On the Joint Control of Multiple Building Systems with Reinforcement Learning

Tianyu Zhang*
University of Alberta
tzhang6@ualberta.ca

Gaby Baasch*
University of Victoria
gaby.baasch@gmail.com

Omid Ardakanian
University of Alberta
ardakanian@ualberta.ca

Ralph Evins*
University of Victoria
revins@uvic.ca

ABSTRACT

Commercial buildings are comprised of multiple mechanical and electrical systems that work in tandem to provide a healthy, safe, and comfortable environment for occupants. These systems have complex interactions with each other, and consume a large amount of energy. In this paper, we apply three model-free deep reinforcement learning algorithms to jointly control HVAC and blind systems in a multi-zone test building, in scenarios with and without automatic dimming of the lights in response to daylight levels. The control agents are trained through interactions with a building simulator that generates traces for the movement of occupants. We investigate the three-way trade-off between energy use, thermal comfort, and visual comfort, and discuss how the joint control of the building systems could provide a better trade-off compared to when they are controlled separately. We compare the performance of the proposed control algorithms assuming the availability of occupancy data with two spatial resolutions, and confirm through experiments that a better trade-off can be achieved should zone-level occupancy information become available. Incorporating zone-level occupancy information, we show that 11.0% and 31.8% more energy can be saved respectively in heating and cooling seasons over existing rule-based baselines that control the same building systems.

CCS CONCEPTS

• **Computing methodologies** → **Simulation evaluation**; • **Theory of computation** → **Reinforcement learning**.

KEYWORDS

Deep Reinforcement Learning, Building Controls, Energy Efficiency.

ACM Reference Format:
Tianyu Zhang, Gaby Baasch, Omid Ardakanian, and Ralph Evins. 2021. On the Joint Control of Multiple Building Systems with Reinforcement Learning. In *The Twelfth ACM International Conference on Future Energy Systems (e-Energy '21), June 28–July 2, 2021, Virtual Event, Italy*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3447555.3464855

*Both authors contributed equally to this research.

e-Energy '21, June 28–July 2, 2021, Virtual Event, Italy

ACM ISBN 978-1-4503-8333-2/21/06…$15.00
https://doi.org/10.1145/3447555.3464855

1 INTRODUCTION

Commercial buildings consume a significant amount of energy worldwide. The commercial sector in the United States used approximately 9.3 quadrillion British thermal units (Btu) in 2019, accounting for 12% of the country's delivered energy consumption [3]. About 40% of this energy is used for space heating, cooling, ventilation, and lighting by two major building systems: Heating, Ventilation, and Air Conditioning (HVAC) and lighting. In Canada, a country with a colder climate, these two systems account for more than 70% of the commercial building energy use [5]. This energy is primarily used to enhance Indoor Environmental Quality (IEQ), which is determined by several factors including air quality, and thermal and visual comfort.

Extensive research has been done in the past few decades to explore how to save on the energy used by building systems, while maintaining thermal and visual comfort of occupants. In particular, various rule-based, model-based, and model-free control techniques have been employed to obtain energy-efficient operation policies for HVAC, lighting, and blind systems. Rule-based techniques are based on a set of control rules defined by the facilities manager. Model-based techniques take advantage of a physics-based or data-driven dynamic model that explains the state evolution (e.g., heat transfer, air flow, occupant movement), whereas model-free techniques aim to learn a control policy through interactions with building systems or a simulated environment. Model-free techniques are more promising when the goal is to control multiple building systems with complex interactions that cannot be precisely modelled [11].

Despite the tremendous progress toward energy-efficient control of building systems, there are several important questions that are yet to be addressed. We outline these research questions below:

(a) **How does the joint control of building systems affect the whole-building energy use?** Due to the complex interactions between building systems, the control decisions made in one system could affect the performance of the other ones. For example, closing blinds in an overheated zone may reduce the energy use of the HVAC system during the day, but this comes at the price of increasing the energy use of the lighting system because lights must be turned on to satisfy the visual comfort requirement. Dimming lights, on the other hand, reduces the amount of energy used for lighting but it may also change the HVAC energy consumption as it influences the internal heat gain. Modelling interactions between building systems in addition to the uncertainty of the environment is indeed a difficult task. To contain complexity, related work either controls a single building system [6, 22, 39], neglecting the interplay between this system and the other systems, or considers the interactions between two or more systems in a single zone [7, 8, 11]. To our knowledge, there is no work that quantifies the amount

of energy that can be saved in a multi-zone building when building systems are controlled jointly.

(b) **What are the best trade-offs between energy use, thermal comfort, and visual comfort?** The trade-off between energy use and thermal comfort has been widely studied in the context of optimal HVAC control. However, there is little work that navigates the three-way trade-off between energy saving, thermal comfort, and visual comfort. This is a barrier to the deployment of the control techniques in real buildings as the facilities manager cannot easily trade energy savings for extra comfort (and vice versa). Ideally, they should be able to tweak some weight parameters to make trade-offs within Pareto-efficient choices.

(c) **Will incorporating zone-level occupancy information noticeably change the performance of a control policy?** To make possible higher energy and cost savings without compromising comfort, most control techniques incorporate occupant presence or count information at the building level. This makes sense because estimating the number of occupants in each zone is difficult without having a number of sensors installed there [16, 37]. Should this information become available, the thermal and visual discomfort can be calculated for each individual occupant that is present in a given zone. However, it is unclear whether incorporating high spatial resolution occupancy data could help achieve a better trade-off between energy consumption, thermal comfort, and visual comfort.

(d) **How does the performance of a given control policy vary across seasons?** The outside air temperature can affect the performance of optimal HVAC control algorithms, so studies often train agents for more than one season. For the joint control of building systems, the difference between seasons can become even more prominent. For example, to improve thermal comfort and reduce energy use, blinds should be open during the day in winter to heat up the building, and vice versa in the summer. Understanding how the control performance varies across seasons and whether there is a specific model-free control algorithm that outperforms others in all cases requires a comprehensive evaluation of building controls in both heating and cooling seasons. This has not been explored in previous work and will provide insight into control strategy selection.

To address these questions, this paper studies the joint control of HVAC, lighting, and blind systems in a five-zone test building modelled in EnergyPlus [9]. We use state-of-the-art deep reinforcement learning (RL) algorithms to determine the optimal control policies for HVAC and blinds, while a daylight auto-dimming strategy is used for lighting control. These algorithms are suitable for this problem because they can handle large state and action spaces, and learn the complex interactions between multiple building systems[1]. We make three specific contributions:

- We utilize three model-free RL algorithms to train agents that can jointly control the supply air temperature and blind angle setpoints for every zone in our test building. These include two actor-critic algorithms, namely Proximal Policy Optimization (PPO) and Soft Actor Critic (SAC), and a Q-learning algorithm, called Branching Dueling Q-Network (BDQN). We evaluate these algorithms in different scenarios in terms of the achieved reward, convergence speed, and stability, following the guidelines provided in [35].
- We investigate the three-way trade-off between energy consumption, thermal comfort, and visual comfort. We discuss the best weight factors for the terms in the reward function; these weights will allow for maximizing energy savings while keeping thermal and visual discomfort below specified thresholds.
- We compare the performance of these algorithms with existing baselines in heating and cooling seasons with building-level and zone-level occupancy information. We show that the energy use would be further reduced if we knew the occupancy state of all zones in a building. This highlights the importance of monitoring or estimating the occupancy state of every zone through multimodal sensor fusion.

[1]Code is available at https://github.com/sustainable-computing/COBS-joint-control

2 RELATED WORK

Numerous attempts have been made to optimally control HVAC, lighting, shading, and other building systems. The control strategies can be broadly divided into three categories: rule-based, model-based, and model-free. Table 1 shows example control strategies from each category. Regardless of which control strategy is adopted, occupancy information can be incorporated in the control loop to achieve an acceptable trade-off between energy savings and occupant comfort.

In the rule-based approach, control rules and schedules are defined by the facilities manager based on their intuition about how the building occupancy varies over time. It is shown in [4] that using static per-zone schedules can considerably reduce the energy consumption of HVAC. In another study [28], it is shown that a rule-based lighting controller can lower the building energy use by up to 12% without negatively affecting the visual comfort of occupants. Rule-based controllers are easy to implement and do not require training complex models, but their performance is highly dependent on the quality of the rules. In practice, the control performance degrades over time with changes in the occupancy schedule and outside air temperature.

In the model-based approach, dynamic models for heat transfer, occupancy, and different components of building systems are utilized in the control loop to minimize the energy use over a time horizon subject to a set of constraints. While a high-order heat transfer model can accurately determine the temperature of every zone in the building, proper identification of this model is difficult. Alternatively, low-order thermal models can be built using a data-driven approach if enough training data is available [17, 40]. These models have proven to be useful for Model Predictive Control (MPC), lowering the energy consumption of the HVAC system while maintaining thermal comfort [25, 34]. Model-based reinforcement learning techniques have recently been proposed to optimize HVAC operation [12, 36]. The basic idea is to learn the system dynamics using a neural network. This neural network is then used to solve an MPC problem. While model-based HVAC control strategies have great performance, explaining interactions between multiple building systems requires more complex models which cannot be easily trained, especially in a building with heterogeneous spaces.

