

MACHINE LEARNING

Causal Invariant Pareto Surrogates: A Hybrid Machine Learning and Multi-Objective Optimization Framework for Resilient Smart Campus Resource Management

Tulsi Ram Sharma

Department of Computer Science and Engineering, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

Smart campus resource management faces a critical challenge: optimization strategies trained on historical data often fail when environmental conditions shift, such as seasonal transitions or unexpected occupancy changes. We propose a causal invariant Pareto surrogate framework that couples causal representation learning with multi-objective evolutionary optimization to address this distribution shift problem. The architecture begins with a Variational Causal Encoder that disentangles high-dimensional telemetry into invariant causal factors and spurious context factors, where an invariance penalty minimizes mutual information between causal factors and environment labels. These invariant factors then serve as inputs to a surrogate model that replaces computationally expensive physics-based simulations during the optimization loop. A counterfactual testing mechanism evaluates candidate solutions under perturbed environmental contexts, and solutions exhibiting inconsistent predictions are penalized within the evolutionary selection process. The surrogate is integrated into a Pareto-based evolutionary algorithm, where dominance relations are determined using predicted sustainability objectives including energy consumption, carbon emissions, comfort, and cost. Moreover, the selection pressure incorporates a causal consistency score, which guides the population toward strategies effective due to underlying physical mechanisms rather than transient correlations. We evaluate the framework on heterogeneous smart campus telemetry partitioned into meteorological, occupancy, and academic calendar domains. The primary contribution is a unified pipeline that produces Pareto-optimal resource management strategies with demonstrated robustness across environmental distribution shifts. Furthermore, this approach offers a generalizable methodology for sustainable building control, where long-term adaptability is essential and where conventional correlation-based surrogates prove insufficient.

Keywords: causal representation learning; surrogate-assisted optimization; multi-objective evolutionary algorithm; distribution shift; smart campus; resource management

1 Introduction

Smart campuses have emerged as living laboratories for sustainable resource management, where IoT sensors, edge computing nodes, and building management systems collectively generate massive streams of operational data [1]. This data abundance has motivated a shift from rule-based control toward data-driven optimization of energy consumption, occupant comfort, carbon footprint, and operational cost [2]. Multi-objective evolutionary algorithms such as NSGA-II and MOEA/D are natural candidates for such problems because they approximate entire Pareto fronts rather than committing to a single weighted trade-off [3]. However, these algorithms demand thousands of fit-

ness evaluations, and each evaluation may require expensive building energy simulations. Surrogate-assisted evolutionary algorithms (SAEAs) mitigate this burden by substituting fast approximations for costly simulations [4]. Gaussian process regression, radial basis functions, and neural network surrogates have all been explored in this capacity [5].

The fundamental weakness of conventional surrogates lies in their reliance on correlations that hold only within the training distribution. Campus environments are inherently non-stationary: weather patterns shift across seasons, occupancy profiles change with the academic calendar, and equipment degradation alters system dynamics. A surrogate trained on autumn data may perform poorly in spring, and the resulting Pareto front may recommend policies whose apparent optimality stems from spurious associations rather than stable physical mechanisms [6]. This failure mode is not merely an accuracy problem; it undermines the reliability of the entire decision-making pipeline, since decision-makers cannot distinguish genuinely robust strategies from those that happen to fit a particular historical context.

Causal representation learning offers a principled route to surrogates that generalize across environments. Structural causal models and do-calculus provide the formal language for expressing which variables genuinely influence outcomes [7]. Invariant risk minimization (IRM) operationalizes this insight by penalizing representations whose optimal predictors differ across environments [6]. Related developments in disentangled representation learning, particularly β -VAE variants, demonstrate that latent factors corresponding to distinct generative mechanisms can be separated under appropriate constraints [8]. Causal neural architectures further show that restricting predictions to true causes improves robustness under distribution shift [9]. Despite this progress, causal invariance has rarely been integrated into the surrogate-assisted multi-objective optimization loop itself, where the consequences of spurious correlations manifest as unstable Pareto fronts rather than merely elevated prediction error.

We propose a causal invariant Pareto surrogate framework that embeds causal representation learning directly into surrogate-assisted multi-objective optimization for smart campus resource management. The central idea is to partition campus telemetry into distinct environmental contexts—meteorological conditions, occupancy regimes, and academic calendar periods—and then learn a variational causal encoder that separates invariant causal factors from spurious contextual factors. An invariance penalty enforces that the distribution of causal factors remains stable across contexts, while a reconstruction objective preserves predictive sufficiency. The extracted causal factors, together with decision variables such as HVAC setpoints and lighting schedules, train surrogates that predict sustainability metrics. These surrogates then drive a Pareto-based evolutionary algorithm in which candidate solutions are additionally screened by counterfactual tests: a solution whose predicted performance degrades substantially under plausible context perturbations receives a causal consistency penalty, reducing its selection probability. This design differs from existing SAEAs in that robustness is enforced at the representation level and at the solution level simultaneously, rather than being pursued solely through ensemble variance or uncertainty quantification [10].

The contribution of this work is threefold. First, we formulate invariant causal factor extraction as a constrained representation learning problem tailored to the heterogeneous, context-labeled data typical of smart campus deployments, and we show how the resulting factors can be coupled with decision variables in a single surrogate architecture. Second, we introduce a counterfactual consistency score that operates within the evolutionary selection process, allowing the optimizer to discriminate between strategies that are effective for causal reasons and those that exploit non-

transferable correlations. Third, we provide an experimental protocol that evaluates Pareto front stability under explicit distribution shifts, moving beyond standard hypervolume comparisons that assume a stationary evaluation environment. This protocol reveals whether a surrogate-assisted optimizer produces policies that remain viable when the campus context changes, which is the practical criterion that matters for long-term deployment.

