
HVAC-DPT: A Decision Pretrained Transformer for HVAC Control

Anaïs Berkes

Department of Computer Science & Technology
University of Cambridge
United Kingdom
amcb6@cam.ac.uk

Abstract

Building operations consume approximately 40% of global energy, with Heating, Ventilation, and Air Conditioning (HVAC) systems responsible for up to 50% of this consumption [1, 2]. As HVAC energy demands are expected to rise, optimising system efficiency is crucial for reducing future energy use and mitigating climate change [3]. Existing control strategies lack generalisation and require extensive training and data, limiting their rapid deployment across diverse buildings. This paper introduces HVAC-DPT, a Decision-Pretrained Transformer using in-context Reinforcement Learning (RL) for multi-zone HVAC control. HVAC-DPT frames HVAC control as a sequential prediction task, training a causal transformer on interaction histories generated by diverse RL agents. This approach enables HVAC-DPT to refine its policy in-context, without modifying network parameters, allowing for deployment across different buildings without the need for additional training or data collection. HVAC-DPT reduces energy consumption in unseen buildings by 45% compared to the baseline controller, offering a scalable and effective approach to mitigating the increasing environmental impact of HVAC systems.

1 Introduction and related work

Advanced controllers have the potential to significantly reduce HVAC energy consumption [4], but most buildings continue to rely on inefficient, rule-based systems. Although various model-based, data-driven, and learning-based HVAC control strategies have been proposed [4, 5], it remains a significant challenge to scale these methods across diverse building types. Model Predictive Control is limited by its reliance on precise and building-specific models, while RL requires extensive training, lasting months or years [6], which often leads to suboptimal performance and occupant discomfort during the learning phase [7]. RL also suffers from severe sample inefficiency, demanding significant amounts of sensor data and requiring retraining for each new building. Even with transfer learning, significant data collection and customisation is still needed to address the variability in building structures and thermal dynamics between the buildings used for training and new target buildings [8, 9].

The transformer architecture [10] has been widely adopted in key areas of machine learning. One major feature of transformers is in-context learning, which makes it possible for them to adapt to new tasks after extensive pretraining [11]. Recent research, such as the Decision-Pretrained Transformer (DPT) by Lee *et al.* [12] and Algorithm Distillation by Laskin *et al.* [13], effectively uses transformer-based in-context learning for sequential decision-making. These methods predict actions based on a query state and historical environment dynamics without the need for weight updates after the initial pretraining phase. Additionally, recent work demonstrates that transformers pretrained on diverse datasets can generalise to new RL tasks in-context, offering a promising approach for extracting

generalist policies from offline RL data [14, 15, 16]. Nevertheless, the application of in-context RL to HVAC control remains unexplored.

In response, we introduce HVAC-DPT, a pretrained decision transformer that uses in-context RL to optimise HVAC systems across multiple building zones without requiring prior data or additional training for new buildings. HVAC-DPT overcomes the limitations of existing control methods by enabling scalable, data-efficient, and generalisable deployment across diverse building types, removing the need for retraining and pre-deployment data collection in new environments. In a year-long evaluation using EnergyPlus [17], HVAC-DPT reduced HVAC energy consumption by 45% compared to baseline operations, demonstrating its transformative potential to reduce the carbon footprint of building operations.

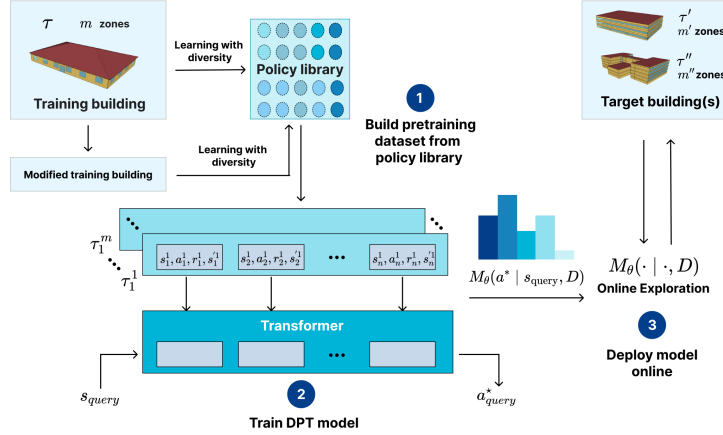


Figure 1: Schematic overview of the proposed methodology.

