then**
5. Select the multi-zone HVAC control action $Setpt_z(t)$ by ε-greedy policy
6. **else**
7. Continue execute action $Setpt_z(t)$
8. **end if**
9. R(t) is obtained according to i.e. (7), $s(t+1)$ is observed from the environment
10. Store the transition ( $s(t), Setpt_z(t), r(t), s(t+1)$ ) in the replay buffer $R$
11. Collect a mini-batch of transitions ( $s^{(i)}(t), Setpt_z{}^{(i)}(t), r^{(i)}(t), s^{(i)}(t+1)$ ) with the size B from replay buffer $R$
12. Set $Q'^{(i)}_{*}(s^{(i)}(t), Setpt_z{}^{(i)}(t); \omega) = r^{(i)}(t+1) + \gamma \max Q'^{(i)}(s^{(i)}(t), Setpt_z{}^{(i)}(t); \omega')$
13. Perform a gradient descent step on i.e."(11)", with respect to the network parameter $\omega$
14. Every delayed policy update $U$ steps reset $Q^{'} = Q$
15. **End for**
16. **End for**

reward and the following state are observed. The transition $(s(t), Setpt_z(t), r(t), s(t+1))$ is saved in a replay buffer and will be utilized for method training in the future. When a sufficient number of transitions is gathered, a mini-batch of transitions is randomly picked to minimize the error between the target $Q^{'}(s,a;\omega^{'})$ network Q value and the output $Q(s,a;\omega)$ network Q value, as indicated by line 11.

$$Loss = \frac{1}{B}\sum_{i=1}^{B}\left[Q^{'}_{*}\left(s^{(i)}(t), Setpt_z{}^{(i)}(t);\omega\right) - Q\left(s^{(i)}(t), Setpt_z{}^{(i)}(t);\omega^{'}\right)\right] \quad (11)$$

The loss functions are used to update the neural network parameter $\omega$. As shown in Equation (11), the loss function is defined as the MSE between target Q value and the current Q value. With delayed policy update $U$, the weighting parameter $\omega$ of the $Q(s,a;\omega)$ network is copied to update the weighting parameter $\omega^{'}$ of the $Q^{'}(s,a;\omega^{'})$ network after the

**Table 4**
DNN structure applied in ED-DQN and DQN algorithm.

| Algorithm | DQN |
|---|---|
| Size of input | 8 |
| Size of output | 1331 |
| No. of hidden layers | 2 |
| Activation function | ReLU |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Discount factor | 0.99 |
| Buffer size | 10000 |
| Delayed policy update | 2 |
| $\varepsilon^{max}$ | 0.99 |
| $\varepsilon^{min}$ | 0.1 |
| $\varepsilon^{decay}$ | 0.00003 |

**Table 5**
DNN structure applied in ED-DDPG and DDPG algorithm.

| Algorithm | ED-DDPG |
|---|---|
| Critic network | 1 → 11→ 64→ 32 |
| Actor network | 1 → 8→ 64→ 32 |
| No. of hidden layers | 2 |
| Activation function | Actor: Tanh; Critic: ReLU |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Discount factor | 0.9 |
| Buffer size | 10000 |
| TAU | 0.005 |

loss function is determined. $B$ represents the size of the minibatch, while $i$ represents the index of the sample within the current minibatch.

## 6. Experiments

Firstly, we present the results of the convergence analysis of ED-DQN, which is based on real-world data. Secondly, we conduct ablation experiments to demonstrate the effectiveness of event definition. Thirdly, we compare ED-DQN with traditional RL methods and other methods to fully validate its superior performance in balancing energy consumption and thermal comfort. Finally, we demonstrate the generalizability of our method by testing it on different thermal environments.

### 6.1. Experiment setting

#### 6.1.1. Algorithm parameter settings

Table 4 presents parameters used in ED-DQN, which has a discrete action space. The action space is discretized in steps of 0.5 °C to cover the range of setpoints from 23 °C to 28 °C per zone, resulting in 11 possible actions per zone and a total of 1331 action combinations for 3-zone HVAC systems. On the other hand, Table 5 presents parameters used in ED-DDPG. The critic network takes both state and action variables as input and outputs the value function, while the actor-network only receives state variables, and outputs an action, represented as a vector containing setpoints for each zone. The activation function used in the output layer of the actor-network is tanh, outputting the value ranged in [−1, 1].

#### 6.1.2. Experiment parameter settings

We choose [23,26] as the thermal comfort zone, with 26 as the optimal setpoint $th^{best}$. Therefore, for Equation (8), the optimal value of $R^{comfort}(t^{'})$ is 0.5. We assigned different thermal comfort levels to variables a, b, c, and d, corresponding to 0.4, 0.3, 0.2, and 0.1, respectively. A higher level indicates a higher thermal comfort level. In order to avoid frequent control, the penalty for switching the HVAC systems is set to 0.5.

Considering the balance of energy consumption and thermal comfort, the reward weights are adjusted by price and room occupancy using dynamic weights. Table 6 shows the different weights for the three function rooms.

#### 6.1.3. Benchmark method

To evaluate the effectiveness of the proposed ED-DQN method, we designed three benchmark cases for comparison without using an RL agent. The specific benchmark cases are as follows:

a) Fixed setpoint case: The setpoint is always the optimum temperature of 26 °C to avoid the thermal discomfort.
b) Rule-based case (RBC): For RBC1, the setpoint is set to 27 °C when the room is unoccupied, and in other cases, the temperature is adjusted based on the number of occupants, as shown in Table 7. For RBC2, by setting the temperature to 28 °C during low peak price periods and 24 °C during high peak price periods, a pre-cooling effect can be achieved to energy saving.

**Table 6**
Dynamic Weights of energy consumption and thermal comfort.

| Weighting parameters | Occupied, price is high | Occupied, price is low | Unoccupied |
|---|---|---|---|
| α | 0.5 | 1 | 0.1 |
| β | 1 | 1 | 1 |

**Table 7**
RBC.

| Weekday/Weekend | Schedule | Room 1 | Room 3 | Room 5 |
|---|---|---|---|---|
| Weekday | 0:00–7: 00, 20:00–24:00 | 25 °C | 26 °C | 27 °C |
| | 7:00–8:00, 18:00–20:00 | 27 °C | 27 °C | 24 °C |
| weekend | 0:00–7: 00, 20:00–24:00 | 25 °C | 26 °C | 27 °C |
| | 7:00–20:00 | 27 °C | 27 °C | 24 °C |