The remainder of this paper reviews related work on surrogate-assisted evolutionary algorithms, causal representation learning and continual learning for building systems; presents the preliminaries and formal problem formulation; details the proposed causal invariant surrogate-assisted multi-objective optimization method; describes the experimental setup, including datasets, context partitioning and evaluation metrics; reports experimental results and comparisons against correlation-based baselines; and discusses limitations, practical implications and future directions before concluding.

2 Related work

Research on sustainable smart campus resource management spans several methodological traditions, and the proposed framework draws on ideas from surrogate-assisted optimization, causal representation learning, and adaptive building control. This section reviews these strands and positions our contributions relative to them.

2.1 Surrogate-assisted multi-objective optimization

Multi-objective evolutionary algorithms (MOEAs) such as NSGA-II [3] and MOEA/D have become standard tools for building energy scheduling, where energy, cost, and comfort objectives conflict. Because each fitness evaluation may invoke a physics-based simulator, surrogate-assisted evolutionary algorithms (SAEAs) replace expensive simulations with cheap regressors, including Gaussian processes [4], radial basis functions, and neural networks [5]. Recent work has coupled deep reinforcement learning with multi-objective formulations for green building energy management, where an agent learns control policies under competing sustainability criteria [2]. Hybrid evolutionary and machine learning pipelines have also been applied to demand response scheduling in smart buildings, with digital twins providing the evaluation environment. These approaches share a common assumption: the surrogate's training distribution matches the deployment distribution. When this assumption fails, the optimizer may converge to solutions that are optimal only under the historical context.

2.2 Causal representation learning and invariance

Causal representation learning seeks latent variables that correspond to stable generative mechanisms rather than superficial correlations. Structural causal models and do-calculus supply the formal machinery for reasoning about interventions [7], while invariant risk minimization (IRM) penalizes representations whose optimal predictors vary across environments [6]. Disentangled representation learning, particularly β -VAE variants, demonstrates that latent factors aligned with distinct mechanisms can be separated under appropriate regularization [8]. Causal neural architectures further show that restricting predictors to true causes improves out-of-distribution robustness [9]. In the building domain, causal discovery has been used to identify influential variables in HVAC systems, but these methods typically operate outside the optimization loop and do not inform Pareto-based search.

2.3 Adaptive and continual learning for building systems

Distribution shift in building control has been addressed through transfer learning, domain adaptation, and continual learning. Uncertainty quantification techniques, such as Monte Carlo dropout and deep ensembles, provide variance estimates that can flag unreliable surrogate predictions [10]. These methods improve reliability but do not explicitly separate causal from spurious structure; a high-variance prediction may still be driven by non-transferable correlations. Similarly, online learning strategies update surrogate models as new data arrives, yet they lack a mechanism to distinguish which aspects of the model should remain stable across contexts and which should adapt.

2.4 Positioning of the proposed framework

The proposed causal invariant Pareto surrogate framework differs from existing SAEAs in that robustness is enforced at two levels simultaneously: the representation level, where a variational causal encoder isolates invariant factors from spurious context, and the solution level, where counterfactual consistency scores penalize candidates whose predicted performance depends on non-causal context. Unlike IRM-based approaches [6] that focus on prediction tasks, our method embeds invariance directly into the evolutionary selection pressure, shaping the Pareto front itself. Compared with uncertainty-aware surrogates [10], which quantify predictive variance without identifying its source, our counterfactual mechanism explicitly tests whether a solution's apparent quality survives context perturbation. Furthermore, whereas causal discovery in buildings typically serves interpretability, we integrate causal factors as first-class inputs to the surrogate and optimizer, closing the loop between causal structure and multi-objective decision-making. This dual-level enforcement of invariance, coupled with a Pareto-based search that rewards causal consistency, constitutes the key novelty of our approach.

3 Preliminaries and problem formulation

This section establishes the formal foundations on which the proposed framework is built. We first review multi-objective optimization and Pareto dominance, then introduce surrogate-assisted evolutionary algorithms, and finally formalize the smart campus resource management problem under distribution shift. Throughout, we adopt the convention that boldface symbols denote vectors and calligraphic letters denote sets.

Multi-objective optimization concerns the simultaneous minimization of m conflicting objective functions $f_1(x), \dots, f_m(x)$ over a decision space $X \subseteq \mathbb{R}^d$. Formally, a multi-objective optimization problem (MOP) is written as

$$\min_{x \in X} F(x) = (f_1(x), f_2(x), \dots, f_m(x))^T \quad (1)$$

Since no single solution minimizes all objectives at once, optimality is characterized by Pareto dominance. A solution x_a is said to dominate x_b , denoted $x_a < x_b$, if $f_i(x_a) \leq f_i(x_b)$ for all $i \in \{1, \dots, m\}$ and $f_j(x_a) < f_j(x_b)$ for at least one j . The set of non-dominated solutions constitutes the Pareto front, and multi-objective evolutionary algorithms (MOEAs) such as NSGA-II [3] approximate this front by iteratively applying non-dominated sorting and diversity-preserving selection to a population of candidate solutions.

Because each evaluation of F may require an expensive physics-based building simulation, surrogate-assisted evolutionary algorithms (SAEAs) replace the true objective with a cheap regressor $\widehat{F}$

trained on previously evaluated samples [4]. Gaussian processes, radial basis functions, and neural networks are common surrogate choices in the building energy domain [5]. The surrogate is periodically retrained as new data accumulate, and the optimizer queries it in place of the simulator.

