CaberNet: Causal Representation Learning for Cross-Domain HVAC Energy Prediction

Kaiyuan Zhai
rickzky1001@gmail.com
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Jiacheng Cui
jiachengc648@gmail.com
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Zhehao Zhang
zhang.zheha@northeastern.edu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Junyu Xue
junyuxue@outlook.com
Southern University of Science and Technology
Shenzhen, China

Yang Deng
yang2.deng@connect.polyu.hk
The Hong Kong Polytechnic University
Hong Kong

Kui Wu
wkui@uvic.ca
University of Victoria
Victoria, Canada

Guoming Tang*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
guomingtang@hkust-gz.edu.cn

Abstract

Cross-domain HVAC energy prediction is essential for scalable building energy management, particularly because collecting extensive labeled data for every new building is both costly and impractical. Yet, this task remains highly challenging due to the scarcity and heterogeneity of data across different buildings, climate zones, and seasonal patterns. In particular, buildings situated in distinct climatic regions introduce variability that often leads existing methods to overfit to spurious correlations, rely heavily on expert intervention, or compromise on data diversity. To address these limitations, we propose CaberNet, a causal and interpretable deep sequence model that learns invariant (Markov blanket) representations for robust cross-domain prediction. In a purely data-driven fashion and without requiring any prior knowledge, CaberNet integrates i) a global feature gate trained with a self-supervised Bernoulli regularization to distinguish superior causal features from inferior ones, and ii) a domain-wise training scheme that balances domain contributions, minimizes cross-domain loss variance, and promotes latent factor independence. We evaluate CaberNet on real-world datasets collected from three buildings located in three climatically diverse cities, and it consistently outperforms all baselines, achieving a 22.9% reduction in normalized mean squared error (NMSE) compared to the best benchmark. Our code is available at https://github.com/SusCom-Lab/CaberNet-CRL.

Keywords

Explainable AI, Causal Inference, Markov Blanket, Self-Supervised Learning

1 Introduction

Heating, Ventilation, and Air Conditioning (HVAC) systems are widely used in residential, commercial, and industrial buildings, accounting for a substantial portion of total energy consumption [42]. In office buildings, this figure can reach approximately 40% [3]. Accurate prediction of HVAC energy usage is therefore critical for enhancing automated control systems [5] and advancing sustainable building management practices [32]. Nowadays, with the proliferation of smart meters and IoT sensors, large volumes of building-level data are readily available, offering significant opportunities for data-driven predictive modeling.

Recent approaches have shown promising results in modeling HVAC systems under the independent and identically distributed (i.i.d.) assumption, using methods ranging from traditional machine learning [7, 25] and deep learning [2, 6, 20, 28] to physics-informed methods [4, 24] and transfer learning [11, 33, 39]. However, a fundamental limitation of most of these methods is their reliance on Empirical Risk Minimization (ERM), which learns purely from statistical correlations and often captures domain-specific patterns rather than underlying *causal mechanisms*[1]. For instance, a model may learn a temperature-load correlation that holds during summer but fails to generalize to winter conditions. Such spurious associations deviate from the true causal drivers of energy consumption, leading to poor transferability to new domains with different spatial or temporal contexts. Consequently, achieving robust generalization with these approaches would require training on abundant data from diverse domains [35], while transfer learning also requires a non-trivial amount of target-domain data for fine-tuning [11, 33]. These dependencies significantly limit the feasibility and practicality of deploying existing methods in real-world scenarios.

*corresponding author

[1]The *causal mechanism* refers to a stable process in which causes give rise to effects. Mathematically, this is often represented as $P(\text{effect} \mid \text{cause})$.

This challenge motivates the development of models that can be trained on buildings with available data and applied directly to others, a setting known as *domain generalization (DG)*. In practice, however, this is difficult due to the high diversity inherent in buildings and their operating conditions [12]. Some sources of variation are directly observable and can be measured in datasets, such as local climate, interior layout, and building materials. By contrast, many others are latent and hard to record, like occupant behavior and maintenance quality, and thus often absent from datasets, complicating the learning of robust, domain-invariant representations.

Causal Machine Learning (CML) has recently demonstrated strong potential for tackling out-of-distribution (OOD) challenges in domain generalization, with successful applications in areas such as computer vision [21, 22, 27] and HVAC system modeling [8, 13, 15]. The core premise of CML is that: *robust generalization under distribution shift requires uncovering the latent mechanisms that generate the data, rather than fitting to surface-level correlations [35].* In this view, predictors should depend on invariant causes (X_{S^*}) of the target variable (Y), ensuring that the conditional distribution $P(Y \mid X_{S^*})$ remains stable across environments. Accordingly, the CML pipeline aims to i) identify and extract features that capture stable, mechanism-level structures shared across domains, consistent with the principle of Invariant Risk Minimization (IRM) [1], and ii) attenuate the influence of domain-variant spurious features [23].

Despite this promise, existing CML methods still face practical limitations. A common line of work relies on prior-knowledge-based causal interventions, e.g., by applying the do-operator to image styles or backgrounds that are known to be unrelated to the classification label [21, 22]. Other approaches embed more specialized expert knowledge into the model [8, 13]. Yet such strategies are infeasible in domains where prior knowledge is limited or uncertain, and also undermine generalizability to new problem settings. In the context of HVAC systems, for instance, even basic questions, such as *how* and *to what extent* occupant behaviour affects HVAC energy consumption, are difficult to specify a priori. A second line of work attempts to sidestep this issue through data filtering, selectively training on samples from similar distributions [15, 39]. While this can reduce the domain gap, it also sacrifices distributional diversity. This loss of diversity is counterproductive for domain generalization, as contrasting environments are essential for discovering causal features. Furthermore, such filtering is particularly problematic when data are already scarce, as it further diminishes the limited information available for learning.

In this work, we therefore revisit the challenge of cross-domain HVAC energy prediction through the lens of causal representation learning (CRL). We propose CaberNet (**Ca**usal **Ber**noulli **Net**work), a novel framework that learns causal representations directly from raw data without requiring prior knowledge or discarding valuable samples. Concretely, our approach integrates three key innovations, where we i) introduce a global feature gate that assigns consistent importance weights across domains, thereby providing interpretability of which raw features drive energy consumption, ii) develop a pure data-driven self-supervised Bernoulli regularization that softly partitions features into higher and lower importance groups, without the need for prior knowledge or labels, and iii) design domain-wise training objectives that reweight per-domain difficulty, penalize cross-domain loss variance, and encourage independence of latent factors. In combination with the task loss, CaberNet guides learning toward invariant, mechanism-level relations rather than domain-specific shortcuts or correlations, enabling robust out-of-domain generalization.

Our contributions are summarized as follows:

- *A causal representation learning framework.* We introduce a unified framework that integrates causal principles with deep sequence modeling, enabling the disentanglement of raw features into Markov blanket causal representations for robust cross-domain prediction under distribution shifts.
- *Self-supervised feature selection.* We propose a novel regularization method that combines ℓ_1 sparsity with Bernoulli-entropy minimization to automatically identify stable causal features without supervised labels.
- *Invariant domain-wise training.* We design a training strategy that aggregates per-domain losses with difficulty weights, penalizes cross-domain loss dispersion, and promotes latent factor independence.
- *Empirical validation and explainability.* We demonstrate that CaberNet achieves a 22.9% reduction in NMSE over the state-of-the-art benchmarks on real-world data from six office floors across three cities, while offering high explainability consistent with HVAC domain knowledge.

2 Causal Perspective on Energy Prediction

2.1 Rationale

Causal inference aims to uncover and model the cause-effect relationships among variables, enabling predictions that remain valid under distribution shifts [29]. Unlike purely statistical methods that rely on correlations, causal approaches seek to identify stable relationships that reflect the underlying data-generating process, thereby improving generalization across unseen domains.

This distinction is critical for energy prediction in HVAC systems. Consider the relationship between outdoor temperature and energy consumption: a spike in temperature during summer increases cooling demand, raising energy usage; a drop in temperature during winter increases heating demand, also raising energy usage. While a correlation exists, the effect of temperature on HVAC energy demand is domain-variant. A model that learns only this seasonal correlation may fail to generalize, as it learns a domain-specific shortcut rather than the underlying causal mechanism. A robust model must instead identify the invariant causes whose relationship with energy consumption remains stable.

Guided by this rationale, we base our approach on three key concepts from causal reasoning.

Guideline 1. Causal Invariance Assumption [30, 31]. The foundation of our approach is the principle that robust prediction requires identifying a stable causal core. Invariant Causal Prediction (ICP) formalizes this by seeking a subset of predictors $S^* \subseteq \{1, \ldots, p\}$ whose conditional relationship with the target variable Y remains stable across multiple environments $e \in \mathcal{E}$. This invariant set S^* (or invariant causal predictors) is crucial for ensuring generalizability under distributional shifts.

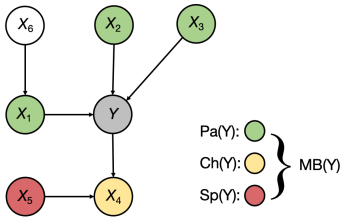


Figure 1: Illustration of Markov blanket (Pa: parent; Ch: child; Sp: spouse).

