NILMBENCH2026: A Benchmark for Energy Disaggregation

Aayush Kuloor*
Indian Institute of Technology, Gandhinagar
Gandhinagar, Gujarat, India
aayush.kuloor@iitgn.ac.in

Anurag Singh*
IIT Gandhinagar
Gandhinagar, Gujarat, India
anurag.s@iitgn.ac.in

Harsh Dhru*
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
harsh.dhru@iitgn.ac.in

Nipun Batra†
IIT Gandhinagar
Gandhinagar, Gujarat, India
nipun.batra@iitgn.ac.in

Abstract

Non-Intrusive Load Monitoring (NILM) aims to decompose aggregate household power signals into appliance-level estimates. Despite many innovations, progress is hindered by the lack of a reproducible benchmark for evaluating models under varying compute budgets and at resolutions critical for real-time (1 min) and utility-scale (15 min) applications. We present **NILMBench2026**, a large-scale benchmark that systematically evaluates sixteen models across regression accuracy, event detection capabilities, computational cost, and generalization on multiple public datasets. Our analysis reveals two key insights: (1) Model performance is highly context-dependent, varying with appliance type and temporal resolution. (2) More critically, most models fail to generalize across buildings-especially when test appliances exhibit power characteristics unseen during training. While some architectures strike a better trade-off between accuracy and efficiency, poor cross-building generalization remains the key bottleneck for real-world adoption. **NILMBench2026** offers a reproducible platform and actionable insights to steer future NILM research toward robust, deployable solutions.

CCS Concepts

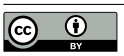
• **Computing methodologies** → *Machine learning algorithms.*

Keywords

Energy disaggregation, Non-intrusive load monitoring

ACM Reference Format:
Aayush Kuloor, Anurag Singh, Harsh Dhru, and Nipun Batra. 2026. NILMBENCH2026: A Benchmark for Energy Disaggregation. In *The 13th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3744256.3812587

*These authors contributed equally to this research.
†Corresponding author.



BuildSys '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2012-3/26/06
https://doi.org/10.1145/3744256.3812587

1 Introduction

Non-Intrusive Load Monitoring (NILM), or energy disaggregation, is the task of decomposing a household's aggregate power signal into appliance-level consumption estimates. By providing fine-grained feedback on individual appliance usage, NILM has emerged as a key enabler of energy efficiency. Prior studies show that such feedback can help households reduce electricity consumption by up to 15% [7, 9].

Pioneered by Hart [13], early NILM relied on combinatorial optimization [11] and edge detection [12]. Recently, the field has undergone a paradigm shift toward data-driven deep learning methods. Deep learning has fueled a surge of innovative architectures, ranging from Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to more recent Transformer-based models [2], achieving significant gains in disaggregation accuracy and modeling capacity.

Despite this rapid innovation, reproducibly comparing diverse NILM models remains a significant challenge. Foundational toolkits such as NILMTK [4, 5] standardized datasets and pipelines, yet recent model comparisons emphasize accuracy, overlooking compute-efficiency metrics like FLOPs, parameters, and inference time. Moreover, evaluations typically focus on a single temporal resolution, limiting applicability across real-world use cases including real-time feedback (1-minute) and utility-scale planning (15-minute). To address these limitations, we introduce **NILMBench2026**, a deployment-aware benchmark that evaluates sixteen models across three axes: accuracy, efficiency, and generalization, using standard datasets UK-DALE [18], REDD [20], REFIT [23].

As part of **NILMBench2026**, we modernize and extend the NILMTK ecosystem to support scalable, reproducible NILM research. We migrated the entire suite of legacy models from deprecated frameworks to modern PyTorch and introduced containerization. This infrastructure does not just enable our benchmark; it standardizes the codebase for future research, lowering the barrier to entry and ensuring that reported improvements are due to architectural advances, not implementation disparities. To ensure reproducibility across environments, we provide containerized workflows via Docker[1] and adopt the uv[2] package manager for isolated installations. We also augment the benchmark suite

[1]https://www.docker.com/
[2]https://docs.astral.sh/uv/

Table 1: Comparison against prior benchmarks. We introduce systematic evaluation of efficiency, multi-resolution support, and cross-domain generalization.

Feature	NILMTK '14 [4]	Contrib '19 [5]	Ours
Deployability	✗	✗	**✓(Docker + uv)**
Models	2	9	**16**
Resolutions	Variable	1-min	**1-min & 15-min**
Efficiency	✗	✗	**✓(FLOPs/Time)**
Cross Building Generalization	✗	✓	✓
Cross Dataset Generalization	✗	✗	✓
Stack	Python 2.7	TensorFlow 1.x	**PyTorch + uv**

with five contemporary architectures, Temporal Convolutional Network (TCN) [3], Convolutional Long Short-Term Memory (ConvLSTM) [27], Multi-State Dual CNN (MSDC) [14], Reformer [19], and NILMFormer [24], broadening architectural diversity and reflecting state-of-the-art design trends in temporal modeling.

Our evaluation surfaces two central insights. First, model performance is highly context-dependent, varying substantially with appliance type and time resolution. Second, and more critically, most models fail to generalize across buildings when test appliances exhibit power characteristics unseen during training. This generalization failure remains a fundamental obstacle to real-world deployment, especially in low-resource or diverse environments.

This paper makes the following contributions: (1) We introduce **NILMBench2026**, a large-scale, reproducible benchmark evaluating sixteen models on three datasets at two resolutions. (2) We analyze these models along the key axes of regression accuracy (MAE), event detection (F1-score), efficiency, and generalization. (3) We modernize the NILMTK ecosystem with a unified PyTorch implementation and containerized setup. (4) We provide insights into the generalization failures that are the core barrier to real-world NILM deployment. To support reproducibility, we provide the source code and benchmark suite on our GitHub repository.[3].

To situate our contribution, Table 1 contrasts NILMBench2026 with foundational benchmarks [4, 5]. While prior works standardized data parsers and baselines, they overlooked critical deployment constraints. We bridge this gap by introducing the first rigorous evaluation of computational efficiency, utility-scale resolutions, and cross-domain generalization, all supported by a modernized, reproducible software stack.

2 Background

2.1 The NILM Problem

Non-Intrusive Load Monitoring (NILM) estimates the power of N appliances, $X_i = (x_{i,1}, \ldots, x_{i,T})$, from the aggregate signal $Y = (y_1, \ldots, y_T)$, modeled as:

$$y_t = \sum_{i=1}^{N} x_{i,t} + \epsilon_t \tag{1}$$

Here, y_t is the total draw at time t, $x_{i,t}$ is appliance i's usage, and where ϵ_t captures both measurement noise and the contribution of unmetered appliances. Crucially, the distribution of $x_{i,t}$ is not stationary; it varies significantly across households due to device heterogeneity. A washing machine in House A may have a fundamentally different power signature than one in House B. This inherent variance turns NILM from a standard inverse problem into a complex domain adaptation challenge, which static training sets often fail to capture.

2.2 Overview of NILM Research

Established by Hart in the 1980s, early NILM approaches were event-based and used combinatorial optimization [13]. The field later shifted to probabilistic methods like Factorial Hidden Markov Models (FHMMs) in the 2000s. A major paradigm shift occurred with the introduction of deep learning by Kelly and Knottenbelt [17], which led to the widespread adoption of Convolutional and Recurrent Neural Networks (CNNs, RNNs) [10, 22]. Recently, Transformer-based models have further advanced the state of the art by effectively modeling long-range dependencies in energy data [24].

2.3 The NILMTK Ecosystem

The Non-Intrusive Load Monitoring Toolkit (NILMTK) [4], released in 2014, became the de facto framework for reproducible NILM research by standardizing data formats, parsers, and evaluation metrics. Our work leverages the NILMTK Experiment API [5] to systematically evaluate all sixteen models.