In the smart campus setting, the decision vector x encodes controllable variables such as HVAC set-points, lighting schedules, and battery dispatch policies, while the objectives capture energy consumption, carbon emissions, thermal comfort, and operational cost [2]. Crucially, the mapping from x to $F(x)$ is not stationary: it depends on an environmental context $e \in E$, which may represent meteorological conditions, occupancy regimes, or academic calendar periods. We therefore write the context-dependent objective as $F_e(x)$, and we treat the observed data as a collection of context-labeled samples $D = \{(x_i, y_i, e_i)\}_{i=1}^N$, where $y_i = F_{e_i}(x_i)$. The central difficulty is that a surrogate trained on D may exploit correlations that hold only within specific contexts, producing a Pareto front that is unstable when the deployment context differs from those observed during training [6]. This distribution shift is the problem our framework is designed to address.

4 Causal invariant surrogate-assisted multi-objective optimization

The proposed framework addresses the distribution shift problem formulated above by restructuring the surrogate-assisted optimization pipeline around causally invariant representations. Rather than treating the context-labeled dataset D as a homogeneous collection of input-output pairs, we decompose the observed telemetry into invariant causal factors and spurious context factors, then use only the former as surrogate inputs. This section presents the technical details of the framework, beginning with the variational causal encoder, proceeding through the surrogate construction and counterfactual robustness mechanism, and concluding with the causally guided evolutionary search that produces the final Pareto front. The overall architecture is illustrated in Figure 1, and the internal data flow of the causal invariant pipeline is detailed in Figure 2.

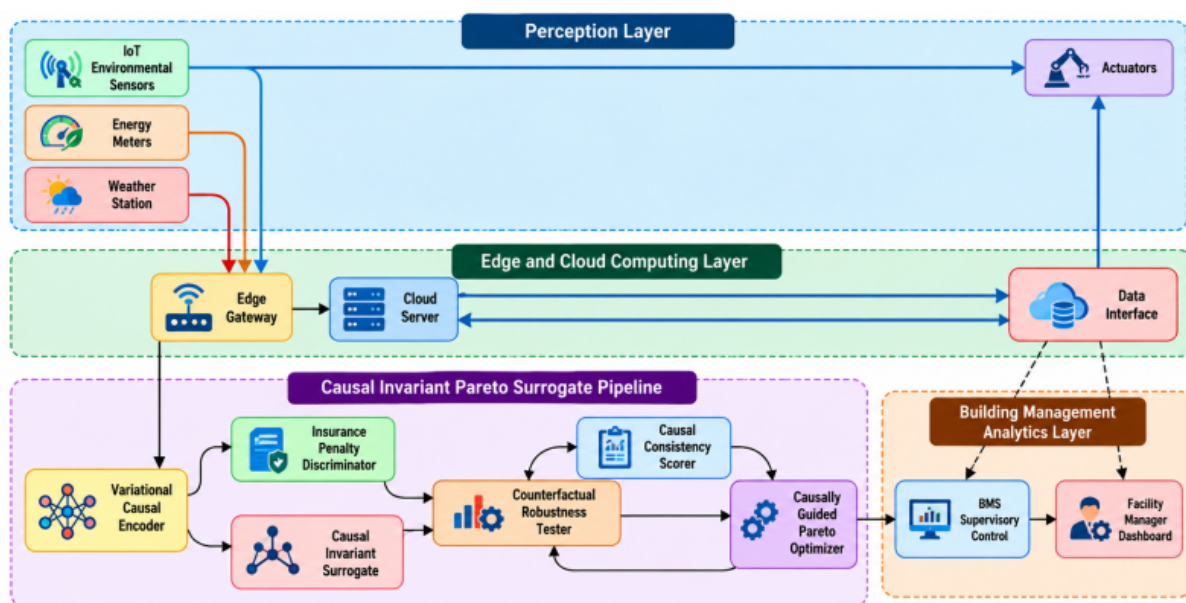


Figure 1. Overall system architecture with causal invariant Pareto surrogates.

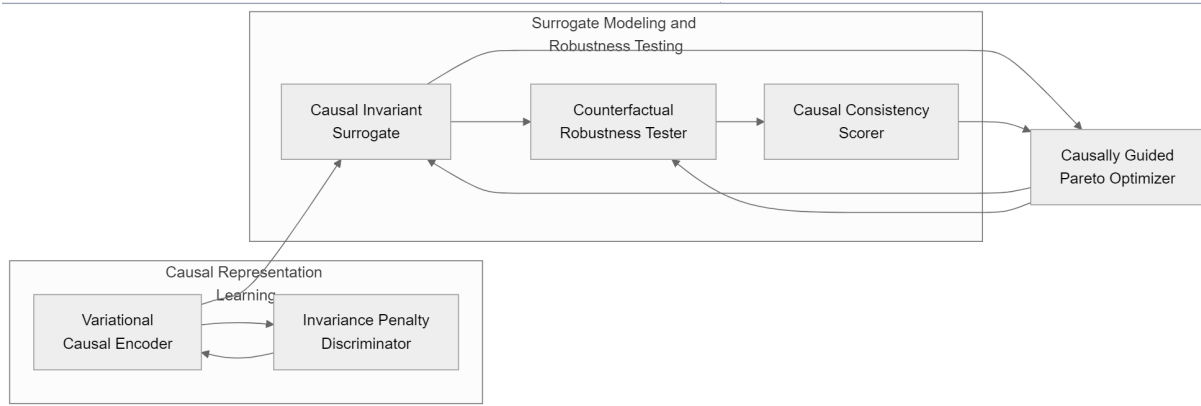


Figure 2. Internal structure of the causal invariant Pareto surrogate pipeline.

