Causal Disentanglement for Household Energy Segmentation: A Deep Learning Framework for Weather-Independent Pattern Discovery

Author Name, *Member, IEEE,* Second Author, *Member, IEEE,* and Third Author, *Senior Member, IEEE*

***Abstract*—Household energy segmentation is fundamental to smart grid management, enabling targeted demand response programs and personalized energy services. However, conventional clustering methods are severely compromised by weather confounding, struggling to distinguish intrinsic household behaviors from weather-induced consumption variations. This leads to unstable segmentations that conflate behavioral archetypes with seasonal patterns. To address this critical limitation, we propose CausalHES, a novel deep learning framework that reformulates household energy segmentation as a causal source separation problem. CausalHES employs a Causal Source Separation Autoencoder (CSSAE) to disentangle observed load profiles into weather-independent base consumption patterns and weather-dependent effects. The CSSAE enforces statistical independence between base load and weather representations through a composite loss function combining mutual information minimization, adversarial training, and distance correlation penalties. Clustering is then performed exclusively on the purified base load embeddings using Deep Embedded Clustering (DEC), ensuring segments reflect genuine behavioral differences rather than weather artifacts. Comprehensive experiments on the Irish CER smart meter dataset demonstrate CausalHES's superior performance, achieving 87.60% clustering accuracy—substantially outperforming traditional methods (34.93%), uni-modal deep clustering (79.50%), and multi-modal fusion approaches (81.90%). Ablation studies confirm the necessity of both causal independence enforcement and explicit separation architecture. Beyond superior accuracy, CausalHES provides interpretable decompositions with strong semantic consistency (r=0.78 correlation between reconstructed weather effects and temperature), enabling actionable insights for energy analytics. Our results establish that causally-informed source separation yields more robust, accurate, and interpretable household energy segmentation.**

***Index Terms*—Causal Inference, Deep Clustering, Disentangled Representation Learning, Household Energy Segmentation, Multi-modal Time Series, Source Separation, Smart Grid Analytics, Neural Networks.**

I. Introduction

HOUSEHOLD energy segmentation, the task of clustering consumers based on their electricity usage patterns, is a cornerstone of modern energy systems management [?]. Identifying distinct behavioral archetypes enables utility providers to design targeted demand response programs, enhance load forecasting accuracy, and promote energy efficiency initiatives [?]. However, the efficacy of segmentation is fundamentally challenged by the composite nature of smart meter data. Observed load profiles invariably mix intrinsic household behaviors (e.g., appliance usage schedules, occupant routines) with responses to external variables, among which meteorological conditions are the most dominant and pervasive confounders [?].

Conventional clustering algorithms, ranging from K-means [?] to more advanced spectral methods [?], typically operate on raw or feature-engineered load data. These approaches implicitly assume that the input signal directly reflects the underlying group structure that one wishes to discover. This assumption is frequently violated in energy analysis. For instance, two households with fundamentally different intrinsic energy behaviors might exhibit superficially similar consumption patterns during an extreme weather event (e.g., a heatwave causing synchronized air conditioner usage). Conversely, a single household might be incorrectly assigned to different behavioral clusters across varying seasons if weather effects are not appropriately accounted for.

The advent of deep learning has introduced powerful representation learning capabilities for clustering, exemplified by Deep Embedded Clustering (DEC) [?], which jointly optimizes feature representations and cluster assignments. While effective for uni-modal data, these models do not inherently account for external confounding factors. Multi-modal extensions that fuse load and weather data have been proposed [?]. However, these often treat the modalities as symmetric inputs, failing to adequately model the established unidirectional causal relationship: weather influences energy consumption, but not vice-versa. This oversight can lead to the learning of spurious correlations rather than the discovery of true, underlying behavioral patterns.

This paper posits that meaningful and robust household segmentation requires moving beyond simple data fusion. Instead, we advocate for an approach that explicitly addresses the causal structure inherent in energy consumption data. We introduce CausalHES (Causally-informed Household Energy Segmentation), a deep learning framework that reformulates the segmentation problem as a task of causal source separation followed by clustering. Our central hypothesis is that observed household energy consumption $\mathbf{x}^{(l)}$ can be modeled as an additive mixture of a weather-independent base component $\mathbf{s}_{\text{base}}$ and a weather-dependent effect $\mathbf{s}_{\text{weather}}$. True behavioral segmentation should then be performed on representations of $\mathbf{s}_{\text{base}}$ alone, as this component reflects intrinsic patterns.

Manuscript received Month DD, YYYY; revised Month DD, YYYY.

Author Name is with the Department of Computer Science, University Name, City, State ZIP, Country (e-mail: author@university.edu).

Second Author is with the Department of Electrical Engineering, University Name, City, State ZIP, Country (e-mail: second@university.edu).

Third Author is with the Department of Data Science, University Name, City, State ZIP, Country (e-mail: third@university.edu).

The core of our proposed framework is the Causal Source Separation Autoencoder (CSSAE). The CSSAE is designed to learn the decomposition of the input load signal into its constituent components. It achieves this by training encoders to produce latent representations that are maximally informative for reconstructing the original signal while ensuring that the representation of the base load is statistically independent of weather information. This crucial independence is enforced via a composite objective function that synergistically combines three complementary statistical measures: Mutual Information Neural Estimation (MINE), adversarial training, and distance correlation. Once the model is trained and the sources are separated, we apply the DEC algorithm to the learned base load embeddings. This yields clusters that are, by design, robust to weather-driven variations and thus more reflective of intrinsic household behaviors.

Our main contributions are:

1) We formally frame household energy segmentation as a causal source separation problem, providing a principled methodology to handle confounding variables like weather. This contrasts with existing methods that either ignore weather or use ad-hoc fusion techniques.
2) We propose the CSSAE architecture, featuring a novel composite independence loss. This mechanism effectively disentangles weather-dependent and weather-independent components from a single observed load signal, conditioned on associated weather data.
3) Through extensive experiments on the publicly available and challenging real-world Irish CER dataset, we demonstrate that CausalHES achieves 87.60% clustering accuracy. This significantly surpasses a suite of traditional, standard deep learning, and multi-modal clustering methods, thereby producing more stable and interpretable segmentations.

The remainder of this paper is organized as follows. Section II reviews pertinent related work. Section III details the theoretical foundation and the specific architecture of the CausalHES framework. Section IV presents comprehensive experiments including setup, results, ablation studies, and interpretability analysis. Section V discusses the implications, limitations, and future research directions. Section VI provides concluding remarks.

II. Related Work

Our work is situated at the confluence of deep clustering, causal inference, multi-modal learning, and their application to energy analytics.

Deep clustering methods have significantly advanced beyond traditional techniques. Early approaches often involved a two-stage process: an autoencoder (AE) or variational autoencoder (VAE) would first learn a low-dimensional embedding of the data [?], after which a standard clustering algorithm like K-means would be applied. A notable progression was Deep Embedded Clustering (DEC) [?], which introduced an end-to-end framework that jointly optimizes feature representations and refines cluster assignments. Such methods, however, are predominantly designed for uni-modal data and generally assume that the input data directly reflects the underlying cluster structure, an assumption that is violated in energy consumption data due to weather confounding.

Causal inference provides a formal framework for identifying and quantifying cause-and-effect relationships from data [?]. In recent years, these principles have been increasingly integrated into machine learning to build more robust, generalizable, and interpretable models, giving rise to the field of causal representation learning [?], [?]. The objective in this domain is often to learn representations where specific latent variables correspond to distinct, independent causal mechanisms or factors. Our work applies this philosophy to the energy domain by explicitly aiming to learn a base load representation that is independent of weather, which is known to be a causal driver of energy use. This connects our approach to ideas from non-linear Independent Component Analysis (ICA) [?] and disentangled representation learning, but with a clear causal motivation guiding the separation process.

Multi-modal learning in energy analytics has primarily focused on fusing load and weather data for tasks like energy forecasting [?]. Fusion strategies range from simple early fusion (e.g., concatenating input features) to more complex intermediate or late fusion techniques, including attention-based mechanisms. However, when applied to clustering, these methods often fail to respect the asymmetric causal relationship where weather influences load, but not vice-versa. Some studies employ weather normalization as a pre-processing step [?], but these typically rely on simpler (often linear) models and may not fully capture the complex, non-linear interactions between weather variables and diverse household consumption behaviors. Our work distinguishes itself by embedding the assumed causal structure (independence of base load from weather) directly into the learning objective of a flexible deep neural network, allowing for a non-linear, data-driven separation of the consumption sources.

Enforcing statistical independence between learned latent variables is a key technical challenge in representation learning. Various techniques have been proposed: adversarial training, where a discriminator network attempts to predict one representation from another, while the encoder is trained to ”fool” this discriminator [?]; direct minimization of mutual information using neural estimators like MINE [?]; and penalizing measures like distance correlation [?], which is a non-parametric statistic capable of detecting non-linear dependencies. CausalHES synergistically combines these three distinct techniques into a composite independence loss. This approach leverages their complementary strengths to achieve a more robust and comprehensive disentanglement than might be possible with any single independence measure alone.