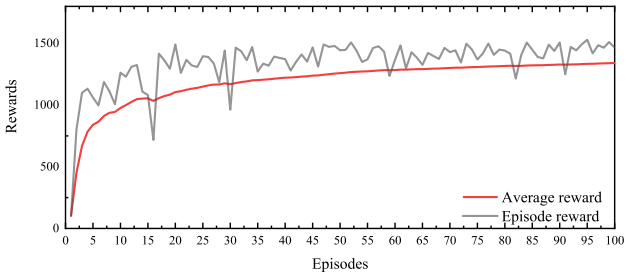



**Fig. 6.** Convergence of the ED-DQN and other methods.

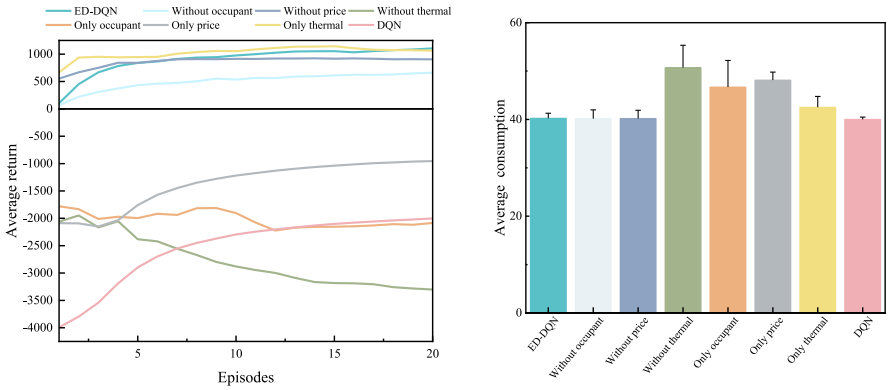
### 6.2. Performance of the event-based HVAC control method

#### 6.2.1. Convergence of the ED-DQN

Fig. 6 shows the convergence of the ED-DQN method, where the reward obtained in each episode gradually increases. The X-axis represents the number of episodes, while the Y-axis represents the rewards earned per episode. Due to the exploration operation and the varying system parameters in each episode, the rewards fluctuate considerably. Nevertheless, as the training advances, the results improve and the fluctuations diminish. To better present the trend of reward variation, we provide the average reward for 100 episodes. As can be seen, the average reward gradually increases and becomes more stable over time, eventually converging after approximately 70 episodes.

#### 6.2.2. Performance under different event designs

Fig. 7 shows the results of the ablation experiments we conducted on events. Since the ED-DQN involves multiple events, different events may lead to different performances, even when using the same system parameters. We gradually removed three essential factors to evaluate the effectiveness of events, namely price changes, room occupancy, and thermal comfort. In Fig. 7(a), we present the average rewards obtained under different events. The X-axis represents the number of episodes and the Y-axis represents the average reward obtained when using different events. In the first 18 episodes, optimizing solely for thermal comfort resulted in the highest average reward. However, this method exhibits a decreasing trend in average reward after the 15th episode, whereas ED-DQN continues to improve and eventually outperforms it. When price changes or room occupancy are not considered, the performance is still good, although the average reward is not as high as that of ED-DQN. Furthermore, In Fig. 7(b), the X-axis represents the different events, while the Y-axis represents the average energy consumption generated over twenty episodes when utilizing these events. Additionally, the error




(a) average return (b) average energy consumption

**Fig. 7.** ED-DQN performance under different events. (a)Compare in average return; (b) Compare in average energy consumption.

bars indicate the highest energy consumption. It can be seen that considering only thermal comfort requires more energy consumption. When thermal comfort events are not considered, the learning effect is the worst, leading to high energy consumption and a significant decrease in the average return. On the other hand, when events are not used (DQN method), the energy consumption is comparable to our method, but the average return and learning speed are significantly lower than our method. Additionally, in the first episode, ED-DQN achieved a reward of approximately 100 and completed the control in about 122 s, whereas DQN obtained a reward of approximately −4000 and took around 172 s to complete. These results demonstrate that ED-DQN outperforms DQN in terms of both running speed and learning speed.

Fig. 8 shows the distribution of indoor temperatures in three different function rooms under various events. There are approximately 1500 data for each room, with the x-axis representing indoor temperature and the y-axis representing the frequency of the occurrence. The two red lines indicate the range of thermal comfort. To better observe the impact of HVAC control, we chose the month of July for comparison, as this is a time when outdoor temperatures are typically high. The experimental results indicate that the HVAC control is sensitive to the inclusion of all three influencing factors. Excluding any one of these factors can decrease the effectiveness of HVAC control and lead to uneven indoor air temperature distribution. Maintaining thermal comfort is particularly critical for building management, and excluding it has a significant impact on HVAC control performance. On the other hand, when only one influencing factor was considered, such as in Fig. 8(e) and (f), the HVAC control performance worsened. Fig. 8(g) maintained indoor temperature within the thermal comfort threshold but at the expense of energy consumption and inflexibility in adapting to environmental change. Lastly, the absence of an event-driven method in Fig. 8(h) (DQN method) leads to poor temperature maintenance in two rooms, with room 1 being too hot and room 5 being too cold. Based on the analysis of Figs. 7 and 8, we can conclude that the ED-DQN method has a significant advantage over the DQN method in HVAC control. The inclusion of all three influencing factors is critical for effective indoor temperature control, with thermal comfort being the most influential factor.

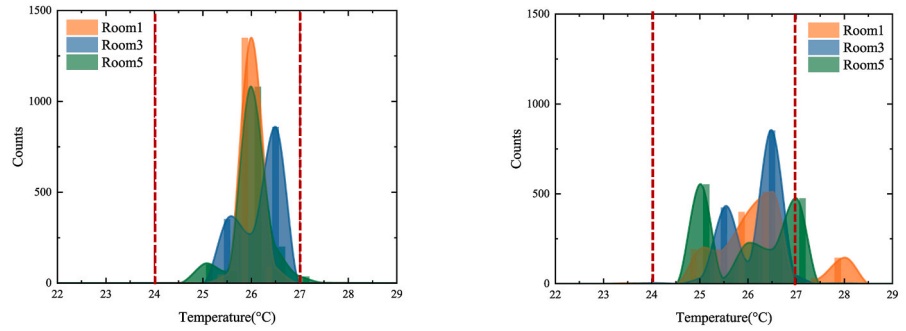
### 6.2.3. *Performance comparison with different methods*

