Reinforcement Learning for Energy-Efficient Residential Heating Control

Ruxiao Chen

Abstract

Optimizing residential heating systems is essential for energy efficiency and decarbonization. Traditional HVAC optimization methods rely on building physics models, which use thermodynamic equations to simulate indoor temperature, humidity, and airflow. These models must account for building materials, insulation properties, external weather conditions, sunlight exposure, and heat exchange through windows and walls. While these models enable accurate control of heating, cooling, and ventilation, they are computationally expensive. These approaches require significant computational resources and struggle to generalize across different buildings due to variations in structure, insulation, and occupancy patterns, making real-time control impractical. This project proposes a reinforcement learning (RL)-based approach to overcome these challenges. By learning control policies directly from data, RL eliminates the need for explicit physics-based models while dynamically adapting to changing conditions. We will implement an RL-based heating control system using Q-learning, which balances sample efficiency and stability, making it suitable for optimizing residential heating operations.

1 Introduction

Residential heating accounts for a significant portion of global energy use, making it a key target for energy efficiency and decarbonization efforts [BPMC20]. Traditional control of residential heating systems often relies on Model Predictive Control (MPC), which uses an explicit thermal model of the building to forecast temperature evolution and solve a receding-horizon optimization problem at each control step [BPMC20, GLW20]. MPC has been shown to be effective in maintaining thermal comfort while reducing energy consumption, especially when accurate building models and forecasts are available.

However, deploying MPC in residential settings presents practical challenges. Each building requires a customized thermal model that captures its specific structural and material properties, such as insulation quality, window-to-wall ratio, and thermal mass [KPP+25]. Developing and calibrating these models is time-consuming and labor-intensive, often requiring expert knowledge and significant historical data. Moreover, solving the optimization problem in real time at each control step imposes a heavy computational burden, particularly under high uncertainty introduced by fluctuating weather conditions and irregular occupant behavior. These limitations hinder the scalability and real-world applicability of MPC across diverse residential buildings.

To address these challenges, we propose a model-free reinforcement learning (RL) approach based on Q-learning. Unlike MPC, Q-learning does not require an explicit thermal model of the building. Instead, it learns an optimal control policy directly from interaction with the environment through trial-and-error [JKHK19]. By continuously updating its Q-values based on observed state transitions and rewards, Q-learning can adapt to the specific thermal characteristics of each building and respond flexibly to dynamic changes. This data-driven control framework significantly reduces the modeling and computational requirements while maintaining control performance, thus offering a scalable and adaptive solution for residential heating optimization.

2 Methods

2.1 RC Model

We adopt a first-order resistance-capacitance (RC) model to represent the thermal dynamics of a residential space [WCL19]. The continuous-time formulation is derived from the energy balance equation:

$$C\frac{dT_{\text{in}}(t)}{dt} = \frac{T_{\text{out}}(t) - T_{\text{in}}(t)}{R} + Q(t), \tag{1}$$

where $T_{\text{in}}(t)$ is the indoor temperature, $T_{\text{out}}(t)$ is the outdoor temperature, $Q(t)$ is the heating or cooling input (in watts), R is the thermal resistance, and C is the thermal capacitance.

To enable discrete-time control, we apply a forward Euler approximation with a fixed sampling time Δt, yielding:

$$T_{t+1} = \left(1 - \frac{\Delta t}{RC}\right) T_t + \frac{\Delta t}{RC} T_{\text{out},t} + \frac{\Delta t}{C} Q_t. \tag{2}$$

For simplicity and to facilitate model-free control, we rewrite the above as:

$$T_{t+1} = aT_t + cQ_t + d, \tag{3}$$

where $a = 1 - \frac{\Delta t}{RC}$ captures the thermal inertia, $c = \frac{\Delta t}{C}$ is the input gain, and $d = \frac{\Delta t}{RC} T_{\text{out},t}$ represents the effect of outdoor temperature. When $T_{\text{out},t}$ is assumed constant or treated as noise, d can be approximated as a constant offset. This compact form enables integration with reinforcement learning algorithms without explicit modeling of physical parameters.