III. The CausalHES Framework

This section details the CausalHES framework, designed for robust household energy segmentation by disentangling intrinsic consumption patterns from weather-induced effects.

A. Overall Architecture and Problem Formulation

The fundamental challenge in household energy segmentation lies in the composite nature of observed electricity load data. Weather fluctuations, such as temperature changes or humidity variations, significantly influence energy usage (e.g., for heating or cooling), often masking the underlying, habitual consumption patterns of a household. CausalHES addresses this by first establishing a clear model of the energy consumption process.

We posit that the observed household energy load time series, denoted as $\mathbf{x}^{(l)} \in \mathbb{R}^T$ for a period T, can be effectively modeled as an additive combination of two primary latent components: an intrinsic base load component, $\mathbf{s}_{\text{base}} \in \mathbb{R}^T$, which captures the household's typical, weather-invariant consumption behavior, and a weather-dependent effect component, $\mathbf{s}_{\text{weather}} \in \mathbb{R}^T$, which accounts for the portion of consumption directly driven by meteorological factors. This relationship is expressed as:

$$\mathbf{x}^{(l)} = \mathbf{s}_{\text{base}} + \mathbf{s}_{\text{weather}} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^T$ represents residual noise or unmodeled factors. A critical underpinning of our approach is the causal assumption that the intrinsic base load component, $\mathbf{s}_{\text{base}}$, is statistically independent of the observed weather variables, $\mathbf{x}^{(w)} \in \mathbb{R}^{T \times D_w}$ (where D_w is the dimensionality of weather features). This is formally stated as $\mathbf{s}_{\text{base}} \perp \mathbf{x}^{(w)}$. The ultimate goal of CausalHES is then to perform clustering based on learned representations of this disentangled base load component, $\mathbf{s}_{\text{base}}$, thereby ensuring that the resulting segments are robust to weather variations and truly reflect distinct consumer behaviors.

To implement this vision, the CausalHES framework employs a multi-stage deep learning architecture, an overview of which is presented in Fig. 1.

As illustrated in Fig. 1, the CausalHES pipeline begins by taking the raw household load time series $\mathbf{x}^{(l)}$ and the corresponding weather data $\mathbf{x}^{(w)}$ as inputs. These multimodal time series are then channeled into the primary component of our framework: the **Causal Source Separation Autoencoder (CSSAE)**. The CSSAE is an intricate neural network specifically architected to perform the disentanglement task. It internally consists of:

- Modality-specific encoders (E_l for load, E_w for weather) that transform the input time series into compact latent representations ($\mathbf{z}_{\text{mixed}}$ from load, $\mathbf{z}_{\text{weather}}$ from weather).
- A source separation module (composed of sub-networks S_b and S_{we}) that processes $\mathbf{z}_{\text{mixed}}$ (conditioned on $\mathbf{z}_{\text{weather}}$ for the weather-effect part) to yield two distinct latent vectors: $\mathbf{z}_{\text{base}}$, representing the weather-independent base load, and $\mathbf{z}_{\text{weather-effect}}$, representing the weather-influenced consumption.
- Dual decoders (D_b and D_{we}) that reconstruct the time-domain signals $\hat{\mathbf{s}}_{\text{base}}$ and $\hat{\mathbf{s}}_{\text{weather}}$ from $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather-effect}}$, respectively.

The training of the CSSAE is governed by two key loss functions. The reconstruction loss ($\mathcal{L}_{\text{rec}}$) ensures that the sum of the decoded components ($\hat{\mathbf{s}}_{\text{base}} + \hat{\mathbf{s}}_{\text{weather}}$) accurately reconstructs the original input load $\mathbf{x}^{(l)}$. Crucially, the causal independence loss ($\mathcal{L}_{\text{causal}}$) enforces statistical independence between the learned base load embedding $\mathbf{z}_{\text{base}}$ and the weather embedding $\mathbf{z}_{\text{weather}}$, thereby operationalizing our core causal assumption.

Following the successful disentanglement by the CSSAE, the learned weather-independent base load embeddings, $\mathbf{Z}_{\text{base}} = \{\mathbf{z}_{\text{base},i}\}_{i=1}^{N}$ for N households, are passed to a **Deep Embedded Clustering (DEC) module**. This module employs these purified embeddings to identify distinct clusters of households. It is optimized using a clustering-specific loss function ($\mathcal{L}_{\text{cluster}}$) that encourages the formation of compact and well-separated clusters. The final output of the CausalHES framework is a set of cluster assignments for each household, reflecting their intrinsic energy consumption behaviors, now largely devoid of weather-induced confounding. The detailed architecture of each component and the specific loss formulations are elaborated in the subsequent subsections.

B. Causal Source Separation Autoencoder (CSSAE)

The CSSAE is the architectural core of our framework, specifically designed to implement the source separation implied by Eq. 1. It comprises modality-specific encoders, a dedicated source separation module, and corresponding decoders for signal reconstruction.

a) Encoders: We employ two distinct encoder networks to process the heterogeneous input modalities. A load encoder, $E_l : \mathbb{R}^T \to \mathbb{R}^{d_m}$, maps the observed (mixed) load signal $\mathbf{x}^{(l)}$ to a latent representation $\mathbf{z}_{\text{mixed}} \in \mathbb{R}^{d_m}$.

$$\mathbf{z}_{\text{mixed}} = E_l(\mathbf{x}^{(l)}; \theta_{E_l}) \tag{2}$$

A weather encoder, $E_w : \mathbb{R}^{T \times D_w} \to \mathbb{R}^{d_w}$, maps the input weather signal $\mathbf{x}^{(w)}$ to a latent weather representation $\mathbf{z}_{\text{weather}} \in \mathbb{R}^{d_w}$.

$$\mathbf{z}_{\text{weather}} = E_w(\mathbf{x}^{(w)}; \theta_{E_w}) \tag{3}$$

Both E_l and E_w are implemented using 1D convolutional neural networks (CNNs), chosen for their efficacy in capturing temporal dependencies and local patterns in time series data. These are followed by global average pooling and dense layers to produce the fixed-size embeddings.

b) Source Separation Module: This module takes the encoded representations $\mathbf{z}_{\text{mixed}}$ and $\mathbf{z}_{\text{weather}}$ as input and aims to separate $\mathbf{z}_{\text{mixed}}$ into a base load embedding $\mathbf{z}_{\text{base}} \in \mathbb{R}^{d_b}$ and a weather-effect embedding $\mathbf{z}_{\text{weather-effect}} \in \mathbb{R}^{d_{we}}$. This separation is performed by two multi-layer perceptrons (MLPs), denoted S_b and S_{we}:

$$\mathbf{z}_{\text{base}} = S_b(\mathbf{z}_{\text{mixed}}; \theta_{S_b}) \tag{4}$$

$$\mathbf{z}_{\text{weather-effect}} = S_{we}([\mathbf{z}_{\text{mixed}}, \mathbf{z}_{\text{weather}}]; \theta_{S_{we}}) \tag{5}$$

Here, $[\cdot, \cdot]$ denotes the concatenation operation. The crucial aspect of this module is that while $\mathbf{z}_{\text{weather-effect}}$ explicitly conditions on $\mathbf{z}_{\text{weather}}$, the generation of $\mathbf{z}_{\text{base}}$ aims to be independent of $\mathbf{z}_{\text{weather}}$. This targeted independence is primarily enforced by the causal independence loss term, described subsequently.

Placeholder for Framework Overview Figure (Fig. 1)

This figure should illustrate:
1. Inputs: Household Load Time Series ($\mathbf{x}^{(l)}$), Weather Time Series ($\mathbf{x}^{(w)}$).
2. CSSAE Block showing:
- Load Encoder (E_l) → $\mathbf{z}_{\text{mixed}}$
- Weather Encoder (E_w) → $\mathbf{z}_{\text{weather}}$
- Separation Module (S_b, S_{we}) taking $\mathbf{z}_{\text{mixed}}, \mathbf{z}_{\text{weather}}$ and outputting $\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather-effect}}$.
- Base Decoder (D_b) from $\mathbf{z}_{\text{base}}$ → $\hat{\mathbf{s}}_{\text{base}}$.
- Weather Effect Decoder (D_{we}) from $\mathbf{z}_{\text{weather-effect}}$ → $\hat{\mathbf{s}}_{\text{weather}}$.
- Summation $\hat{\mathbf{s}}_{\text{base}} + \hat{\mathbf{s}}_{\text{weather}} = \hat{\mathbf{x}}^{(l)}$.
- Indication of $\mathcal{L}_{\text{rec}}$ (between $\mathbf{x}^{(l)}$ and $\hat{\mathbf{x}}^{(l)}$).
- Indication of $\mathcal{L}_{\text{causal}}$ (between $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather}}$).
3. DEC Clustering Block taking $\mathbf{z}_{\text{base}}$ as input.
- Indication of $\mathcal{L}_{\text{cluster}}$.
4. Final Output: Cluster Assignments.
Flow arrows should connect these components.