Table 1: A representative subset of related work for different occupancy levels, control points and methods. For similar control scenarios, more recent studies were chosen.

	Thermal Zones	Occupancy	Control Variables	Control Method
[19]	4 Zones	Building-level (Binary)	Temp. setpoint	Rule-based
[28]	20 Zones	Room-level (Count)	Lights (on/off) Blinds (angle)	Rule-based
[25]	1 Zone	N/A	Temp. setpoint	MPC
[31]	3 Zones	Building-level (Estimated Count)	Temp. setpoint	Model-based
[14]	4 Zones	Building-level (Binary)	Temp. setpoint	MPC
[7]	1 Zone	Building-level (Binary)	HVAC (on/off) Window (open/closed)	Q-learning
[8]	1 Zone	Building-level (Binary)	Lights (on/off) Blinds (step changes in angle)	Q-learning
[23]	1 Zone	Not mentioned	HVAC (heating/cooling power) Window (open/closed) Door (open/closed)	Deep Q-learning
[11]	1 Zone	Building-level (Count)	HVAC setpoint Light (dimming level) Blinds (angle) Windows (open pct.)	BDQN
[26]	1 Zone	N/A	HVAC on/off	DDPG
[15]	1 Zone	N/A	Temp. setpoint Humidity setpoint	DDPG
[6]	5 Zones	Room-level (Binary)	SAT setpoint	PPO

Building systems can also be controlled using a model-free approach. In recent years, model-free reinforcement learning algorithms have been applied to address the optimal control of the HVAC system [33]. Instead of relying on a built-in thermal model, they provide the opportunity for trial-and-error learning through direct interactions with building systems or an external simulated environment. There are three main types of model-free RL algorithms, namely Q-learning (value-based), actor-critic, and policy gradient methods. The Q-learning algorithm updates action values (i.e., Q-values) for each state based on the observation. It is generally more sample efficient than other model-free RL algorithms. Actor-critic methods are also adopted to control the HVAC system. They learn the control policy as well as the Q-values to update the control policy. Policy gradient algorithms are considered the least sample efficient model-free RL algorithms, yet there are usually more stable than the other RL algorithms. Of the 77 papers that applied RL to building controls and were reviewed in [33], three-quarters (59) used valued-based methods, some (12) used actor-critic methods, and a few (3) used policy gradient approaches. (The remaining 3 were model-based approaches.) Despite the large number of reinforcement learning algorithms that are used in the building control domain, they are seldom compared in terms of their performance, stability, and convergence speed.

Joint control of building systems: The whole building energy consumption can be further reduced when building systems are jointly controlled compared to when only HVAC is controlled [11]. Still, the joint control of multiple building systems is challenging because it increases dimensions of state space and action space, and makes it harder to learn an optimal policy due to complex interactions between systems. Previous work focuses on zone-level energy optimization through the joint control of lights and blinds [8], HVAC and windows [7, 10], and all these four systems [11, 21]. The problem space becomes increasingly large as more zones are included; none of these studies address the joint control of building systems in a multi-zone building.

Incorporating occupancy information: Building occupancy is one of the main factors that determine its energy consumption. According to the 2012 Commercial Buildings Energy Consumption Survey (CBECS) [2], a building that was occupied for all 168 hours in a week consumed 46% more energy per square foot than a building that was occupied for only 80 hours in a week. Incorporating more occupancy information can help to optimize control policies. For example, Turley et al. [30] evaluate the energy efficiency and human comfort with different occupancy patterns using MPC, and shows that incorporating the number of occupants in every room is essential for higher energy savings. Unfortunately, such occupancy data is hard to collect at scale, so most previous studies incorporate binary occupancy information (occupied/vacant) at building or floor-level. In this paper, we study both building-level and zone-level occupancy schedules, and examine if it makes sense economically to install occupancy sensors in every zone.

Novelty of this work: As shown in Table 1, no previous work covers the following comprehensive exploration of this research area:

- We control the window blinds, lights[2] and HVAC in a commercial building.
- We focus on a multi-zone building with 5 zones.
- We explore how reward parameters must be tuned to navigate the three-way trade-off between energy use, thermal comfort, and visual comfort.
- We provide a comprehensive comparison of RL-based and rule-based control strategies.
- We evaluate the performance of each control algorithm with both building-level and zone-level occupancy information.

3 BACKGROUND OF BUILDING SYSTEMS

Commercial buildings are controlled by mechanical and electrical systems, such as HVAC, lighting, and shading. The HVAC system consumes a considerable amount of energy to provide heated, cooled, and conditioned air to occupants, thereby maintaining comfortable and healthy indoor conditions. This along with the fact that buildings can store heat due to their thermal mass and have different spatio-temporal occupancy patterns makes the HVAC control problem important and nontrivial. Lights, on the other hand, are often controlled using a reactive strategy because illuminance will change immediately after a control policy is implemented.

Figure 1 offers a generic visualization of a typical HVAC system for a medium size office building. It consists of a centralized air handling unit (AHU) that moves conditioned air through the building via a duct system. In the AHU, the outside air and the return air from zones are mixed together. The mixed air is then heated or cooled to a specified temperature before it is pushed through the

[2]Our RL agents do not control lighting directly. Rather, rule-based (auto) dimming is utilized for lighting control.

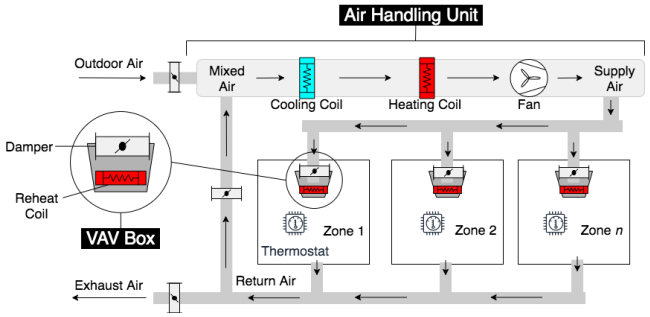


Figure 1: A diagram of an air handling unit (AHU) feeding a number of variable air volume (VAV) systems at terminal zones. The components depicted here match the components of the HVAC system that we consider in this work.

duct system by a fan. In larger office buildings, multiple AHUs may be required. Variable air volume (VAV) systems are often used in office buildings because they allow for zone-specific control with a single AHU. A terminal VAV box exists in each zone. It is responsible for controlling the amount of supply air by opening and closing a damper. A reheat coil may be present in the VAV box to heat the supply air in the zone to the desired temperature. This allows each zone to have its own thermostat with a unique temperature preference, which is often expressed by a thermal comfort range in which no corrective action needs to be taken by the VAV controller. The HVAC can be controlled at the system level, e.g., using the supply air temperature (SAT) setpoint [6], or at the zone level using the thermostat temperature setpoints [11] or mass flow rate setpoint. The HVAC energy consumption is the total energy consumed by VAV systems, AHU heating and cooling coils, and the fan.

Auxiliary building systems, such as lighting and blinds, also have a large effect on occupant comfort and whole-building energy use [8]. The lighting system affects the building energy use, the visual comfort of occupants, and to a lesser extend the thermal comfort of occupants as lights produce heat when they are on. It may consist of dimmable or non-dimmable lights which are located in different building spaces. Dimmable lights are normally controlled using a reactive strategy. They can be dimmed linearly between the maximum and minimum light outputs according to the available daylight measured at some point in the zone. In a simulation, the daylight illuminance is calculated based on cloud Blinds are usually mounted on the inside of windows and consist of a series of equally-spaced slats that are oriented horizontally. The blind controller can change the slat angle from 0 to 180 degrees. By controlling the blind angle, it is possible to change the ratio of direct and diffuse solar radiation passing through the blind. Opening or closing blinds thus changes the amount of heat gain and the illuminance level, thereby affecting both visual and thermal comfort conditions.

Lights, blinds, and HVAC systems have complex interactions. As explained, opening blinds during the day will influence the interior daylight illuminance, providing natural lighting and heating up the zone due to the solar radiation. If the zone temperature goes above the desired zone temperature, the HVAC system will supply more cool air to the zone, affecting the total energy consumption of the HVAC system. On the other hand, switching on the lights

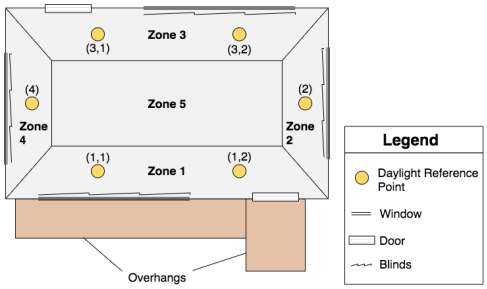


Figure 2: The layout of the medium office building studied in this work, including the daylighting reference points.