2.2 Q-learning

We implement a model-free Q-learning algorithm to learn a heating control policy under the RC dynamics described in Section II-2. The system state is defined as the indoor temperature, discretized into 41 bins from 10°C to 30°C with 0.5°C resolution. The action space includes seven discrete control inputs: $\{-2.0, -1.0, -0.5, 0.0, 0.5, 1.0, 2.0\}$, representing heating or cooling power levels.

The environment evolves according to the discrete-time model:

$$T_{t+1} = aT_t + cQ_t + d, \tag{4}$$

where $a = 0.9$, $c = 0.5$, and $d = 3.0$. At each step, the agent receives a scalar reward based on the squared error from the target temperature $T_{\text{ref}} = 22.0°\text{C}$:

$$r_t = -(T_t - T_{\text{ref}})^2. \tag{5}$$

The Q-table has shape $[41, 7]$, initialized to zeros. Action selection follows an ε-greedy strategy with ε initialized to 1.0 and decayed by 0.995 per episode, down to a minimum of 0.01. The Q-values are updated using the Bellman equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)], \tag{6}$$

with learning rate $\alpha = 0.01$ and discount factor $\gamma = 0.90$.

Training is performed over 1000 episodes, each consisting of 5 time steps. At the beginning of each episode, the indoor temperature is randomly initialized between 15°C and 25°C. After training, the learned policy is evaluated over a 70-step simulation starting from 17°C, and the resulting temperature and control input trajectories are recorded for analysis [WWZ17].

2.3 Model Predictive Control

We implement a standard Model Predictive Control (MPC) scheme to serve as a benchmark. The system follows the same discrete-time RC dynamics:

$$T_{t+1} = aT_t + cQ_t + d, \tag{7}$$

with $a = 0.9$, $c = 0.5$, and $d = 3.0$ as obtained from model identification. At each control step, MPC solves a finite-horizon optimization problem to minimize the cumulative deviation from a reference temperature $T_{\text{ref}} = 22.0°\text{C}$ over a horizon of $H = 24$ steps.

The optimization variables are the control input sequence $Q = \{Q_0, \ldots, Q_{H-1}\}$ and predicted temperatures $T = \{T_1, \ldots, T_H\}$. The cost function is defined as:

$$\min_{Q} \sum_{k=1}^{H} (T_k - T_{\text{ref}})^2 + \lambda_Q \sum_{k=0}^{H-1} Q_k^2, \tag{8}$$

where $\lambda_Q = 0.0$ weights the control effort penalty. The constraints include:

- System dynamics: $T_{k+1} = aT_k + cQ_k + d$
- Control bounds: $Q_{\min} \leq Q_k \leq Q_{\max}$, with $Q_{\min} = -5.0$, $Q_{\max} = 5.0$
- Comfort constraints: $T_{\min} \leq T_k \leq T_{\max}$, with $T_{\min} = 20.0$°C, $T_{\max} = 24.0$°C

Only the first control input Q_0 is applied at each step (receding horizon). The optimization is re-solved at every time step using the current temperature T_t as the initial condition. The solver used is CVXPY with default settings. The system is simulated for 50 steps from an initial temperature of 17.0°C.

3 Experiments

3.1 Comparison With Baseline

To evaluate the effectiveness of the proposed Q-learning approach, we compare its performance against a standard Model Predictive Control (MPC) baseline. The objective is to assess each method's ability to regulate indoor temperature under the same dynamic model and constraints. MPC represents a widely used model-based strategy that relies on accurate system identification and online optimization, while Q-learning offers a model-free, data-driven alternative. This comparison highlights the trade-offs between model fidelity, computational demand, and control performance.

All experiments are conducted in a simulated environment governed by the discrete-time RC model:

$$T_{t+1} = aT_t + cQ_t + d,$$

with fixed parameters $a = 0.9$, $c = 0.5$, and $d = 3.0$. The target indoor temperature is set to $T_{\text{ref}} = 22.0$°C. Control inputs are bounded between -5.0 and 5.0, and the acceptable comfort range is [20.0°C, 24.0°C].

Both controllers are tested over a 50-step closed-loop simulation starting from $T_0 = 17.0$°C. The MPC controller uses a prediction horizon of $H = 24$ and solves a constrained quadratic program at each step. The Q-learning controller is trained over 1000 episodes, each consisting of 5 interaction steps. After training, the learned policy is applied in a closed-loop simulation using greedy action selection.