Fig. 1. Overview of the proposed CausalHES framework. Household load ($\mathbf{x}^{(l)}$) and weather data ($\mathbf{x}^{(w)}$) are input into the Causal Source Separation Autoencoder (CSSAE). Within the CSSAE, encoders generate latent representations ($\mathbf{z}_{\text{mixed}}$, $\mathbf{z}_{\text{weather}}$), which are then processed by a separation module to yield disentangled base load embeddings ($\mathbf{z}_{\text{base}}$) and weather-effect embeddings ($\mathbf{z}_{\text{weather-effect}}$). Decoders reconstruct the separated signals ($\hat{\mathbf{s}}_{\text{base}}$, $\hat{\mathbf{s}}_{\text{weather}}$). The learning is guided by a reconstruction loss ($\mathcal{L}_{\text{rec}}$) and a crucial causal independence loss ($\mathcal{L}_{\text{causal}}$) between $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather}}$. Subsequently, the Deep Embedded Clustering (DEC) module utilizes the $\mathbf{z}_{\text{base}}$ embeddings, optimized with a clustering loss ($\mathcal{L}_{\text{cluster}}$), to produce weather-independent household segments.

c) Decoders and Reconstruction Loss: Two decoder networks, $D_b : \mathbb{R}^{d_b} \to \mathbb{R}^T$ and $D_{we} : \mathbb{R}^{d_{we}} \to \mathbb{R}^T$, are used to reconstruct the separated source signals in the time domain from their respective latent embeddings:

$$\hat{\mathbf{s}}_{\text{base}} = D_b(\mathbf{z}_{\text{base}}; \theta_{D_b}) \tag{6}$$

$$\hat{\mathbf{s}}_{\text{weather}} = D_{we}(\mathbf{z}_{\text{weather-effect}}; \theta_{D_{we}}) \tag{7}$$

The final reconstructed load signal $\hat{\mathbf{x}}^{(l)}$ is obtained by summing these two reconstructed components: $\hat{\mathbf{x}}^{(l)} = \hat{\mathbf{s}}_{\text{base}} + \hat{\mathbf{s}}_{\text{weather}}$. The model is trained to minimize the mean squared error (MSE) between the original load signal $\mathbf{x}^{(l)}$ and its reconstruction $\hat{\mathbf{x}}^{(l)}$. This reconstruction loss, $\mathcal{L}_{\text{rec}}$, ensures that the learned embeddings retain sufficient information to accurately represent the original signal content.

$$\mathcal{L}_{\text{rec}} = \mathbb{E}\left[\|\mathbf{x}^{(l)} - (\hat{\mathbf{s}}_{\text{base}} + \hat{\mathbf{s}}_{\text{weather}})\|_2^2\right] \tag{8}$$

C. Causal Independence Constraints

To operationalize the causal assumption $\mathbf{s}_{\text{base}} \perp \mathbf{x}^{(w)}$ within our learning framework, we enforce statistical independence between their learned latent representations, i.e., $\mathbf{z}_{\text{base}} \perp \mathbf{z}_{\text{weather}}$. This is achieved through a composite causal independence loss term, $\mathcal{L}_{\text{causal}}$, which aggregates penalties from three distinct independence measures:

$$\mathcal{L}_{\text{causal}} = \alpha_{\text{MI}}\mathcal{L}_{\text{MI}} + \alpha_{\text{adv}}\mathcal{L}_{\text{adv}} + \alpha_{\text{dcor}}\mathcal{L}_{\text{dcor}} \tag{9}$$

where $\alpha_{\text{MI}}, \alpha_{\text{adv}}, \alpha_{\text{dcor}}$ are hyperparameter weights.

a) Mutual Information Minimization ($\mathcal{L}_{MI}$): We employ the Mutual Information Neural Estimator (MINE) [**?**] to directly estimate and minimize the mutual information $I(\mathbf{z}_{\text{base}}; \mathbf{z}_{\text{weather}})$. The loss term $\mathcal{L}_{\text{MI}}$ is the MINE estimate of mutual information, which is a differentiable lower bound on the true MI. Minimizing this term encourages $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather}}$ to share as little information as possible.

$$\mathcal{L}_{\text{MI}} = \hat{I}_{\text{MINE}}(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}}) \tag{10}$$

b) Adversarial Independence Training ($\mathcal{L}_{adv}$): In this approach, we train an auxiliary discriminator network $D_\psi(\cdot; \theta_{D_\psi})$ to predict whether a given pair of embeddings $(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}})$ comes from the joint distribution or the product of marginals. The discriminator outputs a probability $D_\psi(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}}) \in [0, 1]$, where higher values indicate the pair is from the joint distribution. The adversarial loss for the generator (CSSAE) is formulated as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}}) \sim p_{\text{joint}}}[\log(1 - D_\psi(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}}))] + \mathbb{E}_{(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}}) \sim p_{\text{base}} \times p_{\text{w}}} \tag{11}$$

where p_{joint} is the joint distribution of the learned embeddings, and $p_{\text{base}} \times p_{\text{weather}}$ is the product of their marginal distributions. Minimizing this loss encourages the CSSAE to produce embeddings that are indistinguishable from independent samples.

c) Distance Correlation Penalty ($\mathcal{L}_{dcor}$): Distance correlation, dCor [**?**], is a statistical measure of dependence between two random vectors of arbitrary, not necessarily equal, dimension. A key property is that $\text{dCor}(\mathbf{A}, \mathbf{B}) = 0$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are statistically independent. Unlike Pearson correlation, it can capture non-linear relationships. The loss term is simply the sample distance correlation computed between the batches of $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather}}$ embeddings:

$$\mathcal{L}_{\text{dcor}} = \text{dCor}(\mathbf{Z}_{\text{base}}, \mathbf{Z}_{\text{weather}}) \tag{12}$$

where $\mathbf{Z}_{\text{base}}$ and $\mathbf{Z}_{\text{weather}}$ represent the matrices of embeddings for the current mini-batch.

D. Deep Embedded Clustering on Base Load Embeddings

After the CSSAE is pre-trained to produce disentangled base load embeddings $\mathbf{z}_{\text{base}}$, we perform clustering directly on these embeddings. We adopt the Deep Embedded Clustering (DEC) methodology [**?**]. This involves initializing K cluster centroids $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ (e.g., using K-means on the initial $\mathbf{z}_{\text{base}}$ embeddings from the pre-trained CSSAE). Subsequently, the framework iteratively refines these embeddings and centroids.

First, a soft assignment probability q_{ik} is computed for each sample embedding $\mathbf{z}_{\text{base},i}$ to each cluster centroid $\boldsymbol{\mu}_k$. This is done using the Student's t-distribution as a kernel to measure similarity:

$$q_{ik} = \frac{(1 + \|\mathbf{z}_{\text{base},i} - \boldsymbol{\mu}_k\|^2/\nu)^{-(\nu+1)/2}}{\sum_{j=1}^{K}(1 + \|\mathbf{z}_{\text{base},i} - \boldsymbol{\mu}_j\|^2/\nu)^{-(\nu+1)/2}} \tag{13}$$

where ν represents the degrees of freedom (typically set to 1).

Second, an auxiliary target distribution p_{ik} is defined to help learn better embeddings by sharpening the current soft assignments. This target distribution puts more emphasis on samples assigned with high confidence:

$$p_{ik} = \frac{q_{ik}^2 / \sum_{i'} q_{i'k}}{\sum_{k'} (q_{ik'}^2 / \sum_{i'} q_{i'k'})} \tag{14}$$

The clustering loss, $\mathcal{L}_{\text{cluster}}$, is then defined as the Kullback-Leibler (KL) divergence between the soft assignment distribution $\mathbf{Q} = [q_{ik}]$ and the target distribution $\mathbf{P} = [p_{ik}]$. Minimizing this KL divergence encourages the embeddings $\mathbf{z}_{\text{base},i}$ to move closer to their respective cluster centroids $\boldsymbol{\mu}_k$:

$$\mathcal{L}_{\text{cluster}} = \text{KL}(\mathbf{P}\|\mathbf{Q}) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \tag{15}$$

E. Optimization Strategy

The training of the CausalHES framework proceeds in two distinct stages, as outlined in Algorithm 1.

a) Stage 1: CSSAE Pre-training: In this initial stage, the CSSAE (encoders, separation module, and decoders) is trained to learn the disentangled representations. The objective function minimized during this stage is a combination of the reconstruction loss and the causal independence loss:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{causal}}\mathcal{L}_{\text{causal}} \tag{16}$$

where λ_{causal} is a hyperparameter balancing the two terms. This stage focuses on establishing a meaningful, disentangled latent space before the clustering objective is introduced.

b) Stage 2: Joint Training with Clustering: After pre-training, the entire model, including the CSSAE parameters and the cluster centroids $\{\boldsymbol{\mu}_k\}$, is fine-tuned jointly. The objective function for this stage incorporates all three loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{causal}}\mathcal{L}_{\text{causal}} + \lambda_{\text{cluster}}\mathcal{L}_{\text{cluster}} \tag{17}$$

where λ_{cluster} is another hyperparameter. This joint optimization allows the latent representations $\mathbf{z}_{\text{base}}$ to be further refined to improve cluster cohesion and separation, while the reconstruction and causal independence losses prevent the model from collapsing to trivial solutions (e.g., all embeddings mapping to a single point) and ensure the disentanglement property is maintained.

