HVAC-SPICE: Value-Uncertainty In-Context RL with Thompson Sampling for Zero-Shot HVAC Control

Anaïs Berkes
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
amcb6@cam.ac.uk

Vincent Taboga
Mila
Université de Montréal
Montreal, QC, Canada
vincent.taboga@mila.quebec

Donna Vakalis
Mila
Université de Montréal
Montreal, QC, Canada
donna.vakalis@olympian.org

David Rolnick
Mila
McGill University
Montreal, QC, Canada
david.rolnick@mila.quebec

Yoshua Bengio
Mila
Université de Montréal
Montreal, QC, Canada
yoshua.bengio@mila.quebec

Abstract

Urban buildings consume 40% of global energy, yet most rely on inefficient rule-based HVAC systems due to the impracticality of deploying advanced controllers across diverse building stock. In-context reinforcement learning (ICRL) offers promise for rapid deployment without per-building training, but standard supervised learning objectives that maximise likelihood of training actions inherit behaviour-policy bias and provide weak exploration under the distribution shifts common when transferring across buildings and climates. We present **SPICE** (**S**ampling **P**olicies **I**n-**C**ontext with **E**nsemble uncertainty), a novel ICRL method specifically designed for zero-shot building control that addresses these fundamental limitations. SPICE introduces two key methodological innovations: (i) a propensity-corrected, return-aware training objective that prioritises high-advantage, high-uncertainty actions to enable improvement beyond suboptimal training demonstrations, and (ii) lightweight value ensembles with randomised priors that provide explicit uncertainty estimates for principled episode-level Thompson sampling. At deployment, SPICE samples one value head per episode and acts greedily, resulting in temporally coherent exploration without test-time gradients or building-specific models. We establish a comprehensive experimental protocol using the HOT dataset to evaluate SPICE across diverse building types and climate zones, focusing on the energy efficiency, occupant comfort, and zero-shot transfer capabilities that are critical for urban-scale deployment.

1 Introduction and related work

Advanced controllers have the potential to significantly reduce HVAC energy consumption Drgoňa et al. [2020], 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 Drgoňa et al. [2020], Wang and Hong [2020], 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 GS et al. [2023], which often leads to suboptimal performance and occupant discomfort during the learning phase Zhang et al. [2021]. 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 Zisman et al. [2023], Zhang et al. [2022].

The transformer architecture Vaswani [2017] 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 Sinii et al. [2023]. Recent research, such as the Decision-Pretrained Transformer (DPT) by Lee *et al.* Lee et al. [2024] and Algorithm Distillation by Laskin *et al.* Laskin et al. [2022], 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 Zhang et al. [2024], Mukherjee et al. [2024], Lin et al. [2023], Dai et al. [2024]. Nevertheless, the application of in-context RL to HVAC control remains unexplored. Prior work on HVAC-DPT explored decision-pretrained, model-free test-time control [Berkes, 2024], but this method did not address fundamental limitations that prevent effective deployment: behaviour-policy bias that constrains improvement beyond training demonstrations, and insufficient exploration under distribution shifts.

Our contributions. We introduce **SPICE** (**S**ampling **P**olicies **I**n-**C**ontext with **E**nsemble uncertainty), the first in-context RL method specifically designed for building control that addresses these critical gaps. Our key contributions are: (i) a novel propensity-corrected, return-aware training objective that mitigates behaviour-policy bias through advantage-weighted learning with epistemic uncertainty, enabling improvement beyond training demonstrations; (ii) explicit value uncertainty estimation via lightweight ensembles with randomised priors that enables principled, temporally coherent exploration through Thompson sampling without requiring test-time gradients; and (iii) a comprehensive experimental framework designed to validate zero-shot transfer capabilities across building types and climate zones, with evaluation protocols specifically targeting the energy efficiency and comfort metrics critical for urban deployment.