This is highly relevant to HVAC systems. While different buildings (domains e) exhibit vastly different joint distributions of features X (e.g., due to climate or occupancy differences), the underlying physical and operational mechanisms governing energy consumption are mainly stable. This means the relationship between the true causal drivers X_{S^*} and energy usage Y is invariant:

$$P\left(Y^e \mid X^e_{S^*}\right) = P\left(Y^{e'} \mid X^{e'}_{S^*}\right), \quad \forall e, e' \in \mathcal{E}$$

while in general,

$$P\left(Y^e \mid X^e\right) \neq P\left(Y^{e'} \mid X^{e'}\right), \quad \forall e, e' \in \mathcal{E}$$

as the full set of predictors X may include domain-variant, spurious factors whose effects differ across environments.

Therefore, our goal is to discover S^* rather than simply rely on all available predictors.

Guideline 2. Markov Blanket Prediction [14, 40]. The Markov blanket (MB) of a target variable Y, denoted as $\mathrm{MB}(Y)$, is the minimal set of variables that renders Y conditionally independent of all other variables in the system [29]. As illustrated by the causal graph in Figure 1, $\mathrm{MB}(Y)$ consists of: i) the parents of Y (its direct causes, X_1, X_2, X_3), ii) the children of Y (its direct effects, X_4), and iii) the spouses of Y (other direct causes of Y's children, X_5).

Formally, for any variable O outside the blanket and distinct from Y (X_6 in Figure 1):

$$Y \perp\!\!\!\perp O \mid \mathrm{MB}(Y), \tag{1}$$

where $\perp\!\!\!\perp$ means "is independent of". This guideline means that $\mathrm{MB}(Y)$ contains all necessary information to predict Y and blocks all influence from extraneous variables.

While using only the parents of Y captures the true causal mechanism $P(Y \mid \mathrm{Pa}(Y))$, it may discard valuable predictive signals from its children and their other causes (spouses). For instance, while carrying umbrellas does not cause rain, observing umbrellas can improve the prediction of rain. Therefore, conditioning on the full $\mathrm{MB}(Y)$ could reduce the conditional variance of Y:

$$\mathrm{Var}(Y \mid \mathrm{MB}(Y)) \leq \mathrm{Var}(Y \mid \mathrm{Pa}(Y)), \tag{2}$$

with equality holding only if the children and spouses provide no additional information about Y (see proof in Appendix A).

Conditioning on all variables can be effective in i.i.d. settings. However, in the presence of domain shifts, a critical distinction emerges: while the $\mathrm{MB}(Y)$ is the *minimal sufficient set* for prediction, the full set of variables X may contain spurious correlations. Thus, in an i.i.d. setting, $\mathrm{Var}(Y \mid X) \leq \mathrm{Var}(Y \mid \mathrm{MB}(Y))$ may hold due to these additional correlations. Yet, under a distribution shift, these correlations can become unstable and lead to severe performance degradation. Therefore, the $\mathrm{MB}(Y)$ represents the optimal trade-off, as it harnesses more predictive signal than the parent set alone while avoiding the inclusion of spurious, domain-specific noise that plagues the full feature set. This makes it an ideal basis for building robust, cross-domain predictors.

Guideline 3. Independence of Causal Representation Assumption [22]. To learn a disentangled and compact latent representation, we adopt the principle that the underlying causal mechanisms are independent. We posit that the learned latent factors $Z = [Z_1, \ldots, Z_d]$ are independent in the marginal sense:

$$p(Z) = \prod_{i=1}^{d} p(Z_i) \tag{3}$$

and that this representation is causally sufficient for prediction [29], meaning it captures all information from X relevant to Y, i.e., $Y \perp\!\!\!\perp X \mid Z$, This assumption encourages the elimination of redundant latent dimensions and supports compact models. Our ablation studies (Section 6.2.2) confirm that stable performance is achievable even with a small hidden dimension d, validating the efficacy of this approach.

Remark. **Tension with Markov blanket**. Markov blankets include the co-parents/spouses of Y's children; such variables are not, in general, mutually independent (e.g., X_5 and X_4 are not independent, as shown in Figure 1). Therefore, enforcing strong factor independence in Z can conflict with a fully learned Markov blanket representation. In practice, we balance this trade-off by treating the independence principle as a regularizing prior rather than a strict requirement. The strength of the independence regularizer $\mathcal{L}_{indy}$ (see Eq. 11) is tuned relative to other terms. This allows the model to retain essential predictive information from the Markov blanket's dependent components while still discouraging unnecessary entanglement and promoting a compact latent space.

2.2 Problem Formulation

We consider the problem of cross-domain HVAC energy consumption prediction. For each domain $e \in \mathcal{E}$ (e.g., a distinct building or one of its floors), the data consist of multivariate time series:

$$\mathcal{D}^e = \{(X^e_{t-w:t-1}, Y^e_t)\}_{t=w}^{n_e}, \quad X^e_t \in \mathbb{R}^p, \; Y^e_t \in \mathbb{R}$$

where w is the size of the time window, X^e_t denotes the observed features (e.g., weather, indoor conditions, etc.) at time t, and Y^e_t is the corresponding HVAC energy consumption.

The fundamental challenge is distribution shift: the marginal distribution of features varies significantly across domains due to differences in climate, occupancy patterns, and building properties, i.e., $P^e(X) \neq P^{e'}(X)$ for $e \neq e'$. Consequently, a predictor trained to minimize empirical risk on a source domain often fails to generalize, as it may leverage these spurious, domain-specific correlations.

Our core assumption, grounded in causal reasoning, is that while the inputs X may change, the underlying *causal mechanism* relating the true drivers X_{S^*} to the target Y remains invariant:

$$P^e\left(Y \mid X^e_{S^*}\right) = P^{e'}\left(Y \mid X^{e'}_{S^*}\right), \quad \forall e, e' \in \mathcal{E}$$

where $X_{S^*} \subseteq X$ denotes the causality-related features that govern Y across environments.

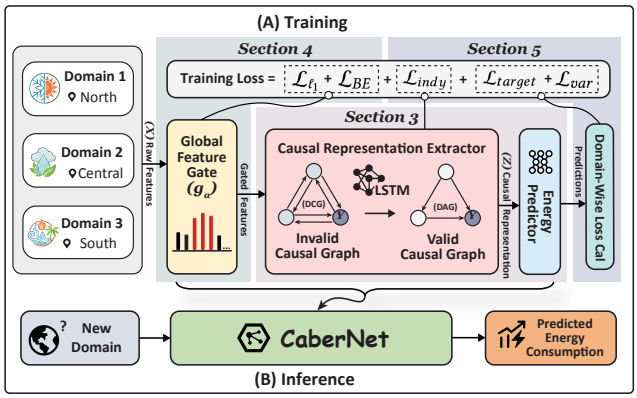


Figure 2: CaberNet framework. The entire process is end-to-end, and the training loss is computed as a weighted sum of all individual losses. Task: training on the source domain and generalizing to the target domain.

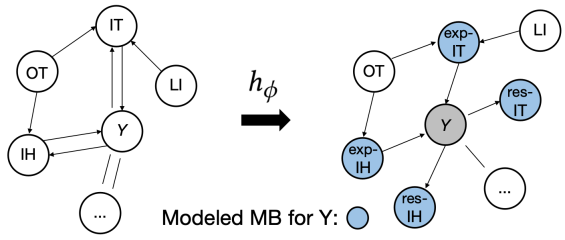


Figure 3: Illustration of reconstructed Markov blanket for Y (IT: indoor temperature, OT: outdoor temperature, IH: indoor humidity, LI: light intensity, exp: explanatory, res: response, MB: Markov blanket). Note that the modelled MB is our target causal representation.

Goal. Our objective is to learn a model comprising: i) a *causal representation extractor* $Z = h_\phi(X)$ (Section 3.1), and ii) an *energy predictor* $\hat{Y} = m_\theta(Z)$ (Section 3.2) trained over source domains $\mathcal{E}_{\text{train}}$, validated over $\mathcal{E}_{\text{val}}$, with the goal of generalizing to unseen target domains $\mathcal{E}_{\text{test}}$ without retraining. Formally, the model is trained on source domains $\mathcal{E}_{\text{train}}$ to minimize the validation loss, and subsequently evaluated on the unseen test domains $\mathcal{E}_{\text{test}}$, i.e.,

$$\min_{\theta,\phi} \mathbb{E}_{e\in\mathcal{E}_{\text{test}}}\left[\ell\left(m_\theta(h_\phi(X^e)), Y^e\right)\right]$$

where $\ell(\cdot,\cdot)$ is a suitable loss, e.g., normalized mean squared error.

The key difficulty in addressing the problem lies in ensuring the composed function $m_\theta \circ h_\phi$ captures the invariant causal relation rather than domain-specific associations. To achieve this, we translate the aforementioned three causal guidelines from Section 2.1 into concrete model design choices and regularization strategies, forcing the learner to ignore spurious correlations and focus on the true causal drivers of energy consumption.

3 CaberNet Framework

The overall structure of CaberNet is shown in Figure 2. In this section, we provide an overview of this framework by introducing its main modules and their roles.