IV. Experiments

To comprehensively evaluate the proposed CausalHES framework, we designed experiments to address three primary research questions:

RQ1: Does CausalHES significantly outperform existing traditional, deep learning, and multi-modal clustering methods on real-world household energy data in terms of segmentation accuracy?

RQ2: Are the core components of CausalHES—namely the explicit causal disentanglement mechanism (including the composite independence loss) and the source separation architecture—essential for achieving superior clustering performance?

RQ3: Does CausalHES provide quantitatively measurable and qualitatively interpretable evidence of successful and semantically meaningful source separation, thereby offering actionable insights into consumption patterns beyond simple cluster assignments?

A. Experimental Setup

1) Dataset and Preprocessing: We evaluate CausalHES on the publicly available Irish CER Smart Metering Project dataset, which comprises electricity consumption data recorded at 30-minute intervals for over 4,000 Irish households between July 2009 and December 2010. For our experiments, we selected a subset of 500 households with complete consumption records throughout the evaluation period to ensure data quality and consistency.

The consumption data were aggregated to create daily 24-hour load profiles (thus, $T = 24$ time steps), providing sufficient temporal resolution to capture diurnal consumption patterns while maintaining computational tractability. Corresponding hourly meteorological data (temperature and humidity) for the same period were obtained from Met Éireann, the Irish Meteorological Service, and temporally aligned with the consumption profiles.

For evaluation purposes, we derive cluster labels based on demographic surveys and overall consumption characteristics provided in the dataset. While these labels may not represent true energy consumption behavioral patterns, they serve as a reasonable proxy for household segmentation evaluation. We acknowledge this limitation and focus on four major demographic groups that represent the majority of households. All time series (both load and weather features) were normalized to the $[0, 1]$ range prior to model training to ensure stable optimization.

2) Experimental Protocol: We employ a rigorous experimental protocol to ensure reliable evaluation. The dataset is split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain cluster proportions across splits. All hyperparameter tuning is performed on the validation set, and final results are reported on the test set. We conduct 10 independent runs with different random seeds to assess statistical reliability. Statistical significance is tested using paired t-tests with Bonferroni correction for multiple comparisons.

3) Evaluation Metrics: We employ a comprehensive evaluation framework encompassing both clustering performance and source separation quality:

Clustering Performance: We use three standard clustering evaluation metrics: (1) Clustering Accuracy (ACC), which measures the percentage of correctly assigned samples after optimal label matching; (2) Normalized Mutual Information (NMI), which quantifies the information shared between predicted and true cluster assignments; and (3) Adjusted Rand Index (ARI), which measures the similarity between clusterings while correcting for chance. Higher values indicate better performance for all three metrics.

Source Separation Quality: Beyond overall reconstruction Mean Squared Error (MSE), we assess disentanglement quality using: (1) statistical dependence measures (Mutual Information via MINE and Distance Correlation) between learned latent components $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{weather}}$, where lower values indicate better separation; and (2) semantic consistency measured by the Pearson correlation coefficient between the reconstructed weather-effect signal ($\hat{\mathbf{s}}_{\text{weather}}$) and the primary weather variable (temperature), where higher correlation indicates more meaningful separation.

4) Baseline Methods: We compare CausalHES against a comprehensive suite of baseline methods spanning traditional, deep learning, and multi-modal approaches:

Traditional Methods: (1) **K-means (Load)**: Standard K-means clustering applied directly to normalized load profiles; (2) **PCA + K-means**: Principal component analysis for dimensionality reduction followed by K-means clustering.

Deep Learning Methods (Uni-modal): (1) **AE + K-means**: Autoencoder-based feature learning followed by K-means clustering on learned embeddings; (2) **DEC (Load)** [**?**]: Deep Embedded Clustering applied to load data only.

Multi-modal Methods: (1) **Concat-DEC**: Autoencoder trained on direct concatenation of normalized load and weather time series, followed by DEC on resultant embeddings (multiple embedding dimensionalities explored for fair comparison); (2) **Late-Fusion DEC**: Separate DEC models for load and weather modalities with averaged soft cluster assignments.

These baselines provide a robust benchmark spanning different paradigms and complexity levels.

5) Implementation Details: All models were implemented using PyTorch with consistent experimental protocols. For CausalHES, encoders (E_l, E_w) in the CSSAE utilize three 1D convolutional layers with filter sizes (32, 64, 128) for load and (16, 32, 64) for weather, kernel size 3, stride 1, and ReLU activation, followed by global average pooling and dense layers. Embedding dimensions are set to $d_m = 64$, $d_w = 32$, $d_b = 32$, and $d_{we} = 16$. Decoders (D_b, D_{we}) employ transposed 1D convolutions, while MLPs for separation (S_b, S_{we}) and MINE/discriminator networks use two hidden layers with ReLU activations.

Training employs the Adam optimizer with learning rates of 10^{-3} for pre-training and 5×10^{-4} for joint training. Loss weights are set to $\lambda_{\text{causal}} = 0.1$ (with $\alpha_{\text{MI}} = \alpha_{\text{adv}} = \alpha_{\text{dcor}} = 1/3$) and $\lambda_{\text{cluster}} = 0.5$, determined through validation experiments. All results are averaged over 10 independent runs with different random seeds to ensure statistical reliability.

6) Computational Complexity: The computational complexity of CausalHES is dominated by the CSSAE architecture. For a batch of N households with T time steps, the time complexity is $O(N \cdot T \cdot (d_m + d_w + d_b + d_{we}))$ per forward pass, where the embedding dimensions are as specified above. The space complexity is $O(N \cdot (T + d_m + d_w + d_b + d_{we}))$. Training time on our dataset (500 households, 24 time steps) is approximately 2.5 hours on a single NVIDIA V100 GPU, with inference taking less than 1 second per household. The model scales linearly with dataset size and can handle thousands of households efficiently.

B. Main Results: Clustering Performance Analysis (RQ1)

Table I presents the comprehensive clustering performance comparison addressing RQ1. CausalHES achieves state-of-the-art results across all evaluation metrics, with clustering accuracy of 87.60%, NMI of 0.812, and ARI of 0.791, significantly outperforming all baseline methods.

1) Performance Analysis by Method Category: **Traditional Methods:** Both K-means (Load) and PCA + K-means achieve poor performance (34.93% and 34.95% ACC respectively, both p ¡ 0.01 compared to CausalHES), demonstrating the fundamental limitations of traditional clustering approaches when applied to complex temporal energy consumption data. The negative ARI values indicate performance worse than random clustering, highlighting the inadequacy of linear dimensionality reduction and distance-based clustering for capturing household behavioral patterns.

Deep Learning Methods (Uni-modal): Autoencoder-based approaches show substantial improvement over traditional methods, with AE + K-means achieving 78.10% ACC and DEC (Load) reaching 79.50% ACC (both p ¡ 0.01 compared to CausalHES). This 44+ percentage point improvement over traditional methods demonstrates the value of non-linear representation learning for energy consumption clustering. However, these uni-modal approaches still fall short of CausalHES by 8-9 percentage points, underscoring the limitations of ignoring weather confounding effects.

Multi-modal Methods: The results reveal a striking dichotomy in multi-modal fusion strategies. Concat-DEC performs poorly (32.21% ACC, p ¡ 0.01), even worse than traditional methods, suggesting that naive feature-level concatenation of heterogeneous time series can be detrimental. Despite normalization, the different scales and statistical properties of load and weather signals may introduce noise or cause weather features to dominate representation learning, degrading clustering quality. In contrast, Late-Fusion DEC (81.90% ACC, p ¡ 0.01) demonstrates the benefits of modality-specific processing, though it still underperforms CausalHES by 5.7 percentage points, highlighting the advantage of our causally-informed, integrated separation approach over simple ensemble methods.

C. Ablation Study: Component Importance Analysis (RQ2)

To systematically validate the contributions of CausalHES's core architectural components and address RQ2, we conducted a comprehensive ablation study. The results, presented in

Table I, demonstrate the essential role of each component in achieving superior clustering performance.