2 Problem definition

HVAC systems consist of one or more air handling units (AHUs) and variable air volume (VAV) systems, as illustrated in Figure 3. Optimising HVAC control can be framed as a sequential decision-making problem, where an agent interacts with the building, adjusts controls (e.g., VAV actuators), and receives rewards to learn control policies.

While a single agent could manage the entire building, this approach limits policy adaptability to buildings with different state-action spaces, such as those with varying numbers of VAV systems. Consequently, we model HVAC control as a multi-agent reinforcement learning (MARL) task, where each agent controls a single zone, enabling independent management across zones, similar to the approach in [6].

The multi-agent Markov decision process is defined as a tuple $(N; S; A_i; R_i; T; H)$, where N denotes the number of agents, S is the state space, A_i is the action space for agent i , $R : S \times A \rightarrow \Delta(R)$ is the reward function, $T : S \times A \rightarrow \Delta(S)$ is the transition function, and H is the horizon.

The state space S , observed by all agents, includes six sensor readings, detailed in Table 1. Each agent’s action space A_i corresponds to the minimum damper position in their VAV system, ranging from 0 (closed) to 1 (fully open). The reward R_i for each agent is the negative energy consumption of the VAV system during the transition from state s to s' . The transition function T is determined by EnergyPlus. Each episode has a length of H . Agents continuously control the VAV systems, but

Table 1: State variables and actions for each agent.

State	Unit
Zone mean temperature	°C
Zone mean humidity	%
Zone occupancy	Binary
Outdoor temperature	°C
Solar radiation	W
Hour of the day	Integer
Action	Unit
VAV minimum damper position	%

the control problem is modelled as episodic, with one month at 15-minute intervals constituting an episode.

3 HVAC-DPT

The method, illustrated in Figure 1, consists of three steps and builds upon the method presented by Lee *et al.* [12]: (1) A dataset $\mathcal{B}$ of RL agent interactions is collected after training a policy library of diverse RL agents in N buildings. (2) A transformer model is trained to predict action labels based on a query state and the in-context dataset of interactions D sampled from $\mathcal{B}$. (3) Once trained, HVAC-DPT can be deployed online in a new building by querying it for predictions of the optimal action in different states.

Algorithm 1 HVAC-DPT

```

1: // Dataset Generation
2: Initialise empty dataset  $\mathcal{B}$ 
3: for  $i$  in  $[N]$  do
4:   Sample training building  $\tau \sim \mathcal{T}_{\text{pre}}$ 
5:   Build policy library: train diverse RL policies  $\pi_{\tau}^i$  for all zones  $i \in [m]$ 
6:   Sample interaction dataset  $D \sim \mathcal{D}_{\text{pre}}(\cdot; \tau)$  from all  $\pi_{\tau}^i$ 
7:   Sample  $s_{\text{query}}^i \sim \mathcal{D}_{\text{query}}$  and  $a^* \sim \pi_{\tau}^i(\cdot | s_{\text{query}})$ 
8:   Add  $(s_{\text{query}}^i, D, a^*)$  to  $\mathcal{B}$ 
9: end for
10: // Pretraining Phase
11: Pretraining Phase
12: Initialise model  $M_{\theta}$  with parameters  $\theta$ 
13: while not converged do
14:   Sample  $(s_{\text{query}}^i, D, a^*)$  from  $\mathcal{B}$ 
15:   Predict  $\hat{p}_j(\cdot) = M_{\theta}(\cdot | s_{\text{query}}^i, D_j)$ 
16:   Compute MSE loss with respect to  $a^*$  and backpropagate to update  $\theta$ 
17: end while
18: // Online Deployment
19: Initialize  $D^i = \{\}$  for all zones
20: Sample target building  $\tau' \sim \mathcal{T}_{\text{test}}$ 
21: for ep in max_eps do
22:   for  $h$  in  $[H]$  do
23:      $s_1 = \text{reset}(\tau')$ 
24:     for zone  $i$  in  $N_{\text{zones}}^{\tau'}$  do
25:        $a_h^i \sim M_{\theta}(\cdot | s_h^i, D^i)$ 
26:        $s_{h+1}^i, r_h^i = \text{step}(\tau', a_h^i)$ 
27:       Add  $(s_1^i, a_1^i, r_1^i, \dots)$  to  $D^i$ 
28:     end for
29:   end for
30: end for=0