3.1 Causal Representation Extractor

Structural causal models (SCMs) are typically represented over observable variables within a directed acyclic graph (DAG) [29]. However, directly applying this framework to HVAC systems is problematic due to the presence of cyclic relationships among observables. For example, indoor temperature and AC energy consumption influence each other in a feedback loop: a higher indoor temperature causes increased AC energy use, which in turn lowers the indoor temperature.

We therefore adopt a decomposition-reconstruction view. Our causal representation extractor h_ϕ maps raw, cyclical time-series data into a latent representation Z where the relationships are disentangled and can be represented by a valid SCM (Figure 2). For instance, the single observable "indoor temperature" can be decomposed into distinct latent factors: an *explanatory* factor that influences energy usage (exp-IT $\rightarrow Y$ and a *response* factor that is affected by it ($Y \rightarrow$ res-IT), as illustrated in Figure 3. Such decomposition breaks the cycle and enables causal interpretation.

Our approach diverges from traditional Markov blanket discovery methods [16, 18], which perform a combinatorial search-and-prune procedure over raw features. Instead, we implicitly bias the learned representation Z towards the Markov blanket of Y by leveraging its predictive advantage (see Guideline 2 and Eq. 2). The reconstructed latent causal graph in Section 6.3.3 further confirms that the learned representation indeed aligns with a Markov blanket structure. In essence, while prior work selects a subset of raw features, our method transforms all features into a latent space where the representation itself aligns with the Markov blanket.

To handle the temporal dependence, we implement h_ϕ using a Long Short-Term Memory (LSTM) network [10]:

$$Z = h_\phi(X_{t-w:t-1}),$$

where Z is the latent causal representation, and is then passed to the energy predictor m_θ. Note that our framework is *model-agnostic*, and the LSTM here serves as a capable backbone but could be replaced with other sequence models; we adopt LSTM primarily due to the moderate dataset size, which may not support training heavier architectures like Transformers.

3.2 Energy Predictor

The absolute scale of HVAC energy consumption can vary significantly across domains due to intrinsic factors such as building floor area, HVAC system capacity, and peak occupancy levels. A model that fails to account for these baseline differences may struggle with systematic biases in prediction.

To address this, we explicitly model these domain-specific properties. For each domain, we compute a vector s of descriptive statistics from the training portion of the energy target Y:

$$s = [\,\mu,\ \sigma,\ q_1,\ q_3\,]^\top \in \mathbb{R}^4,$$

where μ is the mean, σ the standard deviation, and q_1, q_3 the first and third quartiles, respectively. This vector s comprehensively captures the central tendency, spread, and distribution shape of energy consumption for a domain.

This vector is then processed by a small Multi-Layer Perceptron (MLP), which acts as a scale encoder, to generate a domain-specific scaling factor γ and an intercept β, i.e.,

$$(\gamma,\ \beta)\ =\ \mathrm{MLP}(s).$$

With these parameters, we then apply an affine adjustment to the causal representation Z, aligning its scale and offset with the target domain before the final prediction:

$$\tilde{Z}\ =\ \gamma \cdot Z\ +\ \beta, \qquad \hat{Y}\ =\ \mathrm{Linear}(\tilde{Z})$$

The above design yields two benefits:

- *Disentanglement.* It decouples domain-invariant causal mechanisms (learned by h_ϕ) from domain-specific base energy characteristics, avoiding distortion.
- *Generalization.* By normalizing the input to the final predictor across domains, the energy predictor mitigates scale differences, allowing the causal representation to focus on feature-level rather than magnitude-level modeling.

4 Self-Supervised Feature Selection

A primary challenge in domain generalization is identifying which features encode stable, causal mechanisms versus those that form spurious, domain-specific correlations. To address this without reliance on prior knowledge, we cast feature selection as a self-supervised binary classification problem. The effectiveness of this component is validated by the ablation study contrasting CaberNet vs. CaberNet-SIRM in Section 6.2.1.

4.1 Global Feature Gate

We introduce a global, feature-wise gate that assigns each of the input features a latent Bernoulli activation variable S_i, where the probability $f_i = \sigma(\alpha_i)$ is a learnable parameter. This gate is designed to be sample-agnostic, providing consistent feature importance scores across domains. The gate's distribution is shaped by two complementary regularizers: i) an ℓ_1 *sparsity* term for encouraging overall sparsity, and ii) a *Bernoulli entropy* term that polarizes importance scores, effectively performing a soft, differentiable feature selection. Coupled with the prediction task loss, this framework automatically partitions features into lower-importance ("inferior causal") and higher-importance ("superior causal") groups in a purely data-driven manner, as illustrated in Figure 4.

The global feature gate aims to classify the raw features into *inferior causal features* (inf-causal) and *superior causal features* (sup-causal), with both groups contributing to the representation but the latter exerting stronger influence.

To achieve this, we introduce a global gate parameter vector $\alpha \in \mathbb{R}^p$, with p indicating the input dimension. A single, sample-agnostic weight vector g is derived via a softmax activation:

$$g\ =\ \mathrm{softmax}(\alpha) \in (0,1)^p, \quad \sum_{i=1}^{p} g_i = 1 \tag{4}$$

Crucially, this gate vector g is global, i.e., it is independent of the input sample X and remains fixed across all instances and domains. This ensures that the feature importance is consistent, reflecting the intrinsic property of the feature itself rather than its specific value in a given context. To prevent the gate from being influenced by the arbitrary scale of different features, we standardize each feature to a standard normal distribution. This ensures that the learned weights g reflect genuine predictive importance rather than being correlated with feature magnitudes.

The following sections detail the design of the two complementary regularizers: the ℓ_1 sparsity (for overall sparsity) and the Bernoulli entropy (for feature polarization).

4.2 ℓ_1 Sparsity

To encourage a compact representation and suppress non-causal features, we apply an ℓ_1 penalty on the feature importance scores. Specifically, we map the gate parameter α through a sigmoid function to obtain a vector of importance $f = \sigma(\alpha) \in (0,1)^p$. The ℓ_1 sparsity term is defined as the sum:

$$\mathcal{L}_{\ell_1} = \|f\|_1 = \sum_{i=1}^{p} \sigma(\alpha_i). \tag{5}$$

We adopt the sigmoid function here because the softmax output g (from Eq. 4) has a constant ℓ_1 norm of 1 by construction, which would invalidate an ℓ_1 penalty on g. The $\mathcal{L}_{\ell_1}$ penalty suppresses the weights of uninformative features, driving them toward zero. This process acts as an adaptive "trial" mechanism, which automatically identifies and down-weights features that are likely to be inf-causal.

4.3 Bernoulli Entropy

The ℓ_1 penalty encourages sparsity but does not necessarily force features to be fully ON or OFF, leaving feature importance probabilities f_i insufficiently distinct. To resolve this ambiguity and push the model toward a clear feature selection, we introduce Bernoulli entropy as the second regularizer of the importance distribution.

We define the Bernoulli probability for each feature as $f = \sigma(\alpha) \in (0,1)^p$, where $f_i = \Pr(S_i = 1)$ for a hypothetical Bernoulli variable S_i. Note that this is a conceptual device used only to define the entropy, and no actual sampling is performed during training, ensuring the process remains differentiable.

Since there are no ground-truth labels indicating whether a feature is inf-causal or sup-causal, the partitioning (classification) must be learned in a *self-supervised* manner. To this end, we minimize the sum of Bernoulli entropies (BE):

$$\mathcal{L}_{BE}\ =\ \sum_{i=1}^{p} H(\mathrm{Bern}(f_i))\ =\ -\sum_{i=1}^{p} \Big[f_i \log f_i + (1-f_i)\log(1-f_i) \Big], \tag{6}$$

in which the loss reaches its maximum when $f_i = 0.5$ and attains its minimum as f_i approaches 0 or 1.

Theoretical justification. The polarizing effect of $\mathcal{L}_{BE}$ can be understood by analyzing its properties. For a single dimension, define the scalar entropy function $r(f) = H(\mathrm{Bern}(f))$ for $f \in (0,1)$. Its first and second derivatives are:

$$r'(f) = \log \frac{1-f}{f}, \qquad r''(f) = -\frac{1}{f(1-f)} < 0,$$

The negative second derivative proves $r(f)$ is strictly concave, with a unique maximum at $f = \frac{1}{2}$ and minima at the boundaries $f \to 0$ and $f \to 1$.

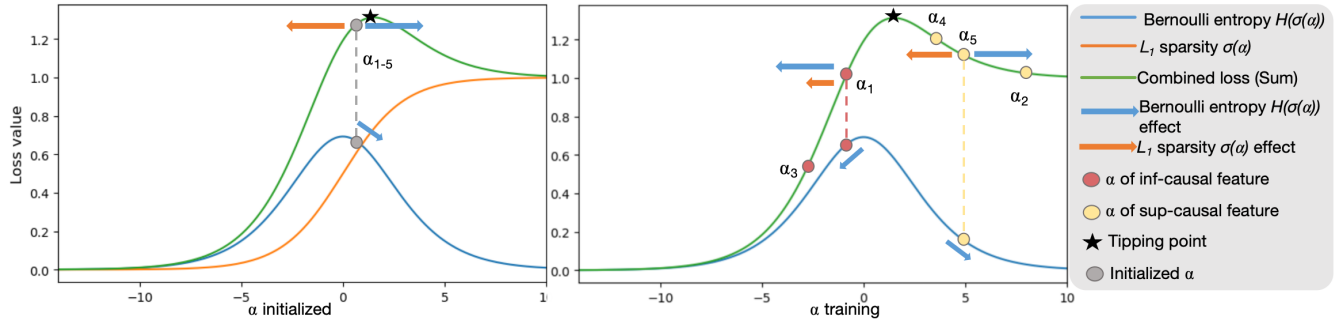


Figure 4: Conceptual diagram of the self-supervised Bernoulli regularization process. The figure illustrates the effect of the regularization on the evolution of α values for each feature immediately after initialization (left), contrasted with the effect on α values of features that have been partitioned after training (right). inf-causal: inferior causal, sup-causal: superior causal.