3 Benchmark Design and Goals

Our benchmark provides a systematic and reproducible framework to evaluate NILM models across dimensions critical for real-world deployment: accuracy, efficiency, temporal resolution, and generalization. The goal of this work is to build a unified benchmark that reflects the operational constraints and practical trade-offs inherent in NILM.

3.1 Benchmarking Philosophy

Our benchmark is guided by three core principles. First, we prioritize **real-world relevance** by evaluating both high-resolution (1-minute) scenarios, crucial for detailed user feedback [6], and low-resolution (15-minute) scenarios that align with standard utility-scale smart grid operations [16, 25]. Second, to ensure **comprehensive coverage**, we evaluate a diverse suite of sixteen models ranging from foundational baselines to current **state-of-the-art (SOTA) architectures**, spanning recurrent, convolutional, attention-based, and hybrid families. Finally, all experiments are grounded in **reproducibility** through a modernized machine learning pipeline built on the NILMTK Experiment API [5], ensuring our results are verifiable and can serve as a stable baseline for future work.

3.2 Datasets and Appliances

We use three public datasets spanning two countries and different grid infrastructures: (i) **REDD** (USA) [20], (ii) **UK-DALE** (UK) [18], and (iii) **REFIT** (UK) [23]. These were selected based on three technical requirements: building diversity, temporal continuity, and open-access standardization. We prioritized multi-residence datasets to support robust cross-building and cross-dataset evaluation (Tasks T2 and T3), as single-building repositories cannot validate generalization. Furthermore, we required high temporal integrity; we excluded alternative datasets due to significant gaps in

[3] https://github.com/nilmtk/nilmtk-contrib

"mains" power data that disrupt the training of sequence-based models. To ensure broad reproducibility, we focused on well-maintained hosts with permissive licenses, bypassing large-scale repositories that involve restrictive access fees. The selected datasets provide a robust testbed varying in grid infrastructure (110/230V) and appliance labeling quality, referring to the accuracy of the ground-truth sub-metered data, while our use of the NILMTK framework ensures all data is processed in a consistent manner.

Table 2: Summary of candidate NILM datasets. The benchmark selection is restricted to open-access repositories with multi-building diversity and high temporal continuity necessary to validate generalization across Tasks T1, T2, and T3.

Dataset	Country	Buildings	Duration	Appliances
REDD [20]	USA	6	3–19 days	10–20
UK-DALE [18]	GBR	5	655 days	5–54
REFIT [23]	GBR	20	2 years	9–21
AMPds	CAN	1	2 years	19+
iAWE	IND	1	73 days	10+
BLUED	USA	1	8 days	19+
DRED	NLD	1	6 months	10+
PecanStreet	USA	1000+	Varies	Varies

We exclude datasets such as AMPds, iAWE, BLUED, and DRED because their single-building nature precludes the evaluation of cross-building generalization (Task T2). Although the PecanStreet repository offers significant scale, it is omitted as the full dataset is not freely available to the research community.

We focus on the following appliances: **(i) Fridge** (always-on, periodic), **(ii) Microwave and Kettle** (short, bursty), **(iii) Washing Machine and Dish Washer** (long, multi-state) and **(iv) Television** (non-periodic, dynamic power consumption) These were chosen as: (i) they are present in most homes across datasets, (ii) their signals are reasonably clean and consistently labeled, and (iii) they span a range of NILM difficulty.

3.3 Temporal Resolutions

We evaluate each model at two resolutions: **1-minute**, which captures high-resolution NILM and user feedback [6]; and **15-minute**, which aligns with smart grid and utility timescales [16, 25].

3.4 Tasks and Evaluation Scenarios

We use the notation BX to refer to Building X from the corresponding dataset. Specific buildings were chosen for the given tasks because of **(i) consistent and long duration of data** and **(ii) good labeling quality**. We define three key generalization tasks:

T1: Intra-Building Generalization. Train and test on disjoint time segments from the same home to evaluate temporal generalization within a building. For UK-DALE, REDD, and REFIT, we train on 30 days from B1 and test on a held-out 1 week from B1. This building was selected based on the presence of consistent data for ≥ 4 target appliances and recording duration ≥ 30 days.

T2: Cross-Building Generalization. Train and test on different homes within the same dataset to assess generalization to unseen buildings. For UK-DALE, we train on B1, B2 and test on B4. For REDD, we train on B1, B2, B3 and test on B6. For REFIT, we train on B2, B3 and test on B4.

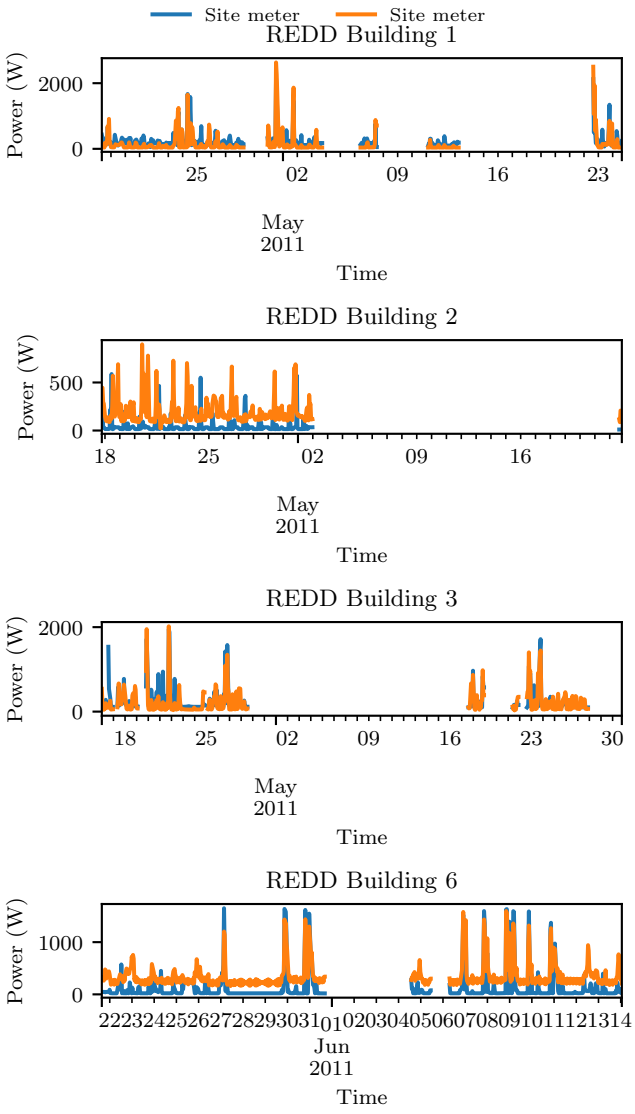


Figure 1: Mains power profiles across REDD buildings demonstrating the temporal gaps and recording inconsistencies that justify the use of a static train-test split for Task T2.

While cross-validation is generally preferred, we use a static split for Task T2 because the buildings in our dataset have disjoint temporal coverage. As illustrated in the mains power traces in Fig. 1, specific test residences such as B6 lack recorded data or active appliance signatures during our standardized training window. By selecting a fixed evaluation window where high-quality ground truth is available for the test set, we ensure the benchmark remains statistically valid despite the underlying data sparsity that would otherwise invalidate a full rotation-based cross-validation.

T3: Cross-Dataset Generalization. Train and test on different homes across datasets to evaluate geographic and domain shift. For REDD → REFIT, we train on REDD B1, B2, B3 and test on REFIT

B2. For REFIT → REDD, we train on REFIT B2, B5, B6 and test on REDD B1.