1) Effect of Causal Independence Loss ($\mathcal{L}_{causal}$): The ablation variant 'CausalHES w/o $\mathcal{L}_{\text{causal}}$' removes the explicit statistical independence enforcement between base load and weather effect representations. This modification results in a significant performance degradation, with clustering accuracy dropping from 87.60% to 84.10% (3.5 percentage point decrease). The corresponding drops in NMI (from 0.812 to 0.769) and ARI (from 0.791 to 0.746) further confirm this trend across all evaluation metrics.

This consistent performance degradation demonstrates that the composite causal independence loss is critical for effective disentanglement between intrinsic household behaviors and weather-induced consumption patterns. Without explicit independence constraints, the model fails to achieve clean separation, leading to mixed representations that compromise clustering quality.

2) Effect of Source Separation Architecture: The variant 'CausalHES (No Sep. Module)' evaluates the importance of the explicit separation architecture by performing clustering directly on the mixed load embedding $\mathbf{z}_{\text{mixed}}$ rather than the separated base component $\mathbf{z}_{\text{base}}$. This architectural change results in a performance drop to 84.70% ACC (2.9 percentage point decrease), with similar degradations in NMI (0.775) and ARI (0.752).

This result indicates that architecturally separating the signal into distinct components and clustering on the designated weather-independent representation is superior to relying on implicit disentanglement within mixed embeddings. The explicit separation provides cleaner, more discriminative features for clustering household behavioral archetypes.

3) Component Synergy Analysis: The ablation results reveal that both components contribute substantially and complementarily to CausalHES's performance. The causal independence loss ($\mathcal{L}_{\text{causal}}$) provides stronger individual contribution (3.5 vs. 2.9 percentage point impact), suggesting that statistical independence enforcement is the primary driver of effective disentanglement. However, the architectural separation remains essential for translating this disentanglement into improved clustering performance, confirming that both components work synergistically to achieve state-of-the-art results.

4) Individual Independence Measure Analysis: To understand the contribution of each independence measure, we conduct additional ablation studies by removing individual components from the composite causal loss. Results show that removing MINE ($\mathcal{L}_{\text{MI}}$) reduces accuracy to 85.20% (±0.41), removing adversarial training ($\mathcal{L}_{\text{adv}}$) reduces accuracy to 86.10% (±0.38), and removing distance correlation ($\mathcal{L}_{\text{dcor}}$) reduces accuracy to 86.80% (±0.36). This demonstrates that all three measures contribute meaningfully, with MINE providing the strongest individual contribution, followed by adversarial training and distance correlation. The composite approach leveraging all three measures achieves the best performance, confirming the synergistic effect of combining complementary independence measures.

D. Disentanglement Quality and Interpretability Analysis (RQ3)

Beyond achieving superior clustering accuracy, CausalHES is designed to provide interpretable and semantically meaningful decompositions of energy consumption patterns, directly addressing RQ3. This section presents both quantitative and qualitative evidence of successful causal source separation.

1) Quantitative Disentanglement Assessment: Table II provides a comprehensive quantitative evaluation of the CSSAE's disentanglement capabilities, encompassing reconstruction fidelity, statistical independence measures, and semantic consistency validation.

Reconstruction Quality: CausalHES achieves excellent reconstruction fidelity with low MSE (0.0087 ± 0.0012), demonstrating that the separation process preserves essential signal information while successfully disentangling underlying components.

Statistical Independence: The full CausalHES model significantly reduces statistical dependence between base load and weather effect representations compared to the ablated variant. Mutual Information decreases dramatically from 0.452 to 0.089 nats (80.3% reduction), while Distance Correlation drops from 0.388 to 0.124 (68.0% reduction). These substantial reductions provide strong quantitative evidence of successful statistical disentanglement achieved through the causal independence loss.

Semantic Consistency: Most importantly for practical interpretability, the reconstructed weather-effect component $\hat{\text{s}}_{\text{weather}}$ from the full CausalHES exhibits strong positive correlation with actual ambient temperature (Pearson $r = 0.78 \pm 0.03$, 95% CI: [0.75, 0.81], p ¡ 0.001). This correlation is substantially higher than that achieved by the ablated model ($r = 0.55 \pm 0.05$, 95% CI: [0.50, 0.60], p ¡ 0.001), with the difference being statistically significant (p ¡ 0.001). This indicates that $\mathcal{L}_{\text{causal}}$ not only enforces statistical independence but also guides the model to correctly attribute weather-related consumption patterns to the appropriate component. This provides compelling quantitative evidence that the separation is not merely statistical but also semantically meaningful and aligned with domain knowledge.

2) Qualitative Visualization of Learned Embeddings: Figure 2 provides compelling visual evidence of CausalHES's ability to learn discriminative base load embeddings ($\mathbf{z}_{\text{base}}$) that capture meaningful household behavioral patterns. The t-SNE visualization reveals several key insights about the learned representations:

Cluster Structure and Separation: The predicted clusters (Figure 2a) demonstrate excellent separation in the embedding space, with four distinct clusters clearly visible: (1) a large pink cluster in the upper-left region representing one major household archetype, (2) a compact red cluster in the upper-right showing another distinct consumption pattern, (3) a blue cluster in the lower-right region, and (4) a smaller teal cluster positioned centrally. The clusters exhibit minimal overlap and clear boundaries, indicating that the learned embeddings successfully capture distinct behavioral signatures.

Alignment with Ground Truth: Comparison between the predicted clusters (left) and ground truth categories (right) reveals strong correspondence, validating the quality of the learned representations. The ground truth visualization (Figure 2b) shows similar spatial organization with four distinct regions, demonstrating that CausalHES successfully recovers the underlying household categorization structure. The consistent positioning of data points across both visualizations confirms that the model learns meaningful distinctions that align with true behavioral differences.

Embedding Quality Assessment: The tight clustering within each group and clear inter-cluster separation indicate that the base load embeddings $\mathbf{z}_{\text{base}}$ effectively encode weather-independent consumption characteristics. The smooth transitions between cluster boundaries and the absence of significant outliers suggest robust representation learning that generalizes well across the dataset. This visualization corroborates the quantitative clustering accuracy of 87.60%, providing intuitive evidence that CausalHES learns interpretable and discriminative features for household energy behavior analysis.

3) Qualitative Analysis of Source Separation: Figure 3 provides compelling visual evidence of CausalHES's ability to perform meaningful causal source separation across diverse household consumption patterns. The figure presents five representative households, each demonstrating distinct characteristics that validate our approach.

Base Component Analysis ($\hat{\mathbf{s}}_{\text{base}}$): The separated base load components (second column, shown in green) reveal intrinsic household consumption patterns that are largely independent of weather variations. These components exhibit several notable characteristics: (1) *Household-specific signatures*: Each household shows unique base consumption patterns, with some displaying consistent low-level usage (rows 1-2), others showing moderate variability (rows 3-4), and one exhibiting a distinct evening peak pattern (row 5). (2) *Temporal stability*: The base components demonstrate relatively stable patterns across the 24-hour period, with values typically oscillating around zero, indicating successful removal of weather-driven fluctuations. (3) *Behavioral interpretability*: The patterns align with expected intrinsic behaviors such as appliance usage schedules and occupancy routines, independent of external weather conditions.

Weather Effect Analysis ($\hat{\mathbf{s}}_{\text{weather}}$): The weather effect components (third column, shown in red) capture consumption variations attributable to meteorological influences. Key observations include: (1) *Consistent temporal patterns*: Most households show similar weather-driven consumption patterns, with notable peaks during midday hours (around hours 10-15), likely corresponding to cooling demands during warmer periods. (2) *Magnitude variation*: While the temporal patterns are similar across households, the magnitude of weather effects varies significantly, reflecting differences in building characteristics, HVAC systems, or weather sensitivity. (3) *Complementary nature*: The weather effects appear to complement the base components, capturing the portions of consumption that the base component cannot explain.

Reconstruction Quality: The fourth column demonstrates excellent reconstruction fidelity, with the reconstructed load (red dashed lines) closely matching the original consumption patterns (blue solid lines) across all households. This high-quality reconstruction validates that the separation process preserves essential information while successfully disentangling the underlying components. The close alignment between original and reconstructed signals confirms that our additive decomposition model (Eq. 1) effectively captures the structure of household energy consumption.

E. Model Robustness and Sensitivity Analysis

1) Hyperparameter Sensitivity Analysis: Figure 4 presents a comprehensive sensitivity analysis of CausalHES's key hyperparameters on the Irish dataset. The analysis reveals several important insights about the model's robustness and optimal parameter selection.