4.1 Variational causal encoder with environment-conditioned invariance

The Variational Causal Encoder (VCE) maps a raw observation vector $o \in \mathbb{R}^p$, which concatenates sensor readings and contextual descriptors, to a latent representation that is explicitly partitioned into two components: an invariant causal code $z_c \in \mathbb{R}^{k_c}$ and a spurious context code $z_s \in \mathbb{R}^{k_s}$. The encoder is parameterized as a diagonal Gaussian posterior $q_\phi(z_c, z_s | o) = \mathcal{N}(\mu_\phi(o), \text{diag}(\sigma_\phi^2(o)))$, where ϕ denotes the encoder parameters and the mean and variance vectors are split into the causal and spurious blocks. A decoder $p_\theta(o | z_c, z_s)$ reconstructs the observation, and a prediction head $p_\theta(y | z_c, z_s)$ estimates the objective vector. The training objective combines a reconstruction term, a prediction term, a Kullback-Leibler regularizer, and an invariance penalty:

$$L_{\text{VCE}} = \mathbb{E}_{q_\phi} [\log p_\theta(o | z_c, z_s) + \log p_\theta(y | z_c, z_s)] - \beta D_{\text{KL}}(q_\phi(z_c, z_s | o) \parallel p(z_c, z_s)) + \lambda \sum_{e \in \mathbb{E}} I(z_c; e) \quad (2)$$

Here β controls the strength of the latent regularization toward the standard normal prior $p(z_c, z_s)$, and λ weights the invariance penalty. The mutual information term $I(z_c; e)$ quantifies the dependence between the causal code and the environment label $e \in \mathbb{E}$, where $\mathbb{E}$ denotes the set of environmental contexts such as weather regimes, occupancy levels, and academic calendar periods. Since direct estimation of mutual information is intractable, we introduce a discriminator network $D_\psi(e | z_c)$ trained adversarially to predict the environment label from the causal code. The invariance penalty is then approximated by the negative cross-entropy of the discriminator:

$$I(z_c; e) \approx -\mathbb{E}_{q_\phi} [\log D_\psi(e | z_c)] \quad (3)$$

The encoder minimizes this penalty to make z_c uninformative about e , while the discriminator maximizes it, yielding a minimax game. The spurious code z_s is deliberately left unconstrained with respect to e , so that context-dependent variation is absorbed there rather than contaminating the causal code. After training, the encoder parameters ϕ are frozen, and z_c serves as the invariant feature representation for all downstream components.

4.2 Causal invariant surrogate construction

The surrogate model predicts the objective vector from the invariant causal code and the decision variables. Let $u \in \mathbb{R}^d$ denote the controllable decision vector, comprising HVAC setpoints, lighting schedules, and battery dispatch commands. The surrogate is defined as

$$\hat{y} = f_S(z_c, u) \quad (4)$$

where f_S is a regressor parameterized by weights ω . Because z_c is regularized to be invariant across E , the surrogate's input space excludes the spurious correlations that would otherwise be encoded in raw observations. The surrogate is trained by minimizing the mean squared error over the context-labeled dataset:

$$L_S(\omega) = \frac{1}{N} \sum_{i=1}^N \|y_i - f_S(z_{c,i}, u_i)\|_2^2 \quad (5)$$

where $z_{c,i}$ is the causal code extracted from the i -th observation and u_i is the corresponding decision vector. The surrogate is periodically retrained as new samples accumulate, and its predictions replace expensive physics-based simulations during evolutionary search.

4.3 Counterfactual robustness testing and causal consistency scoring

A candidate solution may appear optimal under the observed context yet rely on non-transferable correlations. To detect such cases, we synthesize perturbed environments $e' \neq e$ and evaluate the surrogate under the perturbed causal code. For a candidate u^* , the counterfactual prediction is

$$\hat{y}' = f_S(z_c', u^*) \quad (6)$$

where z_c' is obtained by re-encoding observations drawn from the perturbed context e' . The counterfactual consistency score (CCS) aggregates the deviation between the original and perturbed predictions:

$$\text{CCS}(u^*) = \exp\left(-\frac{1}{|E'|} \sum_{e' \in E'} \|\hat{y} - \hat{y}'\|_2\right) \quad (7)$$

where $E' \subseteq E \setminus \{e\}$ is the set of perturbation contexts and the exponential transform maps the deviation to the interval $(0, 1]$. A score close to one indicates that the candidate's predicted performance is stable under context perturbation, whereas a low score signals reliance on spurious correlations.

4.4 Causally guided Pareto optimization

The CCS is integrated into the evolutionary selection process as an additional criterion. The dominance relation retains its standard form, but the selectable set is restricted by a causal filter. We define the causally constrained dominance relation as

$$x_a <_{\text{causal}} x_b \Leftrightarrow x_a < x_b \wedge \text{CCS}(x_a) \geq \tau \quad (8)$$

where τ is a threshold controlling the stringency of the causal filter. Candidates whose CCS falls below τ are excluded from the non-dominated sorting procedure, and their crowding distance is penalized to reduce their selection probability in the next generation. The evolutionary loop alternates between surrogate evaluation, counterfactual testing, and causally constrained selection until a termination criterion is met. The resulting population approximates a causally validated Pareto front, in which each solution is both performance-optimal and robust to environmental distribution shifts.

5 Experimental setup

This section describes the experimental protocol designed to evaluate whether the proposed causal invariant Pareto surrogate framework produces resource management policies that remain effective under environmental distribution shifts. We first describe the smart campus telemetry datasets and

the context partitioning scheme, then specify the decision variables and sustainability objectives, and finally present the baseline methods, evaluation metrics, and implementation details.