Fig. 1 illustrates the closed-loop performance of MPC and Q-learning over a 50-step simulation. Both controllers begin at an initial indoor temperature of 17°C and aim to reach the target setpoint of 22°C while staying within the defined comfort zone of [20°C, 24°C]. In the top subplot, the MPC controller quickly drives the temperature toward the setpoint with minimal overshoot and maintains it steadily within the comfort range. In contrast, the Q-learning agent exhibits slower convergence and noticeable oscillations around the setpoint, especially between steps 15 and 45, indicating reduced stability and control precision.

The middle subplot shows the corresponding control actions. MPC outputs a brief heating phase followed by near-zero control effort, efficiently maintaining thermal comfort. Q-learning, by contrast, applies more aggressive and fluctuating control signals, frequently switching between heating and cooling, which reflects its lack of long-term planning and limited generalization beyond the training distribution.

Bottom subplot of Fig. 1 presents a quantitative comparison between MPC and Q-learning based on four metrics: mean squared error (MSE), mean absolute error (MAE), maximum temperature deviation, and average control effort. MPC achieves a lower MSE (0.45) and maximum error (3.70°C), demonstrating better consistency and tighter regulation around the target temperature. In contrast, Q-learning yields a lower MAE (0.15) but a higher maximum error (2.20°C), indicating that although its average performance is reasonable, it occasionally produces larger deviations. In terms of energy usage, both methods have comparable average control efforts (1.49 for MPC vs. 1.62 for RL), suggesting similar levels of actuation. However, MPC attains this with fewer fluctuations and smoother control profiles.

3.2 Model Transfer Ability Evaluation

Our goal in this section is to assess how well a policy learned via Q-learning in one thermal environment transfers to others with different dynamics. Model-free methods promise adaptability, but their sensitivity to changes in system parameters remains an open question. By comparing transferred Q-learning against MPC under model mismatch, we quantify this generalization capability.

Training setup. The agent is trained exclusively on the "Original Room" RC model with parameters $(a, c, d) = (0.9, 0.5, 3.0)$, then applied this model directly into other settings. The other training setups are the same with last section.