3.5 Evaluation Metrics

- **Mean Absolute Error (MAE):** Defined as the average absolute difference between the predicted power $\hat{y}_t$ and the ground truth power y_t over T time steps: MAE $= \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|$
- **F1-Score:** To assess the model's ability to detect appliance activation states, we compute the F1-score, the harmonic mean of precision and recall. We convert the continuous power signal into binary On/Off states using appliance-specific power thresholds. We utilized the standard NILMTK activation thresholds (e.g., 50W for the fridge, 200W for the microwave). A sensitivity analysis varying the activation thresholds by 20% revealed that while absolute F1-scores fluctuated by an average of 0.05, the relative rankings of the models remained highly stable, confirming the robustness of our event detection evaluation.
- **Floating Point Operations (FLOPs):** A measure of computational complexity that counts the total number of elementary arithmetic operations a model performs in a single forward pass.

All experiments were run on a VM with 24 vCPUs, 96 GB RAM, and an NVIDIA A100 (40 GB) GPU.

3.6 Hyperparameter Tuning

We performed Bayesian optimization using Optuna [1] to ensure fair comparison across architectures. Each model was allocated a fixed budget of 20 trials, optimizing validation MAE. The search space included:

- Sequence length: $[49, 499]$
- Epochs: $[10, 50]$
- Batch size: $\{128, 256, 512\}$
- Learning rate: log-uniform $[10^{-5}, 10^{-2}]$

4 Benchmarking Models

Our benchmark includes a diverse suite of sixteen architectures. We implemented each model according to its respective publication and integrated it into the NILMTK API. We group the models by architectural family and summarize the core mechanism of each.

4.1 Recurrent and Hybrid Architectures

RNN [17] uses a 1D CNN feature extractor followed by a bidirectional LSTM for sequence-to-point prediction. **WindowGRU** [21] is a lightweight recurrent model using a 1D convolution and a bidirectional GRU for efficient online disaggregation. **ConvLSTM** [27] employs a deep 1D CNN stack for hierarchical feature extraction, followed by a BiLSTM. **RNN with Attention** [26] enhances a BiLSTM by incorporating a self-attention mechanism to compute a weighted context vector. **RNN with Attention Classification** [26] combines this regression branch with a parallel CNN-based classification branch, gating the final prediction via element-wise multiplication.

4.2 Fully Convolutional Architectures

Seq2Point [29] is a widely adopted baseline that uses a deep 1D CNN to map an input window to the power value of its center point. **Seq2Seq** [28] uses a fully convolutional encoder-decoder architecture to reconstruct the entire appliance power sequence. **TCN** (Temporal Convolutional Network) [3] uses stacked 1D dilated convolutions to efficiently achieve a large receptive field without recurrence. **ResNet** [15] adapts residual blocks with skip connections for 1D energy signals. Finally, **ResNet with Classification** [26] extends ResNet by adding a parallel CNN branch to predict appliance on/off states, gating the final regression output.

4.3 Transformer-Based Architectures

BERT [8] discretizes the continuous power signal into "energy tokens" processed by a Transformer encoder. **Reformer** [19] replaces standard self-attention with efficient Locality-Sensitive Hashing (LSH) attention to reduce computational cost. **NILMFormer** [24] addresses energy data non-stationarity using instance normalization, processing the window's mean and standard deviation as a "stats token" to learn adaptive denormalization.

4.4 Specialized NILM Frameworks

DAE (Denoising Autoencoder) [17] frames disaggregation as a denoising task, learning to reconstruct clean appliance signals from noisy aggregates. **MSDC** (Multi-State Dual-CNN) [14] models discrete appliance states rather than continuous power, using a dual-branch CNN and a CRF layer to model state transitions.

5 Infrastructure Modernization for NILM Benchmarking

To enable scalable and reproducible NILM evaluation, we modernize NILMTK [4, 5] across three key fronts. First, we re-implement all legacy Keras/TensorFlow models in PyTorch under a unified API, ensuring compatibility with modern tooling and reducing technical debt. Second, we ensure reproducible deployment via Docker and the `uv` package manager, which pin versions and bundle dependencies within isolated environments. Finally, we expand the model library (including NILMFormer) with all architectures conforming to a common interface for training and inference, while modularizing dataset handling via the updated NILMTK Experiment API to support flexible experimentation.

6 Results and Analysis

6.1 Overview of Tasks and Metrics

We evaluate three generalization tasks: intra-building (T1), cross-building (T2), and cross-dataset (T3), using MAE, F1-score, parameters, and FLOPs. Because MAE can mislead on sparse activations, we supplement our quantitative results with a qualitative discussion of model failures to highlight key limitations.

6.2 Task T1: Intra-Building Generalization

This task evaluates how well models learn appliance signatures when trained and tested on data from the same building. This represents a best-case scenario and establishes a performance baseline. The results for T1 are detailed in Tables 3 for UK-DALE, 4 for REDD, and 5 for REFIT.

Overall, most modern architectures perform well in this setting, as they can effectively capture the underlying patterns of the appliance power signatures seen during training. However, certain

Table 3: MAE on UK-DALE for T1 at 1-minute and 15-minute resolutions (mean across three runs). Model performance is directly tied to data resolution, with different architectures excelling at 1-min versus 15-min intervals. Best and second-best per column are in bold and underline, respectively

Model	Fridge		Washing Machine		Microwave		Kettle	
	1m	15m	1m	15m	1m	15m	1m	15m
Mean	42.91	32.09	49.44	49.28	11.21	10.75	34.50	32.73
ConvLSTM	13.82	25.50	6.38	13.55	7.14	8.70	8.77	19.40
Reformer	34.97	25.71	13.16	20.83	10.03	11.15	14.95	20.82
TCN	17.39	24.43	9.72	18.18	7.14	9.05	10.14	20.78
MSDC	15.53	**21.50**	4.07	17.46	6.46	6.27	7.66	**13.46**
NILMFormer	14.23	24.46	**3.01**	**8.96**	**4.72**	7.28	**5.75**	21.00
WindowGRU	20.74	25.17	15.16	23.49	7.99	8.53	12.54	23.44
RNN Att.	16.41	32.01	19.83	21.33	10.54	11.18	45.89	30.60
Seq2Seq	21.48	31.60	8.50	24.60	6.09	9.87	7.62	28.46
Seq2Point	13.99	25.62	4.19	13.62	5.28	9.36	6.40	21.43
RNN	22.20	32.03	16.11	26.36	9.90	10.12	28.83	29.56
ResNet	22.48	31.86	11.88	21.25	7.27	10.77	7.73	30.45
ResNet Cl.	22.29	31.98	14.99	18.66	5.84	**5.10**	17.54	17.70
DAE	25.14	29.73	10.87	26.16	8.43	10.68	9.12	27.36
BERT	24.06	32.45	31.83	57.90	12.21	12.75	32.60	33.99
RNN Att. Cl.	**13.48**	32.02	15.27	18.62	5.99	6.50	10.83	23.72

Table 4: MAE on REDD for T1 at 1-min and 15-min resolutions (mean of 3 runs). Seq2Point excels at high-resolution; NILMFormer leads at 15-min. Best and second-best per column are in bold and underline, respectively