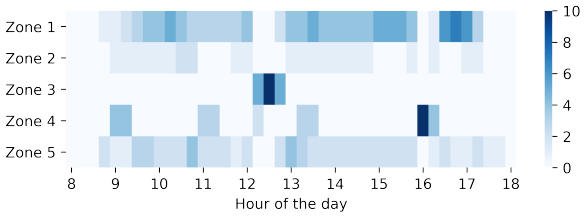


Figure 3: The number of occupants in each zone during working hours.

in a zone will increase illuminance and energy use at the same time. Thus, there are many ways to navigate the three-way trade-off between the energy use, thermal comfort, and visual comfort. While thermal and visual comfort requirements can be satisfied when these systems are controlled independently, this comes at the price of increased energy consumption. The joint control of building systems enables finding a better trade-off between the energy use, and thermal and visual comfort.

4 METHODOLOGY

This paper explores rule-based and RL-based joint control of the building systems for a 5-zone office building in Pittsburgh, Pennsylvania in January (heating season) and July (cooling season). The floor area of this building is 5,000 square feet and it has been used in previous work [6, 28]. The control setpoints that we adjust using different algorithms are supply air temperature setpoint and blind angle setpoint. The building is simulated in the EnergyPlus [9] environment and is controlled via the COmprehensive Building Simulator (COBS) [38] which interacts with EnergyPlus. COBS is used in this work to programmatically execute rule-based control scenarios and to train the RL agents. The office building we control is depicted in Figure 2, and the relevant design details are described in the following sections. We now present our control scenarios, problem formulation, and model-free RL algorithms.

4.1 Simulation Environment

4.1.1 HVAC Design. As shown in Figure 1, the HVAC system is a packaged VAV system with one heating coil and one cooling coil in addition to VAV reheat coils. Similar to [6], we use the SAT setpoint as the HVAC control point. Other VAV setpoints are controlled using a feedback control strategy. The VAV reheat coils are turned off in the cooling season.

Table 2: Control scenarios and corresponding baselines.

HVAC	Blinds	Lighting	Baseline
SAT setpoint	Always open	Not controlled	(1)
SAT setpoint	Always open	Auto dimming	(3)
SAT setpoint	Using the same setpoint	Not controlled	(2)
SAT setpoint	Using the same setpoint	Auto dimming	(4)
SAT setpoint	Using different setpoints	Not controlled	(2)
SAT setpoint	Using different setpoints	Auto dimming	(4)

4.1.2 Occupancy. Two occupancy conditions are considered, namely building-level and zone-level occupancy schedules. The building-level occupancy schedule assumes that all zones are occupied from 8:00 and 18:00. The zone-level occupancy schedule determines the number of occupants that are present in each zone at any given point in time. This helps model the amount of heat emitted by occupants and better assess thermal and visual comfort of the occupants in each zone. We use the COBS platform to generate several zone-level occupancy schedules. Figure 3 illustrates the number of occupants in each zone throughout a day.

4.1.3 Window Blinds. White painted metal blinds are present on the windows in all four perimeter zones. Each slat is 2.5 cm wide and the separation between slats is 1.875 cm. We assume that windows are not motor-operated, hence the blind angle and position can be adjusted without any constraints.

4.1.4 Daylighting. Zonal illuminance values are used for rule-based lighting control and to evaluate the visual comfort of the occupants. They are measured at desk height (76.2cm) using daylighting reference points in EnergyPlus, the positions of which are specified in Figure 2. Zone 5 does not have any daylighting reference points because auto dimming does not occur in zones without windows. We turn on the lights when Zone 5 is occupied and turn them off when occupancy is zero.

4.2 RL Problem Formulation

In this section we describe the Markov decision process (MDP) framework, including state and action spaces, and the reward function. At each time step, the building and its surrounding environment are in some state s_t. The agent exerts a control action a_t to control building systems. This action causes a random state transition to s_{t+1}. The RL agents are trained in six specific scenarios for controlling HVAC, blind and lighting, as outlined in Table 2. Through interactions with the simulated environment, each agent learns an optimal *policy* π, that is a sequence of control actions starting from state s. When blinds are controlled, the agent either learns a policy that adjusts all the blind setpoints in the same way, or a policy that adjusts them independently. Note that lighting is not controlled by the RL agent. Hence, there is either no lighting control or the auto dimming strategy is adopted.

4.2.1 State. The state at time t, denoted by s_t, consists of the following observations: the temperature in each zone including the plenum ($^\circ C$), the number of occupants in each zone (for building-level occupancy schedule, all zones share the same value of 0 or 1, indicating whether the building is occupied), the hour of the day (0-24), the slat angle of the blinds (degrees) in each of the four zones that have windows, the ambient temperature ($^\circ C$), and the site solar radiation (W). In addition to these observations, it contains the ambient temperature and site solar radiation for the next twelve 15-minute time steps. These forecast are assumed to be perfect. Thus, each state consists of 18 observations and 24 predicted values.

4.2.2 Action. The action at time t, denoted by a_t, determines the control decision made in each building system. The action space differs depending on the control scenario and the agent type. The control scenario determines the number of control points while the agent type affects the range of possible actions pertaining to a control point. We always control the SAT setpoint for each control scenario in a range of $[-20, 20^\circ C] + T_{MA}$, where the T_{MA} is the mixed air temperature. The blinds can be controlled with different setpoints, jointly according to the same setpoint, or not controlled at all; in the latter case it is assumed that blinds are not available in the building. The action for each blind is between 0 and 180.

The SAC agent (described next) considers a continuous action space for each control point, while other RL agents consider discrete actions[3]. In particular, we discretize the action for SAT setpoint to 20 and blinds to 18 evenly spaced values. Therefore, in the most complex control scenarios, where we control the blinds using different setpoints, the action space is 5-dimensional for the SAC agent and there are 2,099,520 ($18^4 \times 20$) possible actions for other agents. This large action space makes it difficult to find the optimal policy.

To effectively find the optimal policy with this large action space, we deploy a feature sharing neural network for each agent. That is, instead of having a large number of cells for all possible actions, after a few hidden layers we create multiple branches in the neural network. The number of branches is the same as the number of control points we have in each scenario. For example, we have 5 distinct branches when we have different setpoints for blinds (4 branches for blind setpoints and one for the SAT setpoint). The same idea was used in [11] to reduce the size of neural networks.

4.2.3 Reward. The reward function balances three competing objectives: the total facility energy consumption including both the HVAC system and lights (denoted by E), the occupant thermal comfort (denoted by T_c), and the occupant visual comfort (denoted by V_c). It can be written as follows:

$$R = -\rho_E Norm(E) - \rho_T Norm(T_c) - \rho_V Norm(V_c) \quad (1)$$

where ρ_E, ρ_T and ρ_V are weight factors (reward parameters) that represent the relative importance of different terms in the reward function. These parameters can take values from $\{0.1, 0.4, 0.7, 1.0\}$. We consider all reward functions that are obtained by assigning these values to the parameters in a combinatorial fashion. The $Norm()$ function is defined as:

$$Norm(x) = (x - x_{min})/(x_{max} - x_{min}). \quad (2)$$

It is used to scale each term in the reward function. The process used to calculate E, T_c and V_c is described next.

Energy Consumption: Since both HVAC and lighting systems run on electricity only in our test building, we use the total electricity consumed by HVAC and lighting systems as a measure of the total facility energy use:

$$E = E_{HVAC} + E_L \quad (3)$$

[3]We got better results when we discretized the action space for the other two agents.

where E_{HVAC} is the electricity consumed by the HVAC system and E_L is the electricity consumed by the lights located in zones that have windows[4] (both are expressed in Wh). Note that we do not take into account the energy consumed to operate the blinds since it is negligible compared to the other components.

Thermal Comfort: The occupant thermal comfort is calculated according to the Predicted Mean Vote (PMV) specified by Fanger's model [13], which has been used in building control since the 1960s. The PMV predicts the average vote of a group of people on a 7-point index ranging from +3 = hot to −3 = cold. Both the ISO 7730 standard [1] and ASHRAE [24] recommend maintaining $|PMV|$ below 0.5. Thus, we calculate T_c at a given time step as follows:

$$T_c = \frac{\sum_i O_i \cdot T_{ci}}{\sum_i O_i} \tag{4}$$

where T_{ci} represents the thermal comfort in zone i given by:

$$T_{ci} = \begin{cases} 0, & |PMV_i| \leq 0.5 \\ |PMV_i| - 0.5, & otherwise. \end{cases} \tag{5}$$

PMV_i and O_i indicate respectively the PMV value and occupancy state of zone i. O_i is 1 when zone i is occupied and 0 otherwise.