2 Problem setting

We control a multi–zone building at 15-min intervals using the `Minergym` wrapper. At time t, the observation vector $x_t \in \mathbb{R}^{N_z+6}$ contains: (i) zone air temperatures $\{T_t^i\}_{i=1}^{N_z}$, (ii) outdoor drybulb temperature and relative humidity, (iii) current time of day and day of week, (iv) HVAC electricity and natural gas consumption. For each zone we control two continuous actions:

$$a_t^i = (T_t^{\text{heat},i},\ \Delta_t^i) \in [15, 25] \times [1, 15] \ \ (^\circ\text{C}),$$

with the cooling setpoint computed as $T_t^{\text{cool},i} = T_t^{\text{heat},i} + \Delta_t^i$.

The reward weights comfort and energy use with occupancy-aware temperature penalties:

$$r_t \ = \ -\Big((1+\omega_t)\tfrac{1}{|\mathcal{Z}_c|}\sum_{i\in\mathcal{Z}_c}(T_t^i - 21)^2 \ + \ \beta\ \tfrac{E_t^{\text{hvac}}}{A_{\text{floor}}\cdot 3600}\Big),$$

where $\mathcal{Z}_c$ are controlled zones, β is an energy weight, and ω_t is the occupancy factor derived from time-of-day and day-of-week schedules.

3 SPICE

SPICE addresses key ICRL limitations: behaviour-policy bias and insufficient exploration under deployment shifts. SPICE combines a propensity-corrected, return-aware training objective with explicit value uncertainty for principled exploration without test-time gradients.

Algorithm 1 SPICE: Sampling Policies In-Context with Ensemble Value Uncertainty

// Initialisation
Input: Training dataset $\mathcal{D}$, transformer architecture, new building environment
Initialise transformer trunk f_θ, policy head π_θ, value heads $\{Q_{\phi_k}\}_{k=1}^K$
Initialise and freeze randomised priors $\{p_k\}_{k=1}^K$ (2-layer MLPs, width 64)
Train behaviour model $\hat{\pi}_b$ on $\mathcal{D}$
// Offline Pretraining Phase
while not converged **do**
  Sample minibatch $(s, a, r, s', \mathcal{C})$ from $\mathcal{D}$
  Compute transformer representations: $h = f_\theta(s, \mathcal{C}, G)$
  // Value ensemble update
  for $k = 1$ to K **do**
    Compute $Q_{\phi_k}(s, a) = \tilde{Q}_{\phi_k}(h) + \alpha_{\text{prior}} \cdot p_k(s, a)$
    Compute TD target: $y_k = r + \gamma \max_{a'} \bar{Q}_{\phi_k}(s', a')$
    Update: $\phi_k \leftarrow \phi_k - \eta \nabla_{\phi_k} (Q_{\phi_k}(s, a) - y_k)^2$
  end for
  // Policy update
  Compute ensemble mean: $\overline{Q}(s, a) = \frac{1}{K} \sum_k Q_{\phi_k}(s, a)$
  Compute advantage: $A_\Phi(s, a) = \overline{Q}(s, a) - \max_{a'} \overline{Q}(s, a')$
  Compute uncertainty: $\sigma_\Phi(s, a) = \text{StdDev}_k(Q_{\phi_k}(s, a))$
  Compute weights: $w = \frac{\pi_u(a|s)}{\hat{\pi}_b(a|s)} \cdot \exp(A_\Phi/\alpha_A) \cdot \text{clip}(1 + \lambda_\sigma \sigma_\Phi, 1, c_\sigma)$
  Update policy: $\theta \leftarrow \theta - \eta \nabla_\theta [w \cdot \log \pi_\theta(a \mid s, \mathcal{C}, G)]$
end while
// Zero-Shot Deployment Phase
Initialise empty context $\mathcal{C} = \{\}$ for each building zone
for episode $e = 1, 2, \ldots$ **do**
  Sample value head: $k \sim \text{Uniform}\{1, \ldots, K\}$
  Reset environment: $s_1 \leftarrow$ env.reset()
  for timestep $t = 1$ to H **do**
    $a_t \leftarrow \arg\max_a Q_{\phi_k}(s_t, a \mid \mathcal{C})$
    Execute action: $(r_t, s_{t+1}) \leftarrow$ env.step(a_t)
    Store transition: $(s_t, a_t, r_t) \rightarrow$ episode buffer
  end for
  Compute episode summary: $R_e = \sum_t r_t$, $\text{RtI}_e = R_e - \max_{j<e} R_j$
  Update context: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\text{episode buffer, summary tokens}\}$
  // Maintain rolling window if context size limit reached
end for=0