Model	Fridge		Washing Machine		Microwave		Dish Washer	
	1m	15m	1m	15m	1m	15m	1m	15m
Mean	71.40	58.21	46.48	45.76	31.58	29.62	39.19	38.27
ConvLSTM	18.43	30.54	41.31	31.08	21.04	18.60	18.70	16.14
Reformer	71.77	31.54	**6.52**	38.35	16.77	30.82	11.06	28.10
TCN	15.90	29.30	20.09	24.61	25.50	33.96	11.23	28.02
MSDC	13.88	36.38	21.70	25.93	13.50	34.22	**7.55**	22.18
NILMFormer	15.00	**24.85**	25.24	28.16	**12.27**	25.59	9.80	**10.50**
WindowGRU	23.39	30.49	13.35	**14.48**	14.71	29.61	15.49	28.74
RNN Att.	21.70	56.87	50.16	38.80	34.74	21.41	14.41	26.25
Seq2Seq	21.95	50.47	19.04	33.63	16.63	29.44	11.72	28.12
Seq2Point	**12.09**	31.28	16.11	22.61	12.50	25.63	7.57	15.62
RNN	71.24	57.69	41.26	51.77	20.28	21.21	18.20	35.54
ResNet	18.36	53.04	24.62	27.69	19.50	31.81	8.10	31.47
ResNet Cl.	25.55	48.97	22.79	26.37	17.10	**9.94**	18.32	15.56
DAE	29.36	44.54	25.67	36.13	30.73	31.27	9.93	32.47
BERT	40.48	57.43	79.76	63.83	41.25	31.39	23.46	38.38
RNN Att. Cl.	17.63	50.61	21.42	25.92	17.19	15.58	10.86	15.68

architectural choices give models an edge on specific appliance types. For appliances with sparse, high-power activations like the microwave, ResNet Cl. consistently delivers top-tier performance (Table 5). Its strength comes from its multi-task architecture; the dedicated classification branch learns to identify the appliance's on/off state, and this binary prediction is multiplied with the regression output. This gating mechanism forces the power prediction to zero when the appliance is off, preventing the low-level noise that single-regressor models often produce and making it exceptionally suited for this type of load. For appliances with more complex, multi-state cycles like the washing machine, NILMFormer's ability to handle non-stationary data through its z-normalization mechanism allows it to capture the distinct phases of the cycle more accurately, as seen in Table 3.

The shift from 1-minute to 15-minute data resolution creates a clear performance trade-off. At 1-minute, deep convolutional networks like Seq2Point and TCN demonstrate superior performance, as their stacked 1D and dilated convolutions effectively learn discriminative filters for the sharp, transient signatures of appliance activations. These critical local features are erased by the temporal averaging in 15-minute data, causing their efficacy to decline. Conversely, NILMFormer proves most robust at the 15-minute resolution. It uniquely compensates for the loss of signal detail by using its Transformer architecture's self-attention to model long-range context and integrating exogenous temporal features as powerful priors for when the power signal becomes ambiguous. Hybrid models like ConvLSTM and RNN-Attention represent a middle ground, but their performance remains partially bottlenecked by their initial convolutional layers, which struggle with the smoothed 15-minute signal, making them less effective than NILMFormer's direct contextual reasoning in low-resolution scenarios.

6.3 Task T2: Cross-Building Generalization

This task is a more realistic test of a model's utility, assessing its ability to generalize to an unseen building within the same dataset. As expected, all models exhibit performance degradation compared

Table 5: MAE on REFIT for T1 at 1-minute and 15-minute resolutions (mean across three runs). Best and second-best per column are in bold and underline, respectively

Model	Fridge		Washing Machine		Microwave		Dish Washer	
	1m	15m	1m	15m	1m	15m	1m	15m
Mean	39.91	33.79	34.17	33.93	4.05	3.97	105.68	104.92
ConvLSTM	11.86	17.66	32.29	25.51	3.55	3.25	19.28	35.45
Reformer	12.10	16.95	25.89	26.62	4.45	3.95	**15.96**	37.14
TCN	**10.49**	16.68	24.74	33.23	3.05	3.25	26.27	29.92
NILMFormer	12.47	16.98	22.12	24.71	1.65	3.11	24.74	24.70
WindowGRU	13.03	18.80	23.26	25.26	3.47	3.54	44.05	40.96
RNN Att.	12.48	33.30	46.02	38.03	4.55	6.22	38.60	78.23
Seq2Seq	15.83	33.74	27.06	26.77	2.93	4.06	35.33	54.72
Seq2Point	10.76	**16.10**	21.23	19.45	3.59	1.85	23.70	**22.96**
RNN	13.74	33.62	29.13	32.08	3.67	1.23	25.93	53.80
ResNet	14.67	33.71	**18.99**	32.11	4.04	5.28	27.07	64.45
ResNet Cl.	26.03	33.55	25.09	**18.92**	**0.98**	**1.14**	43.30	61.13
DAE	19.98	33.47	27.43	29.06	2.65	4.48	40.78	111.34
BERT	16.91	30.24	21.81	41.35	1.86	3.57	51.44	109.59
RNN Att. Cl.	39.83	23.57	20.98	22.58	2.13	3.70	77.18	64.05

to T1, highlighting the core challenge of generalizing across different appliance models and household behaviours. The results for T2 are detailed in Tables 6 for UK-DALE, 7 for REDD, and 8 for REFIT.

In this more challenging scenario, models designed specifically to handle data distribution shifts show a clear advantage. NILMFormer demonstrates the strongest generalization for the complex washing machine appliance (Table 6). Its core architectural innovation is using z-normalization (stationarization) and a "TokenStats" mechanism to learn adaptive denormalization via ProjStats. This allows it to adjust its predictions based on the local statistics of the input window, making it less sensitive to the signature variations found in new homes. For the fridge, the RNN Att. Cl. model performs best. Its recurrent nature captures the fridge's periodic cycles, while the attention mechanism allows it to focus on the most salient parts of the signal, and the classification gate ensures robust on/off detection.

Table 6: MAE on UK-DALE for T2. Architectures like NILMFormer that are specifically designed for energy data challenges demonstrate the best generalization to unseen buildings. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Fridge	Washing Machine	Microwave	Kettle	Mean Error
Mean	43.60	31.88	22.96	37.86	34.08
ConvLSTM	20.32	36.77	23.48	29.20	27.44
Reformer	43.60	29.02	21.83	12.81	26.82
TCN	18.99	28.94	22.83	12.60	20.84
MSDC	23.09	23.76	19.41	10.67	19.23
NILMFormer	18.08	**14.53**	17.39	11.67	**15.42**
WindowGRU	22.99	26.68	27.14	18.49	23.83
RNN Att.	21.67	29.91	25.36	31.88	27.20
Seq2Seq	26.70	31.04	19.70	12.67	22.53
Seq2Point	20.06	23.64	19.98	**10.05**	18.43
RNN	43.92	25.87	23.32	42.00	33.78
ResNet	27.66	20.39	22.27	10.33	20.16
ResNet Cl.	24.66	31.54	**16.18**	18.06	22.61
DAE	26.34	31.72	20.55	11.11	22.43
BERT	29.85	25.34	18.59	21.17	23.74
RNN Att. Cl.	**17.96**	36.91	20.10	12.28	21.81

Table 7: MAE on REDD for T2. The results highlight a key generalization challenge: the optimal architecture is highly appliance-dependent, and no single model performs well across all load types. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Fridge	Washing Machine	Dish Washer	Mean Error
Mean	74.64	31.04	13.25	39.64
ConvLSTM	31.50	30.32	2.06	21.30
Reformer	27.69	15.37	1.47	14.84
TCN	28.74	23.15	1.93	17.94
MSDC	30.56	8.91	0.56	13.35
NILMFormer	31.26	**2.64**	0.78	**11.56**
WindowGRU	40.10	7.49	3.88	17.16
RNN Att.	32.69	12.91	0.39	15.33
Seq2Seq	28.05	20.26	0.76	16.36
Seq2Point	33.25	6.59	**0.33**	13.39
RNN	30.71	3.92	3.28	12.64
ResNet	29.37	8.02	1.76	13.05
ResNet Cl.	34.49	7.69	0.89	14.36
DAE	31.88	8.60	1.84	14.11
BERT	37.33	4.68	3.36	15.12
RNN Att. Cl.	**18.39**	55.30	0.35	24.68

6.4 Task T3: Cross-Dataset Transfer

This task represents the most challenging test of generalization, evaluating a model's ability to perform zero-shot transfer from a training dataset in one country (REDD, USA) to a test dataset in another (REFIT, UK). The significant domain shift, caused by differences in grid infrastructure (voltage), appliance models, and consumer habits, leads to a substantial performance drop across all models, as seen in Table 9. This domain shift also reveals architectural limitations, as performance becomes highly inconsistent across appliances; for example, BERT is a top performer on the washing machine but the worst on the dishwasher. Furthermore, the dishwasher proves to be the most challenging load for generalization, with nearly all models struggling to produce a meaningful prediction. No single model performs best across all appliances, reinforcing the difficulty of universal generalization. However, NILMFormer's robust design for non-stationarity again gives it an edge on the complex washing machine, while the simpler, feature-focused architectures of ResNet and TCN prove more resilient for the fridge and dishwasher, respectively.

