Exploring Deep Reinforcement Learning for Holistic Smart Building Control

XIANZHONG DING, University of California Merced School of Engineering, Merced, United States
ALBERTO CERPA, University of California, Merced, Merced, United States
WAN DU, University of California Merced School of Engineering, Merced, United States

In recent years, the focus has been on enhancing user comfort in commercial buildings while cutting energy costs. Efforts have mainly centered on improving HVAC systems, the central control system. However, it's evident that HVAC alone can't ensure occupant comfort. Lighting, blinds, and windows, often overlooked, also impact energy use and comfort. This paper introduces a holistic approach to managing the delicate balance between energy efficiency and occupant comfort in commercial buildings. We present *OCTOPUS*, a system employing a deep reinforcement learning (DRL) framework using data-driven techniques to optimize control sequences for all building subsystems, including HVAC, lighting, blinds, and windows. *OCTOPUS*'s DRL architecture features a unique reward function facilitating the exploration of tradeoffs between energy usage and user comfort, effectively addressing the high-dimensional control problem resulting from interactions among these four building subsystems. To meet data training requirements, we emphasize the importance of calibrated simulations that closely replicate target-building operational conditions. We train *OCTOPUS* using 10-year weather data and a calibrated building model in the EnergyPlus simulator. Extensive simulations demonstrate that *OCTOPUS* achieves substantial energy savings, outperforming state-of-the-art rule-based and DRL-based methods by 14.26% and 8.1%, respectively, in a LEED Gold Certified building while maintaining desired human comfort levels.

CCS Concepts: • **Computing methodologies** → **Planning and scheduling**; **Reinforcement learning**; **Control methods;**

Additional Key Words and Phrases: HVAC, energy efficiency, optimal control, deep reinforcement learning

ACM Reference Format:
Xianzhong Ding, Alberto Cerpa, and Wan Du. 2024. Exploring Deep Reinforcement Learning for Holistic Smart Building Control. *ACM Trans. Sensor Netw.* 20, 3, Article 70 (May 2024), 28 pages. https://doi.org/10.1145/3656043

1 INTRODUCTION

Energy saving in buildings is crucial for society due to the significant energy consumption of buildings, accounting for 32% of global energy consumption and 51% of electricity demand

This work was supported in part by NSF Grant #2239458 and UC National Laboratory Fees Research Program Grant #69763. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.
Authors' addresses: X. Ding and W. Du (Corresponding author) University of California Merced School of Engineering, Merced, CA, United States, 95340; e-mails: xding5@ucmerced.edu, wdu3@ucmerced.edu; A. Cerpa, Electrical Engineering and Computer Science, University of California, Merced, CA, United States, 95340; e-mail: acerpa@ucmerced.edu.

ACM 1550-4859/2024/05-ART70
https://doi.org/10.1145/3656043

worldwide [42, 48]. Advanced control strategies for operating HVAC systems have the potential to improve energy efficiency and human comfort. Especially, in recent years, with the development of the **Internet of Things (IoT)** [61, 62] and smart buildings, massive building sensor data are available for intelligent management [43] and behavior analysis [10]. **Rule-based control (RBC)** is a commonly used method to set actuators in **heating, ventilation, and air-conditioning (HVAC)** systems, such as temperature and fan speed. However, RBC typically relies on static thresholds or simple control loops based on the experience of engineers and facility managers, which may not be optimal and often require adjustments for new buildings during commissioning. These rules are often updated in an ad-hoc manner, based on feedback from occupants or trial and error by HVAC engineers during building operation. Consequently, model-based approaches have emerged to model the thermal dynamics of buildings and implement advanced control algorithms, such as **Proportional Integral Derivative (PID)** [50] and **Model Predictive Control (MPC)** [11], to optimize energy consumption and improve building performance. However, accurately modeling the complex thermal dynamics and accounting for the multitude of influencing factors in buildings can be challenging, leading to simplified models that can handle the data requirements for parameter fitting and computational complexity when solving optimization problems [11]. The intricacies of factors such as weather conditions, occupancy patterns, and building envelope characteristics make it difficult to create precise models. As a result, simplified models are often used, which may not fully capture the complexity of real-world building systems.

To overcome the limitations of model-based methods, alternative model-free approaches based on **reinforcement learning (RL)** have been proposed for HVAC control, including Q-learning [39] and **Deep Reinforcement Learning (DRL)** [63]. RL allows for the learning of an optimal control policy through trial-and-error interactions between a control agent and a building, without explicitly modeling the system dynamics. DRL-based approaches leverage deep neural networks as control agents, enabling them to handle large state and action spaces in building control [63]. Recent works [58, 63] have demonstrated that DRL can achieve real-time control for improving building energy efficiency. However, existing methods have typically focused on single subsystems within buildings, such as the HVAC system [58] or the heating system [63], while disregarding other subsystems that can impact energy consumption and user comfort from a holistic perspective.

In recent times, there has been a growing trend towards equipping buildings with automatically-adjustable windows and blinds. These cutting-edge solutions, such as the intelligent products offered by GEZE [2], incorporate motor-operated windows and blinds that can be controlled to optimize natural ventilation strategies. By automatically adjusting the windows and blinds, buildings can effectively regulate the inflow of fresh air and natural light, reducing the reliance on mechanical systems for HVAC. Furthermore, researchers [18] have been actively investigating the potential of joint control strategies that integrate the HVAC system with other building subsystems, such as blinds [52], lighting [14], and windows [57]. For instance, studies have shown that enabling window-based natural ventilation can result in significant energy savings for HVAC systems. According to research findings, the energy consumed by HVAC can be reduced by 17% to 47% by leveraging window-based natural ventilation strategies [57]. This underscores the importance of considering the synergies between different building subsystems and adopting integrated control approaches to achieve optimal energy performance.

In our study, we propose a comprehensive approach that takes into account all available subsystems within buildings, including HVAC, blinds, windows, and lighting, to achieve specific building energy efficiency and human comfort goals. Modern buildings are complex systems comprising multiple interconnected subsystems that work in tandem to ensure human comfort,

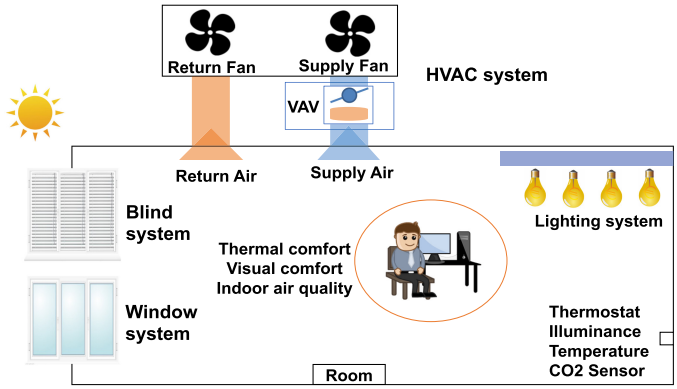

Fig. 1. Four subsystems in a typical building.

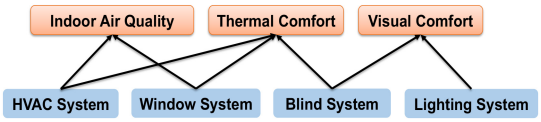

Fig. 2. Relationship between four subsystems and three human comfort metrics.

encompassing thermal comfort, visual comfort, and indoor air quality. As illustrated in Figure 1, various subsystems (HVAC, Blind, Window, and Lighting system) can influence indoor temperature. For instance, the HVAC system plays a crucial role in regulating indoor temperature by adjusting discharge temperature set points at the **Variable Air Volume (VAV)** level. Manipulating blind slats provides control over the entry of external sunlight, which can impact indoor air temperature. Additionally, regulating the window system enables the exchange of indoor and outdoor air, further influencing indoor temperature. These interconnected subsystems need to work harmoniously to achieve optimal energy efficiency and comfort.

To achieve more efficient energy management in buildings, our proposed study focuses on the joint control problem of four subsystems, as depicted in Figure 2. These subsystems play a crucial role in meeting three human comfort metrics. The energy consumption of a building is influenced by the interactions among these subsystems, making it challenging to control them jointly. This is because they may have opposite effects on different human comfort metrics. For instance, opening a window can improve indoor air quality and save energy consumed by the HVAC system for ventilation. However, it may also result in a reduction of indoor temperature in winter or an increase in summer. To mitigate the temperature variation caused by an open window, the HVAC system may need to expend more energy, which could offset the energy savings from natural ventilation.

This paper introduces *OCTOPUS*, a customized DRL-based control system designed to optimize energy management in buildings by controlling four subsystems to meet three human comfort requirements with maximum energy efficiency. *OCTOPUS* leverages the advantages of DRL-based control, including adaptability to new buildings, real-time actuation, and handling of large state spaces. However, there are three main challenges in controlling the four subsystems jointly within a unified framework.

The first challenge is the high-dimensionality of control actions. With a uniform DRL framework, *OCTOPUS* needs to decide on control actions for four subsystems simultaneously and periodically, including HVAC temperature, electric lights brightness, blind slat range, and window

opening proportion. Each subsystem adds one dimension to the action space, resulting in an extremely large set of possible action combinations. To address this, a novel neural architecture is proposed, featuring a shared representation followed by four network branches, one for each action dimension. Additionally, a state value obtained from the shared representation is added to the output of the four branches, allowing for independence in each action dimension while maintaining joint interrelations in the action space.

The second challenge is defining a reward function that captures the trade-offs between energy consumption and human comfort requirements. To tackle this, the optimization problem is formulated as a reward function in the DRL framework. The proposed reward function jointly considers energy consumption, thermal comfort, visual comfort, and indoor air quality, offering better control and flexibility to meet unique user requirements.

The third challenge is the requirement for a large amount of training data. Model-free approaches, including RL techniques, require substantial training data, which may not be readily available from building stakeholders. To overcome this, a calibrated building simulator combined with weather data is used to generate the necessary training data. *OCTOPUS* is trained with 10 years of weather data from distinct locations, Merced, CA and Chicago, IL, to ensure adaptability to different building types and weather profiles. The main contributions of this paper are highlighted as follows:

— This work is the first to utilize DRL in a holistic manner to balance the tradeoff between energy use across four subsystems and three human comforts, as far as our knowledge goes.
— *OCTOPUS* overcomes the challenges posed by the combined joint control of four subsystems with a large action space by employing a unique reward function and a new DRL architecture.
— To address the issue of data training requirements, we adopt a simulation strategy for data generation and put effort into calibrating the simulations to closely match the target building. This allows our system to generate sufficient data within a finite amount of time.
— Extensive experiments show the effectiveness of *OCTOPUS*.

2 MOTIVATION

In this section, we perform a set of preliminary simulations in EnergyPlus [45] to understand the relationships between the different subsystems and their impact on human comfort in a building, as described in Figure 2. This also helps us gain trust that the simulator is functioning correctly, providing intuitive and understandable results. Our goal is to study the effect of different subsystems on three human comfort metrics. We model a single-floor office building of 100 m^2 in Merced, California, equipped with a north-facing single-panel window of 2 m^2 and an interior blind. The simulations are conducted with weather data for the month of October, which represents a shoulder season with slightly cold outdoor temperatures but mostly sunny days, resulting in high solar gain.