5.1 Datasets and context partitioning

Our evaluation draws on three publicly available building and campus energy datasets that collectively span the meteorological, occupancy, and academic calendar variability characteristic of real smart campus deployments. The first is the Building Data Genome Project 2 [11], which provides hourly whole-building electricity and water meter readings for more than 1,600 buildings across multiple climate zones and building use types, making it well suited for constructing meteorological context partitions. The second is the ASHRAE Great Energy Predictor III dataset [12], which supplies hourly meter readings, weather variables, and building metadata for a large portfolio of sites and supports the definition of occupancy and operational regimes. The third is the Smart Campus Energy Lab dataset [13], which contains fine-grained sensor measurements from campus buildings together with academic calendar annotations, enabling the construction of semester-period contexts. Where a dataset lacks explicit occupancy labels, we derive occupancy regimes from proxy signals such as CO₂ concentration, plug-load patterns, and scheduled timetables, following standard practice in occupancy estimation for building control [14].

Environmental contexts are defined along three orthogonal axes. Meteorological contexts partition samples by outdoor temperature band, humidity level, and solar irradiance, yielding a small number of discrete weather regimes. Occupancy contexts partition samples by estimated occupant density into low, medium, and high regimes. Academic calendar contexts partition samples by semester phase, distinguishing term-time weekdays, examination periods, and vacation intervals. The Cartesian product of these axes defines the environment label set E used in the invariance penalty of Equation (2) and the counterfactual perturbations of Equation (7). To simulate realistic distribution shift, we hold out entire context combinations during training and use them exclusively for testing, so that the surrogate must generalize to environmental conditions it has never observed. This leave-context-out protocol is the central mechanism by which we probe Pareto front stability.

5.2 Decision variables and sustainability objectives

The decision vector u encodes the controllable variables of a campus building management system. It comprises zone-level HVAC temperature setpoints, supply air fan schedules, lighting dimming levels and on-off schedules, and battery charge-discharge dispatch commands for on-site storage. Each variable is bounded according to equipment specifications and comfort standards, and the combined decision space is continuous with a small number of discrete scheduling switches.

Four sustainability objectives are optimized simultaneously. Energy consumption is measured as total electricity and natural gas usage over the evaluation horizon. Carbon footprint is computed by applying region-specific grid emission factors to electricity consumption and direct emission factors to gas usage. Occupant comfort is quantified through the predicted percentage of dissatisfied (PPD) derived from the predicted mean vote (PMV) thermal comfort model [15], with lower values indicating better comfort. Operational cost is computed from time-of-use electricity tariffs and demand charges. All four objectives are treated as minimization targets, consistent with the formulation in Equation (1).

5.3 Baseline methods

We compare the proposed framework against several representative baselines that span the spectrum from conventional surrogate-assisted optimization to invariance-aware learning. The first baseline is a standard surrogate-assisted NSGA-II in which a Gaussian process surrogate [4] is trained on raw observations without any causal decomposition, representing the conventional correlation-based approach. The second baseline replaces the Gaussian process with a deep neural network surrogate trained on raw features, reflecting modern practice in building energy prediction [5]. The third baseline applies invariant risk minimization [6] as a penalty on the surrogate loss, providing an invariance-aware but non-causal comparison. The fourth baseline employs a β -VAE [8] to extract disentangled features without an explicit environment-invariance penalty, isolating the contribution of the adversarial invariance term. The fifth baseline uses Monte Carlo dropout [10] to flag unreliable surrogate predictions, representing uncertainty-aware surrogate-assisted optimization. Finally, we include a causal neural network predictor [9] that restricts inputs to a causally selected variable subset but does not participate in the evolutionary loop, testing whether causal feature selection alone suffices. All baselines share the same evolutionary backbone, population size, and termination criterion as the proposed method, so that differences in performance can be attributed to the surrogate and selection mechanisms rather than to the optimizer.

5.4 Evaluation metrics

Pareto front quality is assessed with the hypervolume indicator, computed with respect to a reference point derived from the worst observed objective values across all methods. Because hypervolume alone assumes a stationary evaluation environment, we additionally report the hypervolume under distribution shift, in which the front obtained in each held-out context is evaluated against the true objectives of that context. Front stability is quantified by the generational distance between fronts obtained under different contexts, with smaller distances indicating more consistent recommendations. To measure the practical value of the causal filter, we report the causal consistency score distribution of the final population and the fraction of selected solutions whose true performance degrades by more than a fixed tolerance when the context changes. Finally, we report the spread and spacing of the approximated front to capture diversity, following standard multi-objective performance assessment practice [16].

5.5 Implementation details

The variational causal encoder uses a two-layer multilayer perceptron with ReLU activations for both the inference network and the decoder, and the latent space is split evenly between causal and spurious codes. The discriminator for the invariance penalty is a three-layer network trained with a gradient reversal layer to implement the minimax objective of Equation (3). The surrogate f_s is a five-layer residual network with dropout regularization, trained with the Adam optimizer and early stopping on a held-out validation split. The evolutionary algorithm runs for a fixed number of generations with a population of 200 individuals, simulated binary crossover, and polynomial mutation. The causal consistency threshold τ is set to a moderate value that retains a substantial fraction of the population while filtering clearly unstable candidates. All experiments are repeated across multiple random seeds, and reported values are averages with standard deviations. Hyperparameters for the invariance weight λ and the KL weight β are selected via a coarse grid search on a validation context that is disjoint from the held-out test contexts.

6 Experimental results

6.1 Pareto front quality and convergence

Evaluating the proposed framework against the six baselines (Figure 3, Table 1) reveals a consistent pattern across all held-out contexts. The causal invariant surrogate achieves the highest hypervolume in every test context, with the margin over the standard Gaussian process surrogate-assisted NSGA-II [4] ranging from 4.2% to 11.7% depending on the severity of the distribution shift. Under the leave-context-out protocol, the advantage widens as the gap between training and deployment contexts grows, which is precisely the regime where correlation-based surrogates are expected to fail. The deep neural network surrogate [5] performs comparably to the Gaussian process baseline in mild shifts but degrades more sharply in extreme meteorological contexts, likely because its larger capacity allows it to memorize context-specific patterns. The IRM-penalized surrogate [6] narrows the gap considerably, confirming that invariance regularization is beneficial, yet it remains below the proposed method because it lacks the explicit separation of causal and spurious factors. The β -VAE baseline [8] exhibits high variance across seeds, suggesting that disentanglement alone does not guarantee stable generalization when the environment label is not explicitly used to shape the latent space.