Table 8: MAE on REFIT for T2 (Mean across three runs). Models with a classification component show high stability. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Television	Washing Machine	Microwave	Kettle	Mean Error
Mean	31.22	22.80	**5.22**	36.45	23.93
ConvLSTM	31.36	12.74	7.32	23.53	18.74
Reformer	30.46	15.68	5.41	32.48	21.01
TCN	30.43	12.67	7.21	22.94	18.31
NILMFormer	28.89	**2.13**	7.25	25.10	**15.84**
WindowGRU	**26.20**	25.26	5.83	23.06	20.09
RNN Att.	34.16	24.35	6.98	31.02	24.13
Seq2Seq	36.10	18.28	6.25	28.15	22.20
Seq2Point	36.89	13.97	7.80	**20.15**	19.70
RNN	26.53	24.20	6.59	29.93	21.81
ResNet	31.25	18.86	6.85	23.48	20.11
ResNet Cl.	28.25	14.79	7.15	20.62	17.88
DAE	29.95	19.30	6.57	38.66	23.62
BERT	28.54	7.08	7.65	34.74	19.50
RNN Att. Cl.	26.50	13.98	7.32	28.04	18.96

Bidirectional Transfer Verification: To verify that this generalization failure is not an artifact of directional bias (i.e., specific to REDD → REFIT), we performed the reverse transfer experiment (REFIT → REDD). Our analysis (Table 10) confirms a symmetric performance collapse, with models exhibiting similar degradation in the reverse direction. This indicates that the generalization gap is a fundamental, bidirectional challenge inherent to domain shifts in energy data, rather than a specific quirk of the source dataset.

Table 9: MAE measured for T3. Trained on REDD and tested on REFIT. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Fridge	Washing Machine	Dish Washer	Mean Error
Mean	47.07	47.29	72.03	55.46
ConvLSTM	32.30	22.99	62.20	**39.16**
Reformer	27.84	36.83	62.05	42.24
TCN	29.93	69.28	**61.27**	53.49
NILMFormer	38.68	**18.71**	63.37	40.25
WindowGRU	35.02	45.51	62.82	47.78
RNN Att.	28.68	48.51	61.67	46.28
Seq2Seq	30.45	61.01	61.49	50.98
Seq2Point	34.72	33.87	61.41	43.33
RNN	32.02	41.01	62.17	45.06
ResNet	**26.06**	56.22	61.94	48.07
ResNet Cl.	30.69	66.82	61.58	53.03
DAE	26.19	65.38	62.25	51.27
BERT	35.38	21.90	99.83	52.37
RNN Att. Cl.	31.45	57.62	61.40	50.16

6.5 Failure to Generalize Dynamic Signatures

Beyond magnitude, models also fail to generalize to the temporal dynamics of unseen appliance signatures. Figure 2 provides a stark example of this signature overfitting. The television is a particularly challenging appliance due to its continuous and non-periodic power draw, which varies with on-screen content.

In the same-building scenario (Fig. 2 (a)), the model learns to replicate the noisy, high-frequency signature of the specific television it was trained on. However, when tested on a television in a new

Table 10: MAE measured for T3. Trained on REFIT and tested on REDD. Best and second-best values per column are highlighted in bold and underline, respectively. This mirrors the degradation seen in the forward direction (Table 9), confirming the bidirectional nature of the generalization gap. Best and second-best values are highlighted in bold and underline.

Model	Fridge	Washing Machine	Dish Washer	Mean Error
Mean	63.25	47.56	55.66	55.49
ConvLSTM	48.85	26.70	22.76	32.77
Reformer	63.18	**25.19**	28.82	39.06
TCN	38.13	28.09	22.10	29.44
NILMFormer	36.62	26.54	**15.74**	26.30
WindowGRU	55.88	32.82	25.40	38.03
RNN Att.	**31.71**	30.92	19.66	27.43
Seq2Seq	42.66	26.44	19.97	29.69
Seq2Point	35.54	25.88	16.52	**25.98**
RNN	60.31	33.77	25.19	39.76
ResNet	43.10	27.81	23.45	31.45
ResNet Cl.	46.14	29.41	24.44	33.33
DAE	39.35	32.42	27.91	33.23
BERT	61.77	48.31	28.72	46.27
RNN Att. Cl.	46.05	28.41	24.04	32.83

building (Fig. 2 (b)), its performance collapses. The model misses the power spikes and fails to capture the dynamic nature of the new device. This demonstrates that the model has not learned the general, abstract features of a "television" but has instead memorized the specific electrical fingerprint of a single instance. It acts as a poor filter, unable to adapt to the different frequencies, magnitudes, and patterns of a new device, a failure mode that is a significant barrier to out-of-the-box deployment in new homes.

6.6 The Limitations of MAE for Sparse Events

The cross-building test on the REFIT microwave highlights a key weakness of using MAE as a standalone metric. The microwave's usage is sparse, with brief, high-power spikes followed by long intervals of inactivity. A model can achieve a deceptively low MAE score by correctly predicting zero for these long inactive periods, even if it completely fails to identify the critical high-power activations. This rewards the model for identifying inactivity while masking its failure during active periods.

To provide a balanced view, it is crucial to use event-based metrics that are robust to class imbalance. In the following subsection, we analyze the F1-score, which considers both precision and recall of activation events to offer a practical assessment of performance.

6.7 F1 Score Analysis

To address MAE's limitations on sparse loads, we supplement our analysis using the F1-score. As the harmonic mean of precision and recall, the F1-score provides a balanced assessment of activation event detection, making it robust to the severe class imbalance inherent in NILM data. The results for the cross-building (T2) and cross-dataset (T3) tasks are presented in Tables 11, 12, and 13.

6.8 Insights from F1-Score Evaluation

Performance is Highly Dependent on Appliance Signature. The F1-scores reinforce that model performance is strongly tied to the appliance's electrical signature.

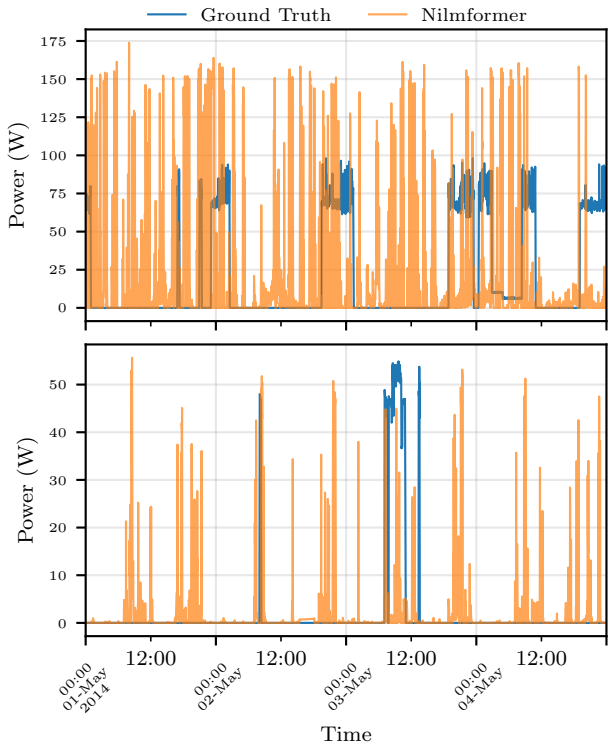


Figure 2: (a, top) T2: NILMFormer's prediction on REFIT television, (b, bottom) T1: NILMFormer's prediction on REFIT television. While the model accurately captures the dynamic signature of the television it was trained on (b, same-building), it fails to identify spikes on an unseen television (a, cross-building). This indicates the model has overfit to the training appliance's specific electrical characteristics, failing to generalize to a new device of the same type.