Effects of HVAC, Blind, and Window Systems on Thermal Comfort. Figure 3 illustrates the impact of three subsystems on thermal comfort, evaluated using the **Predictive Mean Vote (PMV)** index. A PMV value close to zero indicates optimal thermal comfort, with higher positive values indicating higher temperatures (hot) and lower negative values indicating lower temperatures (cold). A detailed explanation of PMV values and ranges can be found in Section 3.4.2. The baseline case (green-solid) represents when all three subsystems are closed, resembling a "fishtank" model where the only impact on the room is from solar gain during the day through the window. When only the blind is open (blue-dashed), the PMV value increases from 1.45 to 1.75, indicating a rise in temperature due to increased solar gain. This effect is most pronounced during the middle

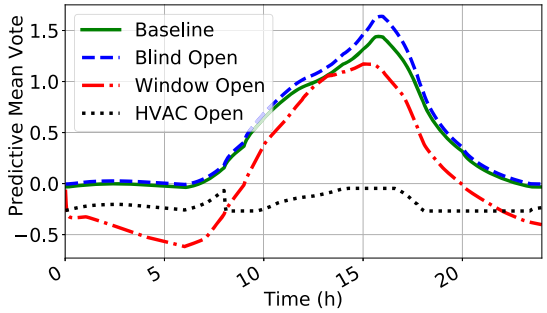

Fig. 3. Thermal comfort, PMV.

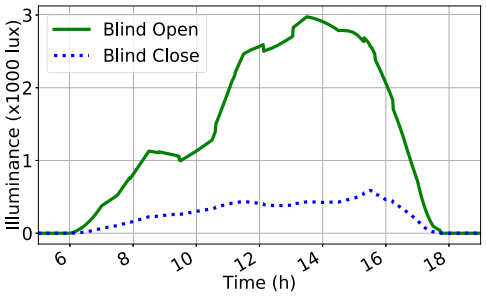

Fig. 4. Visual comfort, Illuminance.

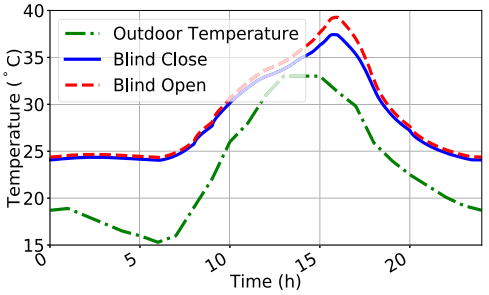

Fig. 5. Temperature effect.

of the day when the sun is at its peak. On the other hand, when the window is open (red-dashed-dot), the PMV value decreases due to the temperature effect, as colder outside air enters the room, resulting in a cooler and more comfortable temperature. The HVAC system (black-dot) maintains the PMV value within an acceptable range (-0.5 to +0.5) by regulating the temperature of the air forced through the room vents. Based on the results in Figure 3, it can be concluded that all three subsystems have a significant impact on thermal comfort.

Effects of Blind system on Visual Comfort and Temperature. Figure 4 depicts illuminance measurements near the window area from 5 am to 7 pm with the blind open and natural light present. Illuminance values of 500-1000 lux or higher are acceptable in most environments. The graph reveals that, for the majority of the day, illuminance values remain within this range when the blind is open, emphasizing the positive impact of open blinds in maintaining adequate natural light levels in the room. In Figure 5, the indoor temperature is compared with the blind open (red-dashed) versus closed (blue-solid), while the outdoor temperature (green-dash-dot) remains lower due to the "fish tank" effect and lack of window ventilation or HVAC system operation during the day. Combining the findings from Figures 4 and 5, it is evident that the blind system can reduce the energy consumed by the lighting system by utilizing natural light, but it may also increase the load on the HVAC system for maintaining indoor temperature. However, during colder outdoor temperatures in winter, sunlight passing through the blind can raise the indoor temperature and save energy used by the HVAC system.

The simulations are carried out to demonstrate the intricate interactions between subsystems and their impact on human comfort. Quantifying the complex relationships among different subsystems and the three human comfort metrics poses a challenge, providing motivation for our work.

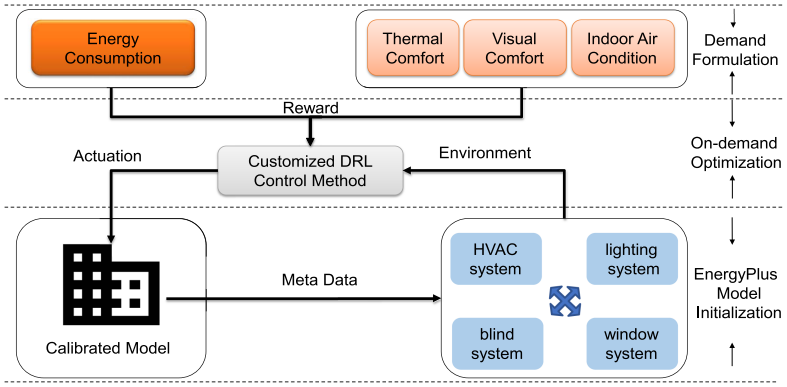

Fig. 6. *OCTOPUS* architecture with four subsystems (including HVAC, lighting, blind and window systems).

3 DESIGN OF OCTOPUS

This section provides a comprehensive description of *OCTOPUS*'s design, encompassing a system overview, DRL-based building control, branching dueling Q-Network, and reward function calculation.

3.1 *OCTOPUS* Overview

The primary objective of *OCTOPUS* is to achieve energy-efficient control of four subsystems in a building while ensuring human comfort. To accomplish this, our goal is to minimize the overall energy consumption E of the building, which includes energy used by heating/cooling coils, water pumps, flow fans in the HVAC system, lights, and motors for blinds and windows adjustment. The energy value E is influenced by the action combination A_s for the four subsystems, which belongs to the set of all possible action combinations A_{all}.

In addition to energy minimization, we aim to maintain human comfort within specific ranges. This is expressed through the following criteria: $P_{min} \leq PMV \leq P_{max}, V_{min} \leq V \leq V_{max}$, and $I_{min} \leq I \leq I_{max}$. The parameter PMV quantifies thermal comfort, V represents visual comfort, and I measures indoor air quality. The consumed energy E and the human comfort metrics (PMV, V, and I) are determined by the current state of all four subsystems, the outdoor weather and the action we are about to take. They can be measured in real buildings or calculated in a building simulator, like EnergyPlus, after the action is executed.

The satisfaction of users' requirements for human comfort necessitates that the achieved results fall within an acceptable range. To represent these ranges, we utilize $[P_{min}, P_{max}]$ for thermal comfort, $[V_{min}, V_{max}]$ for visual comfort, and $[I_{min}, I_{max}]$ for indoor air quality. These ranges can be customized by individual users based on their preferences or established by facility managers in adherence to building standards. Section 3.4 will provide detailed information on calculating the aforementioned parameters (E, PMV, V and I), defining actions A_s, and setting the ranges for human comfort (e.g., $[P_{min}, P_{max}]$).

Our objective is to determine the most optimal action A_s from the set of all possible actions A_{all} for each action interval (15 minutes in our implementation). The chosen actions should sustain the three human comfort metrics within their acceptable ranges throughout the entire control interval while minimizing energy consumption E. To achieve this objective, we have developed a building control system based on DRL, known as *OCTOPUS*. Figure 6 provides an overview of *OCTOPUS*, which comprises three layers: the building layer, control layer, and user demand layer. The building layer encompasses either the physical building or a building simulation model,

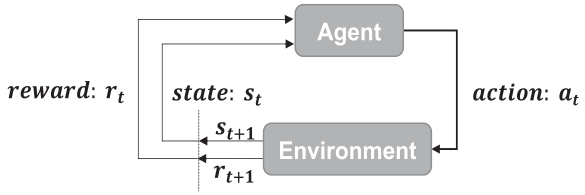

Fig. 7. Reinforcement learning framework.

Table 1. The States in *OCTOPUS*

Name	unit	Name	unit
Outdoor air temperature	°C	Outdoor air relative humidity	kg Water/kg Dry Air
Indoor air temperature	°C	Indoor air relative humidity	kg Water/kg Dry Air
Diffuse solar radiation	W/m^2	Direct solar radiation	W/m^2
Solar incident angle	°	Heating setpoint of the HVAC system	°C
Wind direction	degree from north	Average Predicted Mean Vote (PMV)	N/A
Wind speed	m/s	Cooling setpoint of the HVAC system	°C
Dimming level of lights	%	Window open percentage	%
Blind open angle	°	/	/

along with components for managing sensor data. It supplies sensor data to the control layer and executes the control actions generated by the control layer. The user demand layer quantifies the user's requirements for the three human comfort metrics, and the respective range for each human comfort metric is transmitted to the control layer. The control layer then searches for the optimal control strategy that satisfies the human comfort ranges with the least energy consumption.

3.2 DRL-based Building Control

3.2.1 Basics for DRL and DQN. In a standard RL framework, as shown in Figure 7, an agent learns an optimal control policy by trying different control actions to the environment. In our particular scenario, we employ a building simulation model as the environment due to the extensive data requirements for training the system. Within the framework of DRL, the agent is implemented as a **deep neural network (DNN)**. The interaction between the agent and the environment can be expressed as a tuple $(S_t, A_t, S_{t+1}, R_{t+1})$ for each time step. Here, S_t represents the state of the environment at time t, A_t denotes the control action taken by the agent at time t, S_{t+1} represents the resulting state of the environment after the agent's action, and R_{t+1} denotes the reward received by the agent from the environment. The objective of training the DNN agent is to learn an optimal control policy that maximizes the accumulated reward obtained through different control actions.

3.2.2 State in OCTOPUS. The state serves as the input for the DRL agent during each control step. In this study, the state is constructed as a stack of current and past observations, represented as:

$$S = \{ob_t, ob_{t-1}, \ldots, ob_{t-n}\}, \tag{1}$$

Here, t denotes the current time step, n represents the number of historical time steps considered, and each ob encompasses 15 specific elements shown in Table 1. All these values can be calculated using the EnergyPlus simulation model. To ensure consistency, min-max normalization is applied to each item, converting its value to a range between 0 and 1.

3.2.3 Action in OCTOPUS. The action represents how the DRL agent exerts control over the environment. Given the state, the agent aims to identify the most suitable combinations of actions

for the HVAC, lighting, blind, and window systems to achieve a balance between energy consumption and the three human comfort metrics. When considering these four subsystems, the action can be described as:

$$A_t = \{H_t, L_t, B_t, W_t\}, \tag{2}$$

Here, A_t corresponds to the action combination of the four subsystems at time t. H_t represents the temperature set-point of the HVAC system, which can be selected from 66 different values. L_t denotes the dimming level of the electric lights, while B_t signifies the blind slat angle. The range of blind slat angles can be adjusted from 0° to 180°. Finally, W_t denotes the window's open percentage. In our current implementation, each of these three actuation parameters can be set to 33 values, striking a balance between control granularity and computational complexity.