```

Dataset Generation. The pretraining dataset $\mathcal{B}$ is collected for N training buildings. HVAC-DPT generates a policy library of diverse Proximal Policy Optimisation (PPO) RL agents for the different zones in each training building τ sampled from the distribution over training buildings $\mathcal{T}_{\text{pre}}$. Both policy and environment diversity are used during training, as in [6]. Rollouts of these policies are used to sample an in-context dataset $D = \{s_j, a_j, s'_j, r_j\}_{j \in [n]}$ of transition tuples taken in all zones of τ .

Pretraining. A query state s_{query}^i is sampled for each zone and a label a^* is sampled from an agent in the policy library. The in-context dataset D and query state s_{query} are used to train a model to predict the RL-labeled action a^* via supervised learning. Formally, we train a GPT-2 transformer model M parameterised by θ , which outputs a distribution over actions $\mathcal{A}$, to minimise the expected

loss over samples from the pretraining distribution:

$$\min_{\theta} \mathbb{E}_{P_{\text{pre}}} \sum_{j \in [n]} \ell(M_{\theta}(\cdot | s_{\text{query}}, D_j), a^*). \quad (1)$$

where P_{pre} is the joint pretraining distribution over buildings, in-context datasets, query states and action labels. As we have a continuous $\mathcal{A}$, we set the loss to be the Mean Squared Error (MSE).

Online deployment. The model M_{θ} can be deployed online in an unseen target building τ' by initialising an empty $D^i = \{\}$ for each zone i in N_{zones}^{τ} . HVAC-DPT samples an action $a_h^i \sim M_{\theta}(\cdot | s_h^i, D^i)$ for each zone i at each time-step. D^i is subsequently filled with the interactions $\{s_1^i, a_1^i, r_1^i, \dots, s_H^i, a_H^i, r_H^i\}$ collected during each episode. A key distinction to traditional RL algorithms is that there are no updates to the parameters of M_{θ} . Once deployed, HVAC-DPT simply performs a computation through its forward pass to generate a distribution over actions conditioned on the in-context D^i and query state s_h^i .

4 Results

We used EnergyPlus [17] and COBS [18] to train 100 diverse policies for B_{train} ; further details are provided in Appendix B. Four commonly used controllers were compared [19, 6]:

- (1) The *Baseline* controller, which maintains damper openings at 50%;
- (2) The *Expert* controller, implemented in the EnergyPlus model and designed specifically for each building by HVAC engineers;
- (3) *SARL*, a single agent RL policy that controls all zones' dampers based on interaction with the target building; and
- (4) *MARL*, which controls individual zones using the MARL framework.

Figure 2 demonstrates HVAC-DPT's performance in B_{Denver} , which differs from B_{train} in size and HVAC design, affecting state and action spaces. HVAC-DPT reduces energy consumption by 45% compared to the *Baseline*. HVAC-DPT is only 5% less effective than the *Expert* controller, despite having no prior knowledge of the building. The *SARL* and *MARL* controllers perform 74% and 70% worse, respectively, due to the extensive training required to achieve optimal performance, which can take up to 1,250 years [6]. More details are given in Appendix C

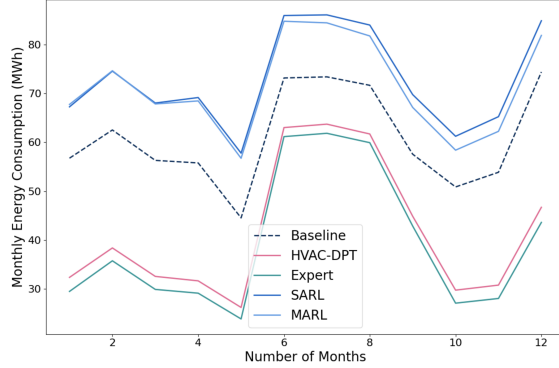


Figure 2: HVAC energy consumption (MWh) of different controllers during the first 12 months of deployment in building B_{Denver} .