Visual Comfort: In this paper, visual comfort is determined using the illuminance rates at the daylighting reference points (see Figure 2). A penalty is applied when the illuminance rates either do not meet or exceed engineering standards for visual comfort. According to the Illuminating Engineering Society of North America, the comfort range for office lighting is between 300 lux and 750 lux [20]. Thus, the visual comfort reward for zone i is given by

$$V_{ci} = \begin{cases} 0 & 300 \leq \mathbb{E}[I_i] \leq 750 \\ 300 - \mathbb{E}[I_i], & \mathbb{E}[I_i] < 300 \\ \mathbb{E}[I_i] - 750, & \mathbb{E}[I_i] > 750, \end{cases} \tag{6}$$

where $\mathbb{E}[I_i]$ is the expected illuminance rate in zone i. The illuminance values, I_i, are obtained from the daylighting reference points labelled in Figure 2. We take the average of the illuminance values of the reference points in each zone and denote it by $\mathbb{E}[I_i]$. Notice that the illuminance value will never fall below 300 lux during the occupancy time as the indoor artificial light can always provide enough illuminance when they are on. Then, the total visual reward V_c is calculated as follows:

$$V_c = \frac{\sum_i O_i \cdot V_{ci}}{\sum_i O_i}. \tag{7}$$

4.3 Deep Reinforcement Learning Algorithms

We use three model-free RL algorithms to control building systems.

Soft Actor-Critic (SAC) is an actor-critic based off-policy maximum entropy RL algorithm with a stochastic actor [18]. It maximizes both the expected reward and the entropy, allowing the agent to explore more widely and simultaneously consider multiple near-optimal policies. It is shown to have stable performance, and be robust to noise and the choice of hyperparameters. The state value function, soft Q-function, and policy are trained by optimizing:

$$\mathcal{J}_V(\psi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi}\left[Q_\theta(s_t, a_t) - \log \pi_\phi(a_t|s_t)\right]\right)^2\right]$$

$$\mathcal{J}_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(s_t, a_t) - R_t + \gamma \mathbb{E}_{s_{t+1} \sim p}[V(s_{t+1})]\right)^2\right]$$

$$\mathcal{J}_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}}\left[\log \pi_\phi\left(f_\phi(\epsilon_t; s_t)|s_t\right) - Q_\theta\left(s_t, f_\phi(\epsilon_t; s_t)\right)\right]$$

where π is the policy, ψ, θ, and ϕ are the parameters for state value function, soft Q-function, and policy, R_t is the reward for the (s_t, a_t) pair, γ is the discount factor, p is the state transition probability, $\mathcal{D}$ is the replay buffer, V is the state value, Q is the state-action value, and f_ϕ is the unbiased gradient estimator.

In this study, we use Adam optimizer with a learning rate of 0.0003. We set the discount factor to 0.99 and consider a batch size of 256. We use a squashed Gaussian policy with two hidden layers and 256 cells in each layer for the actor network. For the critic network, we use a network with two 256-cell hidden layers with the leaky rectified linear unit (ReLU) as the activation function. We use automatic entropy tuning which allows the agent to automatically balance exploitation and exploration.

Proximal Policy Optimization (PPO) is another state-of-the-art policy-gradient algorithm using the actor-critic framework [27]. In PPO, the step size is limited to a trust region [27]. This characteristic enables faster learning, but the agent might be trapped into a sub-optimal policy. PPO optimizes the clipped surrogate objective given by:

$$L(\phi) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\phi)\hat{A}_t,\ \text{clip}\big(r_t(\phi), 1-\epsilon_t, 1+\epsilon_t\big)\hat{A}_t\right)\right]$$

with $r_t(\phi) = \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{\text{old}}}(a_t|s_t)}$, where $\hat{A}_t$ is an estimator of the advantage function at t, ϵ is a hyperparameter that discourages making updates that are far from the current policy, and $\text{clip}\big(r_t(\phi), 1-\epsilon_t, 1+\epsilon_t\big)$ clips the probability ratio between old and new policies within $[1-\epsilon_t, 1+\epsilon_t]$.

In this study, we use two hidden layers with 100 units each layer, utilizing the leaky ReLU activation function for both actor and critic networks. After two hidden layers, the actor network has multiple branches, one for each actuator type. We set the learning rate to 0.0005 and the discount factor to 0.99.

Branching Dueling Q-Network (BDQN) is a branching variant of the dueling double deep Q-network [32]. It is an off-policy algorithm which is shown to outperform various algorithms such as Deep Deterministic Policy Gradient (DDPG) in high dimensional action spaces tasks [29]. For comparison with previous work, we use exactly the same settings that are used in [11]. The Q-value for each branch d and the maximum accumulated reward can be written as:

$$Q_d(s, a_d) = V(s) + \left(A_d(s, a_d) - \frac{1}{n}\sum\nolimits_{a'_d \in \mathcal{A}_d} A_d\left(s, a'_d\right)\right)$$

$$R_d = R + \gamma \frac{1}{N}\sum\nolimits_d Q_d\left(s', \arg\max_{a'_d \subseteq \mathcal{A}_d} Q_d\left(s', a'_d\right)\right)$$

where $\mathcal{A}_d$ is the set of actions that can be taken on branch d, and A_d represents the advantage function.

4.4 Training RL Agents

We split the task into two seasons: winter and summer. The winter season model only uses January data to train and test, and the summer season model only uses July data to train and test. We assume that each episode is one month long and is comprised of

[4] We ignore the electricity consumption of lights in Zone 5, which does not have a window, since we cannot affect this energy consumption.

2,976 15-minute time steps. We use 400 episodes to train the RL agents in each season. EnergyPlus is used to simulate the building environment after each epoch. We use the historical weather data in Pittsburgh to get the outdoor temperature and solar radiation for the current time step and future predictions.

5 EVALUATION METRICS AND BASELINES

We evaluate the RL agents in six different control scenarios and compare their performance with four existing baselines. Four metrics are used for performance evaluation: the total electricity consumption of the month, average thermal comfort over the month, thermal comfort violation rate of the month, and visual comfort violation rate of the month. The thermal comfort violation rate is defined as the percentage of time that the absolute value of PMV averaged over all occupied zones is greater than 0.5 when the building is occupied. We define the visual comfort violation rate similarly.

We consider four rule-based baselines that are implemented in EnergyPlus for each season: (1) HVAC only, (2) HVAC & blinds, (3) HVAC with auto-dimming, and (4) HVAC & blinds with auto-dimming. The performance of the RL agents is compared to the respective rule-based baselines based on the control scenario (see the last column of Table 2).

HVAC: For all baselines, the supply air temperature is controlled by EnergyPlus using the SetpointManager:Warmest/Coldest object that attempts to meet the heating load for multiple zones at a time. Details of the control strategy can be found in the EnergyPlus documentation[5]. In short, the setpoint manager calculates the average SAT that is required to meet the zones' heating/cooling loads based on the supply air mass flow rates, and adjusts the SAT setpoint accordingly.

Blinds: When blind control is included, predefined EnergyPlus programs that are designed to reduce heating and cooling load are used. Specifically, the blinds are closed in the heating season if it is nighttime and the outdoor temperature is below a setpoint. In the cooling season, the blinds are kept open at night, and closed during the day only if the solar radiation on the window exceeds a setpoint. The setpoints were chosen by trying a wide range of values to find the ones that performed the best in terms of the whole-building energy use, and thermal and visual comfort.

Daylighting: When lighting control is included, the Daylighting:Controls object is used so that the overhead lights dim continuously as the daylight illuminance increases[6]. The lights are always turned off during the night with and without auto-dimming.

6 RESULTS

In this section we first evaluate the performance of the four baselines and three RL-based control strategies. We present the trade-offs between whole-building energy consumption, thermal comfort, and visual comfort, and discuss which agent yields better trade-offs for each control scenario. We discuss the best trade-off that can be achieved using each RL algorithm and compare them with rule-based baselines. Finally, for a fixed set of the reward parameters, we explain how incorporating zone-level occupancy information would impact the trade-off curves in both heating and cooling seasons, and compare the control agents in terms of the reward they eventually achieve, their convergence speed, and stability across several random runs.

6.1 A Closer Look at Baseline Strategies

We analyze the performance of the four baselines presented earlier; they are rule-based control strategies that incorporate occupancy information. The black stars in Figure 4 show the performance of these baselines in respective control scenarios in both heating and cooling seasons with different occupancy schedules. To save space, we only discuss the results obtained when zone-level occupancy information is incorporated. Numerical values are provided in Table 3 in the appendix. In the cooling season, using rule-based controllers for HVAC and blinds (Baseline 2) or using auto-dimming in addition to rule-based HVAC control (Baseline 3) reduces the total energy consumption by 12% and 28% compared to Baseline 1 which controls HVAC only. Controlling HVAC and blinds with auto-dimming (Baseline 4) yields 32% more savings than controlling HVAC alone (Baseline 1) and around 5% more savings than controlling HVAC with auto-dimming (Baseline 3). This is because blinds can reduce the solar heat gain during the daytime and provide sufficient natural lighting, thereby lowering the energy use.