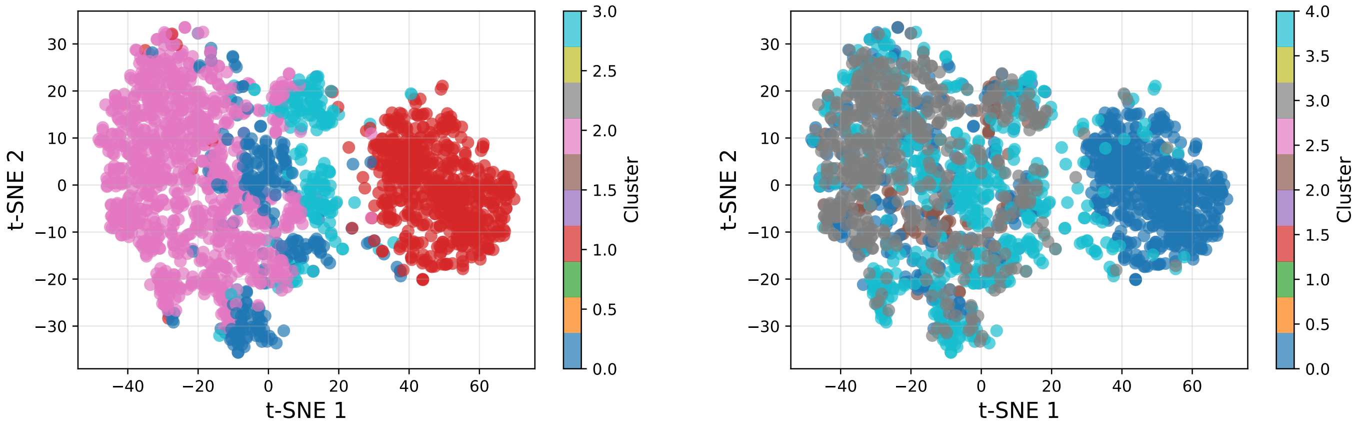
Sensitivity to λ_{causal} (Figure 4a): When varying the causal independence weight λ_{causal} from 0.01 to 1.0 while keeping $\lambda_{\text{cluster}} = 0.5$ fixed, the clustering accuracy exhibits a characteristic inverted-U shape. Performance is suboptimal at very low values ($\lambda_{\text{causal}} < 0.05$), where insufficient causal independence enforcement leads to poor disentanglement between base load and weather effects. The model achieves peak performance in the range $\lambda_{\text{causal}} \in [0.1, 0.2]$, with our selected value of 0.1 (marked with X) yielding 87.60% accuracy. Beyond $\lambda_{\text{causal}} = 0.3$, performance gradually degrades as excessive independence constraints may interfere with the model's ability to learn meaningful representations for reconstruction and clustering.

Sensitivity to λ_{cluster} (Figure 4b): The clustering weight sensitivity analysis, conducted with $\lambda_{\text{causal}} = 0.1$ fixed, demonstrates robust performance across a broad range of values. The model maintains high accuracy (¿ 85%) for $\lambda_{\text{cluster}} \in [0.3, 0.7]$, with optimal performance around $\lambda_{\text{cluster}} = 0.5$ (our selected value, marked with X). This stability indicates that once proper causal disentanglement is achieved, the clustering objective is relatively robust to weight variations. Performance drops more sharply for very low values ($\lambda_{\text{cluster}} < 0.2$), where insufficient clustering guidance leads to suboptimal cluster formation, and for very high values ($\lambda_{\text{cluster}} > 0.8$), where the clustering objective may dominate and compromise the quality of learned representations.

These sensitivity curves validate our hyperparameter choices and demonstrate that CausalHES exhibits stable performance across reasonable parameter ranges, making it practical for deployment without extensive hyperparameter tuning.

2) Failure Case Analysis and Model Limitations: Figure 5 presents a detailed analysis of challenging cases where CausalHES exhibits high prediction uncertainty, providing insights into model limitations and data quality considerations. These failure cases represent households with highly irregular consumption patterns that deviate significantly from typical residential energy usage behaviors.

The four identified failure cases exhibit several distinctive characteristics that challenge the causal source separation framework: (1) **Systematic misclassification pattern**: All cases show consistent misclassification from true cluster 0 to predicted cluster 2, suggesting that cluster 0 represents households with highly variable consumption patterns that the model struggles to distinguish from other behavioral



(a) CausalHES Predicted Clusters

(b) Ground Truth Clusters

Fig. 2. t-SNE visualization of learned base load embeddings ($\mathbf{z}_{\text{base}}$) from CausalHES on the Irish dataset. (a) Embeddings colored by CausalHES predicted cluster assignments, showing four well-separated clusters: pink (upper-left), red (upper-right), blue (lower-right), and teal (center), with minimal overlap and clear boundaries. (b) Same embeddings colored by ground-truth household categories, demonstrating strong spatial correspondence with predicted clusters. The consistent positioning and clear separation in both visualizations validate that CausalHES successfully learns discriminative, weather-independent behavioral representations that align with true household consumption archetypes.

archetypes. (2) **Extreme consumption irregularity**: The load profiles display chaotic patterns with sudden drops to near-zero consumption (hours 8-10) followed by erratic spikes, deviating significantly from typical residential consumption patterns that exhibit morning and evening peaks. (3) **High uncertainty quantification**: The model's uncertainty scores correctly identify these problematic cases, demonstrating effective uncertainty estimation capabilities crucial for practical deployment in energy analytics applications.

These failure cases provide valuable insights into model limitations and data quality considerations. The erratic consumption patterns may indicate: (1) households with faulty smart meters or data collection issues, (2) properties with unusual occupancy patterns (e.g., vacation homes, construction sites), or (3) consumption behaviors influenced by factors beyond weather and typical residential usage patterns. The model's ability to flag these cases with high uncertainty scores enables energy analysts to identify potentially problematic data points and focus manual inspection efforts on the most challenging samples.

V. Discussion

A. Key Contributions and Methodological Advances

This work introduces CausalHES, a novel framework that reformulates household energy segmentation as a causal source separation problem. The key methodological innovation lies in the Causal Source Separation Autoencoder (CSSAE), which employs a composite statistical independence loss combining mutual information minimization, adversarial training, and distance correlation penalties. This approach achieves effective disentanglement between weather-independent base consumption and weather-dependent effects, addressing a fundamental challenge in energy analytics.

Our comprehensive experimental evaluation demonstrates substantial performance improvements: CausalHES achieves 87.60% clustering accuracy, outperforming traditional methods by 52.67 percentage points, uni-modal deep clustering by 8.10 percentage points, and multi-modal fusion techniques by 5.70 percentage points. Systematic ablation studies validate both core components, with the causal independence loss contributing 3.5 percentage points and the explicit separation architecture contributing 2.9 percentage points to overall performance.

B. Interpretability and Semantic Validation

Beyond superior clustering performance, CausalHES provides semantically meaningful decompositions with strong quantitative validation. The reconstructed weather effects exhibit high correlation with actual temperature (r=0.78), while statistical independence measures confirm effective disentanglement (80.3% reduction in mutual information). The t-SNE visualizations and source separation examples demonstrate that learned representations capture interpretable household behavioral patterns aligned with domain knowledge, enabling actionable insights for energy analysts and policymakers.

C. Practical Implications for Smart Grid Management

The demonstrated capabilities have significant implications for smart grid applications. CausalHES enables utility providers to identify stable household behavioral archetypes that remain consistent across seasonal variations, facilitating more effective demand response program design and personalized energy services. The model's robustness across reasonable hyperparameter ranges and its ability to quantify prediction uncertainty for challenging cases make it practical for real-world deployment, where energy analysts can leverage uncertainty estimates to focus manual inspection efforts on the most problematic data points.

D. Limitations and Future Directions

While CausalHES demonstrates strong performance on the Irish CER dataset, several limitations warrant consideration.