Table 11: F1 Score for T2 on UK-DALE [18]. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Fridge	Washing Machine	Microwave	Kettle	Mean F1 Score
Mean	0.60	0.06	0.00	0.02	0.17
ConvLSTM	0.82	0.29	0.08	0.02	0.30
Reformer	0.60	0.19	0.00	0.02	0.20
TCN	0.80	0.23	0.20	0.11	0.34
MSDC	0.75	0.27	0.20	0.43	0.41
NILMFormer	**0.84**	**0.50**	0.10	**0.68**	**0.53**
WindowGRU	0.65	0.09	0.19	0.37	0.33
RNN Att.	0.77	0.14	0.12	0.02	0.26
Seq2Seq	0.67	0.28	**0.27**	0.35	0.39
Seq2Point	0.75	0.30	0.14	0.59	0.45
RNN	0.74	0.15	0.14	0.02	0.26
ResNet	0.66	0.31	0.19	0.38	0.39
ResNet Cl.	0.71	0.24	0.21	0.25	0.35
DAE	0.69	0.24	0.14	0.06	0.28
BERT	0.64	0.08	0.13	0.06	0.23
RNN Att. Cl.	0.81	0.24	0.17	0.30	0.38

- **Periodic Appliances (Fridge):** The fridge, with its consistent cycles, is the easiest to detect. Across all T2 experiments (Tables

11 and 12), most models achieve high F1-scores, often ranging from 0.70 to over 0.90, indicating reliable state detection.

- **Sparse, High-Power Appliances (Microwave and Kettle):** Performance on these event-based loads is far more variable. For instance, in the UK-DALE T2 test (Table 11), **NILMFormer** (0.68) and **Seq2Point** (0.59) excel on the kettle. In the REFIT T2 test (Table 12), **WindowGRU** achieves the highest F1-score for the kettle (0.56). Despite both being sparse and bursty, models perform differently on the microwave versus the kettle due to intra-group variance. Microwaves often exhibit more rigid, square-wave step functions, whereas kettles might have slight ramp-ups or varying baseline durations depending on the water volume, requiring different architectural biases to capture perfectly.
- **Complex, Multi-State Appliances (Washing Machine):** These are by far the most challenging. The consistently near-zero F1-scores suggest that while models might approximate average power (leading to a reasonable MAE), they fail to correctly identify the boundaries of the complex activation cycles.

Table 12: F1 Score for T2 on REFIT [23]. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	Television	Washing Machine	Microwave	Kettle	Mean F1 Score
Mean	0.45	0.00	0.00	0.03	0.12
ConvLSTM	0.52	0.00	0.22	0.42	0.29
Reformer	0.45	0.00	0.00	0.13	0.15
TCN	0.51	0.00	0.27	0.32	0.28
NILMFormer	0.19	0.00	0.00	0.35	0.14
WindowGRU	0.45	0.00	**0.36**	**0.56**	**0.34**
RNN Att.	0.44	0.00	0.24	0.19	0.22
Seq2Seq	0.45	0.00	0.00	0.03	0.12
Seq2Point	0.42	**0.01**	0.06	0.50	0.25
RNN	0.51	0.00	0.00	0.12	0.16
ResNet	0.47	0.00	0.00	0.10	0.14
ResNet Cl.	0.48	0.00	0.00	0.39	0.22
DAE	0.47	0.00	0.02	0.06	0.14
BERT	**0.56**	0.00	0.00	0.03	0.15
RNN Att. Cl.	0.47	0.00	0.17	0.13	0.19

In summary, the F1-score analysis provides a crucial, practical perspective that complements the MAE results. It confirms that MAE can mask critical failures, especially for complex appliances, and reinforces our main paper's conclusion that improving generalization is the most pressing challenge for the NILM community.

The complete failure of the models to predict these power spikes results in a catastrophic breakdown of precision and recall, the two components of the F1 score. In NILM, a true positive is recorded when the model correctly identifies an appliance's activation. As is visually evident, none of the models generate a prediction that corresponds to the high-power ground truth events, leading to zero true positives. Consequently, the recall, which measures the model's ability to find all actual appliance activations, is zero because all events are missed (false negatives). Similarly, precision, which assesses the accuracy of the positive predictions, is also zero because there are no correct positive predictions to be made. An F1 score of 0 is the mathematical outcome when a model fails to identify any true positives, providing a quantitative confirmation of what the plots from Fig. 3 visually demonstrate: a total inability of the models to generalize their learning to the unseen data from a different building. This issue highlights a significant challenge

Table 13: F1 Scores for T3 (REDD → REFIT. Best and second-best values per column are highlighted in bold and underline, respectively.)

Model	Fridge	Washing Machine	Dish Washer	Mean F1 Score
Mean	0.54	0.13	0.15	0.27
ConvLSTM	0.33	0.06	0.00	0.13
Reformer	0.76	0.24	0.00	0.33
TCN	0.72	0.31	0.18	**0.40**
NILMFormer	**0.77**	0.00	0.01	0.26
WindowGRU	0.52	0.07	0.00	0.20
RNN Att.	0.70	0.02	0.01	0.24
Seq2Seq	0.55	0.22	0.00	0.26
Seq2Point	0.59	0.06	0.00	0.22
RNN	0.75	0.24	0.00	0.33
ResNet	0.58	0.22	0.00	0.27
ResNet Cl.	0.58	**0.34**	0.09	0.34
DAE	0.63	0.21	0.19	0.34
BERT	0.58	0.14	**0.31**	0.34
RNN Att. Cl.	0.72	0.27	0.02	0.34

in the field, as variations in electrical systems and appliance models across different locations can prevent models from recognizing even prominent operational signatures.

6.9 Efficiency-Accuracy Trade-off

Our benchmark results (Table 14) challenge the assumption that accuracy scales linearly with compute. Instead, we find a non-monotonic relationship where architectural choice, rather than raw budget, dictates performance. For example, sequential processing makes recurrent models like **RNN** computationally expensive (123.69 GFLOPs), whereas convolutional architectures like **Seq2Seq** and **DAE** are highly efficient. Furthermore, efficiency can be engineered without sacrificing utility, as demonstrated by the aggressive downsampling in **BERT** and sparse state modeling in **MSDC-CRF**, which reduces parameters by 99% compared to standard MSDC.

This disconnect between size and performance is best illustrated by comparing our top performers. While the computationally intensive **NILMFormer** achieves high accuracy, the lightweight **TCN** (with 69.15K parameters) proves equally effective, even on challenging cross-dataset tasks. This shows that architectural inductive bias, that is, how well a model's design matches the signal structure, is a more decisive factor for disaggregation than model size alone.

6.10 Per-Appliance Architectural Trends

Our benchmark shows that optimal architectures vary by appliance due to differing electrical signatures.

Fridge: As a periodic, state-driven appliance, the fridge is well modeled by RNN Att. Cl., which captures cycles, and MSDC, which models 'on', 'off', and 'defrost' states.

Washing Machine and Dishwasher: These multi-stage appliances benefit from models handling long-range dependencies. NILMFormer, with its Transformer design, excels here.

Microwave and Kettle: Defined by brief high-power bursts, they require models sensitive to sparse events. ResNet Classification's gating helps with microwave; Seq2Point CNN effectively handles the kettle's signature.

Television: TVs exhibit continuous, non-periodic, content dependent loads. As shown in Figure 2, signature diversity across models impairs generalization, and no current model handles this reliably.