Controlling blinds and HVAC with a rule-based strategy (Baseline 2) in the heating season also helps reduce the total energy use by 15% over Baseline 1. Yet, unlike the cooling season, adding auto-dimming to Baseline 1 does not reduce the energy use. This is likely because lighting gives off excess energy as heat, hence turning off the lights results in higher heating requirements from the HVAC system. Controlling HVAC and blinds together with auto-dimming (Baseline 4) enables the highest energy savings in the heating season, i.e., 18% reduction in energy use over Baseline 1.

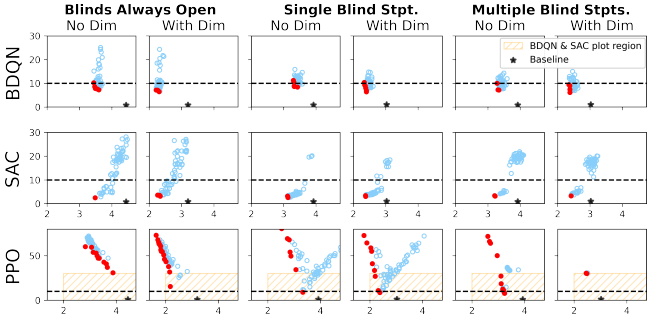
In conclusion, our results show an average energy savings of 26% across both seasons when all three systems are controlled (Baseline 4) compared to when only HVAC is controlled (Baseline 1). This observation motivates the joint control of building systems using more advanced control strategies. In terms of thermal comfort, all baselines were able to meet the ASHRAE PMV requirement. However, their performance is rather poor in terms of visual comfort because the default blind control strategy only closes the blinds at night. As a result, illumination is always high in the perimeter zones.

6.2 Identifying Three-way Trade-offs

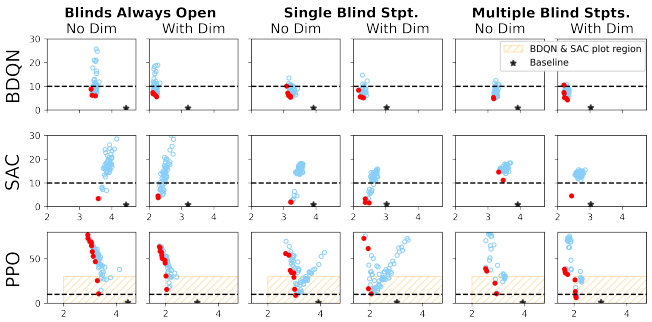
As described in Section 4.2.3, we assess the control performance of RL agents for various combinations of reward parameters $\rho_E, \rho_T, \rho_V \in \{0.1, 0.4, 0.7, 1.0\}$. Figure 4 shows the trade-offs between energy use and thermal comfort offered by the three RL agents in six different scenarios with two types of occupancy schedules. The visual comfort is the third dimension which is not shown in this figure. Each reward parameter setting yields a specific trade-off between the competing objectives, which is depicted by a circle in this figure. The Pareto optimal values are painted in red, and the baseline strategy for each scenario is marked with a black star. Notice that in the cooling season, the result for PPO spreads widely. Therefore, the

[5]Refer to https://bigladdersoftware.com/epx/docs/9-3/input-output-reference/group-setpoint-managers.html#setpointmanagerwarmest

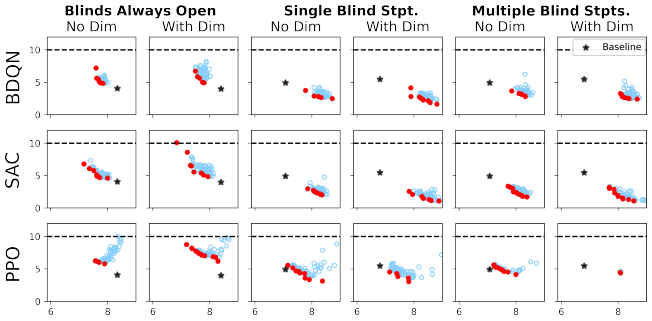
[6]The overhead lights dim linearly when the illuminance increases and stay on with the minimum input power if illuminance surpasses a certain threshold.



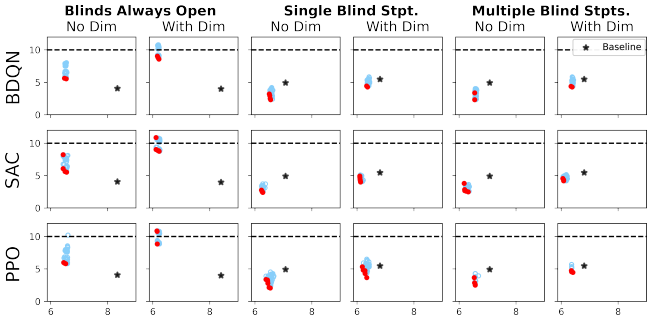
(a) Cooling season with building-level occupancy schedule



(b) Cooling season with zone-level occupancy schedule

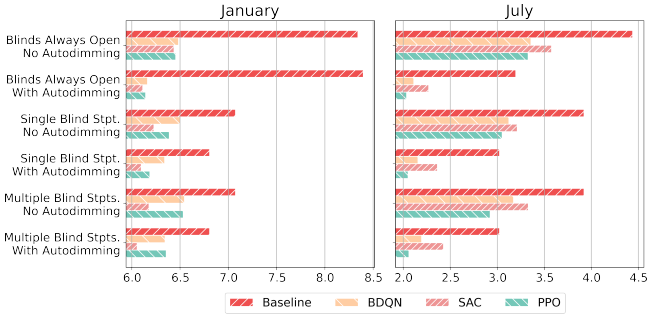


(c) Heating season with building-level occupancy schedule

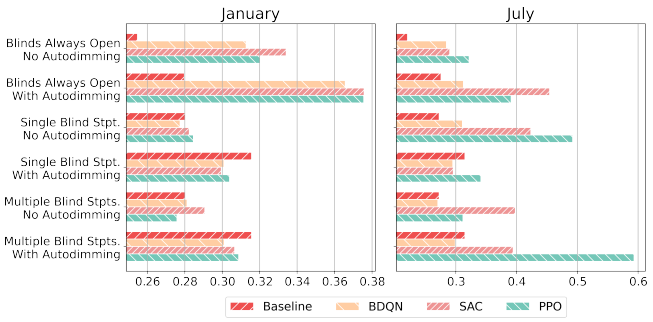


(d) Heating season with zone-level occupancy schedule

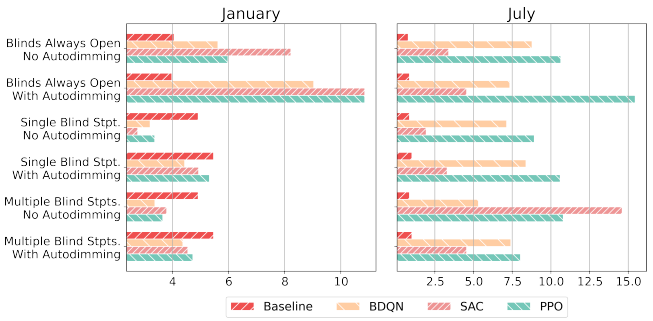
Figure 4: The PMV violation rate (y-axis) versus the monthly electricity consumption in MWh (x-axis) for different reward parameters. Points on the Pareto frontier are colored red and baselines are marked with black stars. The horizontal line shows ASHRAE's threshold (10%) for thermal comfort violation [24].



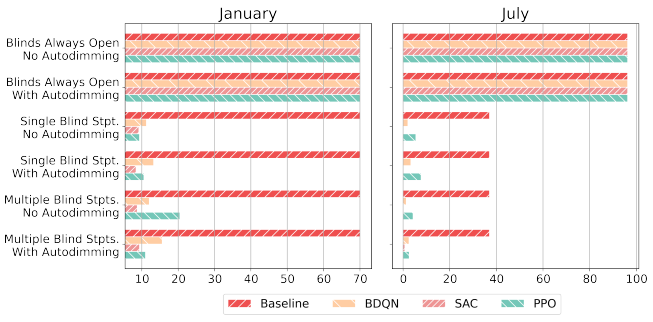
(a) Energy Consumption (MWh)



(b) Thermal Comfort ($|PMV|$)



(c) Thermal Comfort Violation (%)



(d) Visual Comfort Violation (%)

Figure 5: Comparison of different RL agents in different control scenarios using a zone-level occupancy schedule. The results are obtained using the best set of reward parameters for each RL agent. The x-axis is exaggerated.

axis limits for PPO are different from SAC and BDQN. Hatch-filled orange rectangles indicate the axis limits of SAC and BDQN plots on PPO plots. Among the three RL algorithms we considered, PPO seems to be the most sensitive to reward parameters. Nevertheless, we observe that for all three agents it is possible to navigate the three-way trade-offs by tweaking the reward parameters.