As the reward may continue to fluctuate even after convergence (ED-DQN fluctuates only slightly), we select the episode that performs the best after convergence as the result to ensure a more representative outcome. Table 8 compares the performance of ED-DQN with other methods in terms of energy consumption and thermal comfort violation. Specifically, the thermal comfort violation is defined as the ratio of the number of instances exceeding the thermal comfort threshold to the total data volume. The results show that the addition of the event-driven method significantly improves energy consumption and thermal comfort compared to DQN. ED-DQN is able to trade off thermal comfort and energy consumption fairly well, resulting in relatively low energy consumption with very low thermal comfort violations. While DDPG is a popular RL method that is capable of handling continuous control problems, it may not be suitable for our specific design in this paper. DDPG fails to produce desirable outcomes within a limited timeframe, possibly due to incorrect hyperparameter selection or unsuitable environmental reward design, among other factors. Furthermore, the ED-MDP problem we formulated is not compatible with DDPG. ED-DDPG does not handle energy consumption and thermal comfort well, making ED-DQN the preferred method. The benchmark results demonstrate that traditional methods, such as RBC1, are ineffective in achieving the desired thermal comfort level, despite being set to avoid violations. RBC2, which exhibits the lowest energy consumption, has the highest thermal comfort violation. The fixed setpoint method, with a constant setpoint of 26, has the lowest temperature violation but also has the highest energy consumption. These suggest that disengagement from RL control is effective. It can be concluded that ED-DQN, which can be flexibly adapted to unseen physical environments, can provide an economical and comfortable HVAC control. Figs. 9–11 further illustrate the excellence of ED-DQN control.

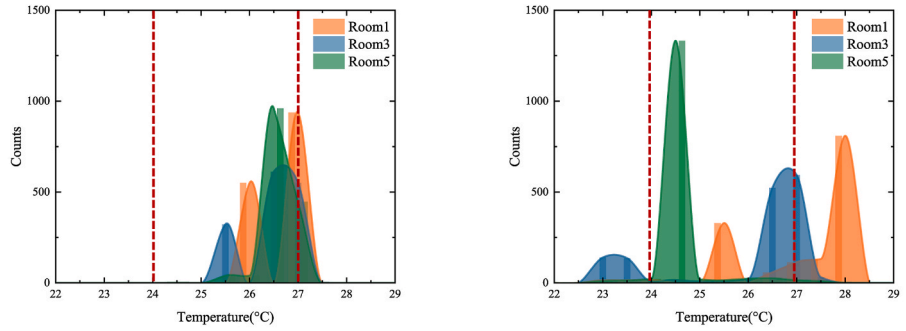
Fig. 9 shows the indoor temperature distribution in Changsha from May to September using different methods. The x-axis represents the percentage of indoor temperature, while the y-axis represents the month. The darker blue indicates lower indoor temperature, while the darker red indicates higher indoor temperature. ED-DQN performs the best with comfortable indoor temperature in all three rooms for most of the time, while DQN performs the worst, with room 1 having a higher temperature from June to August and room 5 having a lower temperature each month. ED-DDPG is better than DDPG in the thermal comfort, but the cooling effect of room 3 is not very effective.

In Fig. 10, the indoor temperatures of the three rooms on July 31 are compared for several methods. The x-axis represents the time of day, and the y-axis represents the corresponding temperature at that time. The three rooms of ED-DQN maintain the best thermal comfort regardless of the outdoor temperature, making the best decision automatically with the room occupancy and price.

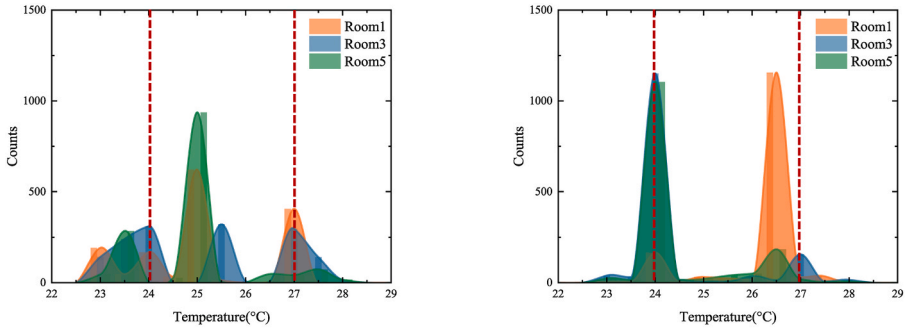
Reducing the decision rate has several benefits, including saving computing resources, reducing response times, minimizing the accumulation of computing errors, and improving the accuracy and stability of the control system. In addition, it helps mitigate equipment damage, reduces repair and replacement costs, and increases the overall reliability and durability of HVAC systems. As shown in Fig. 11, the X-axis



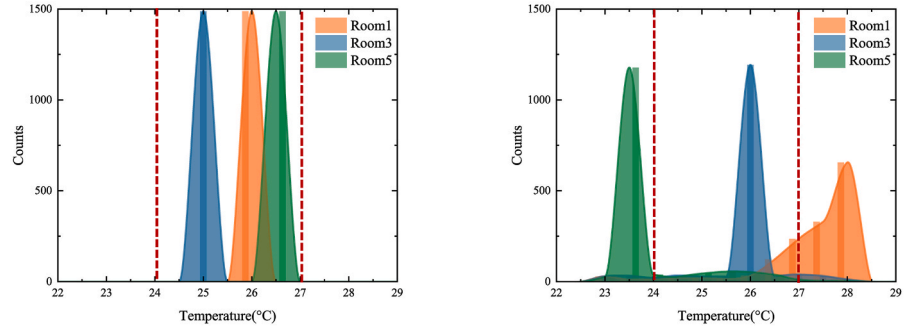

(a) ED-DQN (b) Without occupant




(c) Without price (d) Without thermal




(e) Only occupant (f) Only price




(g) Only thermal (h) DQN

**Fig. 8.** Indoor air temperature distributions of three rooms for different events in July.

**Table 8**
Results of different HVAC control methods in Changsha.

| Control method | ED-DQN | DQN | ED-DDPG | DDPG | RBC1 | RBC2 | Fix setpoint |
|---|---|---|---|---|---|---|---|
| Energy consumption | 40.079 | 40.813 | 41.861 | 39.112 | 39.513 | 38.183 | 40.162 |
| Temperature violation | 6.647% | 49.113% | 17.921% | 18.057% | 17.658% | 64.361% | 6.533% |