According to Equation (2), the total number of possible actions in the action space amounts to 2,371,842 ($66 \times 33 \times 33 \times 33$). Existing DRL architectures like **Deep Q-Network (DQN)** [58] and **Asynchronous Advantage Actor-Critic (A3C)** [63] are not efficient for our problem since the large number of actions necessitates explicit representation within the agent's DNN network. This would significantly increase the number of DNN parameters to be learned, consequently extending the training time [53]. To address this challenge, we employ a novel neural architecture that incorporates a shared representation followed by four network branches, with each branch dedicated to one action dimension.

3.2.4 Reward Function in OCTOPUS. The reward function provides an immediate evaluation of the control effects for each action taken under a specific state. It takes into account both human comfort and energy consumption. To define the reward function, a common approach is to utilize the Lagrangian Multiplier function [30], which converts the constrained formulation into an unconstrained one. The reward function is expressed as:

$$R = -[\rho_1 Norm(E) + \rho_2 Norm(T_c) + \rho_3 Norm(V_c) + \rho_4 Norm(I_c)], \tag{3}$$

In this equation, ρ_1, ρ_2, ρ_3 and ρ_4 represent the Lagrangian multipliers. E represents energy consumption, Tc represents thermal comfort, Vc represents visual comfort, and Ic represents indoor air quality. $Norm(x)$ denotes the normalization process, defined as $Norm(x) = (x - x_{min})/(x_{max} - x_{min})$, which transforms energy and the three human comfort metrics onto the same scale. The reward function combines the objective of minimizing energy consumption with the goal of satisfying the constraints related to human comfort.

The reward function comprises four components. Firstly, there is a penalty for the energy consumption of the HVAC and lighting system. Secondly, there is a penalty associated with the occupants' thermal discomfort. Thirdly, there is a penalty linked to the occupants' visual discomfort. Finally, there is a penalty connected to the occupants' indoor air condition discomfort. Specifically, the reward decreases when the HVAC system consumes more energy or when the occupants experience discomfort concerning the building's thermal, visual, and indoor air conditions. Further details on how to define and formulate energy consumption E, thermal comfort Tc, visual comfort Vc, and indoor air condition Ic are explained in Section 3.4. To address constraints, we may consider the integration of safe deep reinforcement learning [15], which more effectively manages these diverse aspects. This approach, focusing explicitly on thermal, visual, and air quality constraints, is earmarked for exploration in our future work.

3.3 Branching Dueling Q-Network

To address the challenge of high-dimensional actions discussed in Section 3.2.3, *OCTOPUS* employs a novel neural architecture called **Branching Dueling Q-Network (BDQ)**. BDQ is a branching

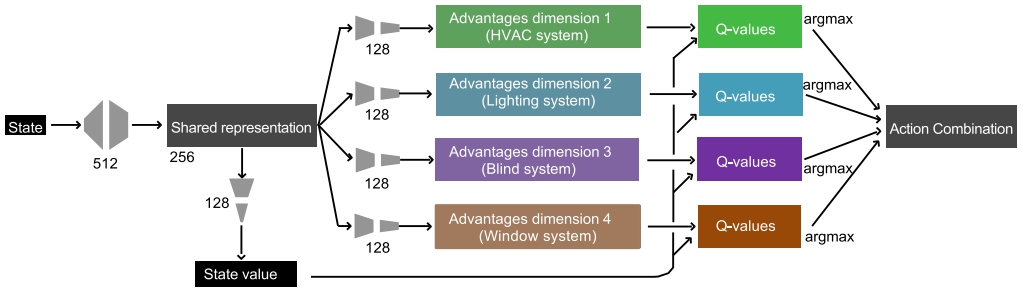

Fig. 8. The specific action branching network implemented for the proposed BDQ agent.

variation of the dueling **Double Deep Q-Network (DDQN)**. It features a shared decision module followed by multiple network branches, each dedicated to a specific action dimension. This architecture enables robust scalability to environments with high-dimensional action spaces and has shown superior performance compared to the **Deep Deterministic Policy Gradient (DDPG)** algorithm, even in the most demanding tasks [51].

In our implementation, we utilize a simulated building model developed in EnergyPlus as the training and validation environment. The BDQ-based agent interacts with the EnergyPlus model, processing the state, which includes building and weather parameters, at each control step. Subsequently, it generates a combined action set for the four subsystems involved in the control process. This approach allows us to effectively handle the complexities of the high-dimensional action problem and achieve optimal control performance in the simulated building environment.

Figure 8 illustrates the action branching network employed by the BDQ agent. Upon receiving a state input, the shared decision module calculates a latent representation, which is utilized for both the state value computation and the output of the network's advantages dimension (as shown in Figure 8) for each branch corresponding to an action dimension. The state value and factorized advantages are then combined through a specialized aggregation layer to produce the Q-values for each action dimension. These Q-values are subsequently utilized to generate a joint-action tuple. The fully connected neural layers' weights are represented by the gray trapezoids, and the figure also depicts the size of each layer (i.e., the number of units).

Training Process: The training process for the BDQ-based control agent is described in Algorithm 1. To start, we initialize a neural network Q with random weights θ, and create another neural network Q^- with the same architecture. The outer "for" loop determines the number of training episodes, while the inner "for" loop handles control at each control time step within an episode. Throughout the training, the most recent transition tuples $(S_t, A_t, S_{t+1}, R_{t+1})$ are stored in the replay memory Λ which is used to generate a mini-batch of samples for training the neural network. The variable A_t stores the control action from the previous step, and S_t and S_{t+1} represent the building state in the previous and current control time steps, respectively. At the beginning of each time slot t, the four actions are updated, and the current state S_{t+1} is obtained. In line 7, the immediate reward R_{t+1} is calculated using Equation 3. Training mini-batches are constructed by randomly selecting transition tuples from the memory.

We compute the target vector and update the weights of the neural network Q using an Adam optimizer at each control step t. For an action dimension $d \in 1, \ldots N$ with n discrete actions, the Q-value of a branch for a state $s \in S$ and an action $a_d \in A_d$ is determined by the common state value $V(s)$ (the output of the shared representation layer in Figure 8) and the corresponding action advantage $A_d(s, a_d)$ of each branch (the output of each advantage dimension in Figure 8). This

ALGORITHM 1: The Training Process of Our BDQ-Based Agent

Input: The range of human comfort metrics and maximum acceptable energy consumption
Output: A trained DRL agent

```
Initialize BDQ's prediction Q with random weights θ;
Initialize BDQ's target Q⁻ with weight θ⁻ = θ ;
for episode =0, 1, . . . , M do
    Obtain the initial state S_t and A_t randomly;
    for control time step t = 0,1, . . . , T do
        Update H_t, L_t, B_t, W_t by the control action, A_t;
        Calculate reward R_{t+1} by Equation (3);
        Obtain current state observation S_{t+1};
        Store (S_t, A_t, S_{t+1}, R_{t+1}) in reply memory Λ;
        Draw mini-batch sample transitions from Λ;
        Calculate the target vector and update weights in neural network Q ;
        Update target network Q_d^-(s, a_d) using Equation (5);
        Perform greedy descent iteratively to tune BDQ by Equation (6).
```

relationship is expressed as:

$$Q_d(s, a_d) = V(s) + \left(A_d(s, a_d) - \frac{1}{n} \sum_{a_d' \in A_d} A_d\left(s, a_d'\right)\right). \tag{4}$$

The target network Q^- is updated with the latest weights of the network Q every c control time steps, where c is set to 50 in our current implementation. Q^- is used to calculate the target value for the next c control steps. We denote y_d as the maximum accumulated reward achievable in the next c steps. It is computed recursively using **temporal-difference (TD)** targets:

$$y_d = R + \gamma \frac{1}{N} \sum_d Q_d^- \left(s', \arg \max_{a_d' \subseteq A_d} Q_d\left(s', a_d'\right)\right), \tag{5}$$

Here, Q_d^- represents the branch d of the target network Q^-, R is the reward function result, and γ is the discount factor. At the end of the inner "for" loop, we compute the following loss function every c control steps:

$$L = \mathbb{E}_{(s,a,r,s')} \sim D\left[\sum_d (y_d - Q_d(s, a_d))^2\right], \tag{6}$$

Here, D denotes a (prioritized) experience replay buffer, and a represents the joint-action tuple $(a_1, a_2, \ldots, a_N)$. The loss function L should decrease as more training episodes are conducted.

3.4 Reward Calculation

In this section, we provide an overview of how we calculate the reward function in Equation (3), which encompasses energy cost E, thermal comfort T, visual comfort V, and indoor air condition I.

3.4.1 Energy Consumption. The energy consumption of a building consists of the heating coil power P_h, cooling coil power P_c, fan power P_f from the HVAC system, and electric light power P_l from the lighting system. The reward function for energy consumption E during a time slot is calculated as:

$$E = (P_h + P_c + P_f + P_l) \tag{7}$$

The heating and cooling coils are responsible for regulating the air temperature in the building, while the fan is responsible for distributing the heated or cooled air to the different zones. The

Table 2. PMV Constants

Parameter	Value	Units
Metabolic Rate	70	W/m^2
Clothing Level	0.5	clo

electric lights are used for general illumination within the zones. These energy components are calculated using the EnergyPlus simulator during both training and evaluation. In our current implementation, we neglect the power consumed by water pumps and motors used for adjusting blinds and windows. This omission is justified by their relatively small contribution compared to the power consumption of the HVAC and lighting systems, and their impact can be safely ignored (accounting for less than 1% of the total energy consumption).

3.4.2 Human Comfort. we establish and explain the measurement of three key human comfort metrics (Thermal Comfort, Visual Comfort and Indoor Air Quality)

Thermal Comfort: In our study, the measurement of thermal comfort is determined by the **PMV (Predicted Mean Vote)** index, calculated using Fanger's equation [23]. The PMV index predicts the average thermal sensation vote on a standardized scale for a large group of individuals. It has been widely adopted and is recommended by organizations such as the **American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE)** [29] and the **International Organization for Standardization (ISO)** [22].

The PMV scale assigns numerical values to thermal sensations, ranging from -3 (cold) to +3 (hot), with 0 indicating a neutral thermal sensation. ASHRAE's thermal comfort index categorizes these sensations as coding values. To ensure optimal thermal comfort, the ISO 7730 standard recommends maintaining the PMV at level 0 with a tolerance of 0.5 [22]. We calculate the reward function for thermal comfort T_c during each time slot using the following formula:

$$T_c = \begin{cases} 0, & PMV \leq P \\ |PMV - P|, & PMV > |P| \end{cases} \tag{8}$$

Here, P represents the threshold value for the PMV, defining the acceptable range of thermal comfort. If the PMV value falls within the range of $[-P, P]$, no penalty is incurred. However, if the PMV value exceeds this range, a penalty is imposed to reflect occupants' dissatisfaction with the building's thermal conditions.