Best trade-offs. To determine the reward parameter setting that yields the 'best' trade-off, we first filter out the parameter settings that result in a PMV violation rate higher than 10% (ASHRAE's threshold [24]). We then choose the parameter setting that minimizes the whole-building energy use among the remaining choices. If the PMV violation rate exceeds 10% for all parameter settings, we choose the parameter setting that minimizes the product of the whole-building energy use and excess discomfort (i.e., the PMV violation rate minus 10%). The trade-off that corresponds to this parameter setting is called the best trade-off. For simplicity, the illumination violation rate is not considered in the process of finding the best trade-off as it is typically in the acceptable range.

Figure 5 provides a comparison between the best trade-offs achieved by each RL agent in different scenarios. Numerical values are provided in Tables 3 and 5 in the appendix. Compared to the baselines, the RL agents can save a significant amount of energy while meeting both thermal and visual comfort requirements in most cases. In scenarios where the blind setpoint is controlled, all agents achieve a significant improvement in visual comfort compared to the baselines in both seasons. This implies that the RL agents are able to learn how to use blinds to limit the amount of glare from sunlight. SAC has the lowest visual comfort violation rate in all scenarios. It is worth mentioning that in the scenario where the SAT setpoint and multiple blind setpoints are controlled with auto-dimming, the best RL agent can reduce the whole-building energy use by 11% in heating season and 31.8% in the cooling season over Baseline 4.

6.3 Incorporating Occupancy Information

We evaluate the control performance using both building-level and zone-level occupancy information. Figures 4a and 4c show the whole-building energy use in cooling and heating seasons along with the thermal comfort violation rate when the RL agents incorporate building-level occupancy information. Figures 4b and 4d show the same result this time assuming that the agents incorporate the zone-level occupancy information. It can be readily seen that better trade-offs can be achieved in the heating season when the control agents incorporate zone-level occupancy information. Specifically, BDQN, SAC, and PPO agents can save respectively 3.3%, 18%, and 14% more energy when they take into account zone-level occupancy information rather than building-level occupancy information. The zone-level occupancy information allows the control agents to meet thermal and visual comfort requirements by conditioning only a subset of zones that are occupied. This reduces the energy consumption in HVAC and lighting systems. Interestingly, incorporating zone-level occupancy information does not appear to offer much in terms of energy savings in the cooling season. We attribute this to the fact that in Pittsburgh less energy is consumed to keep the room temperature within the comfort range in the cooling season than in the heating season. Hence, a smaller amount of energy can be saved by not conditioning the unoccupied zones.

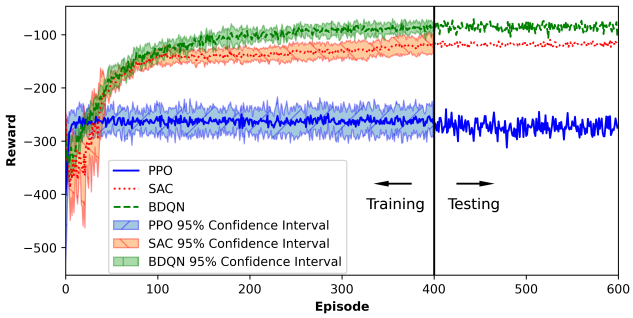


Figure 6: Performance comparison of three RL algorithms on the building control domain. The mean and 95% confidence interval of the episode reward are computed based on 10 independent runs in the cooling season.

Another important observation is that the RL agents cannot always beat the rule-based baselines when they rely on building-level occupancy information. For this reason, we only present the results when a zone-level occupancy schedule is used in the remainder of this section. The performance results for both cases can be found in the appendix (Tables 3-6).

6.4 Performance, Convergence Rate & Stability

Figure 6 illustrates the total reward accumulated in each episode when RL agents control the SAT setpoint and 4 blind setpoints, and lights are auto-dimmed. The episode reward is averaged across 10 runs with different random seeds. The shaded region around the average episode reward depicts the 95% confidence interval. The three RL agents are trained for 400 months and then tested over a period of 200 months in our simulated building. As it can be seen the agents have stable performance in the testing period.

From this figure it is evident that BDQN converges to the highest reward, followed by SAC. Looking at the convergence speed, PPO, SAC, and BDQN agents converge at around 30, 100, and 200 episodes, respectively. SAC and BDQN agents show more stable performance (narrower confidence interval) compared to the PPO agent. They have better sample complexity and can effectively use experiences from previous episodes to update the policy. Unfortunately this means that their running time is higher than PPO and they use more memory. In particular, SAC and BDQN agents finish a run in 38 and 31 hours respectively on a server with Intel Xeon E5-2650 v4 (2.2GHz CPU) and NVIDIA Tesla P100 GPUs, and need around 8GB of memory. For PPO, on the other hand, it takes only 7 hours to run on the same server using 4GB of memory.

While Figure 6 only shows the average reward per episode in the cooling season with zone-level occupancy information and a specific reward parameter setting ($\rho_E = 1$, $\rho_T = 1$ and $\rho_V = 0.4$), we witnessed similar convergence behavior for other reward parameter settings, months, and occupancy schedules.

7 DISCUSSION

We now return to the four research questions raised in the introduction, followed by a discussion of the key differences between the three model-free control strategies. Many of our findings are novel to this work and provide valuable insight for future research.

How does the joint control of building systems affect the whole-building energy use? Before our contribution, the paper that came closest to addressing this question is [11], where the BDQN agent achieved savings compared to rule-based methods, showing the potential of applying model-free RL to the joint control of building systems. However, their baselines included only rule-based HVAC control, even though rule-based blind and lighting strategies have been proven to offer significant savings. For example, in our building the rule-based control of all systems (Baseline 4) reduced the whole-building energy use by 26% on average across both seasons compared to the control of HVAC only (Baseline 1). Our work shows for the first time that RL-based control saves even more energy than rule-based HVAC and blind control, and that incorporating autodimming increases savings even further. Furthermore, we show that this is true even when generalized to the multi-zone scenario.

We provide numerical evidence in Table 3 that motivates the installation of dimmable lights and motorized blinds. Dimmable light can always lower the total energy consumption, especially during the summer, and blinds can slightly reduce the energy use as well. Figure 4 shows a minor improvement in controlling the blinds with a single setpoint over separate setpoints. This can be used to reduce the action space dimension, simplifying the problem.

What are the best trade-offs between energy use, thermal comfort, and visual comfort? The tradeoffs between energy use and thermal comfort are plotted directly in Figure 4. With regards to tuning these two objectives, BDQN and SAC are less sensitive to the reward parameters, whereas PPO is highly sensitive to the reward parameters. Interestingly, Figure 5 shows that all of the RL agents (including PPO) easily improved visual comfort over the rule-based baselines. One way to interpret this is that there is a lot of room for improvement in rule-based blind control strategies, with respect to visual comfort. Overall, visual comfort is relatively easy to optimize without much tuning, but the trade-off between energy use and thermal comfort is more complicated to navigate.

Will incorporating zone-level occupancy information noticeably change the performance of a control policy? As highlighted in Section 6.3, the inclusion of zone-level occupancy offered noticeable energy savings over building-level occupancy in all cases except for SAC in the cooling season. Figure 4 shows that when blinds are included in the heating season, zone-level occupancy is actually required to achieve lower energy use than the rule-based baseline which takes occupancy into account. Based on this result (and the simplicity in aggregating zone-level data up to the building-level) we argue that incorporating zone-level occupancy information to train RL agents is a viable energy reduction strategy.

Our back-of-the-envelope calculation shows that we can save approximately 1.04 MWh in two months (January and July) by incorporating zone-level rather than building-level occupancy information. With extrapolation, the annual energy and cost savings will be respectively 6.24 MWh and $437, assuming a flat rate of 7¢/kWh. This can offset the cost of buying and installing occupancy sensors in the 5 zones.

How does performance vary across seasons? A trend in RL papers for building control is to present results for two seasons and conclude that the agent can find an optimal control policy for both. Our results show that the reality is more complicated. Not only do the energy savings vary across the seasons, but so does the contribution of the building systems to the savings, the potential benefit from fine-grained occupancy data, and the relative performance of different model-free approaches. This is a conundrum for the practitioner who aims to implement RL in real buildings: if the performance varies drastically between seasons, how can one select a generalizable approach? This question warrants attention in future work.

Which RL algorithm works best? We designed a custom control system for multiple building systems using three popular deep reinforcement learning algorithms that can tackle problems with large state and action spaces. BDQN was adopted from previous work [11], where it was shown to have outstanding performance controlling multiple building systems of a single-zone building. To our knowledge, SAC and PPO were not previously applied to control multiple building systems.