Transformer Backbone. We use a causal decision-pretrained transformer that processes multi-episode contexts $\mathcal{C} = \{E_1, \ldots, E_K\}$ of recent trajectories. The transformer trunk f_θ produces shared representations $h_t = f_\theta(s_t, \mathcal{C})$ and supports: (i) a policy head $\pi_\theta(a \mid s_t, \mathcal{C})$, and (ii) an ensemble of K value heads $\{Q_{\phi_k}\}_{k=1}^K$ with independent final layers.

Explicit value uncertainty. Each value head is augmented with a frozen randomised prior function $p_k(s, a)$:

$$Q_{\phi_k}(s, a \mid \mathcal{C}) = \tilde{Q}_{\phi_k}(s, a \mid \mathcal{C}) + \alpha_{\text{prior}} \cdot p_k(s, a), \tag{1}$$

where only $\tilde{Q}_{\phi_k}$ is trainable. Randomised priors preserve hypothesis diversity in sparse data regions.

The ensemble provides: (i) mean $\overline{Q}(s, a) = \frac{1}{K} \sum_k Q_{\phi_k}(s, a)$, (ii) epistemic uncertainty $\sigma_\Phi(s, a) = \text{StdDev}_k[Q_{\phi_k}(s, a)]$, and (iii) advantage $A_\Phi(s, a) = \overline{Q}(s, a) - \max_{a'} \overline{Q}(s, a')$. Each head trains with standard TD targets:

$$\mathcal{L}_{\text{TD}}^{(k)} = \mathbb{E}\left[\left(r + \gamma \max_{a'} \bar{Q}_{\phi_k}(s', a' \mid \mathcal{C}) - Q_{\phi_k}(s, a \mid \mathcal{C})\right)^2\right]. \tag{2}$$

Propensity-corrected, return-aware sequence objective. Standard sequence models learn to reproduce behaviour policy distributions $p(a \mid s, \mathcal{C}, \pi_b)$, which is biased when π_b is suboptimal. We

de-bias using inverse propensities and focus learning where it matters:

$$\mathrm{L}_{\text{SPICE}} = -\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\underbrace{\frac{\pi_u(a \mid s)}{\hat{\pi}_b(a \mid s)}}_{\text{de-bias}} \cdot \underbrace{\exp(A_\Phi(s,a)/\alpha_A)}_{\text{return-aware}} \cdot \underbrace{\text{clip}(1+\lambda_\sigma\,\sigma_\Phi(s,a),\,1,\,c_\sigma)}_{\text{uncertainty-seeking}} \cdot \log \pi_\theta(a \mid s, \mathcal{C})\right] \tag{3}$$

where $\hat{\pi}_b$ is a behaviour model fit to offline data and π_u is a uniform reference policy. This weighting prioritises high-advantage, high-uncertainty actions, which is crucial when training data contains only random and constant policies.

Deployment: episode-level Thompson sampling. At deployment, SPICE samples one value head $k \sim \text{Uniform}\{1, \ldots, K\}$ per episode and acts greedily:

$$a_t = \arg\max_a Q_{\phi_k}(s_t, a \mid \mathcal{C}). \tag{4}$$

This episode-level commitment yields temporally coherent exploration with no test-time gradients, dynamics models, or Bayesian inference.