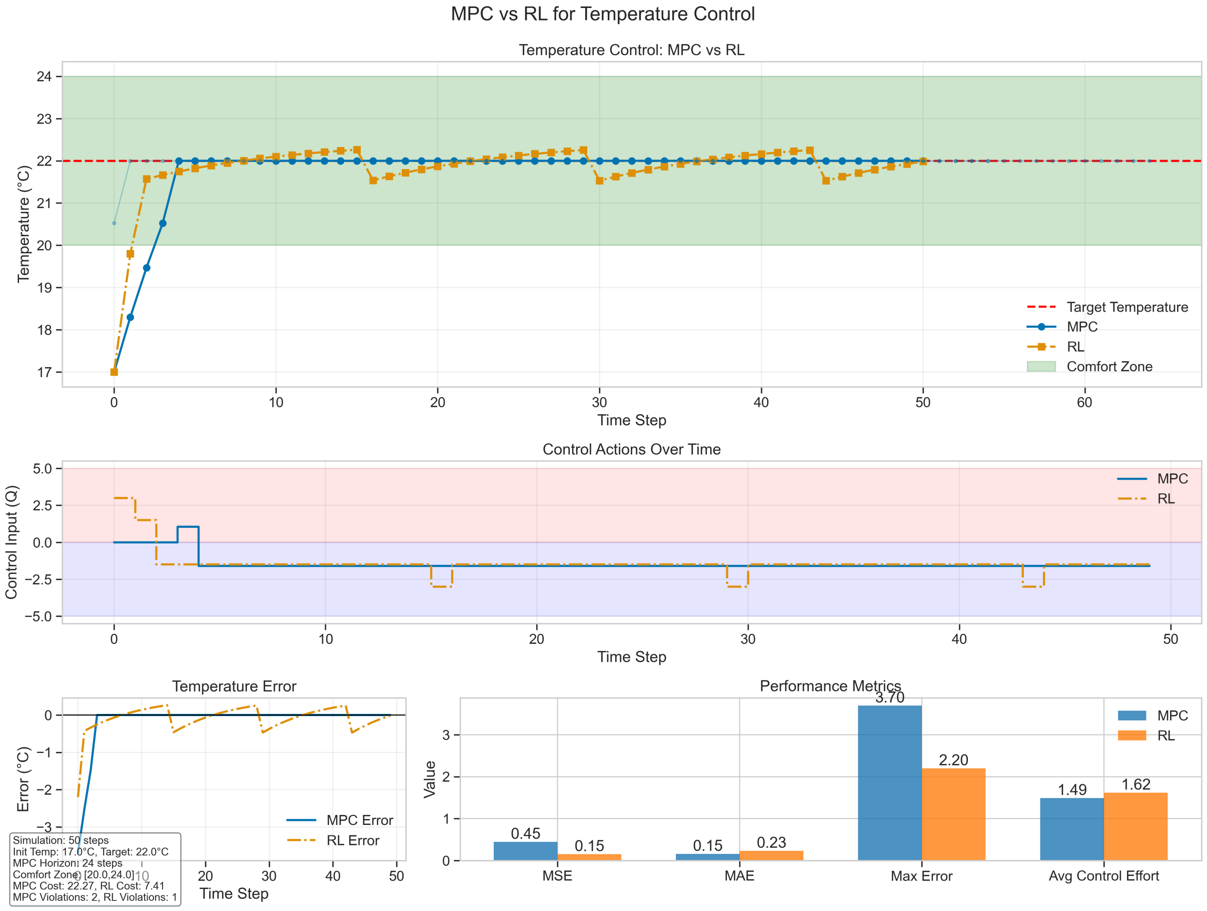

Figure 1: Comparison of Model Predictive Control (MPC) and Qlearning (RL) for temperature regulation. Top-left: temperature trajectories showing the target setpoint and comfort zone. Top-right: control inputs over time. Bottom-left: tracking error curves. Bottom-right: performance metrics (MSE, MAE, maximum deviation, and average control effort).

Evaluation environments. Without further learning, the trained Q-table is deployed in four novel rooms:

Table 1: Evaluation environments with varied thermal dynamics

Environment	a	c	d
High Inertia	0.95	0.40	2.0
Low Inertia	0.80	0.70	4.0
Different Insulation	0.85	0.60	3.5
Extreme Dynamics	0.75	0.90	5.0

Each evaluation is a 70-step closed-loop simulation from $T_0 = 17°\text{C}$, tracking temperature and control actions. Performance metrics (e.g. MSE, comfort violations, control effort) will be computed to compare transferred Q-learning against (i) MPC with the original model and (ii) MPC with perfect model knowledge.

Fig. 2 plots the temperature trajectories when the Q-learning policy—trained in the Original Room—is deployed across four novel environments. In the Original Room (blue), the agent quickly converges to 22 °C and oscillates within ±0.2 °C of the setpoint. In the High Inertia Room (green), the same policy under-heats: temperature rises slowly and drifts up to 24 °C, reflecting the agent's inability to overcome the increased thermal mass. In the Low Inertia Room (red), the policy is too aggressive: it overshoots 22 °C and oscillates with a large amplitude (±1 °C), indicating mismatch with the faster dynamics. The Differently Insulated Room (purple) exhibits a steady-state bias around 21 °C, showing that the policy cannot fully compensate for altered heat losses. Finally, in the Extreme Environment (orange),