We show here that SAC outperforms BDQN in the heating season (in all scenarios except one) with regard to energy savings. Considering thermal comfort, PPO is not able to satisfy the thermal comfort in the cooling season for most cases with average thermal comfort violation rate of 10.8%; SAC exceeds the threshold once and BDQN can always maintain the thermal violation rate under the threshold. Turning our attention to the effort needed to tune reward parameters, PPO is highly sensitive to these parameters, whereas BDQN and SAC are less sensitive to the reward parameters. Also, PPO converges remarkably faster than SAC, and SAC is slightly faster than BDQN. As the requirements might differ from case to case, there is no clear winner among these three RL agents. SAC and BDQN seem to offer more promising results if one can afford the one-time computation cost of training the agents.

8 CONCLUSION

This paper benchmarked multiple model-free reinforcement learning agents and baseline control strategies in a simulated multi-zone building with both zone and building-level occupancy schedules in winter and summer months. We evaluated the effect of controlling different building systems on whole-building energy consumption using different reward parameters, and provided useful insight for practitioners regarding how to make trade-offs within Pareto-efficient choices. Specifically, we showed better trade-offs can be achieved when RL agents rely on zone-level occupancy information rather than building-level occupancy information. We made two important observations when zone-level occupancy information was used by the agents. First, we found that 11.0% and 31.8% more energy can be saved respectively in heating and cooling seasons over existing rule-based baselines that control the same building systems. Second, we found that when lights are dimmed automatically and the RL agent jointly controls HVAC and blinds, the whole-building energy use can be reduced by up to 5.9% and 38.7% respectively in heating and cooling seasons over the case that the RL agent only controls the HVAC system.

In future work, we plan to explore the performance of RL agents when they control more components of building systems, such as the damper position, reheat coil, etc. We will explore the performance of agents in a more complex building with many more zones and investigate whether the agents need to be retrained after several months.

REFERENCES

[1] ISO 7730. 2005. *Ergonomics of the thermal environment-Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria.* International Organization for Standardization.

[2] U.S. Energy Information Administration. 2012. *Commercial buildings energy consumption survey.* U.S. Government Printing Office, Chapter Consumption & Expenditures.

[3] U.S. Energy Information Administration. 2020. *Annual Energy Outlook 2020.* U.S. Government Printing Office, Chapter Commercial Sector Indicators and Consumption.

[4] Omid Ardakanian et al. 2018. Non-intrusive occupancy monitoring for energy conservation in commercial buildings. *Energy and Buildings* 179 (2018), 311–323.

[5] Natural Resources Canada. 2019. *Energy Use Data Handbook, 1990 to 2017.* Natural Resources Canada.

[6] Bingqing Chen et al. 2019. Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19).* ACM, 316–325.

[7] Yujiao Chen et al. 2018. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings* 169 (2018), 195–205.

[8] Zhijin Cheng et al. 2016. Satisfaction based Q-learning for integrated lighting and blind control. *Energy and Buildings* 127 (2016), 43–55.

[9] Drury B. Crawley et al. 2001. EnergyPlus: creating a new-generation building energy simulation program. *Energy and Buildings* 33, 4 (2001), 319–331.

[10] K. Dalamagkidis et al. 2007. Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment* 42, 7 (2007), 2686–2698.

[11] Xianzhong Ding et al. 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 (BuildSys '19).* ACM, 326–335.

[12] Xianzhong Ding et al. 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 (BuildSys '20).* ACM, 50–59.

[13] P.O. Fanger (Ed.). 1970. *Thermal Comfort: Analysis and Applications in Environmental Engineering.* Danish Technical Press, Copenhagen, Denmark.

[14] P. M. Ferreira et al. 2012. Neural networks based predictive control for thermal comfort and energy savings in public buildings. *Energy and buildings* 55 (2012), 238–251.

[15] Guanyu Gao et al. 2019. Energy-Efficient Thermal Comfort Control in Smart Buildings via Deep Reinforcement Learning. (2019). arXiv:arXiv:1901.04693

[16] Shadan Golestan et al. 2018. Data-Driven Models for Building Occupancy Estimation. In *Proceedings of the Ninth International Conference on Future Energy Systems (e-Energy '18).* ACM, 277–281.

[17] Siddharth Goyal et al. 2011. Identification of multi-zone building thermal interaction model from data. In *2011 50th IEEE Conference on Decision and Control and European Control Conference.* IEEE, 181–186.

[18] Tuomas Haarnoja et al. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 1861–1870.

[19] Hao Huang et al. 2015. A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings. *Energy and Buildings* 97 (2015), 86–97.

[20] IES IESNA. 2000. *Lighting handbook.* Illuminating Engineering Society of North America, New York, USA.

[21] D. Kolokotsa et al. 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* 15, 5 (2002), 417–428.

[22] Srinarayana Nagarathinam et al. 2020. MARCO - Multi-Agent Reinforcement Learning Based COntrol of Building HVAC Systems. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems (e-Energy '20).* ACM, 57–67.

[23] Ivars Namatēvs. 2018. Deep reinforcement learning on HVAC control. *Information Technology and Management Science* 21 (2018), 29–36.

[24] American Society of Heating, Refrigerating, Air-Conditioning Engineers, and American National Standards Institute. 2004. *Thermal environmental conditions for human occupancy.* Vol. 55. American Society of Heating, Refrigerating and Air-Conditioning Engineers.

[25] Samuel Privara et al. 2011. Model predictive control of a building heating system: The first experience. *Energy and Buildings* 43, 2 (2011), 564–572.

[26] Zahra Rahimpour et al. 2020. Actor-critic learning for optimal building energy management with phase change materials. *Electric Power Systems Research* 188 (2020), 106543.

[27] John Schulman et al. 2017. Proximal policy optimization algorithms. (2017). arXiv:arXiv:1707.06347

[28] Eric Shen et al. 2014. Energy and visual comfort analysis of lighting and daylight control strategies. *Building and Environment* 78 (2014), 155–170.

[29] Arash Tavakoli et al. 2018. Action Branching Architectures for Deep Reinforcement Learning. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17222

[30] Christina Turley et al. 2020. Development and evaluation of occupancy-aware HVAC control for residential building energy efficiency and occupant comfort. *Energies* 13, 20 (2020), 5396.

[31] Arun Vishwanath et al. 2019. Experimental Evaluation of a Data Driven Cooling Optimization Framework for HVAC Control in Commercial Buildings. In *Proceedings of the 10th ACM International Conference on Future Energy Systems (e-Energy '19).* ACM, 78–88.

[32] Ziyu Wang et al. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, 1995–2003.

[33] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036.

[34] D.A. Winkler et al. 2020. OFFICE: Optimization Framework For Improved Comfort Efficiency. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN).* 265–276.

[35] David Wölfle et al. 2020. A Guide for the Design of Benchmark Environments for Building Energy Optimization *(BuildSys '20).* ACM, 220–229.

[36] Chi Zhang et al. 2019. Building HVAC Scheduling Using Reinforcement Learning via Neural Network Based Model Approximation. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19).* ACM, 287–296.

[37] Tianyu Zhang et al. 2019. ODToolkit: A Toolkit for Building Occupancy Detection. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy '19).* ACM, 35–46.

[38] Tianyu Zhang and Omid Ardakanian. 2020. COBS: COmprehensive Building Simulator. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '20).* ACM, 314–315.

[39] Zimu Zheng et al. 2018. Data Driven Chiller Sequencing for Reducing HVAC Electricity Consumption in Commercial Buildings. In *Proceedings of the 9th International Conference on Future Energy Systems (e-Energy '18).* ACM, 236–248.

[40] D.P. Zhou et al. 2017. Quantitative comparison of data-driven and physics-based models for commercial building HVAC systems. In *American Control Conference (ACC).* IEEE, 2900–2906.

A PERFORMANCE OF RL CONTROL AGENTS

A.1 With the zone-level occupancy schedule

Table 3: RL Agent performance results for different control scenarios using zone-level occupancy schedule.