Thermal comfort is influenced by several factors, which can be classified into personal factors and environmental factors. Personal factors include metabolic rate and clothing level, while environmental factors encompass air temperature, mean radiant temperature, air speed, and humidity. The PMV considers these factors to provide an accurate assessment of thermal comfort. The values for the PMV personal and environmental factors are obtained in real-time from the EnergyPlus simulation model and are shown in Table 2. EnergyPlus uses a node model, representing spaces with a single temperature, ideal for simulations where intra-zone temperature variations are minimal and not critical.

Visual Comfort: Visual comfort in buildings is a critical aspect that considers the appropriate amount of lighting to avoid discomfort caused by insufficient or excessive light levels, including glare. In our study, we utilize the illuminance range as a major metric to assess visual comfort [46]. The illuminance encompasses both natural daylight and electrical lighting sources, making the blind system and lighting system significant subsystems that can influence visual comfort. To quantify visual comfort, we calculate the reward function V_c during each time slot using the following

formulation:

$$V_c = \begin{cases} M_L - F, & F < M_L \\ 0, & M_L \leq F \leq M_H \\ -M_H, & F > M_H \end{cases} \tag{9}$$

Here, F represents the illuminance value, and M_L and M_H denote the lower and upper thresholds, respectively, defining the acceptable range of illuminance. When the illuminance value falls within the range $[M_L, M_H]$, occupants experience visual comfort, and no penalty is incurred. However, if the illuminance value exceeds this range or falls below it, penalties are applied to reflect occupants' dissatisfaction with the building's illuminance condition.

Indoor Air Quality: Carbon dioxide (CO_2) concentration serves as an indicator of **indoor air quality (IAQ)** in buildings, representing the presence of pollutants introduced by the building occupants [21]. While there are other sources of pollution like NOx, **Total Volatile Organic Compounds (TVOC)**, and respirable particles, in this study, we focus on CO_2 concentration as a proxy for IAQ. Efficient ventilation plays a vital role in maintaining satisfactory IAQ within buildings [3]. In our context, ventilation primarily occurs through the HVAC system and window system. To assess indoor air quality, we calculate the reward function I_c during each time slot using the following formulation:

$$I_c = \begin{cases} A_L - C, & C < A_L \\ 0, & A_L \leq C \leq A_H \\ C - A_H, & C > A_H \end{cases} \tag{10}$$

Here, C represents the carbon dioxide concentration value, and A_L and A_H denote the lower and upper thresholds, respectively, defining the acceptable range for CO_2 concentration. When the CO_2 concentration falls within the range $[A_L, A_H]$, occupants experience comfort in terms of indoor air quality, and no penalty is applied. However, if the CO_2 concentration exceeds this range or falls below it, penalties are incurred to reflect occupants' dissatisfaction with the building's indoor air quality. It is important to note that while CO_2 concentration is utilized as a proxy for IAQ in this study, other pollutants and factors contributing to air quality can also be considered in future research to provide a more comprehensive assessment of IAQ.

4 IMPLEMENTATION OF OCTOPUS

This section provides a comprehensive overview of the implementation of *OCTOPUS*, covering various aspects such as platform setup, HVAC modeling and calibration, and *OCTOPUS* training. Each component is explained in detail below.

4.1 Platform Setup

The conceptual flow diagram of our building simulation and control platform is depicted in Figure 9. The building model, representing a LEED Gold Certified Building on our university campus, is created using SketchUp [1]. OpenStudio is utilized to incorporate the HVAC, lighting, blind, and window systems into the building and its zones. For implementing the control scheme, *OCTOPUS*, we employ TensorFlow, an open-source machine learning library for Python.

To enable co-simulation across different models, we utilize the **Building Control Virtual Test Bed (BCVTB)**, which is a Ptolemy II platform [59]. This allows us to integrate the control of various elements such as zone temperature set points, blinds, lighting, and window schedules within EnergyPlus for our building, along with the corresponding weather data. *OCTOPUS* itself is modeled using EnergyPlus version 8.6 [45].

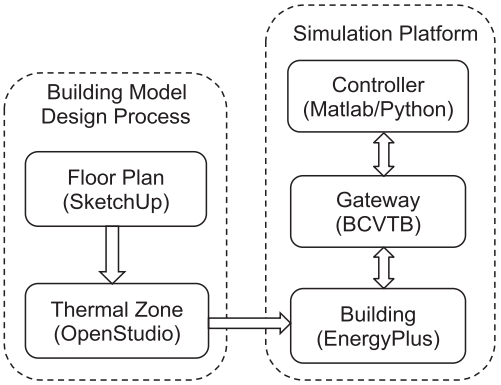

Fig. 9. Workflow of *OCTOPUS*.

To train *OCTOPUS* effectively, we employ 10-year weather data from two different cities: Merced, CA, and Chicago, IL. These cities are selected due to their distinctive weather characteristics. Merced experiences high solar radiation and significant temperature variations, while Chicago is classified as hot-summer humid continental with four distinct seasons. In our training process, we define an "episode" as one iteration of the inner loop in Algorithm 1. By training *OCTOPUS* using this approach and considering diverse weather conditions, we aim to enhance its adaptability and performance in different environmental settings.

4.2 Rule Based Method

We implement a rule-based method based on our current campus building control policy. This policy is initially established during the commissioning phase by a mechanical engineering company and is further optimized by experienced HVAC engineers during the LEED certification process.

The control policy incorporates several key elements. Firstly, we assign different zone temperature setpoints, with each zone having separate heating and cooling setpoints. During the warm-up stage, the heating setpoint is set to 70°F, and the cooling setpoint is set to 74°F. However, we impose limits on these setpoints to ensure energy efficiency and occupant comfort. The cooling setpoint is constrained within the range of 72°F to 80°F, while the heating setpoint is limited between 65°F and 72°F. Secondly, we enforce control restrictions and actuator limits to maintain system stability and prevent extreme adjustments. Specifically, the heating setpoint should not exceed the cooling setpoint minus 1°F. When adjustments are made, both the existing heating and cooling setpoints are shifted by the same amount unless they reach their respective limits. Thirdly, our control system consists of two separate control loops: the Cooling Loop and the Heating Loop. These loops operate continuously to regulate the space temperature and maintain it at the desired setpoint.

By implementing this rule-based approach, which incorporates predefined control policies and constraints, we aim to ensure efficient and effective building temperature control while adhering to the established guidelines and standards.

4.3 HVAC System Description

The HVAC system we modeled is a single duct central cooling HVAC system with terminal reheat, as illustrated in Figure 10. The process starts at the supply fan, located in the **air handler unit (AHU)**, which is responsible for supplying air to the zones. The air from the supply fan first passes through a cooling coil, where it is cooled to the minimum temperature required for the specific zone.

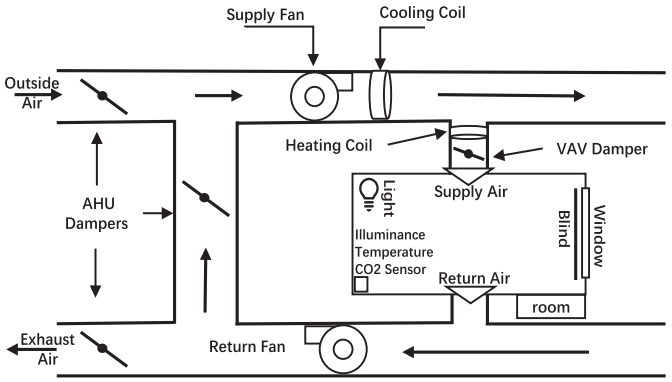

Fig. 10. HVAC single duct VAV terminal reheat layout.

Before the air enters a zone, it passes through a variable air volume (VAV) unit, which controls and adjusts the amount of air flowing into the zone. The VAV unit ensures that the air volume is regulated based on the specific needs of each zone. Terminal reheat is employed to raise the temperature of the air before it is discharged into the zone. This is achieved through a heating coil, which increases the temperature of the air to match the desired discharge setpoint temperature for each zone. Inside the zone, the supplied air is mixed with the existing air, creating a well-mixed environment. To maintain a constant static pressure, a portion of the air is exhausted from the zone. The return air from each zone is mixed in the return duct, and some of it may be directed to an economizer, depending on the system configuration.

By simulating this single duct central cooling HVAC system with terminal reheat, we can accurately model the airflow, temperature regulation, and energy exchange processes to optimize the overall performance and comfort of the building.

4.4 HVAC Modeling and Calibration

The purpose of calibration is to fine-tune the building energy model so that it can accurately replicate the energy consumption patterns observed in the target building. This process involves adjusting various model parameters to align the simulated energy use results with the measured values. The building model calibration process, depicted in Figure 11, consists of four steps.

The first step is to collect real weather data for the desired period from a public weather station. We utilize Dark Sky's API, a publicly available weather website, to gather accurate weather data spanning three months. This real weather data is crucial for ensuring the model accurately reflects the external environmental conditions experienced by the target building.

Next, we replace the default occupancy schedules in the simulation with the actual occupancy schedules obtained from the target building. ThermoSense [12], a system installed in the target building on our campus, allows us to collect detailed occupancy data at the zone level. By incorporating this fine-grained occupancy information into the simulation using EnergyPlus, we can evaluate the building's performance based on precise occupancy patterns.

The third step involves calibrating specific system and control parameters to match those observed in the target building. This calibration process encompasses various considerations, such as selecting the parameters to be calibrated, defining their ranges, and determining the calibration steps within those ranges. In our study, we employ an N-factorial design with five parameters, including infiltration rate, mass flow rate, and heating and cooling setpoints, based on operational experience. By testing different combinations of these parameters, we identify the configuration

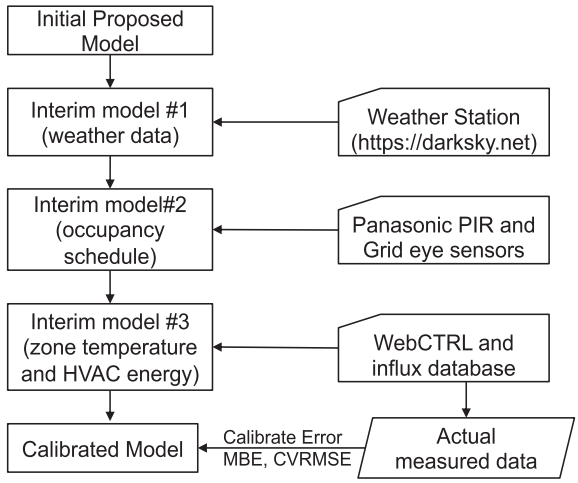

Fig. 11. Building model calibration process.

Table 3. Model Calibration Parameters

Parameter	Range	Adoption
Infiltration Rate	$0.01\ m^3 \sim 0.5\ m^3$	$0.05\ m^3$
Window Type	Single/Double Pane	Single
Window Area	$1m^2 \sim 4m^2$	$2m^2$
Window Thickness	$3mm \sim 6mm$	$3mm$
Fan Efficiency	$0.5 \sim 0.8$	0.7
Blind Type	Interior/Exterior Blind	Interior
Blind Thickness	$1mm \sim 6mm$	$1mm$

that minimizes the calibrated error, ensuring a closer alignment between the simulated and actual building performance. The calibration parameters, along with their ranges and selected values, are presented in Table 3.

