Mazen Bouchur, Nan Li, and Andreas Reinhardt. "Can Attention Improve Sequence-to-Point Load Disaggregation? A Comparative Assessment." In: *Proceedings of the 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*. 2025, pp. 23–33.

Can Attention Improve Sequence-to-Point Load Disaggregation? A Comparative Assessment

Mazen Bouchur
TU Clausthal
Germany
mazenb@ieee.org

Nan Li
TU Clausthal
Germany
nan.li.x2njh@simplelogin.com

Andreas Reinhardt
TU Clausthal
Germany
reinhardt@ieee.org

Abstract

Adding attention mechanisms to neural networks has been widely explored in recent years. Within the realm of Non-Intrusive Load Monitoring (NILM), however, the potential of attention has not been fully investigated to date. While the few existing works found that attention can lead to lower disaggregation errors and/or greater model robustness, what remains unclear is when and under which conditions these gains can be accomplished. We hence conduct a systematic analysis of different attention mechanisms for energy data disaggregation, i.e., Non-Intrusive Load Monitoring (NILM). More specifically, we select three distinct attention mechanisms: Channel Attention (CA), Feed-Forward Attention (FFA), and Self-Attention (SA) and extend them to construct seven different attention modules, which are integrated into the existing sequence-based neural network architecture to develop enhanced models. We evaluate the performance of the seven configurations on the publicly available UK-DALE and REDD datasets under cross-house and cross-dataset settings to quantify both in-distribution accuracy and transfer to unseen data, as is typical in real-world scenarios. Across these controlled comparisons, we found that the effects of attention are strongly context-dependent. That is, they do not only vary with appliance dynamics, but also depend on the dataset characteristics. While some attention configurations showed notable gains for a few appliances, their use led to a lower accuracy for other devices. What is more, accuracy results also differed when trying to disaggregate the same appliance type in two different datasets. Besides discussing these insights in more depth, we further quantify the resource tradeoffs of each attention family. We show that the incurred overhead (in terms of required computation) differs starkly between attention types, yet does not directly correlate with increased performance. In order to cater for a deeper understanding of the potentials and limitations of attention, our work provides design guidance for both practitioners and model designers on choosing attention families under accuracy and efficiency constraints.

CCS Concepts

• **Computing methodologies → Neural networks**.

BUILDSYS '25, November 20–21, 2025, Golden, CO, USA

ACM ISBN 979-8-4007-1945-5/25/11
https://doi.org/10.1145/3736425.3772099

Keywords

Attention Mechanisms, Deep Neural Networks, Energy Disaggregation, Non-Intrusive Load Monitoring

ACM Reference Format:
Mazen Bouchur, Nan Li, and Andreas Reinhardt. 2025. Can Attention Improve Sequence-to-Point Load Disaggregation? A Comparative Assessment. In *The 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BUILDSYS '25), November 20–21, 2025, Golden, CO, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3736425.3772099

1 Introduction

Issues related to energy consumption, monitoring, and management are becoming increasingly important due to the growing demand for energy, which is expected to reach its fastest pace over the course of the years 2025–2027 [17]. Improving the efficiency of energy consumption is a key aspect of sustainable development, especially in light of the rapidly changing geopolitical landscape and its impact on the global energy supply chains [12]. The energy efficiency of consumer appliances in the residential sector is among the key factors that drove down the consumption of this sector to a record low of around 30% in 2023 in the United Kingdom after a steady increase in the four years prior [10]. Behavior changes due to higher energy prices and warmer temperatures have proven to be key drivers of energy consumption, and understanding the behavior of appliances is thus a key aspect of energy management [10]. Governmental policies and climate packages, such as the EU's Energy Efficiency Directive (EED) and the Energy Performance of Buildings Directive (EPBD), are expected to further play a key role in the energy transition.