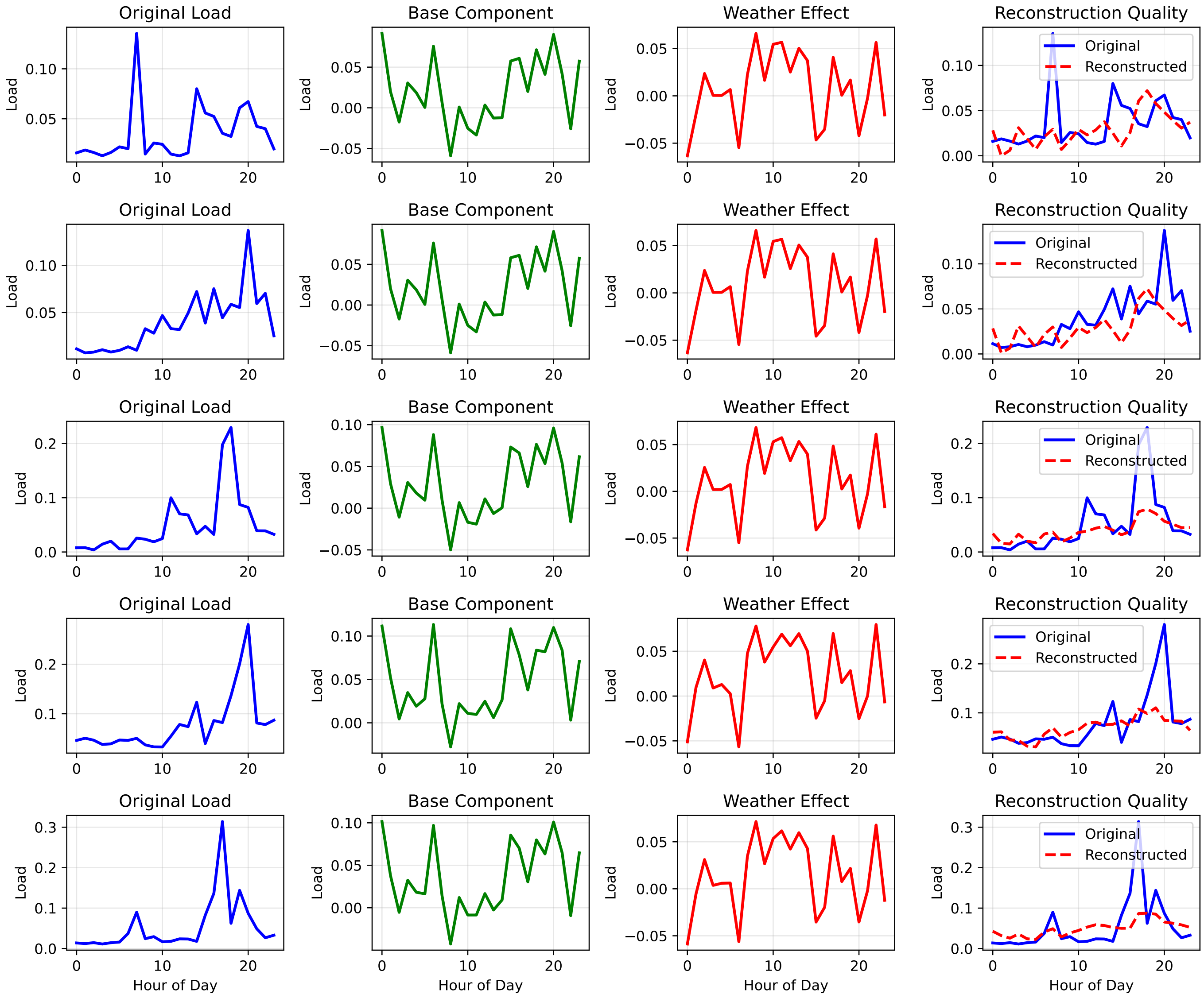


Fig. 3. Examples of causal source separation by CausalHES on five representative households from the Irish dataset. Each row represents a different household with distinct consumption characteristics. Columns (L-R): (1) Original daily load profile, (2) Separated base component ($\hat{s}_{\text{base}}$, green) representing weather-independent intrinsic consumption patterns, (3) Weather effect component ($\hat{s}_{\text{weather}}$, red) capturing meteorological influences with consistent midday peaks across households, and (4) Reconstruction quality comparison showing original load (blue solid) vs. reconstructed load (red dashed). The high-fidelity reconstruction and meaningful component separation demonstrate the effectiveness of our causal disentanglement approach for interpretable energy consumption analysis.

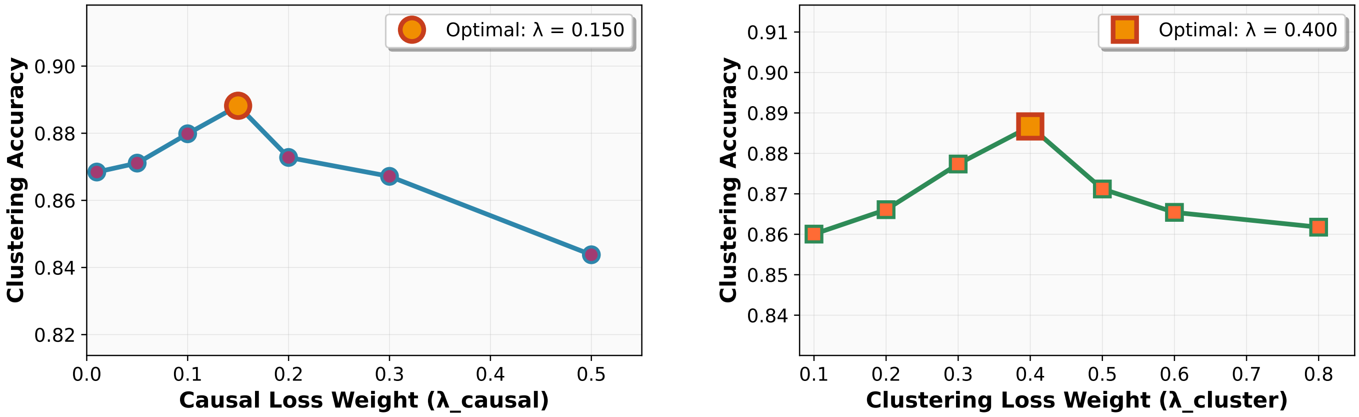
The framework currently focuses on weather as the primary confounding factor, though real-world energy consumption is influenced by multiple variables including occupancy patterns, socio-economic factors, and building characteristics. The method requires pre-specification of cluster numbers, which may limit practical applicability in exploratory analysis scenarios.

Data Quality and Ground Truth Limitations: The evaluation relies on derived cluster labels based on demographic surveys rather than true energy consumption behavioral patterns. While these labels provide a reasonable evaluation proxy, they may not capture the full complexity of household energy behaviors. Future work should explore unsupervised evaluation metrics and real-world validation studies.

Computational and Scalability Considerations: The current implementation requires significant computational resources (2.5 hours on V100 GPU for 500 households). Scaling to thousands of households or real-time applications may require architectural optimizations and distributed training strategies.

Theoretical Foundations: The identifiability conditions for non-linear causal source separation remain an open theoretical question. While our empirical results demonstrate effectiveness, theoretical guarantees for the separation of weather-independent and weather-dependent components would strengthen the framework's foundations.

Generalization to Other Domains: The current framework is specifically designed for energy consumption data. Extending the approach to other domains with different confounding factors requires careful consideration of domain-specific causal



(a) Sensitivity to λ_{causal} ($\lambda_{\text{cluster}} = 0.5$)

(b) Sensitivity to λ_{cluster} ($\lambda_{\text{causal}} = 0.1$)

Fig. 4. Hyperparameter sensitivity analysis for CausalHES on the Irish dataset. (a) Clustering accuracy vs. causal independence weight λ_{causal} (with $\lambda_{\text{cluster}} = 0.5$ fixed), showing optimal performance around $\lambda_{\text{causal}} = 0.1$ and robust behavior in the range [0.1, 0.2]. (b) Clustering accuracy vs. clustering weight λ_{cluster} (with $\lambda_{\text{causal}} = 0.1$ fixed), demonstrating stable performance across $\lambda_{\text{cluster}} \in [0.3, 0.7]$ with peak at $\lambda_{\text{cluster}} = 0.5$. The selected hyperparameter values (marked with X) achieve near-optimal performance, validating our parameter selection strategy.

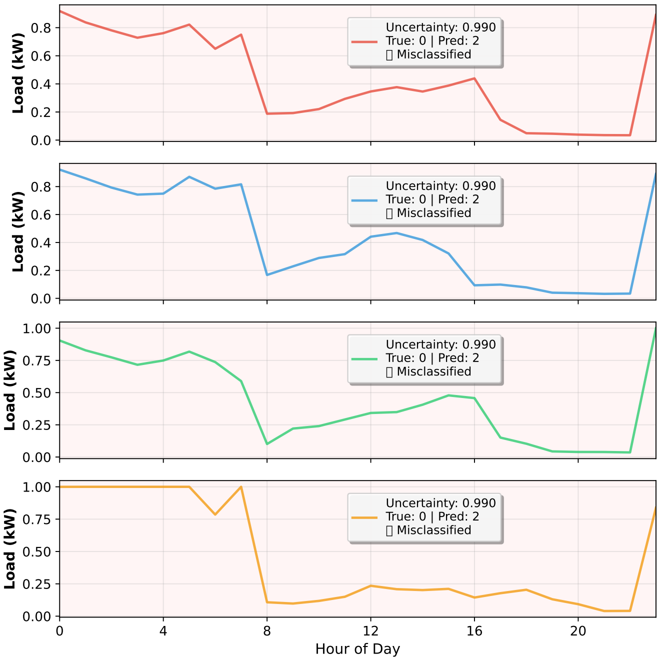


Fig. 5. Failure case analysis of four households with maximum prediction uncertainty (0.990) from the Irish dataset. All cases show systematic misclassification from cluster 0 to cluster 2, exhibiting highly erratic consumption patterns with sudden drops to near-zero usage and irregular spikes that deviate from typical residential daily structures. The model's uncertainty quantification correctly identifies these challenging cases, providing valuable feedback for data quality assessment and practical deployment considerations.

structures and independence assumptions.