		Blinds always open				Single blind setpoint				Multiple blind setpoints			
		No dimming		With Dimming		No dimming		With Dimming		No dimming		With Dimming	
		Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.
Energy (MWh)	Baseline	8.34	4.44	8.4	3.2	7.07	3.92	6.81	3.02	7.07	3.92	6.81	3.02
	BDQN	6.48	3.36	6.16	2.11	6.51	3.12	6.34	2.16	6.55	3.17	6.35	2.2
	SAC	***6.44***	3.58	***6.11***	2.27	***6.23***	3.21	***6.1***	2.36	***6.18***	3.33	***6.06***	2.43
	PPO	6.46	***3.33***	6.14	***2.04***	6.39	***3.05***	6.19	***2.05***	6.53	***2.93***	6.36	***2.06***
Thermal Comfort ($\|PMV\|$)	Baseline	***0.25***	***0.22***	***0.28***	***0.28***	***0.28***	***0.27***	0.32	0.32	***0.28***	***0.27***	0.32	0.32
	BDQN	0.31	0.28	0.37	0.31	0.28	0.31	***0.3***	***0.3***	0.28	0.27	***0.3***	***0.3***
	SAC	0.33	0.29	0.38	0.45	0.28	0.42	0.3	0.3	0.29	0.4	0.31	0.39
	PPO	0.32	0.32	0.38	0.39	0.28	0.49	0.3	0.34	0.28	0.31	0.31	0.59
Thermal Comfort Violation (%)	Baseline	***4.06***	***0.78***	***3.98***	***0.86***	4.92	***0.86***	5.47	***1.02***	4.92	***0.86***	5.47	***1.02***
	BDQN	5.62	8.78	9.04	7.35	3.21	7.14	***4.44***	8.39	***3.38***	5.32	***4.38***	7.41
	SAC	8.22	3.39	10.86*	4.55	***2.77***	1.94	4.93	3.29	3.8	14.61*	4.55	4.55
	PPO	5.98	10.64*	10.86*	15.46*	3.37	8.93	5.32	10.61*	3.66	10.8*	4.73	8.04
Visual Comfort Violation (%)	Baseline	70.16	96.33	70.16	96.33	70.16	37.11	70.16	37.11	70.16	37.11	70.16	37.11
	BDQN	70.16	96.33	70.16	96.33	11.28	2.11	13.26	3.4	12.07	1.39	15.62	2.57
	SAC	70.16	96.33	70.16	96.33	***9.19***	***0.17***	***8.45***	***0.26***	***8.77***	***0.17***	***9.38***	***0.83***
	PPO	70.16	96.33	70.16	96.33	9.38	5.51	10.59	7.77	20.5	4.32	11.03	2.67

* The value exceeds the 10% threshold for thermal comfort violation, which is suggested by ASHRAE.

Table 4: Total facility energy use for the best trade-off offered by each RL algorithm using zone-level occupancy information.

Control Scenario	Baseline Number	Month	Baseline (MWh)	BDQN (MWh)	SAC (MWh)	PPO (MWh)	Best Agent (improvement over baseline)
SAT setpoint Blinds always open	(1)	January	8.34	6.48 (22.3%)	6.44 (22.78%)	6.46 (22.54%)	SAC (22.78%)
		July	4.44	3.36 (24.32%)	3.58 (19.37%)	—	BDQN (24.32%)
SAT setpoint Blinds always open Auto dimming	(3)	January	8.4	6.16 (26.67%)	—	—	BDQN (26.67%)
		July	3.2	2.11 (34.06%)	2.27 (29.06%)	—	BDQN (34.06%)
SAT setpoint Single blind setpoint	(2)	January	7.07	6.51 (7.92%)	6.23 (11.88%)	6.39 (9.62%)	SAC (11.88%)
		July	3.92	3.12 (20.41%)	3.21 (18.11%)	3.05 (22.19%)	PPO (22.19%)
SAT setpoint Single blind setpoint Auto-dimming	(4)	January	6.81	6.34 (6.9%)	6.1 (10.43%)	6.19 (9.1%)	SAC (10.43%)
		July	3.02	2.16 (28.48%)	2.36 (21.85%)	—	BDQN (28.48%)
SAT setpoint Multiple blind setpoints	(2)	January	7.07	6.55 (7.36%)	6.18 (12.59%)	6.53 (7.64%)	SAC (12.59%)
		July	3.92	3.17 (19.13%)	—	—	BDQN (19.13%)
SAT setpoint Multiple blind setpoints Auto-dimming	(4)	January	6.81	6.35 (6.75%)	6.06 (11.01%)	6.36 (6.61%)	SAC (11.01%)
		July	3.02	2.2 (27.15%)	2.43 (19.54%)	2.06 (31.79%)	PPO (31.79%)

– The agent's thermal comfort violation rate exceeds the 10% threshold. Thus, the realized reduction in energy use is not reported.

A.2 With the building-level occupancy schedule

Table 5: RL Agent performance results for different control scenarios using building-level occupancy schedule.

		Blinds always open				Single blind setpoint				Multiple blind setpoints			
		No dimming		With Dimming		No dimming		With Dimming		No dimming		With Dimming	
		Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.	Jan.	Jul.
Energy (MWh)	Baseline	8.34	4.44	8.4	3.2	***7.07***	3.92	***6.81***	3.02	***7.07***	3.92	***6.81***	3.02
	BDQN	7.59	3.47	7.5	2.22	7.78	3.32	7.89	***2.35***	7.85	3.31	8.08	***2.35***
	SAC	***7.17***	***3.49***	***6.84***	2.26	7.85	***3.12***	7.83	2.37	7.72	***3.21***	7.69	2.42
	PPO	7.57	3.87	7.18	***2.18***	7.16	3.33	7.15	2.38	7.22	3.21	8.07	2.47
Thermal Comfort ($\|PMV\|$)	Baseline	***0.25***	***0.22***	***0.28***	***0.28***	0.28	***0.27***	0.32	0.32	0.28	***0.27***	0.32	0.32
	BDQN	0.29	0.28	0.31	0.34	0.23	0.3	0.25	0.3	0.23	0.27	0.24	***0.3***
	SAC	0.3	0.42	0.34	0.33	***0.22***	0.46	***0.23***	***0.24***	***0.22***	0.42	***0.23***	0.46
	PPO	0.29	0.56	0.33	0.38	0.26	0.8	0.27	0.97	0.26	0.29	0.26	0.56
Thermal Comfort Violation (%)	Baseline	***4.06***	***0.78***	***3.98***	***0.86***	4.92	***0.86***	5.47	***1.02***	4.92	***0.86***	5.47	***1.02***
	BDQN	7.19	8.59	6.72	7.19	3.75	8.83	4.14	9.38	3.67	7.27	3.28	9.45
	SAC	6.8	2.5	10.08*	3.59	***2.97***	3.2	***2.58***	3.44	***3.36***	3.36	***3.2***	3.28
	PPO	6.25	30.86*	8.75	15.47	5.55	8.98	4.53	8.67	5.62	9.45	4.45	30.16*
Visual Comfort Violation (%)	Baseline	70.16	96.33	70.16	96.33	70.16	37.11	70.16	37.11	70.16	37.11	70.16	37.11
	BDQN	70.16	96.33	70.16	96.33	12.5	3.52	11.8	4.45	10.86	4.14	10.78	5.31
	SAC	70.16	96.33	70.16	96.33	***8.05***	***0.0***	***8.05***	***0.23***	***8.28***	***0.08***	***8.05***	***1.64***
	PPO	70.16	96.33	70.16	96.33	12.19	5.16	9.22	5.47	15.62	11.95	12.11	7.19

* The value exceeds the 10% threshold for thermal comfort violation, which is suggested by ASHRAE.

Table 6: Total facility energy use for the best trade-off offered by each RL algorithm using zone-level occupancy information.

Control Scenario	Baseline Number	Month	Baseline (MWh)	BDQN (MWh)	SAC (MWh)	PPO (MWh)	Best Agent (improvement over baseline)
SAT setpoint Blinds always open	(1)	January	8.34	7.59 (8.99%)	7.17 (14.03%)	7.57 (9.23%)	SAC (14.03%)
		July	4.44	3.47 (21.85%)	3.49 (21.4%)	—	BDQN (21.85%)
SAT setpoint Blinds always open Auto dimming	(3)	January	8.4	7.5 (10.71%)	—	7.18 (14.52%)	PPO (14.52%)
		July	3.2	2.22 (30.63%)	2.26 (29.38%)	2.18 (31.88%)	PPO (31.88%)
SAT setpoint Single blind setpoint	(2)	January	7.07	7.78 (-10.04%)	7.85 (-11.03%)	7.16 (-1.27%)	PPO (-1.27%)
		July	3.92	3.32 (15.31%)	3.12 (20.41%)	3.33 (15.05%)	SAC (20.41%)
SAT setpoint Single blind setpoint Auto-dimming	(4)	January	6.81	7.89 (-15.86%)	7.83 (-14.98%)	7.15 (-4.99%)	PPO (-4.99%)
		July	3.02	2.35 (22.19%)	2.37 (21.52%)	2.38 (21.19%)	BDQN (22.19%)
SAT setpoint Multiple blind setpoints	(2)	January	7.07	7.85 (-11.03%)	7.72 (-9.19%)	7.22 (-2.12%)	PPO (-2.12%)
		July	3.92	3.31 (15.56%)	3.21 (18.11%)	3.21 (18.11%)	SAC (18.11%)
SAT setpoint Multiple blind setpoints Auto-dimming	(4)	January	6.81	8.08 (-18.65%)	7.69 (-12.92%)	8.07 (-18.5%)	SAC (-12.92%)
		July	3.02	2.35 (22.19%)	2.42 (19.87%)	—	BDQN (22.19%)

— The agent's thermal comfort violation rate exceeds the 10% threshold. Thus, the realized reduction in energy use is not reported.