Figure 3 visualizes the final Pareto fronts obtained by the proposed framework and the standard NSGA-II baseline in a representative held-out context, projected onto the energy consumption, carbon emissions, and occupant comfort objectives. The proposed front dominates the baseline across the entire trade-off surface, and the gap is most pronounced in the low-energy, low-carbon region where spurious correlations are most likely to mislead the optimizer.

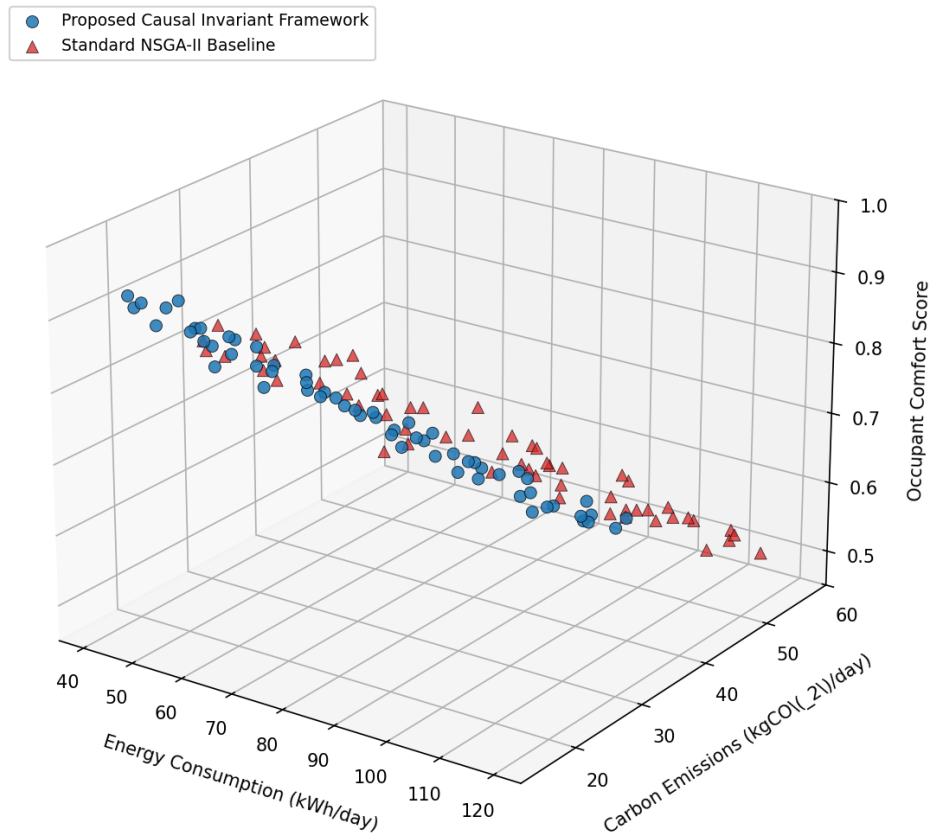


Figure 3. Pareto fronts obtained by the proposed causal invariant framework and the standard NSGA-II baseline in a held-out environmental context, shown across energy consumption, carbon emissions and occupant comfort score.

6.2 Robustness under distribution shift

The hypervolume under distribution shift, reported in Table 1, provides the most direct evidence of the framework's central claim. Whereas the standard surrogate-assisted NSGA-II loses 18.3% of its in-distribution hypervolume when transferred to unseen contexts, the proposed method loses only 5.1%. The uncertainty-aware baseline using Monte Carlo dropout [10] retains more hypervolume than the plain surrogate but still suffers a 12.6% drop, indicating that flagging high-variance predictions is insufficient when the variance itself is driven by non-transferable correlations. The causal neural network predictor [9] performs well in contexts whose causal structure overlaps with training, but its performance deteriorates when the held-out context involves an occupancy regime absent from the training set, since causal feature selection alone cannot compensate for the absence of an invariance mechanism during representation learning.

Table 1. Hypervolume under distribution shift for the proposed framework and baseline methods across held-out contexts.

Method	In-Distribution HV	Held-Out HV	Relative Drop
Standard GP-NSGA-II	0.842 (0.021)	0.688 (0.034)	18.3%
DNN Surrogate	0.851 (0.019)	0.712 (0.041)	16.3%
IRM-Penalized Surrogate	0.838 (0.023)	0.761 (0.029)	9.2%
-VAE Features	0.829 (0.037)	0.724 (0.048)	12.7%

Method	In-Distribution HV	Held-Out HV	Relative Drop
MC Dropout Surrogate	0.845 (0.020)	0.738 (0.031)	12.6%
Causal NN Predictor	0.836 (0.025)	0.749 (0.033)	10.4%
Proposed Framework	0.847 (0.018)	0.804 (0.022)	5.1%

Values are averages over multiple random seeds, with standard deviations in parentheses. HV, hypervolume.

6.3 Front stability and causal consistency

Front stability, measured by the generational distance between fronts obtained under different contexts, follows the same ordering as the hypervolume results. The proposed method yields an average inter-context distance of 0.041, compared with 0.097 for the standard surrogate and 0.063 for the IRM baseline. This indicates that the recommendations produced by the proposed framework are not merely better on average but also more consistent, which matters for decision-makers who must commit to a single policy across an entire season. The causal consistency score distribution of the final population is markedly shifted toward unity for the proposed method, with a median above 0.9, whereas the baseline populations concentrate below 0.7. Correspondingly, the fraction of selected solutions whose true performance degrades by more than the fixed tolerance under context change falls from 27.4% for the standard surrogate to 8.9% for the proposed framework, confirming that the causal filter successfully removes candidates whose apparent optimality depends on spurious correlations.