SPICE addresses critical deployment challenges in building control through four key advantages. Debiased sequence learning enables the method to improve beyond the quality of training demonstrations, which is essential since the training data typically consists of simple baseline controllers rather than optimal policies. Explicit value uncertainty estimation provides robust exploration under the climate and occupancy distribution shifts which are present when deploying to new buildings, ensuring stable performance across diverse operating conditions. The method maintains low operational costs by avoiding computationally expensive dynamics models or gradient updates, making it suitable for deployment on standard building management systems. Finally, the shared transformer trunk with zone-specific contexts enables efficient scaling to multi-zone buildings without proportional increases in model complexity.

4 Experimental Design and Methodology

We design an experimental framework to validate SPICE's zero-shot transfer across building types and climate zones.

Data and Protocol. Using HOT's standardised epJSON–EPW pairs Berkes et al. [2025], we train on *Office Small* buildings across six climates {3B El Paso, 3C San Diego, 4A New York, 4B Albuquerque, 5A Buffalo, 5B Denver} and test zero-shot transfer to *Office Medium* buildings in {4C Seattle, 7 International Falls, 0B Dubai, 3A Atlanta}. This tests generalisation across building types and climate zones critical for urban deployment.

Training Protocol. Dataset $\mathcal{D}$ comprises only *random* and *constant* controllers (1 week per building–weather pair) to simulate realistic scenarios where optimal demonstrations are unavailable. We train $K \in \{5, 7\}$ value heads with frozen randomised priors (2-layer MLPs, width 64) and optimise $\mathcal{L}_{\text{SPICE}}$ with proposal selection for continuous actions.

Evaluation Metrics. We assess three urban deployment criteria: (i) HVAC energy consumption (kWh), (ii) occupant comfort (% time within temperature bands), and (iii) adaptation speed (episodes to 90% performance). Primary evaluation uses Seattle 4C, Office Medium over one week versus Static and ASHRAE baselines.

Expected Results. We will verify whether SPICE achieves: (a) energy reduction while maintaining comfort, (b) robust TMY→AMY climate transfer, and (c) rapid adaptation within 2-3 episodes despite suboptimal training data. Results forthcoming in camera-ready version.

5 Discussion and Conclusion

SPICE's methodological innovations directly address deployment barriers in building control. The propensity-corrected objective and ensemble uncertainty estimation enable zero-shot transfer across

building types and climates while maintaining computational efficiency suitable for standard building management systems.

The implications for urban deployment are substantial: eliminating extensive training phases while providing robust performance across diverse building stock. Future work includes extension to multi-building portfolios and renewable energy integration. While limitations include behaviour model requirements and potential calibration drift, these represent manageable engineering challenges. SPICE demonstrates that uncertainty-aware in-context RL can finally enable scalable intelligent building control. Expected Results: Based on our methodology, we will verify whether SPICE can achieve: (a) significant energy reduction compared to ASHRAE baselines while maintaining occupant comfort standards, (b) robust performance across TMY→AMY climate shifts through episode-level Thompson sampling, and (c) rapid adaptation within 2-3 episodes despite training only on suboptimal random/constant policies via our de-biased, advantage-weighted training objective. Experimental validation is ongoing with results forthcoming in the camera-ready version.

References

Anaïs Berkes. Hvac-dpt: A decision pretrained transformer for hvac control. *arXiv preprint arXiv:2411.19746*, 2024.

Anaïs Berkes, Yoshua Bengio, David Rolnick, and Donna Vakalis. A hot dataset: 150,000 buildings for hvac operations transfer research. In *Proceedings of the 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 171–180, 2025.

Zhenwen Dai, Federico Tomasi, and Sina Ghiassian. In-context exploration-exploitation for reinforcement learning. *arXiv preprint arXiv:2403.06826*, 2024.

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.

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.

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.

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.

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.

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.

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.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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

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.

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.

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.

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.