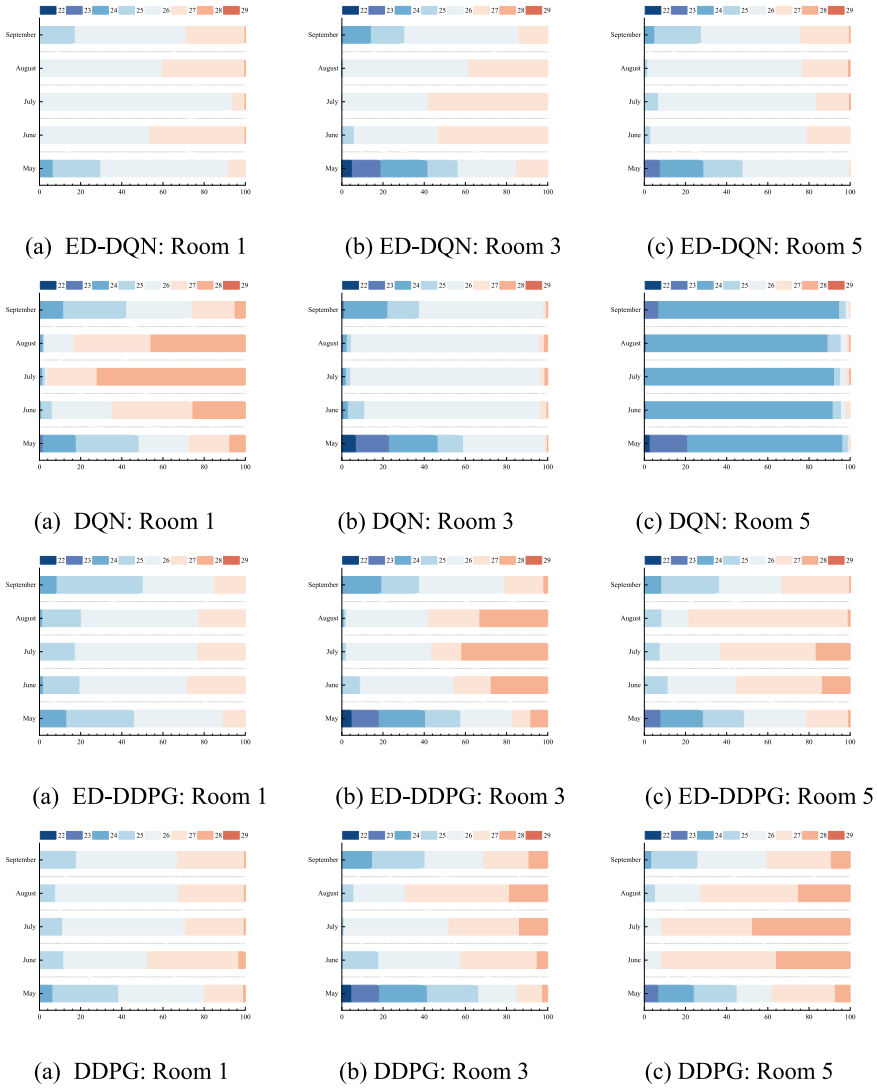



**Fig. 9.** Indoor air temperature distributions of three rooms in different methods from May to September in Changsha.

represents the time step from May to September, and the Y-axis represents the number of decisions of the corresponding method. It is evident that the decision frequency of ED-DQN is much lower than that of DQN.

To simplify the expression, we define the number of decisions in DQN as $D_1$ and the number of decisions in ED-DQN as $D_2$. The decision reduction rate can be calculated as follows:

$$\varphi = \frac{D_1 - D_2}{D_1} \tag{12}$$

Table 9 provides a comparison of the number of action decisions and reduction decision rate between ED-DQN and DQN. The results indicate that ED-DQN outperforms DQN, with reduction decision rates of 87.697%, 77.218%, and 75.077% for the three rooms, respectively.

To fully validate the versatility and robustness of ED-DQN, it was further tested in two new thermal environments. Tables 10 and 11 (related tables and figures in appendix B) compare the energy consumption and thermal comfort violations under various RL methods and three benchmark cases. Similar to the results in Table 8, the fixed setpoint control strategy provides the lowest violation. DDPG achieved relatively low energy consumption, but at the expense of poorer thermal comfort. Although ED-DDPG outperforms DDPG in terms of thermal comfort, it results in relatively higher energy consumption. Among all the RL methods, DQN exhibited the highest thermal comfort violation. Specifically, RBC2, which focuses only on price changes, resulted in the highest thermal comfort violation.

Compared to other reinforcement learning methods, ED-DQN

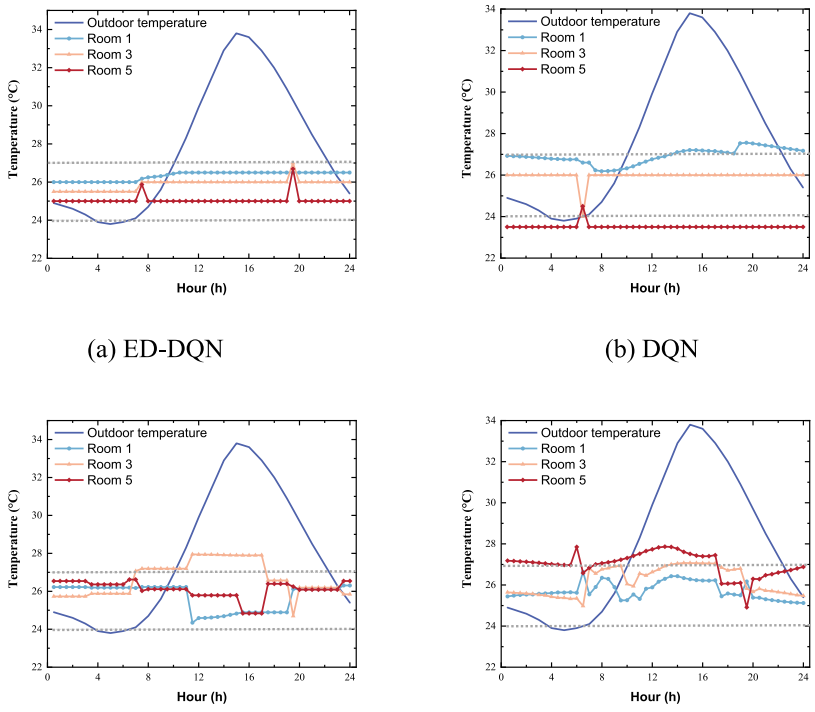



**Fig. 10.** Indoor temperature on July 31 in Changsha. (a) ED-DQN; (b) DQN; (c) ED-DDPG; (d) DDPG.

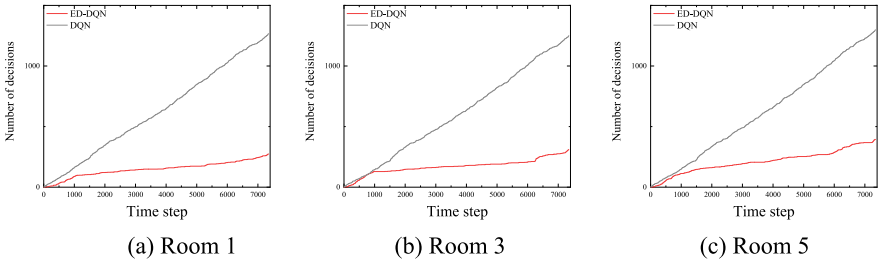



**Fig. 11.** The number of decisions of three rooms in different methods from May to September in Changsha.

**Table 9**
A comparison of the number of decisions from May to September in Changsha.

| Room | DQN ($D_1$) | ED-DQN ($D_2$) | Reduction of decision rate ($\varphi$) |
|---|---|---|---|
| 1 | 1268 | 156 | 87.697% |
| 3 | 1251 | 285 | 77.218% |
| 5 | 1300 | 324 | 75.077% |