In the multivariate case, the total loss $\mathcal{L}_{BE}(f) = \sum i = 1^p r(f_i)$ is separable. Its Hessian is a diagonal matrix with entries $r''(f_i) < 0$, confirming that $\mathcal{L}_{BE}$ is strictly concave on $(0,1)^p$. The loss is uniquely maximized when all $f_i = 0.5$ and minimized only at the vertices of the hypercube, where $f \in 0, 1^p$.

Therefore, minimizing $\mathcal{L}_{BE}$ drives each f_i away from ambiguity and toward a deterministic 0 or 1, effectively performing an *unsupervised binary partition* of the features into {inf-causal, sup-causal}.

4.4 Combined Regularization Dynamics

The full regularization effect emerges from the combination of the ℓ_1 sparsity and Bernoulli entropy terms. Together, they create a biased, bi-stable dynamic that pushes features toward either 0 or 1, with a tipping point determined by the ratio of the regularization strengths. The task loss then determines which features cross this threshold, resulting in a self-supervised partition into inferior and superior causal groups.

The combined regularizer is defined as:

$$\mathcal{L}_{\ell_1 BE}(f) = \lambda_{BE} \underbrace{\sum_{i=1}^{p} H(\text{Bern}(f_i))}_{\mathcal{L}_{BE}} + \lambda_{\ell_1} \underbrace{\sum_{i=1}^{p} f_i}_{\mathcal{L}_{\ell_1}}, \tag{7}$$

where $\mathcal{L}_{BE}$ encourages polarization ($f_i \to \{0,1\}$) and $\mathcal{L}_{\ell_1}$ penalizes overall activation mass.

Initialization and symmetry breaking. To break the symmetry inherent in random initialization (where $f_i \approx \frac{1}{2}$), we initialize all gate parameters α with a small positive bias (e.g., $\alpha = 0.01$). This results in initial probabilities $f_i = \sigma(0.01) \approx 0.5025$, providing a uniform but deterministic starting point slightly above 0.5, which avoids random initial flips and ensures stable optimization.

Gradient dynamics and the tipping point. By the chain rule, the gradient of the combined regularizer with respect to the gate parameters is:

$$\frac{\partial \mathcal{L}_{\ell_1 BE}}{\partial \alpha_i} = \frac{\partial \mathcal{L}_{\ell_1 BE}}{\partial f_i} \cdot \frac{\partial f_i}{\partial \alpha_i} = \underbrace{\left(\lambda_{BE} \log \frac{1-f_i}{f_i} + \lambda_{\ell_1}\right)}_{B(f_i)} \underbrace{f_i(1-f_i)}_{>0}.$$

Hence, the sign of the gradient is governed entirely by the bracket $B(f_i)$. Setting $B(f_i) = 0$ yields a *tipping point*:

$$f_i^\star = \frac{1}{1+\exp(-\lambda_{\ell_1}/\lambda_{BE})} = \sigma\left(\frac{\lambda_{\ell_1}}{\lambda_{BE}}\right) > \tfrac{1}{2}. \tag{8}$$

Because $f_i(1-f_i) > 0$, this tipping point defines the gradient-descent dynamics:

$$\begin{cases} f_i < f_i^\star & \Rightarrow \quad \frac{\partial \mathcal{L}_{\ell_1 BE}}{\partial \alpha_i} > 0 \Rightarrow \alpha_i \downarrow \Rightarrow f_i \downarrow \to 0, \\ f_i > f_i^\star & \Rightarrow \quad \frac{\partial \mathcal{L}_{\ell_1 BE}}{\partial \alpha_i} < 0 \Rightarrow \alpha_i \uparrow \Rightarrow f_i \uparrow \to 1. \end{cases}$$

Thus, the $\mathcal{L}_{\ell_1}$ term shifts the unstable equilibrium from 0.5 (pure Bernoulli entropy) to $f_i^\star > 0.5$, creating a bias where features must "prove their worth" to be activated.

Interaction of the regularizers. The two regularizers interact synergistically: i) The Bernoulli entropy term ($\mathcal{L}_{BE}$) acts as a polarizing force, pushing f_i away from 0.5 toward the extremes; ii) the ℓ_1 sparsity term ($\mathcal{L}_{\ell_1}$) acts as a global suppressing force, providing a constant positive bias in the gradient that raises the activation threshold to $f_i^\star$. This makes it harder for a feature to remain active unless it is strongly beneficial for the task.

Self-supervised selection. The combined effect, in concert with the task loss, facilitates a self-supervised selection process:

- Features that provide strong predictive signals receive gradients from the task loss that help them overcome the raised threshold $f_i^\star$. Once above, the entropy term pushes them further toward 1, cementing them as sup-causal features.
- Features that provide weak signals are unable to overcome the ℓ_1 suppression. They fall below $f_i^\star$ and are pushed by both regularizers toward 0, becoming inf-causal features.

In this way, the model performs an unsupervised binary feature partition guided by the target prediction objective.

5 Domain-Wise Training

To learn representations that generalize across domains, we train across multiple domains by aggregating per-domain losses with difficulty-aware weights and by adding regularizers that align performance and disentangle causal representation. This ensures the

model learns succinct invariant mechanisms rather than domain-specific shortcuts. Its contribution is further substantiated through the ablation study comparing LSTM and Cabernet-SIRM in Section 6.2.1.

5.1 Difficulty-aware Aggregation

A simple average of per-domain losses can be biased towards larger datasets or domains with inherently higher-variance signals, and thus cause the model to overfit to these dominant distributions. To overcome this, we minimize a difficulty-weighted average of per-domain losses, which aims to balance the contribution of each domain regardless of its size or inherent complexity.

Specifically, for each domain e in a mini-batch (with samples from multiple domains), we compute its domain-specific NMSE:

$$L_e \;=\; \mathrm{NMSE}\big(m_\theta(h_\phi(X^e)),\, Y^e\big).$$

We then aggregate these losses into a global target using difficulty-based weights:

$$\mathcal{L}_{target} \;=\; \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} w_e\, L_e, \tag{9}$$

where the weights w_e are designed to counterbalance domain difficulty. We estimate this difficulty score d_e by combining the coefficient of variation (CV) of the target and the mean CV of the inputs:

$$\mathrm{CV}_Y^{(e)} = \frac{\mathrm{Std}(Y^e)}{\mathbb{E}[|Y^e|]}, \mathrm{CV}_X^{(e)} = \frac{1}{p}\sum_{j=1}^{p}\frac{\mathrm{Std}(X^e_{\cdot j})}{\mathbb{E}[|X^e_{\cdot j}|]}, d_e = \tfrac{1}{2}\big(\mathrm{CV}_Y^{(e)} + \mathrm{CV}_X^{(e)}\big).$$

The weight for each domain is then set as the inverse of its difficulty score: $w_e = 1/d_e$. This scheme prevents any single complex or large domain from dominating the objective.

5.2 Cross-domain Variance Regularization

Motivated by the Causal Invariance Assumption (Guideline 1, Section 2.1), we penalize the dispersion of difficulty-weighted domain losses within a batch, nudging the representation toward causal invariance rather than domain-specific fits.

We encourage h_ϕ to capture relationships that are consistently predictive across all training domains, rather than specialized, domain-specific shortcuts. Concretely, we penalize the dispersion of the difficulty-weighted per-domain losses within each batch:

$$\mathcal{L}_{var} \;=\; \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\big(w_e L_e - \bar{L}\big)^2. \tag{10}$$

Driving $\mathcal{L}_{var}$ down nudges the representation toward a mechanism that performs consistently across domains, aligning with the *causal invariance* assumption that the conditional $P(Y \mid X_{S^*})$ should be stable over environments. This encourages the model to capture causal representations rather than domain-specific correlations.

5.3 Independence of Causal Representation

In accordance with the Independence of Causal Representation Assumption (Guideline 3, Section 2.1), we encourage the learned latent representation Z to be composed of statistically independent factors. This promotes disentanglement, improves interpretability, and prevents the model from learning redundant, entangled representations.

Let $Z = h_\phi(X) \in \mathbb{R}^d$ be the latent representation for a batch of data. We compute the sample correlation matrix C of the latent dimensions across the batch:

$$C_{ij} \;=\; \frac{\langle Z_i,\, Z_j\rangle}{\|Z_i\|\,\|Z_j\|}, \quad \text{for } i,j = 1,2,\ldots,d,$$

where Z_i denotes the i-th dimension of Z.