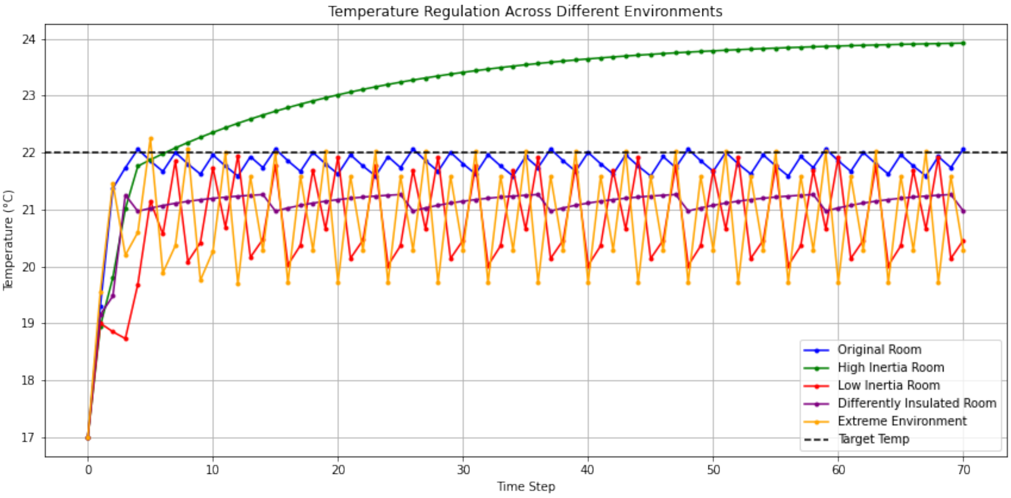

Figure 2: Temperature trajectories of the Q-learning controller trained in the Original Room and evaluated across five different environments with varied thermal dynamics.

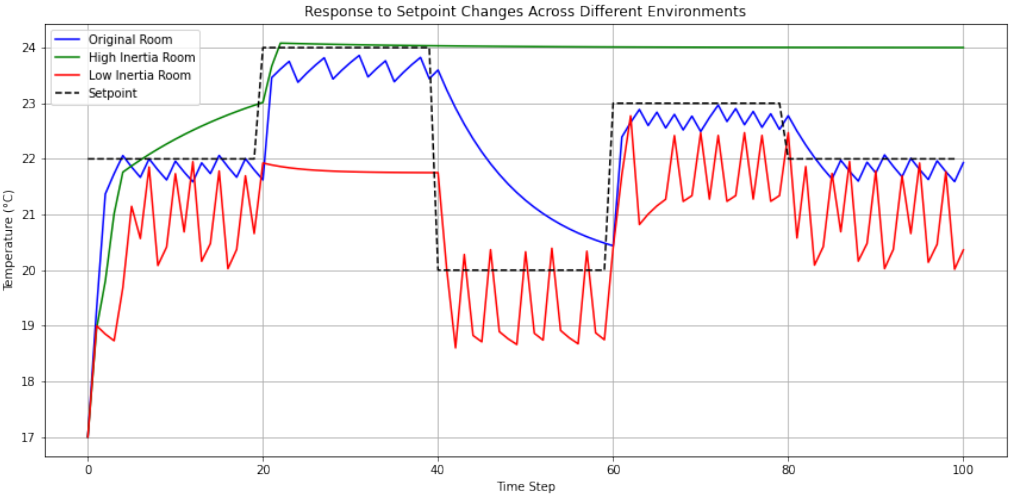

Figure 3: Response of the Q-learning controller to a time-varying setpoint schedule when deployed in three different environments: Original Room, High Inertia Room, and Low Inertia Room.

the trajectory swings widely above and below the setpoint, demonstrating severe instability under drastic parameter shifts. Overall, the Q-learning policy transfers acceptably for mild mismatches but fails to maintain comfort when thermal dynamics change substantially, underscoring the need to compare its robustness against MPC with both nominal and perfect model knowledge.

Fig. 3 illustrates the closed-loop temperature response to a sequence of setpoint shifts at steps 20, 40, 60 and 80. In the Original Room (blue), the learned Q-learning policy tracks each jump with modest overshoot (0.5 °C) and settles within the comfort bounds. In the High Inertia Room (green), the same policy is too sluggish, failing to reach higher targets and remaining above lower setpoints. Conversely, in the Low Inertia Room (red), it overreacts, producing large oscillations around each new setpoint. These results demonstrate that a policy trained under one set of dynamics does not generalize robustly to environments with substantially different thermal inertia.