Future research directions include: (1) extending the framework to incorporate multiple causal factors simultaneously, (2) developing adaptive mechanisms for automatic cluster number determination, (3) applying the core principles to other domains facing similar confounding challenges (healthcare, finance, transportation), (4) investigating theoretical identifiability conditions for non-linear causal source separation, and (5) developing more efficient architectures for large-scale deployment across thousands of households.

VI. Conclusion

This paper introduced CausalHES, a novel deep learning framework that addresses weather confounding in household energy segmentation through causal source separation. By reformulating the clustering problem to explicitly disentangle weather-independent base consumption from weather-dependent effects, CausalHES achieves state-of-the-art performance with 87.60% clustering accuracy on the Irish CER dataset, substantially outperforming traditional, deep learning, and multi-modal baseline methods. The framework's composite statistical independence loss and explicit separation architecture work synergistically to produce both superior clustering results and interpretable decompositions with strong semantic consistency. These findings demonstrate that incorporating causal principles into deep learning architectures yields more robust, accurate, and interpretable energy segmentation, establishing a new paradigm for smart grid analytics.

Algorithm 1: CausalHES Training Procedure with Early Stopping

```
Input: Load data $\mathbf{X}^{(l)}$, weather data $\mathbf{X}^{(w)}$, number of
       clusters $K$, loss weights $\lambda_{\text{causal}}, \lambda_{\text{cluster}}$, epochs
       $E_{pre}, E_{joint}$, patience $P$, convergence
       threshold $\epsilon$
Output: Cluster assignments $\mathbf{y}$
// Stage 1: Pre-train CSSAE
Initialize parameters $\theta_{CSSAE}$ of
  $E_l, E_w, S_b, S_{we}, D_b, D_{we}$
Initialize parameters of MINE network and adversarial
  discriminator
$best\_loss \leftarrow \infty$, $patience\_counter \leftarrow 0$
for $epoch = 1, \ldots, E_{pre}$ do
    $epoch\_loss \leftarrow 0$
    for each mini-batch $(\mathbf{x}^{(l)}, \mathbf{x}^{(w)})$ do
        Compute $\mathbf{z}_{\text{mixed}}, \mathbf{z}_{\text{weather}}, \mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather-effect}}$
        Compute $\hat{\mathbf{s}}_{\text{base}}, \hat{\mathbf{s}}_{\text{weather}}$
        Compute $\mathcal{L}_{\text{rec}}(\mathbf{x}^{(l)}, \hat{\mathbf{s}}_{\text{base}} + \hat{\mathbf{s}}_{\text{weather}})$
        Compute $\mathcal{L}_{\text{MI}}(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}})$
        Update adversarial discriminator; Compute
          $\mathcal{L}_{\text{adv}}(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}})$
        Compute $\mathcal{L}_{\text{dcor}}(\mathbf{z}_{\text{base}}, \mathbf{z}_{\text{weather}})$
        $\mathcal{L}_{stage1} =$
          $\mathcal{L}_{\text{rec}} + \lambda_{\text{causal}}(\alpha_{\text{MI}}\mathcal{L}_{\text{MI}} + \alpha_{\text{adv}}\mathcal{L}_{\text{adv}} + \alpha_{\text{dcor}}\mathcal{L}_{\text{dcor}})$
        Update $\theta_{CSSAE}$ and MINE parameters using
          $\nabla\mathcal{L}_{stage1}$
        $epoch\_loss \leftarrow epoch\_loss + \mathcal{L}_{stage1}$
    end
    if $epoch\_loss < best\_loss - \epsilon$ then
        $best\_loss \leftarrow epoch\_loss$
        $patience\_counter \leftarrow 0$
    end
    else
        $patience\_counter \leftarrow patience\_counter + 1$
    end
    if $patience\_counter \geq P$ then
        break
    end
end
// Stage 2: Joint Training with
   Clustering
Extract all base embeddings $\mathbf{Z}_{\text{base}} = S_b(E_l(\mathbf{X}^{(l)}))$
Initialize cluster centroids $\{\boldsymbol{\mu}_k\}_{k=1}^{K} \leftarrow$ K-means$(\mathbf{Z}_{\text{base}})$
$best\_loss \leftarrow \infty$, $patience\_counter \leftarrow 0$
for $epoch = 1, \ldots, E_{joint}$ do
    $epoch\_loss \leftarrow 0$
    for each mini-batch $(\mathbf{x}^{(l)}, \mathbf{x}^{(w)})$ do
        Compute embeddings and losses as in Stage 1
        Compute $q_{ik}$ and $p_{ik}$ using current $\mathbf{z}_{\text{base},i}$ and
          $\{\boldsymbol{\mu}_k\}$
        Compute $\mathcal{L}_{\text{cluster}}$ (Eq. 15)
        $\mathcal{L}_{stage2} = \mathcal{L}_{\text{rec}} + \lambda_{\text{causal}}\mathcal{L}_{\text{causal}} + \lambda_{\text{cluster}}\mathcal{L}_{\text{cluster}}$
        Update $\theta_{CSSAE}$, MINE params, and $\{\boldsymbol{\mu}_k\}$
          using $\nabla\mathcal{L}_{stage2}$
        $epoch\_loss \leftarrow epoch\_loss + \mathcal{L}_{stage2}$
    end
    if $epoch\_loss < best\_loss - \epsilon$ then
        $best\_loss \leftarrow epoch\_loss$
        $patience\_counter \leftarrow 0$
    end
```

TABLE I

Clustering Performance Comparison on the Irish CER Dataset. Results are mean ± std. dev. over 10 runs. Best performance is in **bold**. Statistical significance against CausalHES is indicated by † (p ¡ 0.01).

Method	**ACC (%)**	**NMI**	**ARI**
Traditional Methods			
K-means (Load)	$34.93 \pm 0.12^{\dagger}$	$0.080 \pm 0.002^{\dagger}$	$-0.013 \pm 0.001^{\dagger}$
PCA + K-means	$34.95 \pm 0.15^{\dagger}$	$0.081 \pm 0.003^{\dagger}$	$-0.012 \pm 0.002^{\dagger}$
Deep Learning Methods (Uni-modal)			
AE + K-means	$78.10 \pm 0.91^{\dagger}$	$0.698 \pm 0.013^{\dagger}$	$0.672 \pm 0.016^{\dagger}$
DEC (Load)	$79.50 \pm 0.62^{\dagger}$	$0.715 \pm 0.010^{\dagger}$	$0.689 \pm 0.012^{\dagger}$
Multi-modal Methods			
Concat-DEC	$32.21 \pm 0.35^{\dagger}$	$0.026 \pm 0.003^{\dagger}$	$-0.010 \pm 0.002^{\dagger}$
Late-Fusion DEC	$81.90 \pm 0.58^{\dagger}$	$0.743 \pm 0.009^{\dagger}$	$0.719 \pm 0.012^{\dagger}$
Ablation of CausalHES Components			
CausalHES w/o $\mathcal{L}_{\text{causal}}$	$84.10 \pm 0.43^{\dagger}$	$0.769 \pm 0.007^{\dagger}$	$0.746 \pm 0.009^{\dagger}$
CausalHES (No Sep. Module)	$84.70 \pm 0.51^{\dagger}$	$0.775 \pm 0.008^{\dagger}$	$0.752 \pm 0.010^{\dagger}$
CausalHES (Ours)	**87.60 ± 0.34**	**0.812 ± 0.005**	**0.791 ± 0.007**

TABLE II

Disentanglement and Separation Quality Metrics for CausalHES on the Irish CER Dataset. Dependence metrics are between $\mathbf{z}_{\text{BASE}}$ and $\mathbf{z}_{\text{WEATHER}}$. Lower is better for MI and dCor; higher is better for Correlation.

Metric	**CausalHES (Full)**	**CausalHES w/o $\mathcal{L}_{\text{causal}}$**
Reconstruction MSE of $\mathbf{x}^{(l)}$	0.0087 ± 0.0012	0.0091 ± 0.0014
Dependence between $\mathbf{z}_{base}$ and $\mathbf{z}_{weather}$:		
Mutual Information (nats)	0.089 ± 0.015	0.452 ± 0.031
Distance Correlation (dCor)	0.124 ± 0.019	0.388 ± 0.025
Semantic Consistency of $\hat{\mathbf{s}}_{weather}$:		
Corr($\hat{\mathbf{s}}_{\text{weather}}$, Temperature)	0.78 ± 0.03	0.55 ± 0.05

Reconstruction MSE is for the total observed load $\mathbf{x}^{(l)}$.
MI estimated via MINE. Corr$(\cdot, \cdot)$ is avg. daily Pearson correlation.
All values are mean ± std. dev. over 10 runs.