We then promote independence by shrinking the off-diagonals of C, penalizing the deviation from the identity:

$$\mathcal{L}_{indy} \;=\; \frac{1}{d(d-1)}\|C - I_d\|_1 \;=\; \frac{1}{d(d-1)}\sum_{i\neq j}|C_{ij}|, \tag{11}$$

where $d(d-1)$ serves as a normalization factor.

This operationalizes the Independence of Causal Representation assumption and discourages redundant entanglement among latent coordinates.

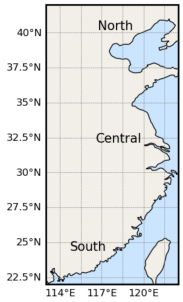
6 Experiments

Datasets. Publicly available datasets that combine indoor sensing with HVAC energy consumption across multiple climate zones remain scarce. Therefore, our experiments are conducted on data we collected via multi-site sensor deployments. The data comprises six office-floor datasets from three commercial buildings located in three Chinese cities, spanning a north–Central–south climate gradient (Figure 5). Indoor sensing is deployed uniformly across the occupied areas. Per timestamp, the feature set includes: indoor temperature, indoor humidity, light intensity, CO_2 concentration, indoor air pressure, and total volatile organic compounds (TVOC). Indoor sensors log at a 5-minute resolution. We also ingest outdoor temperature collected from NASA POWER [26] and resample it to 5-minute resolution via linear interpolation to match the indoor data cadence. The data used in the experiments cover the period from November 2024 to July 2025.

Data processing and feature engineering. A complete pipeline is provided in our code. Because different floors may host varying numbers of devices (Figure 5), we aggregate features using fixed summary statistics, ensuring consistent input dimensionality per dataset, regardless of the device count. To handle missing data, we apply a standardized pipeline (details in our code) for alignment and short-gap filling, while discarding segments that fail integrity checks. We add a binary feature `is_work` that marks work hours/-days (including make-up days) versus off-hours/holidays. Besides, for each domain, we compute a summary of the energy scale (see Section 3.2). The z-score normalization is applied to features. Finally, we window the multivariate time series into length-12 input sequences (the most recent 60 minutes) and predict the next 5-minute AC energy consumption.

Evaluation protocol. To remove the effect of scale differences across domains, we report the normalized mean squared error (NMSE). For a test domain e with ground-truth $\{y_t^e\}_{t=1}^{n_e}$ and predictions $\{\hat{y}_t^e\}_{t=1}^{n_e}$; we define

$$\mathrm{NMSE}^e \;=\; \frac{\sum_{t=1}^{n_e}\big(y_t^e - \hat{y}_t^e\big)^2}{\sum_{t=1}^{n_e}\big(y_t^e\big)^2}.$$



Area	avg. air temp	climate	floor	sensors
North	14.4°C	Temperate monsoon	L–A	49
			L–B	55
Central	16.9°C	Subtropical monsoon	L–A	32
			L–B	37
South	23.1°C	Subtropical monsoon	L–A	27
			L–B	49

Figure 5: Dataset summary and collection sites.

We adopt a leave-one-domain-out protocol, which is standard in domain generalization [17]: Each time, one domain is held out as a test, and the remaining domains form the training pool. Within every training domain, we split 20% of data as validation and use the remaining 80% for fitting; the model checkpoint achieving the best validation NMSE is then evaluated on the held-out test domain.

Implementation details. We train all models with Adam optimizer (learning rate 2×10^{-4}) for 500 epochs. Unlike many of the domain generalization literatures [21, 22], we avoid any data augmentation or causal intervention procedures, as they may inject human-induced biases. Our approach aims to remain as data-driven as possible, relying on the training distributions rather than external knowledge. The hidden size is $d = 64$ and the batch size is 512. All experiments are conducted on a NVIDIA RTX A6000 GPU.

6.1 Performance Comparison

We compare against four representative methods. LSTM [10] serves as a classical deep sequence model and backbone of the CaberNet. LSTM–LiNGAM combines causal discovery with sequence prediction: we first estimate a causal graph from training data using direct-LiNGAM [37], extract the Markov blanket of Y from the discovered graph, and then train an LSTM using only the blanket variables for prediction. We also include two recent benchmarks tailored to HVAC energy prediction: Shift-GRU [19], a GRU regressor trained on inputs pre-aligned by per-feature time lags estimated via Pearson correlation; and STL [39], a transfer-learning LSTM that trains on samples most similar to the target domain. CaberNet (Ours) integrates causal representation learning with sequence modeling. All models are trained using the same protocol (see Section 6). The result is shown in Table 1. We mainly compare with the LSTM-based approach since it also serves as the backbone of CaberNet, ensuring a fair and controlled evaluation.

CaberNet achieves the best NMSE on all six domains (Table 1); averaged over domains it improves by **22.9%** relative to the best benchmark. Per-domain gains over other benchmarks are consistent and substantial, indicating that causal-invariance regularization and soft feature selection improve generalization across buildings and climates. Moreover, CaberNet is more interpretable (see Section 6.3), as its causal representation is both transparent in composition and grounded in causal reasoning. This transparency supports industrial deployment and operator trust, which is especially important in human health–related HVAC applications.

Table 1: Cross Domain Evaluation (bold: best, underline: second, *italic*: equal)

Method	NMSE						
	North		Central		South		Average
	L-A	L-B	L-A	L-B	L-A	L-B	
LSTM	0.114	0.130	0.165	*0.286*	0.368	0.354	0.236
LSTM-LiNGAM	0.064	0.198	0.206	*0.286*	0.381	0.318	0.242
Shift-GRU	0.139	0.190	0.189	0.291	0.289	0.287	0.231
STL	0.114	0.161	0.213	0.363	0.356	0.334	0.257
CaberNet	**0.063**	**0.119**	**0.131**	**0.254**	**0.261**	**0.242**	**0.178**

Shift-GRU achieves the second average NMSE and is especially strong in the South, where it beats plain LSTM, likely because per-feature time-lag alignment mitigates phase shifts from thermal inertia and lag offsets in warm-humid climates. However, Pearson correlation is linear and pairwise, so its shifts can be unreliable under nonlinear or multi-lag dynamics. This limits robustness relative to CaberNet. By contrast, STL performs worst overall: its similarity-based filtering shrinks data diversity and sample size, which is ill-suited when data are scarce and highly variable across domains.

LSTM–LiNGAM is overall comparable to LSTM, and in some cases markedly better (North L-A and South L-B), probably because the discovered blanket retained strong drivers (e.g., *outdoor_temperature*, *is_work*) while discarding weak ones, as shown in Appendix B. However, its performance varies: hard masking can remove useful proxy variables; LiNGAM also assumes a linear, acyclic data-generating process with non-Gaussian errors and no latent confounding [37], which may be too restrictive for the complex and nonlinear nature of HVAC dynamics. Moreover, LiNGAM's binary include/exclude decision is brittle: features that are not strictly causal but still carry a latent causal signal (e.g., proxies, mediators, regime indicators) are dropped entirely. By contrast, CaberNet uses soft gating to retain such partial contributions while appropriately down-weighting them. In this way, the model can still exploit these signals in forming the final causal representation, thereby improving robustness under domain shift. These misspecifications can yield suboptimal blankets and underperformance on other floors. Notably, in Central L-B LiNGAM selected essentially the full feature set as the blanket, making its input identical to LSTM and resulting in identical NMSE. The detailed SCMs constructed by LiNGAM are shown in Appendix B.

6.2 Ablation Study

6.2.1 Regularization. We ablate the Bernoulli regularization $\mathcal{L}_{\ell_1 BE}$ by removing the global feature gate and also dropping the cross-domain variance and independence penalties. The remaining objective aggregates losses *per domain* (equal domain contribution) with the difficulty coefficient on the target loss only. This ablation is called CaberNet-SIRM (simplified invariant risk minimization), which balances risk across environments, but without an explicit invariance regularization; it stands between plain ERM (the LSTM baseline in Section 6.1) and our full CaberNet model.

LSTM (ERM) vs. CaberNet-SIRM (IRM). Compared to the LSTM baseline (Table 1), ERM can occasionally win on a specific

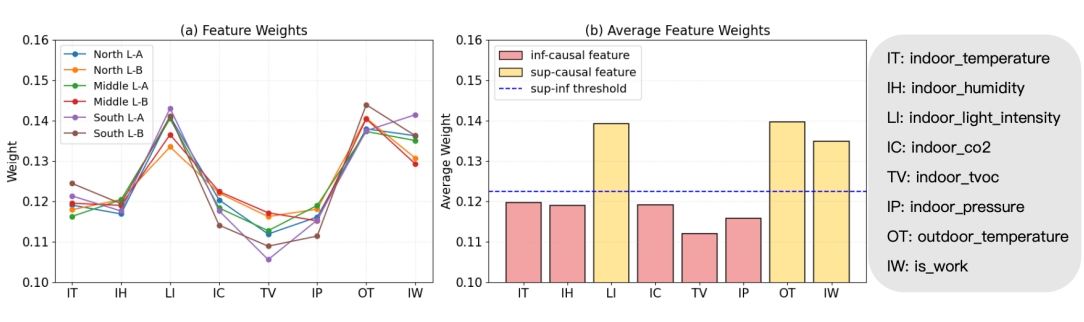


Figure 6: Feature weights (left: test on labeled dataset, right: average over all datasets).