Below are the equations related to the HVAC Single Duct VAV Terminal Reheat:

1) Window Type, Area, Thickness: The heat transfer through windows involves complex calculations. The U-factor (thermal transmittance) is typically calculated using the following equation:

$$Q = U \cdot A \cdot \Delta T$$

where: - Q is the heat transfer, - U is the U-factor, - A is the window area, and - ΔT is the temperature difference.

2) Fan Efficiency: The fan power consumption (P_{fan}) can be calculated using the fan efficiency (η_{fan}):

$$P_{\text{fan}} = \frac{\text{Air Flow Rate} \cdot \text{Pressure Rise}}{\eta_{\text{fan}}}$$

Finally, we compare the calibrated model's performance with the measured zone temperature and energy consumption stored in the building database. This allows us to assess the accuracy of the calibrated model by examining the discrepancies between the simulated and actual data. The entire calibration process, from model setup to error analysis, typically spans approximately one month, ensuring a comprehensive and accurate representation of the target building.

Table 4. Modeling Error after Calibration

	MBE	CVRMSE
February (hourly temperature)	-1.48%	5.32%
March (hourly temperature)	-0.26%	4.95%
April (hourly temperature)	1.20%	5.06%
May (hourly temperature)	0.48%	4.38%
February - May (monthly energy)	-3.83%	12.33%

Table 5. Parameter Settings in DRL Algorithms

$\triangle t_c$	15 m	β_1	0.9
Minibatch Size	64	β_2	0.999
Learning Rate	10^{-4}	Action Dimension	35040
γ	0.99	Action Space	$2.37 * 10^7$

ASHRAE Guideline 14-2002 [28] provides the evaluation criteria for calibrating **Building Energy Model (BEM)**. According to this guideline, both monthly and hourly data can be used for calibration, and **Mean Bias Error (MBE)** and Coefficient of Variation of the Root Mean Squared Error (CVRMSE) are utilized as evaluation metrics. For monthly calibration data, ASHRAE Guideline 14-2002 recommends an MBE of 5% and a CVRMSE of 15% as acceptable thresholds. If hourly calibration data are used, the requirements are slightly relaxed to 10% for MBE and 30% for CVRMSE. In our study, we employ hourly data to calculate the error metrics for average zone temperature, while monthly data is utilized for energy error metrics since energy data is typically available on a monthly basis.

The calibration results for zone temperature and energy consumption are presented in Table 4. It is evident that with the optimal parameter settings, we achieve less than 2% MBE and less than 6% CVRMSE for zone temperature, indicating a high level of accuracy in temperature prediction. Although the CVRMSE for monthly heating and cooling energy demand appears relatively large, both the MBE and CVRMSE values remain within the acceptable range defined by the guideline. This implies that the model can accurately estimate monthly energy consumption, despite the slightly higher variability in energy predictions. By meeting these calibration criteria, our BEM model demonstrates its capability to provide accurate calculations for both average zone temperature and monthly energy consumption, ensuring the reliability of the model outputs.

4.5 *OCTOPUS* Training

OCTOPUS is trained in an offline mode. The 10-year weather data from the two test locations, Merced, CA, and Chicago, IL, is divided randomly into training and testing sets. Eight years of weather data are allocated for training the models, while the remaining two years are reserved for testing and evaluating the model's performance. The parameter settings used in our DRL algorithms are presented in Table 5.

For the implementation of OCTOPUS, we employ the Adam optimizer [32] for gradient-based optimization, utilizing a learning rate of 10^{-4}. The agent is trained using a minibatch size of 64 and a discount factor γ of 0.99, which balances the importance of immediate and future rewards. The target network is updated every 10^3 time steps to enhance stability and convergence.

To introduce non-linearity into the network, we utilize the **rectified linear activation function (ReLU)** for all hidden layers, while the output layers employ linear activation. The shared network module consists of two hidden layers with 512 and 256 units, respectively. Additionally,

each branch in the network includes one hidden layer with 128 units. The weights in the network are initialized using Xavier initialization [26], while the biases are initialized to zero.

In our training process, we incorporate prioritized replay with a buffer size of 10^6 and linear annealing of β, ranging from β_0 = 0.4 to 1 over 2 x 10^6 steps. While the $\epsilon-$greedy policy is commonly used with Q-learning, random exploration in physical, continuous-action domains can be inefficient. To enhance action exploration in our building environment, we sample actions from a Gaussian distribution with a mean at the greedy actions and a small fixed standard deviation throughout training to encourage lifelong exploration. During training, we employ a fixed standard deviation of 0.2, whereas during evaluation, the standard deviation is set to zero. This exploration strategy yielded slightly better performance compared to using an $\epsilon-$greedy policy with a fixed or linearly annealed exploration probability. Each time (action) slot in our system lasts for 15 minutes. After 1,000 episodes, we achieved convergence of our reward function, as explained in Section 5.6.

5 EVALUATION

In this section, we conduct a performance comparison between *OCTOPUS*, the rule-based method, and the latest DRL-based method.

5.1 Experiment Setting

In this section, we compare the performance of three different control methods: the rule-based HVAC control, the conventional DRL-based method (referred to as DDQN-HVAC), and *OCTOPUS*, as introduced in Section 4.2. The rule-based method solely controls the HVAC system, while DDQN-HVAC employs the dueling DQN architecture to control the water-based heating system, as described in [63]. To ensure a fair comparison, we initialize the lights in all experiments, as the rule-based method and DDQN-HVAC do not control the light system. *OCTOPUS*, however, has the capability to dim the lights if the blind is open during the day, and it may also interact with the blind and window system.

To evaluate human comfort, we measure three metrics: PMV, illuminance, and carbon dioxide concentration. The acceptable ranges for these metrics are determined based on building standards and previous research. Specifically, the PMV comfort range is set from -0.5 to 0.5 [5], the illuminance comfort range is set from 500 to 1000 lux [46], and the carbon dioxide concentration comfort range is set from 400 to 1000 ppm [3].

We apply the three control methods to the building model presented in Section 4, simulating a two-month period (January and July) in two different locations with distinct weather patterns. Table 6 presents the human comfort results for the three control methods, as well as their energy consumption. The violation rate is calculated by dividing the time when a human comfort metric falls outside its acceptable range by the total simulated time. It is important to note that future work will explore additional quality of service metrics, such as the magnitude and duration of violations or combinations thereof, to provide a more comprehensive assessment of the control methods' performance.

5.2 Human Comfort

Based on the results presented in Table 6, we observe that all three control methods are effective in maintaining the PMV within the desired range for the majority of the time, as indicated by the low violation rates. However, *OCTOPUS* and DDQN-HVAC exhibit slightly higher average PMV violation rates compared to the rule-based method, with differences of 2.19% and 2.22%, respectively. This can be attributed to the energy-saving approach employed by the DRL-based methods, which aim to set the PMV value close to the boundary of the acceptable range. As shown in Table 6, the

Table 6. Human Comfort Statistical Results for Rule-Based, DDQN-HVAC and *OCTOPUS* Schemes

Location	Method	Metric	PMV		Illuminance (lux)		CO_2 Concentration (ppm)		Energy Consumption (kWh)	
			Jan.	July	Jan.	July	Jan.	July	Jan.	July
Merced	Rule-Based Method	Mean	0.03	-0.25	576.78	646.45	623.61	668.03	1990.99	3583.03
		Std	0.11	0.13	152.54	157.11	120.64	181.22		
		Violation	0	2%	0.94%	0	0.3%	3.629%		
	DDQN-HVAC [63]	Mean	-0.19	0.28	576.78	646.45	625.62	648.01	1859.10	3335.58
		Std	0.21	0.11	152.54	157.11	122.62	120.57		
		Violation	2.99%	4.4%	0.94%	0	0	0.2%		
	OCTOPUS	Mean	-0.31	0.27	587.12	569.88	594.77	612.33	1756.24	2941.46
		Std	0.2	0.10	382.27	75.83	111.59	110.35		
		Violation	5.7%	2.5%	0.26%	0.2%	1.31%	0.33%		
Chicago	Rule-Based Method	Mean	-0.28	-0.15	583.27	637.07	610.26	638.33	3848.61	3309.56
		Std	0.11	0.02	163.96	151.37	63.94	151.37		
		Violation	3.09%	0	1.1%	0	0	0		
	DDQN-HVAC [63]	Mean	-0.32	0.24	583.27	637.07	612.74	649.32	3605.21	3078.67
		Std	0.08	0.07	163.96	151.37	65.09	90.16		
		Violation	3.7%	2.9%	1.1 %	0	0	0		
	OCTOPUS	Mean	-0.4	0.29	598.34	544.09	640.31	633.71	3496.54	2722.03
		Std	0.1	0.11	259.88	55.37	99.85	111.04		
		Violation	4.2%	1.47%	1.6 %	0	1%	1.31%		

average PMV values for *OCTOPUS* and DDQN-HVAC (-0.36 and -0.26, respectively) are closer to the range boundary (-0.5) compared to the rule-based method (-0.13).

In terms of visual comfort and indoor air quality, all three control methods demonstrate a minimal violation rate. For illuminance, *OCTOPUS* and DDQN-HVAC achieve mean illuminance values of 590.69 lux and 610.89 lux, respectively. *OCTOPUS* optimizes energy consumption by leveraging natural light whenever possible. Concerning indoor air quality, the average CO_2 concentration for *OCTOPUS*, DDQN-HVAC, and the rule-based method is 620.28 ppm, 633.92 ppm, and 635.06 ppm, respectively. *OCTOPUS* actively adjusts both the window system and HVAC system to maintain the desired CO_2 concentration level. On the other hand, DDQN-HVAC and the rule-based method rely solely on the HVAC system for control.

Overall, while there are slight differences in performance across the three methods, all of them achieve satisfactory results in maintaining human comfort and adhering to the predefined comfort ranges. *OCTOPUS* showcases its advantage by utilizing natural light and incorporating multiple control strategies, resulting in effective energy savings without compromising comfort.

5.3 Energy Efficiency

The results presented in Table 6 demonstrate that *OCTOPUS* achieves significant energy savings compared to the rule-based control method and DDQN-HVAC, with average energy reductions of 14.26% and 8.1%, respectively. This performance gain is consistent across both cities tested. *OCTOPUS* achieves these energy savings by leveraging the synergies between different subsystems and optimizing their usage.