exhibits more pronounced advantages in terms of thermal comfort. The design of the proposed method prioritizes thermal comfort, ensuring that the indoor environment remains thermally comfortable for the most of the time. Furthermore, our experimental results demonstrate that ED-DQN significantly reduces the number of decisions required, leading to more efficient utilization of computational resources. In contrast to rule-based control (RBC) methods that heavily rely on expert knowledge, ED-DQN offers greater flexibility. While RBC struggles to strike a balance between thermal comfort and energy consumption, ED-DQN effectively weighs these two factors through a process of learning and optimization. By adapting to specific environmental requirements, ED-DQN achieves outstanding performance by minimizing energy consumption while preserving thermal comfort. Therefore, it can be safely concluded that ED-DQN outperforms other methods in terms of both energy consumption and thermal comfort, making it a promising approach for HVAC control.

**Table 10**
Test results of different HVAC control methods in Chongqing.

| Control method | ED-DQN | DQN | ED-DDPG | DDPG | RBC1 | RBC2 | Fix setpoint |
|---|---|---|---|---|---|---|---|
| Energy consumption | **44.020** | 39.749 | 45.671 | **37.118** | 44.968 | 43.773 | 45.689 |
| Temperature violation | **12.263%** | 34.336% | 16.351% | 33.597% | 14.532% | 67.325% | **4.818%** |

**Table 11**
Test results of different HVAC control methods in Shanghai.

| Control method | ED-DQN | DQN | ED-DDPG | DDPG | RBC1 | RBC2 | Fix setpoint |
|---|---|---|---|---|---|---|---|
| Energy consumption | **36.036** | 32.512 | 37.769 | **29.452** | 37.596 | 36.489 | 37.811 |
| Temperature violation | **13.139%** | 33.465% | 16.528% | 29.559% | 15.548% | 62.012% | **4.909%** |

## 7. Conclusion and further work

In this paper, a new framework called ED-MDP and a new method called ED-DQN are proposed for optimizing HVAC control in multi-zone residential buildings. The experimental results demonstrate that the proposed ED-DQN method, compared to traditional methods, achieves a better balance between thermal comfort and energy consumption without requiring an accurate thermodynamic model. Compared to traditional RL methods, ED-DQN can learn faster through non-periodic learning, leading to a significant reduction in the frequency of decisions. Additionally, this method exhibits generalization capabilities for unknown environments.

However, although ED-DQN performs well in HVAC systems, the definition of events relies on prior knowledge, which is not always completely accurate. Therefore, constructing accurate and reliable events is of utmost importance. To achieve this goal, leveraging machine learning and data visualization methods to dynamically mine deep data is an important research direction for the future.

## CRediT authorship contribution statement

**Qiming Fu:** Writing – review & editing, Methodology, Funding acquisition. **Zhu Li:** Writing – original draftWriting – original draft, Software. **Zhengkai Ding:** Software, Investigation, Data curation. **Jianping Chen:** Visualization, Funding acquisition. **Jun Luo:** Project administration, Formal analysis. **Yunzhe Wang:** Visualization, Supervision, Funding acquisition. **You Lu:** Supervision, Resources.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgment

This work was financially supported by the National Key R&D Program of China (No.2020YFC2006602), National Natural Science Foundation of China (No. 62072324, No.62102278), University Natural Science Foundation of Jiangsu Province (No.21KJA520005), Primary Research and Development Plan of Jiangsu Province (No.BE2020026), and Postgraduate Education Reform Project of Jiangsu Province.

## Appendix A

The term “State-space-based mathematical modeling” refers to the development of a mathematical model for a system that is based on the linear relationship between the system’s state variables and input variables [31]. In our study, we focused on a typical three-person apartment layout and determined the number of rooms and internal spaces. By considering the heat transfer regions of the building, we discretized the building into a series of temperature nodes, with walls, ceilings, and floors treated as boundaries. We assumed a multi-story exterior wall divided into n layers and defined corresponding temperature nodes for each layer [29]. The specific equations are presented as Equations. (A.1)- (A.6) [32]:

For the inner surface (node “1″), the equation is as follows

$$\frac{1}{2}S_1\rho_1\kappa_1\frac{dT_1}{dt}=H_1(T_{inner}-T_1)+\frac{tc_1}{\kappa_1}(T_2-T_1)+\sum_i rh_{1,i}\left(T_{1,i}-T_1\right)+h_1 \tag{A.1}$$

For the intermediate node (node “j"), the equation is as follows:

$$\left(\frac{1}{2}S_{j-1}\rho_{j-1}\kappa_{j-1}+\frac{1}{2}S_j\rho_j\kappa_j\right)\frac{dT_j}{dt}=\frac{tc_{j-1}}{\kappa_{j-1}}\left(T_{j-1}-T_j\right)+\frac{tc_j}{\kappa_j}\left(T_{j+1}-T_j\right) \tag{A.2}$$

For the outer surface (node “n+1″), the equation is as follows:

$$\frac{1}{2}S_n\rho_n\kappa_n\frac{dT_{n+1}}{dt}=H_{n+1}(T_{outer}-T_{n+1})+\frac{tc_n}{\kappa_n}(T_n-T_{n+1})+\sum_i rh_{n+1,i}\left(T_{n+1,i}-T_{n+1}\right)+h_{n+1} \tag{A.3}$$

Assuming that the indoor air in each room is a lumped parameter temperature node, the heat balance equation is shown below:

$$S_{inair}\rho_{inair}V_{inair}\frac{dT_{inair}}{dt}=\sum_i H_i f_i(T_i-T_{inair})+S_{inair}\rho_{inair}\frac{r_{AC}\Delta V_{inair}}{3600}(T_{outair}-T_{inair})+h_{conv}+h_{hvac} \tag{A.4}$$

Where the subscript ‘inair’ denotes indoor air and ‘outair’ denotes outdoor air. By dividing the temperature distribution of the building system into a set of state variables, we can effectively describe and analyze the thermal behavior of the building using a state-space method. The heat transfer principle and energy balance principle are employed to establish the heat balance equation for each temperature node. The heat balance equation, which encompasses all temperature nodes within each room, can be reformulated as a set of linear differential equations in matrix form, as follows:

$$C\mathrm{T}=\mathrm{Z}\mathrm{T}+\eta \tag{A.5}$$

Where $C$ represents the heat storage capacity matrix, Z represents the heat flux-flux relationship matrix due to temperature differences between adjacent temperature nodes, and $\eta$ is a vector describing the thermal perturbations at the temperature nodes. T is the temperature vector consisting of

all calculated temperature nodes of the room. Other relevant parameters have been listed in Table A.1.