5 Conclusion

This paper introduces HVAC-DPT, a pretrained decision transformer that uses in-context RL to optimise HVAC systems. Within the first year of deployment in new buildings, HVAC-DPT reduces energy consumption by 45% and 70% compared to baseline operations and RL agents respectively, all without additional training or data collection. This demonstrates HVAC-DPT's ability to generalise effectively across buildings, addressing critical challenges in HVAC control, such as scalability, data dependency, and training efficiency. Future work will validate HVAC-DPT in real-world settings, reinforcing its potential as a widely deployable solution for sustainable building management.

References

- [1] European Commission. In focus: Energy efficiency in buildings, 2026.
- [2] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.

- [3] Mat Santamouris. Cooling the buildings—past, present and future. *Energy and Buildings*, 128:617–638, 2016.
- [4] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L Vrabie, et al. All you need to know about model predictive control for buildings. *Annual Reviews in Control*, 50:190–232, 2020.
- [5] Zhe Wang and Tianzhen Hong. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, 269:115036, 2020.
- [6] Aakash Krishna GS, Tianyu Zhang, Omid Ardakanian, and Matthew E Taylor. Mitigating an adoption barrier of reinforcement learning-based control strategies in buildings. *Energy and Buildings*, 285:112878, 2023.
- [7] Tianyu Zhang, Gaby Baasch, Omid Ardakanian, and Ralph Evins. On the joint control of multiple building systems with reinforcement learning. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, pages 60–72, 2021.
- [8] Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, and Sergey Kolesnikov. Emergence of in-context reinforcement learning from noise distillation. *arXiv preprint arXiv:2312.12275*, 2023.
- [9] Tianyu Zhang, Mohammad Afshari, Petr Musilek, Matthew E Taylor, and Omid Ardakanian. Diversity for transfer in learning-based control of buildings. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*, pages 556–564, 2022.
- [10] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [11] Viacheslav Sinii, Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, and Sergey Kolesnikov. In-context reinforcement learning for variable action spaces. *arXiv preprint arXiv:2312.13327*, 2023.
- [12] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [13] Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.
- [14] Xiangyuan Zhang, Weichao Mao, Haoran Qiu, and Tamer Başar. Decision transformer as a foundation model for partially observable continuous control. *arXiv preprint arXiv:2404.02407*, 2024.
- [15] Subhojyoti Mukherjee, Josiah P Hanna, Qiaomin Xie, and Robert Nowak. Pretraining decision transformers with reward prediction for in-context multi-task structured bandit learning. *arXiv preprint arXiv:2406.05064*, 2024.
- [16] Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.
- [17] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. Energyplus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4):319–331, 2001.
- [18] Tianyu Zhang and Omid Ardakanian. Cobs: Comprehensive building simulator. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 314–315, 2020.
- [19] Ki Uhn Ahn, Deuk-Woo Kim, Hyun Mi Cho, and Chang-U Chae. Alternative approaches to hvac control of chat generative pre-trained transformer (chatgpt) for autonomous building system operations. *Buildings*, 13(11):2680, 2023.

- [20] R. American Society of Heating, A.-C. Engineers. *Energy Standard for Buildings Except Low-Rise Residential Buildings*. ASHRAE Inc, Peachtree Corners, GA, USA, standard 90.1-2019 edition, 2019.