Figure 12 illustrates the daily energy consumption of the three methods in January in Merced. In most days, *OCTOPUS* consumes less energy than the other two methods. However, it's important to note that *OCTOPUS* is not always the best-performing method, particularly in the first half of the month. This can be attributed to the fluctuating outdoor temperature, with an average range of $2°C$ to $13°C$ during this period. Nevertheless, *OCTOPUS* demonstrates clear energy savings in

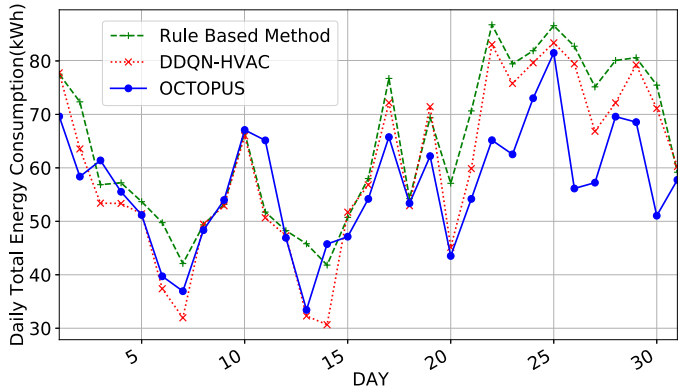

Fig. 12. Daily energy consumption of control methods.

the second half of the month when the temperature range expands from -1°C to 18°C. During this time, *OCTOPUS* effectively utilizes external air by opening the windows for natural ventilation. Comparing the energy savings in July presented in Table 6, *OCTOPUS* outperforms the rule-based method and DDQN-HVAC with energy reductions of 17.6% and 11.7%, respectively. In July, the outdoor temperature ranges from 15°C to 42°C in Merced and from 15°C to 40°C in Chicago. Opening the windows when the temperature falls within the acceptable range allows *OCTOPUS* to save energy consumed by the HVAC system. However, in January, when both cities experience cold weather, the windows remain closed most of the time and contribute less to energy savings.

In summary, *OCTOPUS* consistently achieves substantial energy savings across different months and cities, demonstrating its effectiveness in optimizing energy consumption. The system capitalizes on external factors, such as outdoor temperature and natural ventilation, to further enhance its energy-saving capabilities.

5.4 Performance Decomposition

To analyze the energy-saving contributions of each subsystem, we implement four versions of *OCTOPUS*: *OCTOPUS* with only the HVAC system (OCTOPUS_HVAC), *OCTOPUS* with HVAC system and lighting system (OCTOPUS_HVAC_L), *OCTOPUS* with HVAC system, lighting system, and blind system (OCTOPUS_HVAC_L_B), and *OCTOPUS* with all four subsystems (OCTOPUS_HVAC_L_B_W). Figure 13 illustrates the energy consumption of these versions during different months and at different locations (Merced and Chicago).

Compared to the rule-based method, OCTOPUS_HVAC achieves an additional energy saving of 6.16% by focusing solely on HVAC control. The integration of the lighting system in OCTOPUS_HVAC_L results in a further energy saving of 2.73%. Incorporating the blind system in OCTOPUS_HVAC_L_B contributes an additional energy saving of 1.93%. Finally, by including the window system in OCTOPUS_HVAC_L_B_W, an additional 3.44% energy savings is achieved. Each subsystem demonstrates distinct contributions to energy savings during January and July.

In January, the four subsystems (HVAC, lighting, blinds, and windows) contribute to energy savings as follows: HVAC - 6.16%, lighting - 2.73%, blinds - 1.93%, and windows - 0%. However, in July, these contributions change to HVAC - 5.9%, lighting - 3.31%, blinds - 1.99%, and windows - 6.4%. The window system exhibits the most noticeable difference between the two months, contributing 6.4% to the total energy savings. This variation is attributed to the operational characteristics discussed earlier. In January, the windows remain closed most of the time, limiting their potential

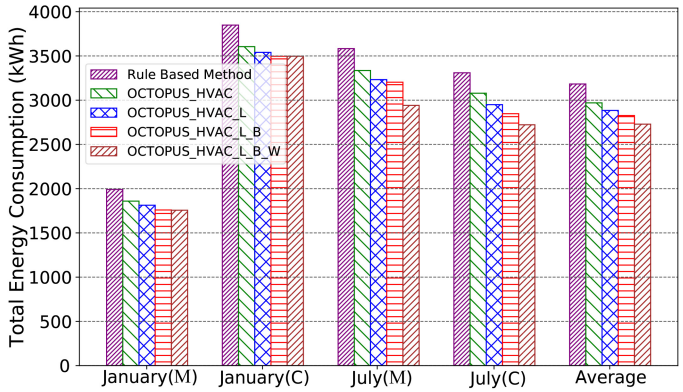

Fig. 13. Performance contribution of each subsystem.

Table 7. Different Parameters for Reward Function in Octopus

Parameter (ρ_1,ρ_2, ρ_3, ρ_4)	PMV		Illuminance (lux)		CO_2 Concentration (ppm)		Energy (kWh)
	Mean	Std	Mean	Std	Mean	Std	
1, 1, 1, 1	−0.36	0.15	587.35	94.52	587.25	101.14	3250.55
5, 1, 1, 1	−0.33	0.16	611.71	131	608.48	175.1	3221.20
10, 1, 1, 1	−0.31	0.16	624.97	189.04	647.77	150.33	3150.62
2, 3, 1, 1	−0.383	0.10	569.88	75.83	636.5	179.46	2941.46
2, 5, 1, 1	−0.481	0.13	689.23	146.66	616.02	177.32	2900.44

for energy savings. Conversely, in July, the utilization of cold outdoor air for cooling purposes significantly reduces the reliance on the HVAC system.

Overall, the analysis demonstrates the distinct energy-saving contributions of each subsystem within *OCTOPUS*. The integration of multiple subsystems progressively enhances energy efficiency, with the window system playing a particularly significant role during months with favorable outdoor conditions.

5.5 Hyperparameters Setting

The hyperparameters in the reward function (Equation (3)) are carefully tuned to strike a balance between energy consumption and human comfort. Table 7 presents the performance results of the trained DRL agents obtained during the hyperparameter tuning experiments. Evaluation metrics include total energy consumption, as well as the mean and standard deviation of PMV, illuminance, and carbon dioxide concentration.

Interestingly, the control performance results obtained from different hyperparameters do not always align with intuitive expectations. One might anticipate that increasing the weight ρ_1 and decreasing ρ_2, ρ_3 and ρ_4 would lead to lower energy consumption while meeting the requirements for thermal comfort, visual comfort, and indoor air quality. However, the results presented in Table 7 demonstrate that increasing the weight of energy does not necessarily result in reduced energy consumption. These counter-intuitive outcomes may arise from the delayed reward problem, whereby the DRL agents become trapped in local optimal areas during training.

Among the five experiments listed in Table 7, the fourth row stands out by achieving a remarkable 17.9% reduction in energy consumption while only slightly compromising the three human

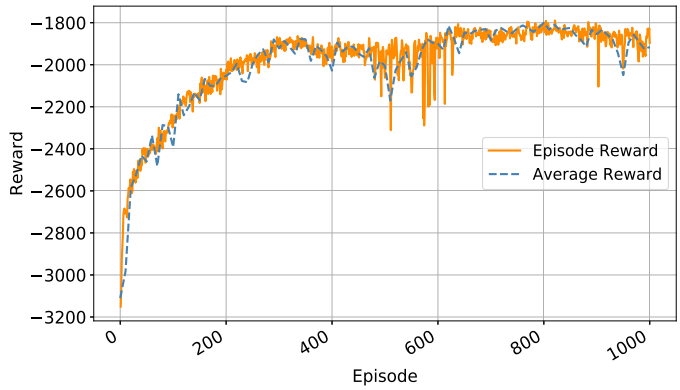

Fig. 14. The convergence of *OCTOPUS*.

comfort quality metrics in the testing model. This experiment achieves a superior balance between human comfort and energy consumption. Consequently, the hyperparameters specified in the fourth row are adopted for the trained agent.

These findings underscore the complexity of optimizing the reward function in DRL-based HVAC control. The non-linear relationship between the hyperparameters and the control performance necessitates a systematic exploration to identify the most effective configuration. The selected hyperparameters strike an effective compromise between energy savings and human comfort, showcasing the practical value of the trained agent.

5.6 Convergence of *OCTOPUS* Training

Figure 14 illustrates the accumulated reward of *OCTOPUS* at each episode during the training process. In this process, we calculate the reward function at every control time step, which occurs every 15 minutes. Consequently, one episode corresponds to 2,880 time steps, representing a month of simulation. The accumulated reward for each episode, depicted as "episode reward" in Figure 14, is the sum of the rewards obtained over the 2,880 time steps.

The results in Figure 14 indicate that the episode reward increases progressively and eventually stabilizes as the number of training episodes rises. Once the episode reward reaches a relatively steady state, it implies that further improvements to the learned control policy are unlikely, signifying convergence of the training process. Notably, the training reward exhibits fluctuations between adjacent episodes due to the large number of time steps within each episode (i.e., 2,880). Because some of these 2,880 time steps are selected randomly using an exploration rate (determined by a Gaussian distribution with a standard deviation of 0.2), the rewards calculated at these steps can vary dynamically. During these time steps, the action generated by the agent is not used, and instead, a random action is chosen to prevent convergence to local minima.

To enhance the stability of the training process, we employ a sliding window of 10 episodes to smooth the episode reward. By averaging the rewards within this window, the resulting average reward in Figure 14 demonstrates improved stability during the training phase. This smoothing technique provides a clearer picture of the training progress by mitigating the effects of temporary fluctuations in the reward values.

6 RELATED WORK

Conventional control of the HVAC system. Model predictive control (MPC) models have been developed for HVAC control. It is a planning-based method that solves an optimal control problem iteratively over a receding time horizon. Some of the advantages of MPC are that it takes into

consideration future disturbances and that it can handle multiple constraints and objectives, e.g., energy consumption and human comfort [11]. However, it can be argued that the main roadblock preventing the widespread adoption of MPC is its reliance on a model [36, 47]. By some estimates, modeling can account for up to 75% of the time and resources required for implementing MPC in practice [49]. Because buildings are highly heterogeneous, a custom model is required for each thermal zone or building under control.

There are two paradigms for modeling building dynamics: physics-based and statistics-based [47]. Physics-based models, e.g., EnergyPlus, utilize physical knowledge and material properties of a building to create detailed representation of the building dynamics. A major shortcoming is that such models are not control-oriented. Nonetheless, it is not impossible to use such models for control [8]. For instance, exhaustive search optimization is used to derive control policy for an EnergyPlus model [64]. Furthermore, a physics-based model requires significant modeling effort, because they have a large number of free parameters to be specified by engineers (e.g., 2,500 parameters for a medium-sized building [31]); and information required for determining these parameters are scattered in different design documents [27].

Statistical models assume a parametric model form, which may or may not have physical underpinnings, and identify model parameters directly from data. Dinh and Kim [20] propose a hybrid control that combines MPC and direct imitation learning to reduce energy cost while maintaining a comfortable indoor temperature. While this approach is potentially scalable, a practical problem is that the experimental conditions required for accurate identification of building systems fall outside of normal building operations [4].

Integration of occupant-driven approaches in building control. In recent years, the delicate balance between energy efficiency and occupant comfort in commercial buildings has been a focal point in building control strategies. While conventional control methods, such as MPC, have shown promise, they often face challenges, particularly in the extensive modeling efforts required for heterogeneous buildings.