Table 2: Ablation Results on Regularization

Method	NMSE						
	North		Central		South		Average
	L-A	L-B	L-A	L-B	L-A	L-B	
CaberNet-SIRM	0.075	0.162	0.134	0.294	0.301	0.268	0.206
CaberNet	**0.063**	**0.119**	**0.131**	**0.254**	**0.261**	**0.242**	**0.178**

Table 3: Ablation Results on Independence of Z

Method	Hidden Size		
	8	16	32
CaberNet-NoIndy	0.237	0.232	0.207
CaberNet	**0.214**	**0.210**	**0.192**

test domain (e.g., *North L-B*). This happens because ERM's sample-proportional training implicitly upweights large/source-like domains; when such a domain is "lucky" and happens to resemble the held-out target distribution, ERM benefits from this alignment. By contrast, CaberNet-SIRM gives each training domain equal influence, foregoing that accidental advantage. This "ERM>IRM" phenomenon is common, especially when ERM is well tuned [9, 34]; nonetheless, *on average* CaberNet-SIRM still outperforms ERM by **12.7%**, indicating that equalizing domain contributions reduces overfitting to domain-specific variations of large domains and improves robustness across diverse targets.

CaberNet vs. CaberNet-SIRM. We see CaberNet has consistent additional gains, yielding a **13.6%** relative improvement. The full model's advantages come from (i) *Bernoulli regularization*: the combined $\mathcal{L}_{\ell_1 BE}$ objective polarizes gate probabilities toward $\{0, 1\}$ and raises the neutrality threshold $f^\star$ (Eq. 8), suppressing weak spurious cues while preserving genuinely predictive signals, thereby yielding stable, parsimonious inputs to m_θ; (ii) the cross-domain variance penalty $\mathcal{L}_{var}$, which steers optimization toward invariant mechanisms rather than domain-specific correlations. Together, these elements deliver the best cross-domain generalization.

6.2.2 Independence of causal representation. We ablate the independence loss $\mathcal{L}_{indy}$ (Eq.11), thereby not enforcing the *Independence of Causal Representation* hypothesis (Guideline 3, Section2.1), while keeping all other components identical. We refer to this variant as CaberNet–NoIndy. To assess compactness, we vary the latent size $d \in \{32, 16, 8\}$ (see Tables 3 and details in Appendix C).

Overall, CaberNet outperforms CaberNet-NoIndy. The independence loss compresses redundancy and allocates capacity to informative factors, so the latent dimensions carries useful and non-overlapping signal. Consequently, when reducing model width, CaberNet degrades more slowly: from hidden size 32 → 16 the relative NMSE increase is about **9.4%** for CaberNet versus 12.1% for CaberNet-noIndy; from 32 → 8 it is about **11.5%** versus 14.5%, respectively. This stability suggests that encouraging factor independence reduces the redundancy in the representation, making the model less sensitive to width constraints. This is consistent with prior research on independence-promoting bottlenecks [22].

6.3 Explainability

6.3.1 Feature weighting. We extract global gate weights g (Eq. 4) that reweight standardized raw features before they enter the causal representation extractor. Rather than hard-masking features as in LiNGAM-style selection, the gate performs *soft* selection: it assigns continuous importances and yields a data-driven split into *sup-causal* (higher weight) and *inf-causal* (lower weight) variables. Softness is desirable because even features with weaker direct effects can carry a latent causal signal or act as proxies that help form a better causal representation Z; they are therefore down-weighted rather than discarded.

Per-domain patterns (Figure 6 left). Across all six domains, the curves exhibit a highly similar ranking of feature importance with only mild dispersion. This cross-domain alignment suggests that the gate captures relationships that are stable rather than domain-specific, consistent with the *causal invariance* assumption (Section 2.1). In other words, the learned importance profile transfers across buildings/floors/cities, which supports our goal of domain-robust prediction.

Averaged importance (Figure 6 right). Two variables, namely *TVOC* and *indoor air pressure*, consistently fall below the sup/inf threshold and are classified as *inf-causal*; this is plausible since they do not directly drive AC operation in these offices. *Outdoor temperature* receives the largest weight, which agrees with building physics and AC control logic. The calendar proxy *is_work* is also strongly weighted, reflecting occupancy-driven operation typical of office buildings. *Light intensity* appears higher than one might expect if viewed only as a heat-gain channel; in practice it functions as a robust proxy for *time-of-day/usage regime* (natural light cycles

Table 4: Top-3 Contributing Features per Representation Dimension (Original and Debiased Jacobian)

Z	1st Feature	2nd Feature	3rd Feature
Original Jacobian			
Z_0	outdoor_temperature	indoor_light	indoor_pressure
Z_1	outdoor_temperature	indoor_pressure	is_work
Z_2	indoor_pressure	outdoor_temperature	indoor_temperature
Z_3	outdoor_temperature	indoor_light	is_work
Z_4	indoor_humidity	outdoor_temperature	indoor_pressure
Debiased Jacobian			
Z_0	outdoor_temperature	indoor_light	indoor_pressure
Z_1	indoor_pressure	outdoor_temperature	indoor_tvoc
Z_2	indoor_pressure	outdoor_temperature	indoor_temperature
Z_3	outdoor_temperature	indoor_light	is_work
Z_4	indoor_humidity	outdoor_temperature	indoor_pressure

and artificial lighting schedules), which is highly informative for AC operation in the absence of an explicit time feature, and thus effectively following the pathway: light → time → AC operation.

By contrast, *indoor temperature* and *indoor humidity* have relatively modest weights: when AC is on, these variables are actively regulated near setpoints that meet human comfort (low variability, limited incremental predictive signal), and when AC is off (e.g., nights), their drift toward outdoor conditions does not immediately translate into consumption since the AC is not operating. Thus, they provide limited incremental predictive value.

6.3.2 Jacobian-Based Feature Contribution. Motivated by previous work [36, 41], we quantify how each raw feature contributes to each dimension of the d-dimensional Z using Jacobians and then *de-bias* magnitudes by removing the scaling effect of the global gate's weights. For clarity, we fix the hidden size to $d = 5$.

Jacobian-based contribution. Let the input window be $X_{t-w:t-1} \in \mathbb{R}^{w\times p}$ with p raw features per time step and w steps, and let the learned representation be $Z \in \mathbb{R}^d$. We define the time-averaged Jacobian (magnitude) with respect to the *original* inputs as

$$J = \frac{1}{w}\sum_{t=1}^{w}\left|\frac{\partial Z}{\partial X_t}\right| \in \mathbb{R}^{d\times p},$$

whose (i, j) entry measures the absolute influence of the j-th raw feature on Z_i across the time window.

To eliminate the confounding influence of the gate weights, we debias the contribution matrix by right-multiplication with $\mathrm{diag}(g)^{-1}$; using the chain rule (see Appendix D),

$$J_{\text{debiased}} = J \cdot \mathrm{diag}(g)^{-1} \tag{12}$$

Why debias? J reflects the dependency actually used by the trained model (post-gating), while J_{debiased} reveals the intrinsic sensitivity of the encoder absent the structural rescaling by g. A large J_{debiased} but small J indicates a feature that is intrinsically informative yet suppressed by the gate, whereas large values in both J_{debiased} and J indicate a feature that is both informative and actively used by the model. This bias is discussed deeply in work [38].

The top-3 contributing features per Z_i are summarized in Table 4, and the full numerical values are provided in Appendix E. We summarize the key observations below.

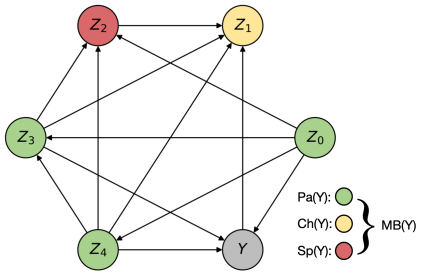


Figure 7: Reconstructed SCM (hidden size: 5).

(1) Sup-/inf-causal alignment. According to Section 6.3.1, we regard *outdoor_temperature*, *indoor_light*, and *is_work* as sup-causal. In Table 4, these three sup-causal features account for **9/15** of the Top-3 slots under the original J and **8/15** under the debiased J_{debiased}. By contrast, the smallest-weight inf-causal feature *indoor_tvoc* appears only **once** (and only in the debiased view). This pattern is consistent with the gate-derived sup/inf split: sup-causal features contribute more to the learned causal representation.

(2) Dominance of outdoor temperature with complementary specialization. Although its magnitude shrinks after debiasing, *outdoor_temperature* remains within the Top-2 for *every* Z_i in both J and J_{debiased}, reflecting its primacy as a physical driver (also see its highest average weight in Figure 6). The other Top-3 entries vary across Z_i, suggesting complementary specialization among latent coordinates: the model preserves the dominant outdoor signal while, under the independence regularizer $\mathcal{L}_{indy}$ (Eq. 11), allocating the remaining capacity to less redundant features so as to balance prediction accuracy and $\mathcal{L}_{indy}$.