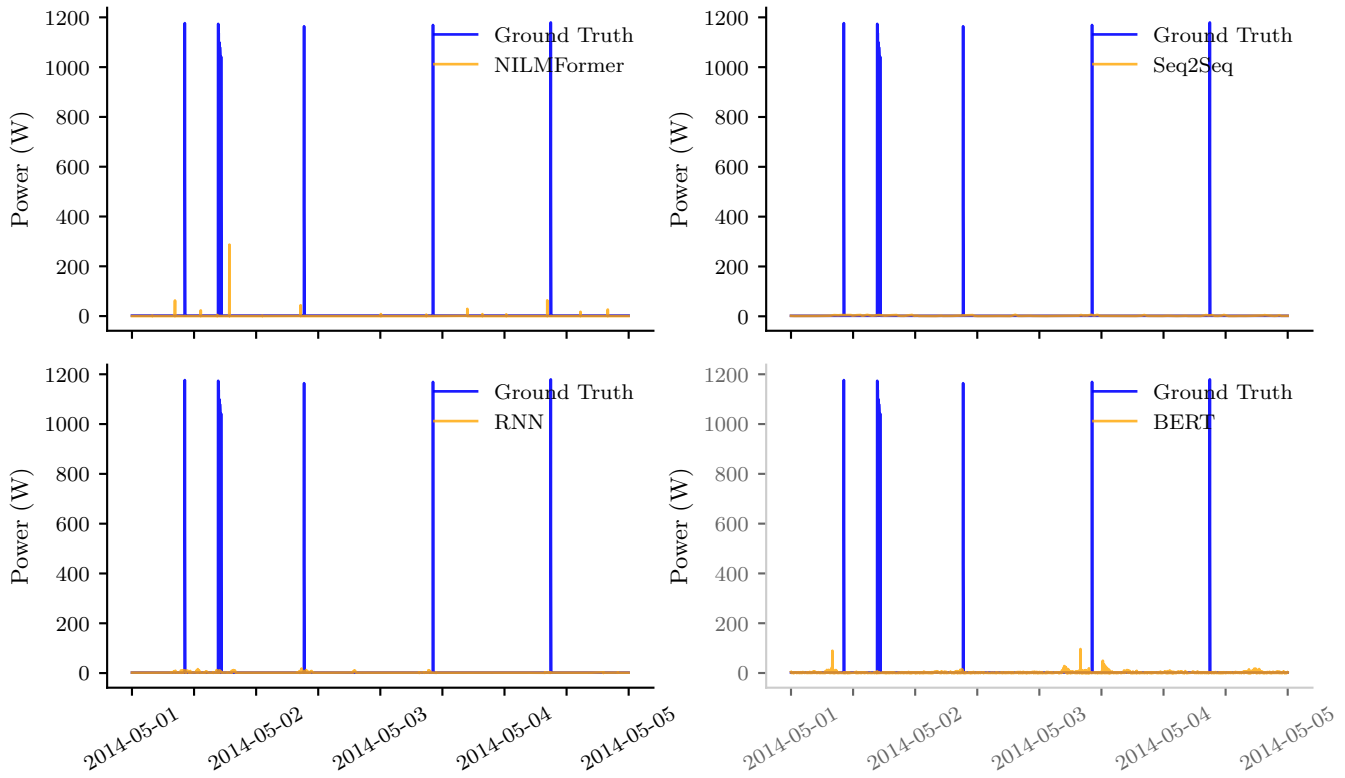


Figure 3: Microwave power predictions from NILM models on T2 on the REFIT [23] dataset. All models fail to capture the high-power activation spikes, resulting in a complete miss of all events and an extremely low F1 score.

Table 14: Computational performance, model size, and inference time for all evaluated models. Best and second-best values per column are highlighted in bold and underline, respectively.

Model	FLOPs (in GFLOPs)	Parameters	Inference Time (in 10^{-4} sec)
ConvLSTM	46.55	483 K	2009
Reformer	90.71	943 K	1125
TCN	20.45	69 K	317
MSDC	25.42	12680 K	395
NILMFormer	134.96	383 K	354
WindowGRU	34.03	427 K	177
RNN Att.	197.65	1330 K	72
Seq2Seq	9.42	447 K	38
Seq2Point	72.02	3620 K	39
RNN	123.69	1270 K	65
ResNet	26.30	669 K	62
Resnet Cl.	106.57	4200 K	89
DAE	9.42	832 K	**25**
BERT	**5.35**	803 K	63
RNN Att. Cl.	349.32	4940 K	5447
MSDC-CRF	7.30	**24** K	57

6.11 Real-World Implications of MAE Differences

To contextualize these metrics, consider the real-world implications of the energy demand range and MAE differences. For instance, consider the Fridge on UK-DALE. The RNN model yields an MAE of ~43W, whereas the RNN Att. Cl. reduces this to ~18W. Over the course of a year (8,760 hours), this 25W reduction in continuous absolute error equates to roughly 219 kWh of corrected energy attribution. At standard residential electricity rates, this translates to significant financial and behavioral implications for consumer feedback systems.

6.12 Summary of Findings

Our benchmark reveals four takeaways for the community:

- **No Single Model Wins:** The best architecture is highly dependent on the appliance's electrical signature. CNN-based models excel at sparse, high-power events, while Transformers are better for complex, multi-state appliances. Our finding that optimal architectures are appliance-dependent suggests that real-world deployment will likely require an ensemble or multi-expert system. While this maximizes accuracy, it linearly increases the parameter footprint and FLOPs. Future smart-meter edge devices will either require specialized AI accelerators to run multiple lightweight models (like TCN) concurrently, or rely on cloud-based processing for heavyweights like NILMFormer.
- **Generalization is the Main Hurdle:** Most models fail to generalize to unseen buildings and datasets. The drop in performance from single-building to cross-building tasks is significant and represents the barrier to real-world adoption.

- **MAE is Misleading for Sparse Events:** Always predicting 'off' yields low MAE but misses all activations. Event metrics are essential.
- **Efficiency-Accuracy Trade-off:** This trade-off is not monotonic, as architectural innovation often outweighs raw computational power. Balanced models like TCN deliver competitive accuracy at a fraction of the cost of heavyweights like NILMFormer, while the most efficient models consistently underperform.

7 Limitations

While our benchmark offers a comprehensive view of the current NILM landscape, we acknowledge specific limitations that contextualize our findings and define boundaries for their interpretation.

Geographic and Grid Bias. Our evaluation relies on three canonical datasets (UK-DALE, REDD, REFIT) sourced exclusively from the UK and USA. Consequently, our findings regarding generalization may not fully transfer to regions with different grid standards (e.g., varying voltage/frequency standards in Asia) or distinct appliance usage patterns found in developing nations.

Appliance Scope. We benchmarked six core appliance types representing common difficulty levels. However, we excluded emerging high-impact loads, such as Electric Vehicle (EV) chargers, heat pumps, and solar inverters, due to the scarcity of high-quality, publicly labeled data for these devices. As electrification accelerates, future benchmarks must expand to include these critical loads.

Fixed Evaluation Splits. Due to the immense computational cost of benchmarking 16 models across multiple datasets, resolutions, and tasks (16 models $\times$ 3 datasets $\times$ 2 resolutions $\times$ 6 appliances = 576 configurations, each run 3 times), we utilized fixed train-test splits rather than k-fold cross-validation. While we mitigated variance by averaging multiple runs, we acknowledge that fixed splits may not capture the full distribution of performance variance across all possible temporal windows.

Offline vs. Online Inference. Our benchmark evaluates models in an offline, batch inference setting. We do not assess metrics critical for real-time, edge-based deployment, such as streaming latency, memory footprint under continuous operation, or model adaptability to concept drift over months or years of operation.