Fig. 4 compares average reward, steady-state error, and trajectory stability for the Q-learning policy—trained in the Original Room—when deployed across five environments. The Original Room yields the highest average reward (≈–0.1), minimal steady-state error (0.10 °C) and low variability (std 0.15 °C), confirming successful training. In the High Inertia Room, the policy under-heats consistently, producing the lowest reward (–2.32), large steady-state error (1.90 °C) but very low variance (std 0.02 °C). The Low Inertia Room exhibits moderate offset (1.50 °C) and pronounced oscillations (std 0.70 °C), reflecting

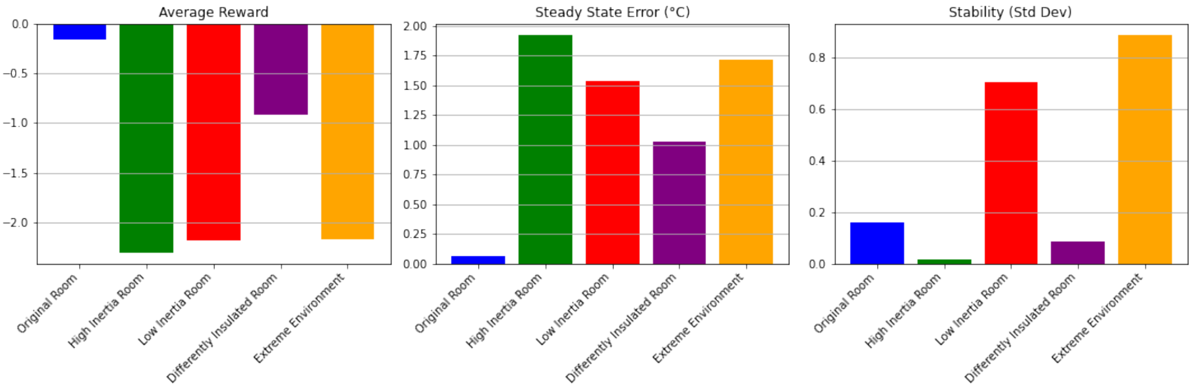

Figure 4: Performance metrics of the Q-learning controller across five different environments, showing average reward, steady-state error, and temperature stability (standard deviation).

excessive responsiveness to faster dynamics. The Differently Insulated Room shows intermediate performance (error 1.03 °C, std 0.08 °C), while the Extreme Environment combines poor reward (−2.20), high error (1.73 °C) and instability (std 0.88 °C). These results demonstrate that transfer performance degrades as the discrepancy in thermal dynamics increases, motivating a direct comparison with MPC under similar model mismatch.

Fig. 6 and Fig. 5 illustrate the behavior of an MPC controller designed using the Original Room model when deployed in environments with mismatched thermal dynamics. Despite lacking awareness of the parameter shift, MPC maintains stable and smooth responses across all test environments. Temperature trajectories remain monotonic, with no oscillations or instability, in contrast to the Q-learning policy, which exhibited significant performance degradation under the same conditions.

However, the setpoint tracking accuracy degrades as model mismatch increases. For example, in the High Inertia Room, MPC underestimates the system's thermal resistance, leading to persistent overshoot. In the Low Inertia Room, insufficient actuation results in steady-state offset. These biases remain moderate, and critically, MPC avoids the instability and control saturation observed in the transferred Q-learning policy. This demonstrates that even with imperfect models, MPC offers better robustness and safer deployment under unmodeled dynamics.

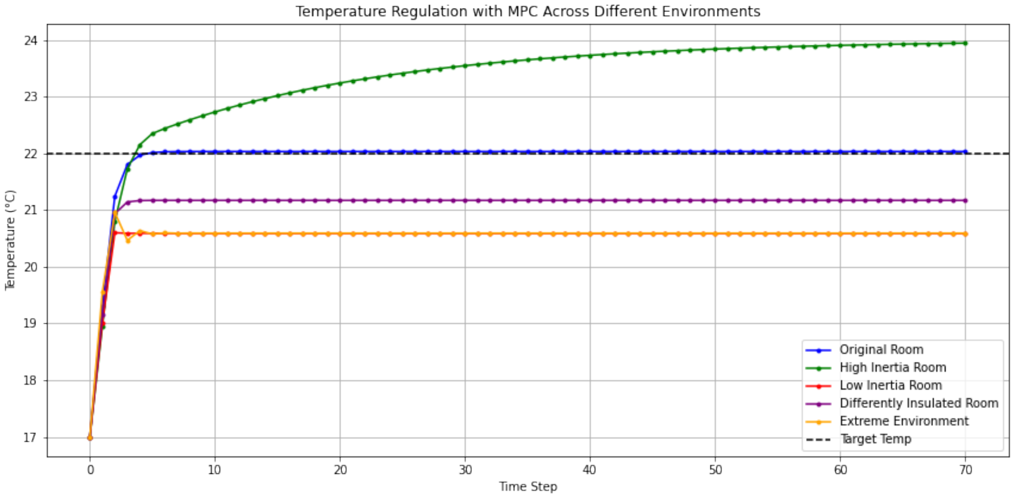

Figure 5: Temperature trajectories of the MPC controller applied to five different environments with varying thermal dynamics.