Figure A.1 depicts the HVAC workflow. During normal operation of the HVAC system, after determining factors such as occupancy and internal heat gain, vector $\eta$ is assigned to the five rooms. Subsequently, the cold loads of R1, R3, R5, and the temperature nodes are calculated, while the temperature nodes of R2 and R4 are directly computed. If the time step exceeds the predefined interval, the result is output; otherwise, the internal gain calculation continues to determine the cold load at the new time step. The cold load of the function room at each time step "j" is calculated as follows [29]:

$$q(j-1) = S_{inair}\rho_{inair}V_{inair}\left(T_{inair}(j-1) - T_{\text{target}}\right) / \Delta t + \sum_i H_{wall}f_i\left(T_{wall,i}(j-1) - T_{\text{target}}\right) + S_{inair}\rho_{inair}\frac{r_{AC}\Delta V_{in}}{3600}\left(T_{outair} - T_{\text{target}}\right) + q_{conv}(j) \tag{A.6}$$

**Table A.1**
Symbol List

| | | | |
|---|---|---|---|
| $S$ | Specific heat capacity, $J/(kg\,k)$ | $q_{conv}$ | convection part of the internal heat gains, $W$ |
| $\rho$ | density, $kg/m^3$ | $H_{wall}$ | convective heat transfer coefficient of the wall, $W/(m^2\,\mathrm{K})$ |
| $\kappa$ | thickness, $m$ | $T_{wall}$ | inner wall surface temperature, $^\circ C$ |
| $H$ | convective heat transfer coefficient, $W/(m^2\,\mathrm{K})$ | $T_{\text{target}}$ | room target control temperature, $^\circ C$ |
| $T_{inner}$ | node temperature of the air close to the inner surface, $^\circ C$ | $h$ | radiative heat gain per area, $W/m^2$ |
| $T_{outer}$ | node temperature of the air close to the outer surface, $^\circ C$ | $h_{conv}$ | heat flux of convective heat transfer, $W/m^2$ |
| $T$ | node temperature, $^\circ C$ | $h_{hvac}$ | heat flux of HVAC, $W/m^2$ |
| $t$ | time, $s$ | $f$ | floor or room inner wall area, $m^2$ |
| $tc$ | thermal conductivity, $W/(m\,\mathrm{K})$ | $V$ | volume, $m^3$ |
| $i$ | room inner wall surface | $r_{AC}$ | room air change rate per hour |
| $rh$ | longwave radiative heat transfer coefficient, $W/(m^2\,\mathrm{K})$ | | |

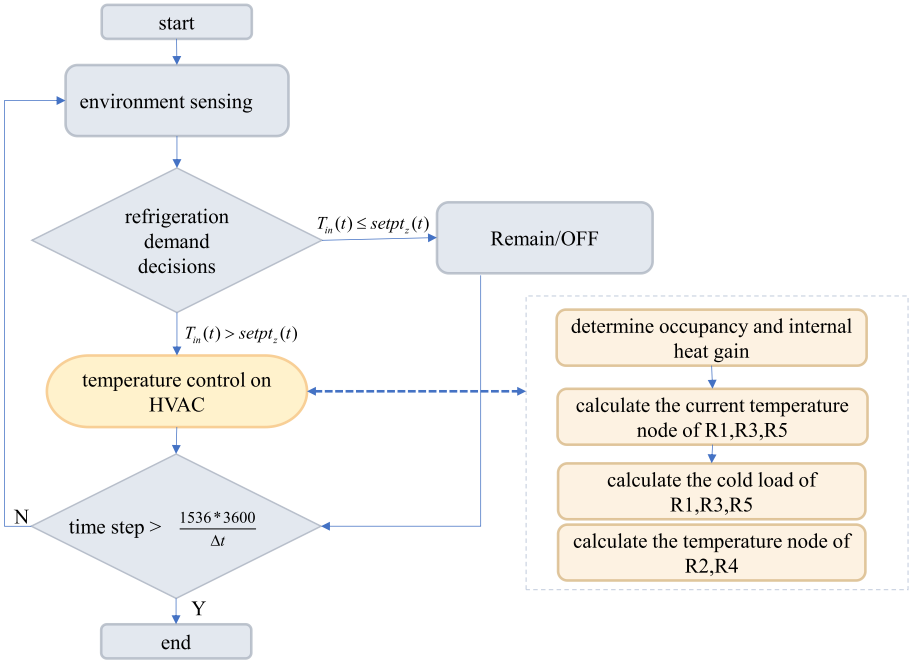



**Fig. A.1.** HVAC operation flow chart

## Appendix B

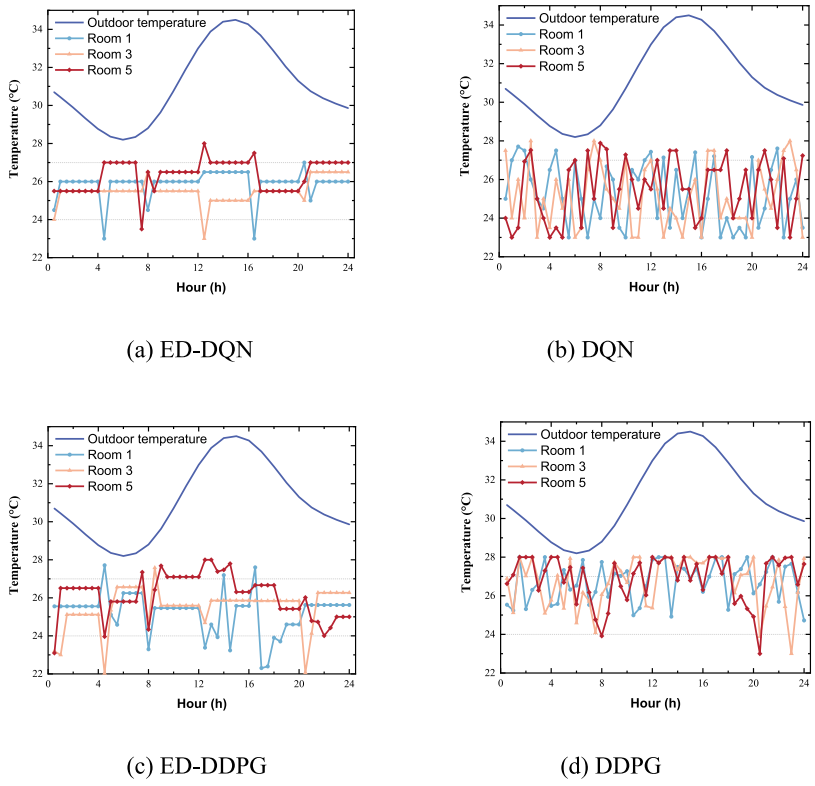



**Fig. B.1.** Indoor temperature on July 31 in Chongqing. (a) ED-DQN; (b) DQN; (c) ED-DDPG; (d) DDPG.

**Table B.1**
A comparison of the number of decisions from May to September in Chongqing.

| Room | DQN ($D_1$) | ED-DQN ($D_2$) | Reduction of decision rate ($\varphi$) |
|---|---|---|---|
| 1 | 6537 | 1679 | 74.315% |
| 3 | 6591 | 2131 | 67.668% |
| 5 | 6580 | 2167 | 67.067% |