8 Future Research Directions

To bridge the generalization gap and handle the inherent non-stationarity of energy data, we propose transitioning from static supervised learning toward more adaptive frameworks. Based on the failure modes identified in our benchmark, we identify five critical paths for future research:

Domain Adaptation and Transfer Learning. Since model performance degrades significantly across datasets (e.g., REDD to REFIT), implementing Unsupervised Domain Adaptation (UDA) is essential. Future work should focus on aligning the latent feature distributions of "source" and "target" buildings, enabling models to adjust to varying grid voltages and appliance signatures without requiring ground-truth labels for every new deployment.

Self-Supervised Pre-training. Models often overfit to specific training signatures. Utilizing Self-Supervised Learning (SSL) on large-scale unlabeled mains data, analogous to masked language modeling in NLP, helps learn abstract load representations, making downstream fine-tuning robust to individual device variations.

Multi-Task Learning with State Classification. Our benchmark demonstrates that models with dedicated classification branches (e.g., ResNet Cl.) are significantly more stable for sparse, high-power loads. Expanding this architecture to include probabilistic state modeling or Multi-State Dual-CNNs could better capture complex, multi-stage appliance cycles (such as washing machines) that currently result in near-zero F1-scores.

Adaptive Denormalization & Metadata Integration. Building on the success of NILMFormer, incorporating local statistics (mean and standard deviation) and exogenous features (e.g., time-of-day or occupancy) as priors can help models distinguish between ambiguous signatures. This is effective at lower temporal resolutions (e.g., 15-minute), where the fine-grained signal shape is lost.

Generative Data Augmentation. To combat the lack of diversity in training sets, Generative Adversarial Networks (GANs) or Diffusion Models could be employed to synthesize "adversarial" appliance signatures. By artificially varying the magnitude and temporal duration of training samples, models can be trained to be resilient to the out-of-distribution power characteristics that currently cause failures in cross-building scenarios.

Community Benchmarking & Leaderboards. Finally, we advocate for the establishment of maintained leaderboards and annual competitions, mirroring the successful models in the Machine Learning (e.g., Kaggle), NLP (e.g., GLUE), and Computer Vision (e.g., ImageNet) communities. A centralized, dynamic leaderboard would shift the incentive structure from achieving marginal accuracy gains on static, well-worn test sets to demonstrating robust generalization on held-out, out-of-distribution benchmarks.

9 Conclusion

In this work, we introduced **NILMBench2026**, a large-scale benchmark to evaluate NILM models across accuracy, efficiency, and generalization. Our key finding is that generalization remains the main barrier to real-world deployment. While many models perform well within a building, their accuracy drops sharply in new homes, due to unfamiliar appliance signatures and interference from co-occurring loads. We also find no universally best model: performance depends heavily on the appliance type. MAE can mislead for sparse-use appliances, hiding failures on critical events. A trade-off exists between accuracy and computational efficiency. NILMBench2026 suggests a shift in research focus, from marginal accuracy gains in controlled settings to developing robust, generalizable models. Bridging this generalization gap is the singular prerequisite for deploying NILM at scale. Future benchmarks must move beyond accuracy to assess robustness, potentially by introducing adversarial perturbations or synthetic domain shifts. NILMBench2026 provides the reproducible platform necessary to launch this new era of reliability-focused research.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. ACM, New York, NY, USA, 2623–2631. doi:10.1145/3292500.3330701

[2] Georgios-Fotios Angelis, Christos Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. 2022. NILM Applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings* 261 (2022), 111951. doi:10.1016/j.enbuild.2022.111951

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint arXiv:1803.01271* (2018).

[4] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. 2014. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In *Proceedings of the 5th International Conference on Future Energy Systems (ACM e-Energy)*. ACM, Cambridge, UK, 265–276. doi:10.1145/2602044.2602051

[5] Nipun Batra, Rithwik Kukunuri, Ayush Pandey, Raktim Malakar, Rajat Kumar, Odysseas Krystalakos, Mingjun Zhong, Paulo Meira, and Oliver Parson. 2019. Towards reproducible state-of-the-art energy disaggregation. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*. ACM, New York, NY, USA, 193–202. doi:10.1145/3360322.3360844

[6] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy* 52 (2013), 213–234. doi:10.1016/j.enpol.2012.08.062 Special Section: Transition Pathways to a Low Carbon Economy.

[7] Sarah Darby. 2006. The Effectiveness of Feedback on Energy Consumption. *Environmental Change Institute, University of Oxford* (2006). Technical Report.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186. doi:10.48550/arXiv.1810.04805

[9] Jon Froehlich, Leah Findlater, and James Landay. 2011. Disaggregated End-Use Energy Sensing for the Smart Grid. In *Proceedings of the Pervasive Computing*. doi:10.1109/MPRV.2010.74

[10] Alon Harell, Stephen Makonin, and Ivan V. Bajić. 2019. Wavenilm: A Causal Neural Network for Power Disaggregation from the Complex Power Signal. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8335–8339. doi:10.1109/ICASSP.2019.8682543

[11] G.W. Hart. 1992. Nonintrusive appliance load monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891. doi:10.1109/5.192069

[12] George W. Hart. 1985. *Prototype Nonintrusive Appliance Load Monitor*. Progress Report 2. MIT Energy Laboratory. Prepared for Electric Power Research Institute under Contract RP 2568-2.

[13] George W. Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891. doi:10.1109/5.192069

[14] Jialing He, Yiwei Liu, Wei Zhang, Hao Wang, and Quanshi Tan. 2023. MSDC: Exploiting Multi-State Power Consumption in Non-Intrusive Load Monitoring Based on a Dual-CNN Model. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*. 5078–5083. doi:10.48550/arXiv.2302.05565

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. doi:10.48550/arXiv.1512.03385

[16] Jana Huchtkoetter and Andreas Reinhardt. 2020. On the Impact of Temporal Data Resolution on the Accuracy of Non-Intrusive Load Monitoring. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*. ACM, Virtual Event, Japan, 270–273. doi:10.1145/3408308.3427974

[17] Jack Kelly and William Knottenbelt. 2015. Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 55–64. doi:10.1145/2821650.2821672

[18] Jack Kelly and William Knottenbelt. 2015. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data* 2, 1 (2015), 150007. doi:10.1038/sdata.2015.7

[19] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*.

[20] J. Zico Kolter and Matthew J. Johnson. 2011. REDD: A public data set for energy disaggregation research. In *Proceedings of the SustKDD workshop on Data Mining Applications in Sustainability*. San Diego, CA, USA.

[21] Odysseas Krystalakos, Christoforos Nalmpantis, and Dimitris Vrakas. 2018. Sliding Window Approach for Online Energy Disaggregation Using Artificial Neural Networks. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM, 1–6. doi:10.1145/3200947.3201011

[22] Myeung-Hun Lee and Hyeun-Jun Moon. 2023. Nonintrusive Load Monitoring Using Recurrent Neural Networks with Occupants Location Information in Residential Buildings. *Energies* 16, 9 (2023). doi:10.3390/en16093688

[23] David Murray, Lina Stankovic, and Vladimir Stankovic. 2016. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Scientific Data* 3, 1 (2016), 160122. doi:10.1038/sdata.2016.122

[24] Adrien Petralia, Philippe Charpentier, Youssef Kadhi, and Themis Palpanas. 2025. NILMFormer: Non-Intrusive Load Monitoring that Accounts for Non-Stationarity. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 4761–4772. doi:10.1145/3711896.3737251

[25] Antonio Ruano, Alvaro Hernandez, Jesus Ureña, Maria Ruano, and Juan Garcia. 2019. NILM Techniques for Intelligent Home Energy Management and Ambient Assisted Living: A Review. *Energies* 12, 11 (2019), 2203. doi:10.3390/en12112203

[26] Hetvi Shastri and Nipun Batra. 2021. Neural Network Approaches and Dataset Parser for NILM Toolkit. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '21)*. ACM, Coimbra, Portugal, 1–4. doi:10.1145/3486611.3486652

[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*. 802–810.

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*. 3104–3112.

[29] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.