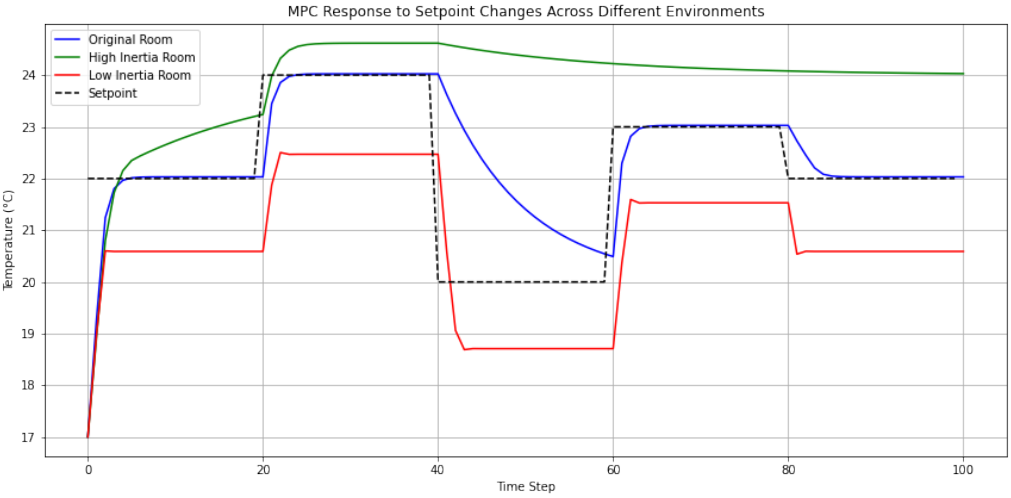

Figure 6: Response of the MPC controller to a time-varying setpoint schedule in three different environments: Original Room, High Inertia Room, and Low Inertia Room.

4 Conclusion

This paper compared model-free Q-learning and model-based MPC for residential temperature control under dynamic and structural variation. While both methods perform well in the nominal environment, only MPC maintains stability and smooth tracking when transferred to unseen thermal dynamics. Q-learning, despite being trained without explicit models, suffers from oscillation, steady-state error, and reduced robustness under mismatched conditions. In contrast, MPC, even with outdated model parameters, consistently produces bounded, monotonic responses. These results highlight the trade-off between adaptability and reliability, and suggest that model-based control remains advantageous for safety-critical applications with uncertain dynamics.

References

[BPMC20] Silvio Brandi, Marco Savino Piscitelli, Marco Martellacci, and Alfonso Capozzoli. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings*, 224:110225, 2020.

[GLW20] Guanyu Gao, Jie Li, and Yonggang Wen. Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning. *IEEE Internet of Things Journal*, 7(9):8472–8484, 2020.

[JKHK19] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, and Jong Wook Kim. Q-learning algorithms: A comprehensive classification and applications. *IEEE Access*, 7:133653–133667, 2019.

[KPP$^+$25] Arash J. Khabbazi, Elias N. Pergantis, Levi D. Reyes Premer, Panagiotis Papageorgiou, Alex H. Lee, James E. Braun, Gregor P. Henze, and Kevin J. Kircher. Lessons learned from field demonstrations of model predictive control and reinforcement learning for residential and commercial hvac: A review, 2025.

[WCL19] Zequn Wang, Yuxiang Chen, and Yong Li. Development of rc model for thermal dynamic analysis of buildings through model structure simplification. *Energy and Buildings*, 195:51–67, 2019.

[WWZ17] Tianshu Wei, Yanzhi Wang, and Qi Zhu. Deep reinforcement learning for building hvac control. In *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6, 2017.