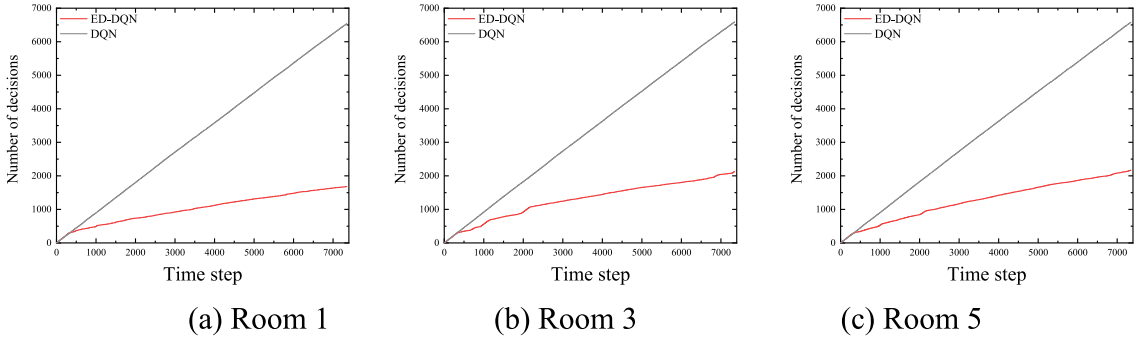



**Fig. B.2.** The number of decisions of three rooms in different methods from May to September in Chongqing.

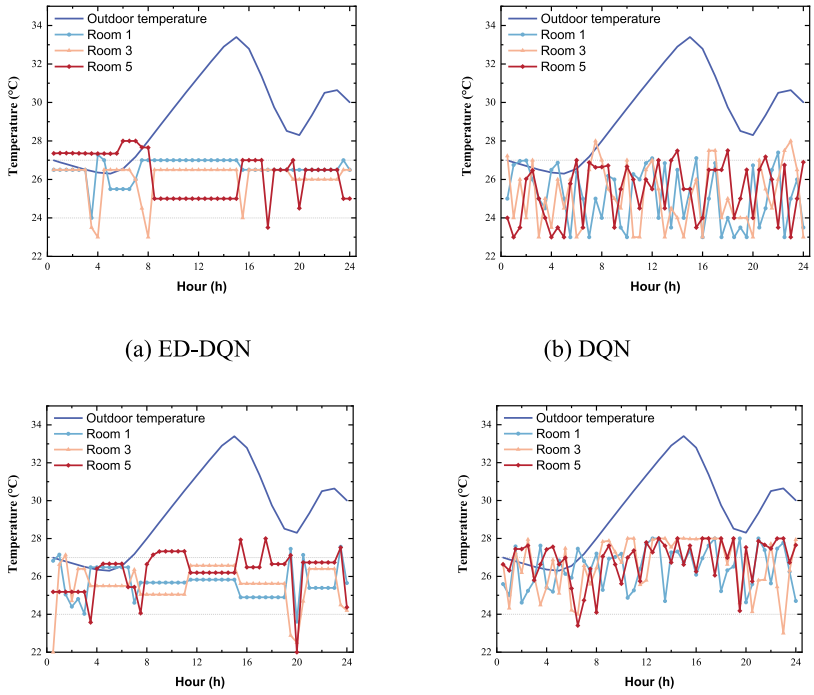



**Fig. B.3.** Indoor temperature on July 31 in Shanghai. (a) ED-DQN; (b) DQN; (c) ED-DDPG; (d) DDPG.

**Table B.2**
A comparison of the number of decisions from May to September in Shanghai.

| Room | DQN ($D_1$) | ED-DQN ($D_2$) | Reduction of decision rate ($\varphi$) |
|---|---|---|---|
| 1 | 6537 | 1723 | 73.642% |
| 3 | 6591 | 2511 | 61.906% |
| 5 | 6580 | 2471 | 62.447% |

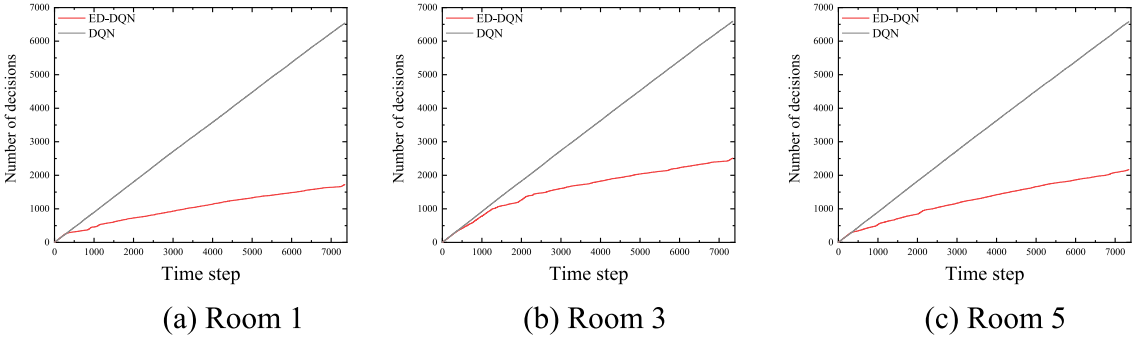



**Fig. B.4.** The number of decisions of three rooms in different methods from May to September in Shanghai.

## References

[1] I. Hamilton, O. Rapf, D.J. Kockat, D.S. Zuhaib, T. Abergel, M. Oppermann, N. Steurer, Global Status Report for Buildings and construction[J], United Nations Environmental Programme, Nairobi, Kenya, 2020.
[2] W. Li, G. Gong, H. Fan, P. Peng, L. Chun, X. Fang, A clustering-based approach for "cross-scale" load prediction on building level in HVAC systems[J], Appl. Energy 282 (2021), 116223.
[3] H. Qi, S. Xiao, R. Shi, M.P. Ward, Y. Chen, W. Tu, Z. Zhang, COVID-19 transmission in Mainland China is associated with temperature and humidity: a time-series analysis[J], Sci. Total Environ. 728 (2020), 138778.
[4] B. Delač, B. Pavković, K. Lenić, D. Maderić, Integrated optimization of the building envelope and the HVAC system in nZEB refurbishment[J], Appl. Therm. Eng. 211 (2022), 118442.
[5] A. Ambroziak, A. Chojecki, The PID controller optimisation module using Fuzzy Self-Tuning PSO for Air Handling Unit in continuous operation[J], Eng. Appl. Artif. Intell. 117 (2023), 105485.