A Additional system details

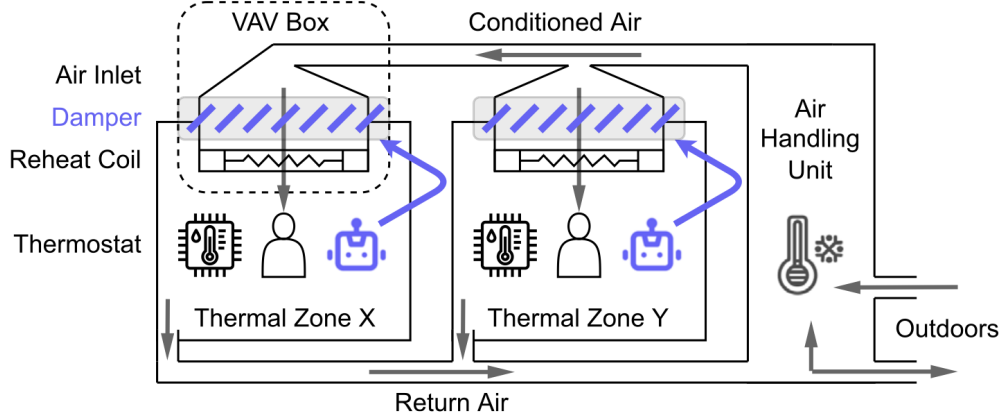


Figure 3: Illustration of an air loop in a multi-zone building equipped with a forced-air heating and cooling system [6].

The controller in EnergyPlus adjusts the AHU and other VAV control points to ensure thermal comfort by regulating the supply air temperature and/or reheat coil power [6].

B Additional experiment details

Dataset Generation The control agents in the policy library are trained following the approach outlined in [6]. We use PPO with a clipping parameter $\epsilon = 0.2$, which constrains policy updates within a trust region to ensure stability. The actor and critic networks are implemented with two hidden layers, each consisting of 64 units, and use the hyperbolic tangent as the activation function. The learning rate is fixed at 0.0003, and the batch size is set to 2,976, corresponding to the length of one episode. The EnergyPlus model, used for simulating building operations, operates with 15-minute time steps, and each episode spans one month. Weather data from January 1991 is employed for training. All policies are trained using PPO within a multi-agent reinforcement learning (MARL) framework for 1,000 episodes, incorporating both environment and policy diversity as described in [6].

Pretraining The HVAC-DPT model was trained using the policy library under the following conditions: a horizon of 2,967 steps, a learning rate of 0.001, and a dropout rate of 0.0. The Transformer model architecture consisted of three layers, with eight attention heads and an embedding dimension of 128. The training process was carried out for 100 trajectories over 118 epochs using the AdamW optimizer with a weight decay of 0.0001. The loss function employed was MSE. The model was evaluated using a test split of 20%, and the training was conducted using the PyTorch framework.

Online Deployment We trained HVAC-DPT on B_{train} , a small office prototype building as defined by the ASHRAE Standard 90.1 [20]. B_{train} is located in Denver, Colorado, and contains five thermal zones, each having an AHU and VAV system. The total floor area of this building is 511.16 m^2 . We used the approach presented in [6] to build a diverse policy library of PPO agents. We analysed HVAC-DPT’s performance on an unseen building B_{Denver} , a medium office prototype. B_{Denver} is located in Denver, Colorado, and contains 15 thermal zones across three floors. Each floor consists of

5 zones and has an AHU and 5 VAV systems. The total floor area of this building is $4,982.19 \text{ m}^2$. We used weather data from the year 2000 and average monthly energy consumption values over 10 runs.

C Additional results

Table 2: Monthly total HVAC energy consumption (in MWh) of different controllers during the first year after deployment in B_{Denver} .

Controller	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Baseline	56.75	62.52	56.30	55.78	44.54	73.15	73.38	71.65	57.57	50.87	53.87	74.38
Expert	29.51	35.77	29.92	29.15	23.88	61.16	61.84	59.92	42.88	27.11	28.07	43.62
MARL	67.74	74.63	67.84	68.46	56.73	84.75	84.42	81.74	67.13	58.39	62.24	81.85
SARL	67.25	74.55	68.03	69.15	57.78	85.92	86.05	83.97	69.75	61.23	65.24	84.85
HVAC-DPT	32.38	38.40	32.58	31.67	26.23	63.01	63.71	61.70	44.88	29.77	30.80	46.71

Table 3: Yearly percentile difference of total HVAC energy consumption compared to HVAC-DPT during the first year after deployment in B_{Denver} .

Controller	$\Delta_{HVAC-DPT}$
Baseline	+ 45.62%
Expert	- 5.78%
MARL	+ 70.56%
SARL	+ 74.12%