The works [9, 34, 35] address key aspects of occupant-building interaction and demand response optimization in microgrids. Baldi et al. [9] propose a switched self-tuning approach to automate occupant-building interaction via smart zoning of thermostatic loads, dynamically regulating set points based on occupancy patterns. Korkas et al. [35] focus on occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources, presenting a novel control algorithm for integrating energy generation, consumption, and occupant behavior. Additionally, Korkas et al. [34] contribute to the field by developing a **distributed demand management system (D-DMS)** for grid-connected microgrids, incorporating feedback actions for improved performance under changing conditions.

These papers provide valuable insights into dynamically managing occupant comfort while optimizing energy efficiency, addressing limitations in conventional control methods. While our work leverages physics-based and statistical models for building dynamics, the incorporation of feedback-driven strategies from these studies could offer further enhancements in real-world building control systems.

Conventional control of multiple subsystems. Blind system should be considered as an integral part of fenestration system design for commercial and office buildings, in order to balance daylighting requirements versus the need to reduce solar gains. The impact of glazing area, shading device properties and shading control on building cooling and lighting demand was calculated using a coupled lighting and thermal simulation module [52]. The interactions between cooling and lighting energy use in perimeter spaces were evaluated as a function of window-to-wall ratio and shading parameters.

The impacts of window operation on building performance was investigated [57] for different types of ventilation systems including natural ventilation, mixed-mode ventilation, and conventional VAV systems in a medium-size reference office building. While the results highlighted the impacts of window operation on energy use and comfort and identified HVAC energy savings with mixed-mode ventilation during summer for various climates, the control for window opening fraction was estimated by experience and is not salable for different kinds of buildings. Kolokotsa et al. [33] propose an energy-efficient fuzzy controller that utilizes a genetic algorithm to manage four subsystems (HVAC, lighting, window, and blind) and ensure optimal human comfort. However, the drawback lies in the genetic algorithm's time-consuming nature, taking several minutes to hours to generate a single control action, making it impractical for real-time implementation in building control systems.

RL-based control of the HVAC system. As a general artificial intelligence technology, deep learning [41] and deep reinforcement learning [17] have been applied in many fields, e.g., Internet of energy [40], unmanned aerial vehicles [56], smart microgrids [38], edge computing [13]. Many works also apply RL for HVAC control [44]. RL control can be a "model-free"control method, i.e., an RL agent has no prior knowledge about the controlled process. RL learns an optimal control strategy by "trial-and-error". Therefore, it can be an online learning method that learns an optimal control strategy during actual building operations. Fazenda et al. [24] investigated the application of a reinforcement-learning-based supervisory control approach, which actively learns how to appropriately schedule thermostat temperature setpoints. However, in HVAC control, online learning may introduce unstable and poor control actions at the initial stage of the learning. In addition, it may take a long time (e.g., over 50 days reported in [24]) for an RL agent to converge to a stable control policy for some cases. Therefore, some studies choose to use an HVAC simulator to train the RL agent offline [63].

Unlike MPC, simulators with arbitrary high complexity can be directly used to train RL agents because of its "model-free" nature. Li and Xia [39] employ Q-learning to regulate HVAC control, while Dalamagkidis et al. [16] introduce a **Linear Reinforcement Learning Controller (LRLC)** that utilizes linear function approximation of the state-action value function to achieve thermal comfort while minimizing energy consumption. However, tabular Q-learning approaches are not well-suited for problems characterized by a large state space, such as the one involving the four subsystems. Le et al. [37, 54] propose a control method of air free-cooled data centers in tropics via DRL. Vazquez-Canteli et al. [55] develop a multi-agent RL implementation for load shaping of grid-interactive connected buildings. Ding et al. [19] design a model-based RL method for multi-zone building control. Zhang and Lam [63] implement and deploy a DRL-based control method for radiant heating systems in a real-life office building. An et al. [6, 7] design a safe MBRL HVAC control approach that can achieve low human comfort violation with a dynamics model trained on a small dataset. Gao et al. [25] propose **deep deterministic policy gradients (DDPGs)**-based approach for learning the thermal comfort control policy. Although the above works can improve the performance of HVAC control, they only focus on the HVAC subsystem.

7 DISCUSSION

Deploying in a Real Building. Although we have developed a calibrated simulation model of a real building on our campus for training and evaluation purposes, we have not yet deployed *OCTOPUS* in an actual building due to the lack of access to an automatic blind and window system. However, we are actively seeking financial support to collaborate with our facility team for a potential upgrade that would enable the deployment of *OCTOPUS*. Our ultimate goal is to deploy *OCTOPUS* in real buildings, as it has been specifically designed for such applications. For

a new building, the deployment process involves creating an EnergyPlus model and calibrating it using real building operation data. Once the *OCTOPUS* control agent is trained using the calibrated simulation model and real weather data, it can be deployed in the building for real-time control. The control agent takes the building's state as input and generates control actions for the four subsystems at a predefined action interval (e.g., every 10 minutes). *OCTOPUS* is capable of providing real-time control, as a single inference takes only 22 ms. In our future work, we plan to deploy *OCTOPUS* in a real building, where it can demonstrate its effectiveness in achieving energy savings while maintaining human comfort. This practical deployment will allow us to validate the performance and benefits of *OCTOPUS* in a real-world setting.

Scalability of *OCTOPUS*. *OCTOPUS* is currently designed to operate in a one-zone building configuration with a single HVAC system, lighting zone, blind, and window. However, real-world buildings, including small homes, typically consist of multiple lighting zones, blinds, and windows, each requiring individual control actions within a subsystem. Scaling OCTOPUS to handle such complex building configurations is a challenge we aim to address in our future work. To overcome the scalability problem, we plan to extend *OCTOPUS* by increasing the number of BDQ branches. Each branch will correspond to a subsystem in each zone of the building. By incorporating multiple branches, *OCTOPUS* will be able to effectively manage and control various subsystems in a building with multiple lighting zones, blinds, and windows. Tackling this scalability issue is crucial to ensure the applicability of *OCTOPUS* in realistic building environments, where the complexity and diversity of subsystems necessitate advanced control strategies. Our ongoing research aims to develop innovative approaches that enable *OCTOPUS* to handle large-scale building configurations while maintaining its efficiency and effectiveness in achieving energy savings and optimizing human comfort.

Building Model Calibration. A crucial element of our architecture relies on utilizing a calibrated building model that closely resembles the target building. This allows us to generate the necessary training data effectively. It is not trivial to have a building model in the EnergyPlus simulator. Achieving an accurate calibrated model is indeed a laborious process that involves extensive trial-and-error experimentation with numerous parameters. Among the vast array of parameters available in EnergyPlus, we leveraged our expertise and sought guidance from experts to identify the most critical parameters and determine a reasonable range of values to explore. It took us approximately four weeks to fine-tune the model to an acceptable level. We want to emphasize that this timeframe reflects the complexity of the calibration process and the dedication required for its successful completion.

Despite our efforts, there is no universal solution, and challenges may arise, particularly when dealing with unconventional building architectures or specialized HVAC systems that cannot be readily replicated in a simulation environment. It is not trivial that this model is well-calibrated. These unique cases may require additional considerations and careful adaptation of the simulation model to ensure accurate representation. We acknowledge that achieving a precise calibration can be demanding and time-consuming, but it is a necessary step to ensure the fidelity and reliability of our training process.

Furthermore, it is not trivial to have 10-year data to make sure the model is well-calibrated. We acknowledge the significance of long-term data in validating our calibrated model. In our study, we recognize the importance of extending our analysis over a 10-year period to ensure the robustness of our findings. This extended duration allows us to capture variations in building performance under diverse conditions, contributing to the overall reliability of our approach.

Accepting Users' Feedback. Certain existing work [60] has introduced a mechanism for users to provide feedback to the control server. This feedback serves as personalized preferences regarding

various human comfort metrics and is taken into account during the control decision-making process. *OCTOPUS* is designed to readily accommodate user feedback, enabling the training of an enhanced agent model with a minor modificationadjusting the calculated comfort values in the reward function based on user feedback. This functionality can be leveraged during the initial training phase or for subsequent updates once the system is deployed. To illustrate, the *OCTOPUS* control agent can undergo incremental training at specific intervals, such as one month. The newly-trained agent can then be employed for real-time control, incorporating the insights gained from user feedback. This iterative approach allows *OCTOPUS* to continuously adapt and improve its control strategies based on user preferences and evolving comfort requirements. By fostering this feedback loop, *OCTOPUS* ensures a more personalized and user-centric control experience in real-world building environments.

The Markov Property in Smart Building Control The Markov Decision Process (MDP) and its inherent Markov property serve as the foundational framework for our study in smart building control. The Markov property, characterized by the independence of future states from past states, plays a pivotal role in shaping the decision-making dynamics within our DRL model. In the context of smart building control, the Markov property simplifies complex decision-making by allowing the DRL agent to focus exclusively on the current state of the environment. This immediate-state dependency aligns seamlessly with the nature of control decisions in smart buildings. For instance, when managing HVAC systems or lighting, actions are primarily influenced by real-time sensor data and environmental conditions, rendering the Markov property a natural fit. By assuming that future states depend only on the present state and action, we streamline the decision-making process, making it computationally tractable for real-time applications.

However, it's important to recognize that this simplification does come with challenges. Building an accurate Markov model for smart building control requires meticulous calibration, typically involving extensive experimentation with simulation models like EnergyPlus. Achieving a precise match between the model and the real-world building dynamics is laborious but crucial for reliable decision-making. In our future work, we aspire to augment the validation of the Markov property by incorporating empirical data to support the assumption of state independence.

8 CONCLUSIONS

This paper introduces *OCTOPUS*, an innovative control system for buildings that leverages deep reinforcement learning (DRL) to effectively manage multiple subsystems (e.g., HVAC, lighting, blinds, windows) while balancing energy consumption and human comfort. Our architecture addresses several challenges, including handling large action states, designing a unique reward function that considers both energy efficiency and comfort, and meeting data requirements by utilizing historical weather data and a calibrated simulator specific to the target building.

To evaluate *OCTOPUS*, we compare its performance against state-of-the-art approaches. We benchmarked it against a rule-based control scheme implemented in a LEED Gold-certified building, as well as a DRL scheme focused on optimized heating from existing literature. The results demonstrate significant energy savings, with *OCTOPUS* achieving a remarkable 14.26% and 8.1% reduction in energy consumption while consistently maintaining, and sometimes even enhancing, human comfort across temperature, air quality, and lighting aspects. By combining advanced DRL techniques, a tailored reward function, and comprehensive simulations, *OCTOPUS* showcases its effectiveness in optimizing building control operations. It offers a promising solution for enhancing energy efficiency and occupant comfort in modern buildings, contributing to sustainable and comfortable living environments.

REFERENCES

[1] 2018. *sketchup*. https://www.sketchup.com

[2] 2019. *GEZE: Products, System Solutions and Services for Doors and Windows*. https://www.geze.com/en

[3] ANSI/ASHRAE Standard 62.1. 2016. Ventilation for acceptable indoor air quality.

[4] Clarence Agbi, Zhen Song, and Bruce Krogh. 2012. Parameter identifiability for multi-zone building models. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE.