[6] Q. Fu, X. Chen, S. Ma, S. Ma, N. Fang, B. Xing, J. Chen, Optimal control method of HVAC based on multi-agent deep reinforcement learning[J], Energy Build. 270 (2022), 112284.
[7] Q. Fu, Z. Han, J. Chen, Y. Lu, H. Wu, Y. Wang, Applications of reinforcement learning for building energy efficiency control: a review[J], J. Build. Eng. 50 (2022), 104165.
[8] A.P. Wemhoff, Calibration of HVAC equipment PID coefficients for energy conservation[J], Energy Build. 45 (2012) 60–66.
[9] S. Baldi, C.D. Korkas, M. Lv, E.B. Kosmatopoulos, Automating occupant-building interaction via smart zoning of thermostatic loads: a switched self-tuning approach [J], Appl. Energy 231 (2018) 1246–1258.
[10] H. Wang, S. Wang, A hierarchical optimal control strategy for continuous demand response of building HVAC systems to provide frequency regulation service to smart power grids, J. Energy 230 (2021), 120741.
[11] M. Bird, C. Daveau, E. O'Dwyer, S. Acha, N. Shah, Real-world implementation and cost of a cloud-based MPC retrofit for HVAC control systems in commercial buildings[J], Energy Build. 270 (2022), 112269.
[12] T. Zeng, P. Barooah, An Adaptive MPC Scheme for Energy-Efficient Control of Building HVAC Systems, 2021, https://doi.org/10.1115/1.4051482.
[13] M.C. Mozer, F.N.N. House, An Environment that Adapts to its Inhabitants[C]// AAAI Spring Symposium on Intelligent Environments, 1988, pp. 10–11, l.
[14] Y. Chen, L.K. Norford, H.W. Samuelson, A. Malkawi, Optimal control of HVAC and window systems for natural ventilation through reinforcement learning[J], Energy Build. 169 (2018) 195–205.
[15] W. Valladares, M. Galindo, J. Gutiérrez, W.C. Wu, K.K. Liao, J.C. Liao, C.C. Wang, Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm[J], Building and Environment 155 (2019) 105–117.
[16] G. Gao, J. Li, Y. Wen, Deepcomfort: energy-efficient thermal comfort control in buildings via reinforcement learning[J], IEEE Internet Things J. 7 (9) (2020) 8472–8484.
[17] X. Fang, G. Gong, G. Li, L. Chun, P. Peng, W. Li, X. Chen, Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building HVAC system[J], Appl. Therm. Eng. 212 (2022), 118552.
[18] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, F. Li, Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning[J], Appl. Energy 281 (2021), 116117.
[19] J. Wang, J. Hou, J. Chen, Q. Fu, G. Huang, Data mining approach for improving the optimal control of HVAC systems: an event-driven strategy[J], J. Build. Eng. 39 (2021), 102246.
[20] J. Wang, T. Zhao, Event-driven online decoupling control mechanism for the variable flow rate HVAC system based on the medium response properties[J], Build. Environ. 218 (2022), 109104.
[21] Q.S. Jia, J. Wu, Z. Wu, X. Guan, Event-based HVAC control—a complexity-based approach[J], IEEE Trans. Autom. Sci. Eng. 15 (4) (2018) 1909–1919.
[22] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, X. Guan, A review of deep reinforcement learning for smart building energy management[J], IEEE Internet Things J. 8 (15) (2021) 12046–12063.
[23] R.S. Sutton, A.G. Barto, Reinforcement Learning: an Introduction, MIT Press[J], Cambridge, MA, 1998, 22447.
[24] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, D. Hassabis, Human-level control through deep reinforcement learning[J], Nature 518 (7540) (2015) 529–533.
[25] W. Jin, Q. Fu, J. Chen, Y. Wang, L. Liu, Y. Lu, H. Wu, A novel building energy consumption prediction method using deep reinforcement learning with consideration of fluctuation points[J], J. Build. Eng. 63 (2023), 105458.
[26] B. Gu, Y. Sung, Enhanced DQN framework for selecting actions and updating replay memory considering massive non-executable actions[J], Appl. Sci. 11 (23) (2021), 11162.
[27] L. Luo, N. Zhao, Y. Zhu, Y. Sun, A* guiding DQN algorithm for automated guided vehicle pathfinding problem of robotic mobile fulfillment systems[J], Comput. Ind. Eng. (2023), 109112.
[28] A. Iqbal, M.L. Tham, Y.C. Chang, Double deep Q-network-based energy-efficient resource allocation in cloud radio access network[J], IEEE Access 9 (2021) 20440–20449.
[29] J. Deng, R. Yao, W. Yu, Q. Zhang, B. Li, Effectiveness of the thermal mass of external walls on residential buildings for part-time part-space heating and cooling using the state-space method[J], Energy Build. 190 (2019) 155–171.
[30] China Meteorological Bureau, Tsinghua University, China Standard Weather Data for Analyzing Building Thermal conditions[S], China Architecture and Building Press, Beijing, 2005.
[31] Y. Jiang, State-space method for the calculation of air-conditioning loads and the simulation of thermal behavior of the room[J], Build. Eng. 88 (1982) 122–141.
[32] D. Yan, J. Xia, W. Tang, F. Song, X. Zhang, Y. Jiang, DeST — an integrated building simulation toolkit Part I: fundamentals[J], Build. Simulat. 1 (2) (2008) 95–110.
[33] J. Wang, G. Huang, Y. Sun, X. Liu, Event-driven optimization of complex HVAC systems[J], Energy Build. 133 (DEC) (2016) 79–87.
[34] Z. Xu, G. Hu, C.J. Spanos, S. Schiavon, PMV-based event-triggered mechanism for building energy management under uncertainties[J], Energy Build. 152 (2017) 73–85.
[35] Z. Wu, Q.S. Jia, X. Guan, Optimal Control of Multiroom HVAC System: an Event-Based Approach[J], IEEE Transactions on Control Systems Technology, 2015.