(3) Role of *indoor_light* and *is_work*. These two sup-causal proxies appear fewer times than *outdoor_temperature* in Top-3 lists, which is consistent with their indirect action on energy (time-of-day/usage regime indicators) discussed in Section 6.3.1. They are likely distributed across multiple Z_i and partially decomposed, rather than concentrated in a few coordinates. Importantly, Table 7 in Appendix shows that they maintain high–mid ranks overall, even not often in the Top-3, surpassing most inf-causal features.

6.3.3 Causal Discovery: SCM Reconstruction. Based on Section 6.3.2, we reconstruct an SCM over the latent coordinates Z and energy consumption Y. We use the same direct-LiNGAM [37] algorithm as in Section 6.1 to model causality. The reconstructed SCM is shown in Figure 7. We next interpret the reconstructed SCM using the contribution matrix above (Table 4).

(1) Markov blanket alignment. Z accurately represents the Markov blanket of Y (Figure 7), aligning with our objective that the causal representation extractor h_ϕ learns Markov blanket representations (Section 3.1) and the *Guideline 2: Markov Blanket Prediction* (Section 2.1).

(2) High in-degree aggregators. In Figure 7, Z_1 and Z_2 exhibit higher in-degree (multiple incoming arrows), indicating they act as aggregators of upstream signals. Consistently, their debiased Jacobian J_{debiased} is topped by indoor variables (Table 4; *indoor_pressure* for both), which aligns with the view that they are primarily *affected* nodes rather than pure drivers.

(3) Indoor thermodynamic response captured by Z_2. Among all coordinates, *indoor_temperature* appears in the Top-3 *only* for Z_2 (both J and J_{debiased}), suggesting that Z_2 encodes more the building's indoor thermal response. The reconstructed SCM shows arrows from other latents into Z_2; notably, Z_0, Z_3, and Z_4 are strongly associated with *outdoor_temperature* in Table 4. This is consistent with the physical pathway

$$\text{outdoor} \;\rightarrow\; \text{building dynamics} \;\rightarrow\; \text{indoor},$$

i.e., outdoor-driven temperature latents feeding into the indoor-response temperature latent.

(4) Parents of Y. Edges into Y originate from latents whose Top-3 include *outdoor_temperature* and *indoor_humidity*/*pressure* (e.g., Z_0, Z_3, Z_4; see Table 4), which is plausible: outdoor thermal load and indoor moisture/airmass proxies are primary physical drivers of AC energy. Latents associated with regime proxies (*indoor_light*, *is_work*; e.g., Z_3) tend to contribute either directly or indirectly, matching their proxy nature discussed in Section 6.3.1.

7 Conclusion

We presented CaberNet, a causal and interpretable representation learning framework for cross-domain energy prediction. The method couples a global feature gate with self-supervised Bernoulli regularization and domain-wise objectives for bias learning toward invariant, mechanism-level relations. Evaluated on real-world datasets collected from buildings in different climate zones, CaberNet achieves state-of-the-art out-of-domain performance, improving average NMSE by 22.9% over the best benchmark. Moreover, its interpretable latent representations align with domain knowledge, providing transparency critical for real-world energy optimization. Overall, CaberNet overcomes a fundamental limitation in current approaches to cross-domain HVAC energy prediction and helps develop more adaptive, efficient, and automated HVAC control strategies across diverse built environments.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[2] Paul Boadu Asamoah and Ekundayo Shittu. 2025. Evaluating the performance of machine learning models for energy load prediction in residential HVAC systems. *Energy and Buildings* 334 (2025), 115517.

[3] Australian Government Department of Climate Change, Energy, the Environment and Water. 2013. HVAC Energy Breakdown Factsheet. https://www.environment.gov.au/system/files/energy/files/hvac-factsheet-energy-breakdown.pdf. Accessed: 2025-09-10.

[4] Yongbao Chen, Qiguo Yang, Zhe Chen, Chengchu Yan, Shu Zeng, and Mingkun Dai. 2023. Physics-informed neural networks for building thermal modeling and demand response control. *Building and Environment* 234 (2023), 110149.

[5] Aniruddh Chennapragada, Divya Periyakoil, Hari Prasanna Das, and Costas J Spanos. 2022. Time series-based deep learning model for personal thermal comfort prediction. In *Proceedings of the thirteenth ACM international conference on future energy systems*. 552–555.

[6] Francesco Giuseppe Ciampi, Andrea Rega, Thierno ML Diallo, Francesco Pelella, Jean-Yves Choley, and Stanislao Patalano. 2024. Energy consumption prediction of industrial HVAC systems using Bayesian Networks. *Energy and Buildings* 309 (2024), 114039.

[7] Xue Cui, Minhyun Lee, Choongwan Koo, and Taehoon Hong. 2024. Energy consumption prediction and household feature analysis for different residential building types using machine learning and SHAP: Toward energy-efficient buildings. *Energy and Buildings* 309 (2024), 113997.

[8] Patrick Nzivugira Duhirwe, Jack Ngarambe, and Geun Young Yun. 2024. Causal effects of policy and occupant behavior on cooling energy. *Renewable and Sustainable Energy Reviews* 206 (2024), 114854.

[9] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434* (2020).

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Ali Hooshmand and Ratnesh Sharma. 2019. Energy predictive models with limited data using transfer learning. In *Proceedings of the tenth ACM international conference on future energy systems*. 12–16.

[12] Shushan Hu, Jiale Wang, Cathal Hoare, Yehong Li, Pieter Pauwels, and James O'Donnell. 2021. Building energy performance assessment using linked data and cross-domain semantic reasoning. *Automation in Construction* 124 (2021), 103580.

[13] Jiajing Huang, Naghmeh Ghalamsiah, Abhidnya Patharkar, Ojas Pradhan, Mengyuan Chu, Teresa Wu, Jin Wen, Zheng O'Neill, and Kasim Selcuk Candan. 2024. An entropy-based causality framework for cross-level faults diagnosis and isolation in building HVAC systems. *Energy and Buildings* 317 (2024), 114378.

[14] Bo Jiang, Yuang Wei, Ting Zhang, and Wei Zhang. 2024. Improving the performance and explainability of knowledge tracing via Markov blanket. *Information Processing & Management* 61, 3 (2024), 103620.

[15] Fuyang Jiang and Hussain Kazmi. 2025. What-if: A causal machine learning approach to control-oriented modelling for building thermal dynamics. *Applied Energy* 377 (2025), 124550.

[16] Waqar Khan, Lingfu Kong, Sohail M Noman, and Brekhna Brekhna. 2023. A novel feature selection method via mining Markov blanket. *Applied Intelligence* 53, 7 (2023), 8232–8255.

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*. 5542–5550.

[18] Yang Li, Kevin B Korb, and Lloyd Allison. 2021. Markov blanket discovery using minimum message length. *arXiv preprint arXiv:2107.08140* (2021).

[19] Huiheng Liu, Yanchen Liu, Xun Guo, Huijun Wu, Huan Wang, and Yanni Liu. 2023. An energy consumption prediction method for HVAC systems using energy storage based on time series shifting and deep learning. *Energy and Buildings* 298 (2023), 113508.

[20] Huiheng Liu, Yanchen Liu, Huakun Huang, Huijun Wu, and Yu Huang. 2024. Energy consumption dynamic prediction for HVAC systems based on feature clustering deconstruction and model training adaptation. In *Building Simulation*, Vol. 17. Springer, 1439–1460.

[21] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14052–14060.

[22] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8046–8056.

[23] Hongbo Ma, Jiacheng Wei, Guowei Zhang, Xianguang Kong, and Jingli Du. 2024. Causality-inspired multi-source domain generalization method for intelligent fault diagnosis under unknown operating conditions. *Reliability Engineering & System Safety* 252 (2024), 110439.

[24] Zhihao Ma, Gang Jiang, and Jianli Chen. 2024. Physics-informed ensemble learning with residual modeling for enhanced building energy prediction. *Energy and Buildings* 323 (2024), 114853.

[25] Ritwik Mohan and Nikhil Pachauri. 2025. An ensemble model for the energy consumption prediction of residential buildings. *Energy* 314 (2025), 134255.

[26] NASA Prediction Of Worldwide Energy Resources (POWER). 2025. NASA POWER Data Access Viewer. https://power.larc.nasa.gov/. Accessed: 2025-07-10.

[27] Toan Nguyen, Kien Do, Duc Thanh Nguyen, Bao Duong, and Thin Nguyen. 2023. Causal inference via style transfer for out-of-distribution generalisation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1746–1757.

[28] Zhongjun Ni, Chi Zhang, Magnus Karlsson, and Shaofang Gong. 2024. A study of deep learning-based multi-horizon building energy forecasting. *Energy and Buildings* 303 (2024), 113810.

[29] Judea Pearl. 2009. *Causality*. Cambridge university press.

[30] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78, 5 (2016), 947–1012.

[31] Niklas Pfister, Peter Bühlmann, and Jonas Peters. 2019. Invariant causal prediction for sequential data. *J. Amer. Statist. Assoc.* 114, 527 (2019), 1264–1276.

[32] Bruno Cristino Pinheiro and Paulo Henrique de Mello SantAna. 2025. AHP-based decision making to selecting energy-efficient air conditioning equipment in a commercial building. *Energy and Buildings* 329 (2025), 115281.