Load disaggregation, also known as Non-Intrusive Load Monitoring (NILM), enables appliance-specific energy monitoring by observing only the total energy consumption of a household. This not only offers deeper insights into the origins of energy use, but also facilitates its optimization, strategic load scheduling, and demand management [34]. The advantage of load disaggregation lies in its ability to reduce the number of energy meters installed, thereby decreasing wiring complexity and enhancing retrofitting capabilities. In recent years, deep learning techniques have been dominating the field of load disaggregation, and have shown promising results in terms of accuracy, generalization to unseen data, ease of deployment, as well as economic considerations [8]. The Sequence-to-Point (S2P) model is a sequence-based model that has shown promising results in load disaggregation [41]. Its success can be attributed to the ability of sequence-based models to capture temporal dependencies and the effectiveness of Deep Neural Networks

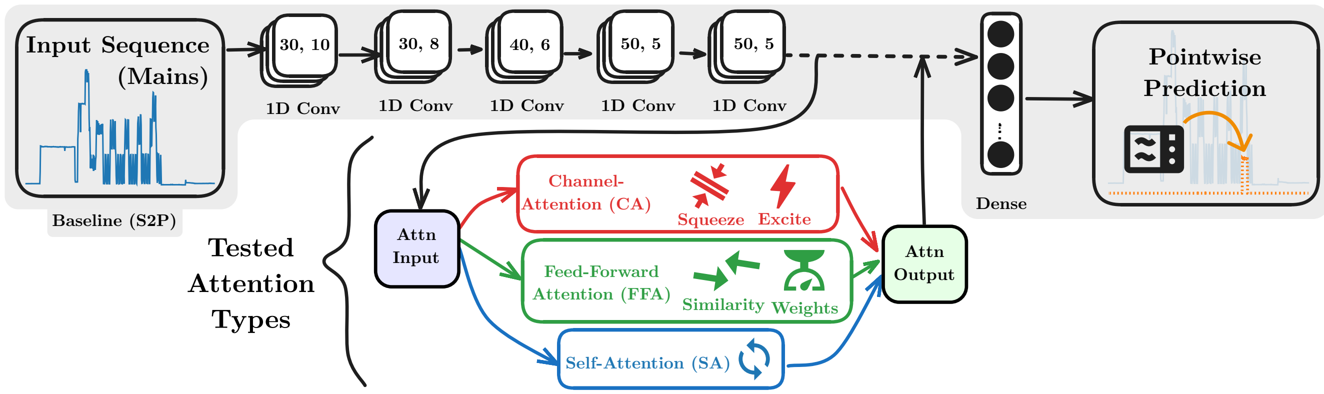


Figure 1: Different attention mechanisms are integrated into the Sequence-to-Point (S2P) model between the last convolutional layer and the dense layer. The considered attention types and corresponding implementation details are provided in Section 3 and Figure 2, respectively. The numbers in the boxes indicate the number of filters and kernel sizes in each convolution layer.

(DNNs) in learning complex patterns from data. A number of studies have shown diverse extensions of the S2P model, including the integration of attention mechanisms, to improve the accuracy of load disaggregation [5, 40]. Still, no systematic comparison of different attention mechanisms at structural level has been conducted to date, as most of the works in this area evaluate a single attention design within a custom architecture, making gains from one attention type hard to compare across different attention families.

Attention in deep learning has become a key component of most contemporary neural network architectures, with applications in areas such as Natural Language Processing (NLP), Speech Recognition, and Computer Vision (CV) [4]. The NILM problem can be considered as a long sequence regression or classification problem. When processing long sequences, attention mechanisms offer significant advantages by allowing the model to dynamically focus on the parts of the input that are highly relevant to the task while ignoring irrelevant information [37]. Selective attention enables the model to concentrate on features critical to the current output while suppressing irrelevant ones, thereby effectively enhancing the model's accuracy [4].

In light of these characteristics, we study the impact of attention in the context of NILM. Rather than proposing a new end-to-end model or seeking to identify a *one-size-fits-all* configuration, we present a framework that systematically incorporates and evaluates attention mechanisms into the established S2P model, as depicted in Fig. 1. We conjecture that integrating the *right* type of attention will enable the NILM algorithm to more effectively capture essential features from time series data, thereby enhancing the accuracy of load monitoring [16, 41]. Motivated by this premise, the study seeks to explore the foundational aspects of the attention mechanism and rigorously evaluate its performance under diverse and complex load types, as well as in low sampling rate conditions. Specifically, experiments will be conducted on real-world datasets to determine whether the attention mechanism can improve the prediction accuracy of NILM models under these conditions and if the introduced overhead is justified by the performance gains.