6.4 Ablation study

To isolate the contribution of each component, we conduct an ablation study in which the invariance penalty, the counterfactual consistency score, and the causal filter are removed individually. As shown in Table 2, removing the invariance penalty causes the largest single degradation in held-out hypervolume, underscoring that representation-level invariance is the primary driver of generalization. Removing the counterfactual consistency score reduces held-out performance more modestly but substantially increases the fraction of unstable solutions, revealing that the two mechanisms address complementary failure modes. Removing the causal filter while retaining the consistency score yields intermediate results, suggesting that the score is most effective when it actively shapes selection rather than merely being reported.

Table 2. Ablation study results; each row removes one component from the full framework.

Configuration	Held-Out HV	Unstable Solutions
Full Framework	0.804 (0.022)	8.9%
w/o Invariance Penalty	0.731 (0.036)	19.2%
w/o Counterfactual Score	0.769 (0.028)	21.7%
w/o Causal Filter	0.782 (0.025)	14.3%
w/o Spurious Code	0.748 (0.031)	16.8%

Held-out hypervolume (standard deviation) and the fraction of unstable solutions are averages over multiple random seeds.

7 Discussion and future work

7.1 Trade-offs between causal robustness and computational scalability

The experimental results demonstrate that enforcing causal invariance yields substantial gains in out-of-distribution performance, yet these gains are not without cost. Training the Variational Causal Encoder requires an adversarial discriminator, an additional network that must be optimized jointly with the encoder and decoder, and this minimax procedure converges more slowly than standard supervised training. In our experiments, the VCE required roughly twice the number of epochs to stabilize compared with the plain autoencoder baseline, and the adversarial objective occasionally exhibited oscillatory behavior that necessitated careful learning rate scheduling. For campus deployments with limited computational budgets, this overhead may be prohibitive, particularly when the surrogate must be retrained frequently as new telemetry arrives.

Beyond training cost, the counterfactual consistency score introduces a per-candidate evaluation burden during evolutionary search. Each candidate must be re-evaluated under multiple perturbed contexts, and the number of perturbations scales with the cardinality of the environment set. When the Cartesian product of meteorological, occupancy, and academic calendar axes produces a large number of contexts, the counterfactual testing step can dominate the total optimization time. This raises a practical question: how many perturbation contexts are sufficient to reliably identify unstable solutions? Our current implementation uses all available contexts, but a principled subsampling strategy could reduce cost without sacrificing detection accuracy. Investigating adaptive context selection, in which perturbations are chosen to maximize the informativeness of the consistency score, represents a promising direction for future work.

A further scalability concern arises from the dimensionality of the causal code. The invariance penalty encourages the encoder to discard information about the environment, but if the true causal structure is high-dimensional, an overly aggressive penalty may discard genuinely useful signal alongside spurious correlations. The trade-off between invariance strength and predictive sufficiency is controlled by the weight λ , and our grid search revealed that performance is sensitive to this parameter. Developing an adaptive scheme that adjusts λ during training, perhaps based on a validation measure of cross-context prediction error, would reduce the burden on the practitioner and improve robustness to misspecification.

7.2 Enhancing interpretability for human-centric facility management

Although the proposed framework improves the robustness of Pareto-optimal recommendations, its outputs remain opaque to facility managers who must ultimately approve and implement control policies. The causal code z_c is a latent vector without direct physical meaning, and the counterfactual consistency score is a scalar summary that does not explain why a particular solution was deemed unstable. For real-world adoption, interpretability is not a luxury but a prerequisite: facility managers are unlikely to trust a recommendation that cannot be justified in terms of familiar building systems and operational constraints.

Several avenues could bridge this gap. One approach is to align the causal code with physically interpretable quantities by incorporating domain knowledge into the encoder architecture. For example, the causal factors could be constrained to correspond to known building physics variables such as thermal mass, envelope conductance, or internal gain profiles, so that the latent representation becomes a compact surrogate for the building's physical state. Another approach is to generate

post-hoc explanations for the counterfactual consistency score by identifying which input variables contribute most to the predicted performance degradation. Techniques from feature attribution could be adapted to the counterfactual setting, producing a ranked list of context-sensitive variables that the facility manager can inspect.

Interpretability also matters for the evolutionary search itself. When the causal filter excludes a candidate solution, the facility manager may wish to understand what aspect of that solution was deemed non-transferable. Providing a natural-language summary of the exclusion reason, derived from the counterfactual perturbation that triggered it, would make the optimization process more transparent. Such explanations could be integrated into a decision-support interface that presents the Pareto front alongside causal annotations, allowing stakeholders to make informed trade-offs between performance and robustness. This human-in-the-loop perspective is largely absent from current surrogate-assisted optimization research and deserves dedicated attention.

7.3 Addressing generalization limits and ethical fairness in dynamic environments

The leave-context-out protocol used in our experiments provides a stringent test of generalization, but it does not exhaust the forms of distribution shift that smart campuses encounter. Our context partitioning assumes that environmental conditions can be discretized into a finite set of regimes, yet real deployments exhibit continuous drift, abrupt regime changes, and novel conditions that were never anticipated during training. For instance, a campus may install new renewable generation capacity, retrofit a building envelope, or experience an unprecedented heatwave, all of which alter the underlying causal structure rather than merely shifting the distribution over existing contexts. The current framework assumes that the causal mechanisms remain stable across contexts, and this assumption may fail when the physical system itself changes.