[33] Fabian Raisch, Thomas Krug, Christoph Goebel, and Benjamin Tischler. 2025. GenTL: A General Transfer Learning Model for Building Thermal Dynamics. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*. 322–333.

[34] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2020. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761* (2020).

[35] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[36] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[37] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR* 12, Apr (2011), 1225–1248.

[38] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415* (2024).

[39] Zhuoqun Xing, Yiqun Pan, Yiting Yang, Xiaolei Yuan, Yumin Liang, and Zhizhong Huang. 2024. Transfer learning integrating similarity analysis for short-term and long-term building energy consumption prediction. *Applied energy* 365 (2024), 123276.

[40] Naiyu Yin, Hanjing Wang, Yue Yu, Tian Gao, Amit Dhurandhar, and Qiang Ji. 2024. Integrating Markov Blanket Discovery Into Causal Representation Learning for Domain Generalization. In *European Conference on Computer Vision*. Springer, 271–288.

[41] Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. 2024. Opti-CAM: Optimizing saliency maps for interpretability. *Computer Vision and Image Understanding* 248 (2024), 104101.

[42] Hongliang Zhang and Shang Xu. 2025. Climate Change, Air Conditioning Adoption, and Household Electricity Use: Evidence from the Northwestern United States. *Environmental and Resource Economics* 88 (2025), 1–33.

A Proof of Equation: Lower Conditional Variance with the Markov Blanket

Assume all variables $X_1, \ldots, X_n$ have finite second moments. Since $\mathrm{Pa}(Y) \subseteq \mathrm{MB}(Y)$, apply the law of total variance with nested conditioning:

$$\begin{aligned}\mathrm{Var}\big(Y \mid \mathrm{Pa}(Y)\big) &= \mathbb{E}\Big[\mathrm{Var}\big(Y \mid \mathrm{MB}(Y)\big) \,\Big|\, \mathrm{Pa}(Y)\Big] + \\ &\qquad \mathrm{Var}\Big(\mathbb{E}\big[Y \mid \mathrm{MB}(Y)\big] \,\Big|\, \mathrm{Pa}(Y)\Big) \\ &\geq \mathbb{E}\Big[\mathrm{Var}\big(Y \mid \mathrm{MB}(Y)\big) \,\Big|\, \mathrm{Pa}(Y)\Big]\end{aligned}$$

Taking expectation over $\mathrm{Pa}(Y)$ yields

$$\mathbb{E}_{\mathrm{Pa}(Y)}\Big[\,\mathrm{Var}\big(Y \mid \mathrm{Pa}(Y)\big)\,\Big] \;\geq\; \mathbb{E}_{\mathrm{Pa}(Y)}\Big[\,\mathbb{E}\big[\mathrm{Var}(Y \mid \mathrm{MB}(Y)) \,\big|\, \mathrm{Pa}(Y)\big]\,\Big]$$

Hence,

$$\mathbb{E}\left[\mathrm{Var}(Y \mid \mathrm{Pa}(Y))\right] \geq \mathbb{E}\left[\mathrm{Var}(Y \mid \mathrm{MB}(Y))\right]$$

Hence, conditioning on the larger information set $\mathrm{MB}(Y)$ (parents, children, and co-parents) cannot increase the residual uncertainty about Y, and under standard regularity conditions, this is stated pointwise (a.s.) as

$$\mathrm{Var}\big(Y \mid \mathrm{MB}(Y)\big) \;\leq\; \mathrm{Var}\big(Y \mid \mathrm{Pa}(Y)\big).$$

B Causal Graphs Extracted by LSTM-LiNGAM

We report the Markov blankets and causal structures discovered by LSTM-LiNGAM for each building-floor dataset.

Table 5: Causal structures discovered by LiNGAM. Abbreviations: IT = indoor_temperature, IH = indoor_humidity, IL = indoor_light_intensity, IC = indoor_co2, IP = indoor_pressure, IV = indoor_tvoc, OT = outdoor_temperature, W = is_work.

Dataset	Parents	Children	Spouses	Markov Blanket
North L-A	–	{IC, OT}	{W, IP, IL, IH}	{IL, IC, IH, W, IP, OT}
North L-B	{IT, IL, IC, IP, OT, W}	–	–	{IT, IL, IC, OT, W, IP}
Central L-A	–	{OT}	{IP, IL}	{IP, OT, IL}
Central L-B	{IV, IP, W}	{IT, IH, IC, OT}	{W, IV, OT, IP, IT, IL}	{W, IV, IC, OT, IP, IL, IT, IH}
South L-A	{IT, IL, IP}	{IH, IV}	{IV, IL, W, IP, IC, IT}	{IH, IL, W, IT, IP, IC, IV}
South L-B	{IL, IP}	{IH, OT}	{IV, IT, IP, OT, IL}	{IV, IT, IH, IP, OT, IL}

C Detail of indy loss ablation study

Table 6: Performance of CaberNet-NoIndy vs. original CaberNet (NMSE).

Area	L	32		16		8		Average	
		no-indy	indy	no-indy	indy	no-indy	indy	no-indy	indy
North	L-A	0.145	0.081	0.238	0.129	0.205	0.143	0.207	0.192
	L-B	0.149	0.142	0.162	0.162	0.161	0.159		
Central	L-A	0.135	0.138	0.160	0.147	0.161	0.135	0.232	0.210
	L-B	0.263	0.238	0.278	0.263	0.323	0.307		
South	L-A	0.299	0.285	0.279	0.279	0.280	0.261	0.237	0.214
	L-B	0.251	0.270	0.277	0.282	0.289	0.278		

D Derivation of the Debiased Jacobian

We debias the contribution matrix to remove the influence of weights g. The encoder applies a global, sample-agnostic gate $g = \mathrm{softmax}(\alpha) \in (0,1)^p$ (Eq. 4) and forms $\tilde{X}_t = g \odot X_t$. By the chain rule,

$$\frac{\partial Z}{\partial X_t} \;=\; \frac{\partial Z}{\partial \tilde{X}_t}\,\frac{\partial \tilde{X}_t}{\partial X_t} \;=\; \frac{\partial Z}{\partial \tilde{X}_t}\,\mathrm{diag}(g). \tag{13}$$

Let $A_t := \left|\frac{\partial Z}{\partial \tilde{X}_t}\right|$. Since $g_j \geq 0$ and $\mathrm{diag}(g)$ rescales columns, we have the columnwise identity

$$\left|A_t \,\mathrm{diag}(g)\right| \;=\; A_t \,\mathrm{diag}(g),$$

hence

$$J \;=\; \frac{1}{T}\sum_{t=1}^{T}\left|\tfrac{\partial Z}{\partial \tilde{X}_t}\right| \;=\; \frac{1}{T}\sum_{t=1}^{T} A_t \,\mathrm{diag}(g) \;=\; \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} A_t\right)}_{\frac{1}{T}\sum_{t=1}^{T}\left|\frac{\partial Z}{\partial \tilde{X}_t}\right|} \mathrm{diag}(g).$$

Therefore, the *debiased* Jacobian that removes the structural rescaling by g is obtained by right-multiplication with $\mathrm{diag}(g)^{-1}$:

$$\boxed{J_{\text{debiased}} \;=\; J\,\mathrm{diag}(g)^{-1} \;=\; \frac{1}{T}\sum_{t=1}^{T}\left|\tfrac{\partial Z}{\partial \tilde{X}_t}\right|}$$

Remark. For clarity, we assume h_ϕ to be linear when deriving Eq. (13). In reality, since the network contains non-linear activation functions, the mapping is only piecewise linear. Consequently, the debiasing may not be perfectly accurate in a global sense, but it still provides a reasonable correction that mitigates the bias to a large extent.

E Weighted derivatives of causal representation

Table 7: Jacobian-based influence of raw features on each representation dimension (Original vs. Debiased)

Original Jacobian								
Z	outdoor_temperature	indoor_light	indoor_pressure	is_work	indoor_temperature	indoor_co2	indoor_humidity	indoor_tvoc
Z_0	0.03127	0.00991	0.00802	0.00748	0.00736	0.00688	0.00473	0.00270
Z_1	0.00496	0.00251	0.00472	0.00347	0.00195	0.00279	0.00065	0.00312
Z_2	0.00147	0.00063	0.00242	0.00064	0.00100	0.00045	0.00048	0.00087
Z_3	0.02558	0.01579	0.00568	0.01150	0.00968	0.00681	0.00348	0.00397
Z_4	0.00147	0.00041	0.00063	0.00042	0.00052	0.00040	0.00249	0.00051
Debiased Jacobian								
Z_0	0.22649	0.07026	0.06907	0.05490	0.06183	0.05714	0.04049	0.02416
Z_1	0.03596	0.01777	0.04068	0.02544	0.01638	0.02316	0.00554	0.02789
Z_2	0.01062	0.00448	0.02083	0.00473	0.00837	0.00373	0.00408	0.00778
Z_3	0.18528	0.11191	0.04891	0.08439	0.08127	0.05659	0.02972	0.03542
Z_4	0.01063	0.00288	0.00544	0.00310	0.00436	0.00328	0.02129	0.00455