[5] Refrigerating American Society of Heating and Air-Conditioning Engineers. Standard 55. 2017. Thermal environmental conditions for human occupancy.

[6] Zhiyu An, Xianzhong Ding, and Wan Du. 2024. Go beyond black-box policies: Rethinking the design of learning agent for interpretable and verifiable HVAC control. *arXiv preprint arXiv:2403.00172* (2024).

[7] Zhiyu An, Xianzhong Ding, Arya Rathee, and Wan Du. 2023. CLUE: Safe model-based RL HVAC control using epistemic uncertainty estimation. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 149–158.

[8] Ercan Atam and Lieve Helsen. 2016. Control-oriented thermal modeling of multizone buildings: Methods and issues: Intelligent control of a building system. *IEEE Control Systems Magazine* (2016).

[9] Simone Baldi, Christos D. Korkas, Maolong Lv, and Elias B. Kosmatopoulos. 2018. Automating occupant-building interaction via smart zoning of thermostatic loads: A switched self-tuning approach. *Applied Energy* 231 (2018).

[10] Guneet Bedi, Ganesh Kumar Venayagamoorthy, and Rajendra Singh. 2020. Development of an IoT-driven building environment for prediction of electric energy consumption. *IEEE Internet of Things Journal* 7, 6 (2020), 4912–4921.

[11] Alex Beltran and Alberto E. Cerpa. 2014. Optimal HVAC building control with occupancy prediction. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-efficient Buildings*. 168–171.

[12] Alex Beltran, Varick L. Erickson, and Alberto E. Cerpa. 2013. ThermoSense: Occupancy thermal based sensing for HVAC control. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*. 1–8.

[13] Xianfu Chen, Honggang Zhang, Celimuge Wu, Shiwen Mao, Yusheng Ji, and Medhi Bennis. 2018. Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning. *IEEE Internet of Things Journal* 6, 3 (2018), 4005–4018.

[14] Zhijin Cheng, Qianchuan Zhao, Fulin Wang, Yi Jiang, Li Xia, and Jinlei Ding. 2016. Satisfaction based Q-learning for integrated lighting and blind control. *Energy and Buildings* (2016).

[15] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. 2018. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757* (2018).

[16] Konstantinos Dalamagkidis, Denia Kolokotsa, Konstantinos Kalaitzakis, and George S. Stavrakakis. 2007. Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment* (2007).

[17] Xianzhong Ding and Wan Du. 2022. DRLIC: Deep reinforcement learning for irrigation control. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 41–53.

[18] Xianzhong Ding, Wan Du, and Alberto Cerpa. 2019. OCTOPUS: Deep reinforcement learning for holistic smart building control. In *Proceedings of the 6th ACM International Conference on Systems for Energy-efficient Buildings, Cities, and Transportation*. 326–335.

[19] Xianzhong Ding, Wan Du, and Alberto E. Cerpa. 2020. MB2C: Model-based deep reinforcement learning for multi-zone building control. In *Proceedings of the 7th ACM International Conference on Systems for Energy-efficient Buildings, Cities, and Transportation*. 50–59.

[20] Huy Truong Dinh and Daehee Kim. 2021. MILP-based imitation learning for HVAC control. *IEEE Internet of Things Journal* 9, 8 (2021), 6107–6120.

[21] Steven J. Emmerich and Andrew K. Persily. 2001. *State-of-the-art Review of CO2 Demand Controlled Ventilation Technology and Application*. Citeseer.

[22] P. O. Fanger. 1984. Moderate thermal environments Determination of the PMV and PPD indices and specification of the conditions for thermal comfort. *ISO 7730* (1984).

[23] Poul O. Fanger. 1970. Thermal comfort. analysis and applications in environmental engineering. *Thermal Comfort. Analysis and Applications in Environmental Engineering*. (1970).

[24] Pedro Fazenda, Kalyan Veeramachaneni, Pedro Lima, and Una-May O'Reilly. 2014. Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems. *Journal of Ambient Intelligence and Smart Environments* 6, 6 (2014), 675–690.

[25] Guanyu Gao, Jie Li, and Yonggang Wen. 2020. DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning. *IEEE Internet of Things Journal* (2020).

[26] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[27] Bo Gu, Semiha Ergan, and Burcu Akinci. 2014. Generating as-is building information models for facility management by leveraging heterogeneous existing information sources: A case study. In *Construction Research Congress 2014: Construction in a Global Network*.
[28] ASHRAE Guideline. 2002. Guideline 14-2002, measurement of energy and demand savings. *American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, Georgia* (2002).
[29] Jeff S. Haberl, David E. Claridge, and Charles Culp. 2005. ASHRAE's guideline 14-2002 for measurement of energy and demand savings: How to determine what was really saved by the retrofit. (2005).
[30] Kazufumi Ito and Karl Kunisch. 2008. *Lagrange Multiplier Approach to Variational Problems and Applications*. Siam.
[31] Omer Tugrul Karaguzel and Khee Poh Lam. 2011. Development of whole-building energy performance models as benchmarks for retrofit projects. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*. IEEE.
[32] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[33] D. Kolokotsa, G S. Stavrakakis, K. Kalaitzakis, and D. Agoris. 2002. Genetic algorithms optimized fuzzy controller for the indoor environmental management in buildings implemented using PLC and local operating networks. *Engineering Applications of Artificial Intelligence* (2002).
[34] Christos D. Korkas, Simone Baldi, and Elias B. Kosmatopoulos. 2018. Grid-connected microgrids: Demand management via distributed control and human-in-the-loop optimization. In *Advances in Renewable Energies and Power Technologies*. Elsevier, 315–344.
[35] Christos D. Korkas, Simone Baldi, Iakovos Michailidis, and Elias B. Kosmatopoulos. 2016. Occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources and energy storage. *Applied Energy* 163 (2016), 93–104.
[36] Devanshu Kumar, Xianzhong Ding, Wan Du, and Alberto Cerpa. 2021. Building sensor fault detection and diagnostic system. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 357–360.
[37] Duc Van Le, Rongrong Wang, Yingbo Liu, Rui Tan, Yew-Wah Wong, and Yonggang Wen. 2021. Deep reinforcement learning for tropical air free-cooled data center control. *ACM Transactions on Sensor Networks (TOSN)* (2021).
[38] Lei Lei, Yue Tan, Glenn Dahlenburg, Wei Xiang, and Kan Zheng. 2020. Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids. *IEEE Internet of Things Journal* 8, 10 (2020), 7938–7953.
[39] Bocheng Li and Li Xia. 2015. A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 444–449.
[40] Lin Lin, Xin Guan, Yu Peng, Ning Wang, Sabita Maharjan, and Tomoaki Ohtsuki. 2020. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. *IEEE Internet of Things Journal* 7, 7 (2020), 6288–6301.
[41] Miaomiao Liu, Xianzhong Ding, and Wan Du. 2020. Continuous, real-time object detection on mobile devices without offloading. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 976–986.
[42] Oswaldo Lucon, Diana Ürge-Vorsatz, A. Zain Ahmed, Hashem Akbari, Paolo Bertoldi, Luisa F. Cabeza, Nicholas Eyre, Ashok Gadgil, L. D. Harvey, and Yi Jiang. 2014. Buildings. (2014).
[43] Daniel Minoli, Kazem Sohraby, and Benedict Occhiogrosso. 2017. IoT considerations, requirements, and architectures for smart buildings-energy optimization and next-generation building management systems. *IEEE Internet of Things Journal* 4, 1 (2017), 269–283.
[44] Zoltan Nagy, June Y. Park, and J. Vazquez-Canteli. 2018. Reinforcement learning for intelligent environments: A tutorial. *Handbook of Sustainable and Resilient Infrastructure* (2018).
[45] U.S. Department of Energy. 2016. *EnergyPlus 8.6.0*. https://energyplus.net/
[46] David Christopher Pritchard. 2014. *Lighting*. Routledge.
[47] Samuel Privara, Jiří Cigler, Zdeněk Váňa, Frauke Oldewurtel, Carina Sagerschnig, and Eva Žáčeková. 2013. Building modeling as a crucial part for building predictive control. *Energy and Buildings* 56 (2013), 8–22.
[48] Hamid Rajabi, Zhizhang Hu, Xianzhong Ding, Shijia Pan, Wan Du, and Alberto Cerpa. 2022. MODES: Multi-sensor occupancy data-driven estimation system for smart buildings. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. 228–239.
[49] Peter Rockett and Elizabeth Abigail Hathway. 2017. Model-predictive control for non-domestic buildings: A critical review and prospects. *Building Research & Information* (2017).
[50] Wai Wai Shein, Yasuo Tan, and Azman Osman Lim. 2012. PID controller for temperature control with multiple actuators in cyber-physical home system. In *2012 15th International Conference on Network-Based Information Systems*. IEEE, 423–428.

[51] Arash Tavakoli, Fabio Pardo, and Petar Kormushev. 2018. Action branching architectures for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[52] Athanassios Tzempelikos and Andreas K. Athienitis. 2007. The impact of shading design and control on building cooling and lighting demand. *Solar Energy* (2007).
[53] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
[54] Duc Van Le, Yingbo Liu, Rongrong Wang, Rui Tan, Yew-Wah Wong, and Yonggang Wen. 2019. Control of air free-cooled data centers in tropics via deep reinforcement learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*.
[55] Jose R. Vazquez-Canteli, Gregor Henze, and Zoltan Nagy. 2020. MARLISA: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*.
[56] Chao Wang, Jian Wang, Jingjing Wang, and Xudong Zhang. 2020. Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards. *IEEE Internet of Things Journal* 7, 7 (2020), 6180–6190.
[57] Liping Wang and Steve Greenberg. 2015. Window operation and impacts on building energy consumption. *Energy and Buildings* 92 (2015), 313–321.
[58] Tianshu Wei, Yanzhi Wang, and Qi Zhu. 2017. Deep reinforcement learning for building HVAC control. In *Proceedings of the 54th Annual Design Automation Conference 2017*. 1–6.
[59] Michael Wetter. 2011. Co-simulation of building energy and control systems with the building controls virtual test bed. *Journal of Building Performance Simulation* (2011).
[60] Daniel A. Winkler, Alex Beltran, Niloufar P. Esfahani, Paul P. Maglio, and Alberto E. Cerpa. 2016. FORCES: Feedback and control for occupants to refine comfort and energy savings. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1188–1199.
[61] Kang Yang, Yuning Chen, Xuanren Chen, and Wan Du. 2023. Link quality modeling for LoRa networks in orchards. In *Proceedings of the 22nd ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN'23)*.
[62] Kang Yang and Wan Du. 2022. LLDPC: A low-density parity-check coding scheme for LoRa networks. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys'22)*.
[63] Zhiang Zhang and Khee Poh Lam. 2018. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In *Proceedings of the 5th Conference on Systems for Built Environments*. 148–157.
[64] Jie Zhao, Khee Poh Lam, and B. Erik Ydstie. 2013. EnergyPlus model-based predictive control (EPMPC) by using MATLAB/SIMULINK and MLE+. (2013).

Received 8 September 2023; revised 31 January 2024; accepted 24 March 2024