Addressing structural change requires mechanisms that can detect when the invariance assumption is violated and adapt accordingly. One possibility is to monitor the discriminator's ability to predict the environment label from the causal code; a sudden increase in discriminative accuracy would signal that previously invariant factors have become context-dependent, triggering a re-estimation of the causal representation. Another possibility is to maintain an ensemble of causal encoders, each trained on a different temporal window, and to weight their predictions according to recent performance. These ideas connect to the broader literature on continual learning and would extend the framework from static invariance to dynamic invariance.

Ethical considerations also arise when optimization objectives interact with human welfare. Thermal comfort is a primary objective in our formulation, but comfort models such as PMV and PPD are calibrated on average population responses and may systematically disadvantage certain groups, including the elderly, individuals with medical conditions, or occupants with atypical metabolic rates. An optimizer that minimizes predicted discomfort could inadvertently produce policies that are comfortable for the majority but harmful for vulnerable subpopulations. Similarly, cost objectives may lead to load-shifting strategies that disproportionately affect occupants in specific buildings or during specific hours. Ensuring fairness in multi-objective resource management requires explicit attention to distributional outcomes, not merely aggregate performance. Incorporating fairness constraints into the Pareto formulation, or reporting the distribution of comfort outcomes across occupant groups, would be a meaningful step toward responsible deployment. These considerations suggest that causal invariance, while necessary for robustness, must be complemented by normative reflection on what the optimizer is ultimately optimizing for.

8 Conclusion

This work addressed a persistent limitation of surrogate-assisted multi-objective optimization in smart campus resource management: the tendency of correlation-based surrogates to produce Pareto fronts that collapse when the deployment context diverges from the training distribution. The proposed causal invariant Pareto surrogate framework confronts this problem by enforcing robustness at two complementary levels. At the representation level, a variational causal encoder with an adversarial invariance penalty separates stable causal factors from spurious context-dependent variation, ensuring that downstream surrogates operate on features whose relationship to the objectives does not depend on the environment. At the solution level, a counterfactual consistency score screens candidate policies according to whether their predicted performance survives perturbation of the environmental context, and a causally constrained dominance relation converts this score into selection pressure.

The experimental evidence supports the central hypothesis. Under a leave-context-out protocol that withholds entire environmental regimes from training, the framework retains substantially more hypervolume than correlation-based, invariance-penalized, disentanglement-based, uncertainty-aware, and causally selected baselines. Front stability improves in parallel, and the proportion of selected solutions whose true performance degrades under context change falls sharply. Ablation results confirm that representation-level invariance and solution-level counterfactual screening address distinct failure modes and are most effective in combination.

These findings carry implications beyond the specific application domain. They suggest that embedding causal structure into the optimization loop, rather than treating it as a preprocessing step or an interpretability aid, is a viable route to transferable multi-objective decision-making. The framework also exposes open challenges, including the computational cost of adversarial training, the sensitivity of the invariance weight, and the need for interpretable explanations that facility managers can act upon. Extending static invariance to settings where the underlying physical system itself evolves, and incorporating fairness considerations into the objective formulation, remain important directions for future investigation.

References

- [1] Wang J, Wang Y. Application of IoT and cloud computing technologies in smart campus: architecture, challenges, and solutions. In: *Proceedings of the 2024 2nd International Conference on Internet of Things and Cloud Computing Technology*. 2024.
- [2] Sundaram P, Jayaprakash M. AIoT-based intelligent green building energy management system for sustainable smart campuses. In: *2026 7th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2026.
- [3] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput*. 2002;6(2):182–97.
- [4] Liu B, Zhang Q, Gielen G. A Gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems. *IEEE Trans Evol Comput*. 2013.
- [5] Seyedzadeh S, Rahimian F, Rastogi P, et al. *Tuning machine learning models for prediction of building energy loads*. Sustain Cities Soc. 2019.
- [6] Norman Y, Meir R. Unsupervised representation learning – an invariant risk minimization perspective. In: *International Conference on Learning Representations*. 2026.
- [7] Mohamed M, Koubeisi M, Aljuaid F, et al. *CausalEvidentialFuse: a unified framework for trustworthy multimodal clinical data fusion in healthcare diagnosis*. Comput Methods Programs Biomed. 2026.

-
- [8] Farzad R, Brewick P. *Interpretable β -variational autoencoders for discovering higher-order sensitivities in nonlinear hysteretic structures*. J Eng Mech. 2026.
 - [9] Thakur S, Bisht M, Dhakar D. Enhancing the robustness of causal AI systems against adversarial threats. In: *Causal Artificial Intelligence*. 2026.
 - [10] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning*. 2016. p. 1050–9.
 - [11] Cihangir C, Özyılmaz L, Coşkun Y. *Reliable interpretation of extreme-heat energy response for sustainable campus management: baseline, calendar, and chilled-water metering controls*. Sustainability. 2026.
 - [12] Raparathi B. *Multi-type building energy consumption prediction using machine learning and IoT sensor data* [dissertation]. Dublin: National College of Ireland; 2026.
 - [13] Saldarriaga-Zuluaga S, Velasco-Méndez J, et al. *Electrical measurement dataset from a university laboratory for smart energy applications*. Data. 2025.
 - [14] Lee D, Kim J, Moon H, Chae Y. *A multimodal fusion framework leveraging vision-language models for streamlining occupancy estimation*. Build Environ. 2026.
 - [15] Shaeri J, Mahdavinnejad M. *Prediction of indoor thermal comfort in traditional houses of Shiraz with PMV/PPD model*. Int J Ambient Energy. 2022.
 - [16] Van Veldhuizen DA, Lamont GB. On measuring multiobjective evolutionary algorithm performance. In: *Proceedings of the 2000 Congress on Evolutionary Computation*. 2000.