In this paper, we make the following key contributions:

- We design and integrate *seven* attention-based extensions of the S2P model, spanning three distinct attention mechanisms (Channel-, Feed-Forward-, and Self-Attention).
- We establish a controlled evaluation protocol that keeps data splits and training settings aligned, in order to enable a fair comparison of attention families independent of architecture-specific factors. We evaluate these models on the UK-DALE and REDD datasets across five appliance types.
- We quantify accuracy and efficiency jointly and show that effects are highly context dependent.
- We analyze resource overhead (model size and training time) for each attention variant, highlighting tradeoffs in speed and complexity.
- We provide empirically grounded guidelines for selecting appropriate attention mechanisms in NILM, based on appliance characteristics and deployment constraints.

2 Related Work

Recent advances in DNNs have significantly improved NILM performance, moving beyond earlier methods that relied heavily on manual setup, expert knowledge, and event-based feature extraction [13, 16]. Early NILM approaches also suffered from a limited adaptability to multi-state or continuously variable appliances. By contrast, DNN-based approaches capture complex appliance behaviors more effectively and generalize better to unseen environments [20].

Among DNN architectures used in NILM, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) models have shown promising results. For instance, [3] introduced an extended Denoising Autoencoder (DAE) with additional CNN and pooling layers, while [36] proposed a parallel regression-classification network. [14] and [24] used RNNs for effective disaggregation, and LSTM-based approaches were explored in [19] and [31]. Other state-of-the-art techniques such as Transfer Learning [11], Generative Adversarial Networks (GAN) [1], and Federated Learning [7] have also been explored in NILM.

The S2P model [41] is a CNN-based architecture that predicts only the midpoint of an input window of mains readings, assuming the appliance state at the center is correlated with both preceding and following power signals. An analysis of the feature maps learned by the convolutional networks (specifically, the output of the final convolutional layer) reveals that the network is capable of automatically extracting critical features from the input signal, such as change points, power levels, and usage durations of appliances. This sequence-based single-value prediction technique improves accuracy and is widely adopted as a baseline in many NILM studies [30].

Attention in the context of Machine Learning (ML) was first introduced for machine translation tasks [2]. Since then, it has become an essential component of neural architectures, with a wide range of applications such as multimedia content description (e.g., image caption generation, video description generation, speech recognition) [6], and computer vision (e.g., image classification, object detection) [38]. This mechanism enables the network to selectively focus on different parts of the input, emphasizing important information while downplaying less relevant details.

Attention mechanisms, especially Self-Attention (SA), have recently garnered interest for their ability to capture temporal and contextual relationships in power data. For example, [5] and [40] introduced SA to better model long-range dependencies, while [27] harnessed parallel multi-scale attention to combine global, local, and token-level features. Feed-Forward Attention (FFA) has also been employed, such as in [28], to enhance on/off state detection. In [25] attention modules have been integrated into both bidirectional LSTM and CNN-based S2P models, demonstrating improved disaggregation accuracy for appliances with fixed operation cycles. Their approach highlights the benefits of attention in focusing on key temporal regions, particularly in CNN architectures where the attention-enhanced S2P model achieved over 10% error reduction. In [35], the authors propose MSANet, a sequence-to-sequence model that replaces full self-attention with dilated-window attention to mitigate the quadratic complexity, using three parallel branches to capture multi-scale features. MSANet has been shown to accomplish MAE and F1 improvements over CNN+, LSTM+, and BERT4NILM. Recently NILM-LANN, a lightweight CNN with a Squeeze and Excitation (SE) attention block achieved competitive classification accuracy while targeting edge deployment [23]. Other methods incorporate attention alongside multi-task learning [9] or embed it into transformers [26].

Although attention mechanisms have proven useful across many NILM studies, most publications explore only a single variant and do not examine alternative designs. Consequently, the literature contains few direct comparisons among different attention types. To the best of our knowledge, no previous work has methodologically evaluated the effects of multiple attention families and their common extensions in order to measure their individual impact. This research gap leaves practitioners with limited guidance when selecting the most suitable attention mechanism for a particular appliance or deployment scenario.

3 Attention Mechanisms

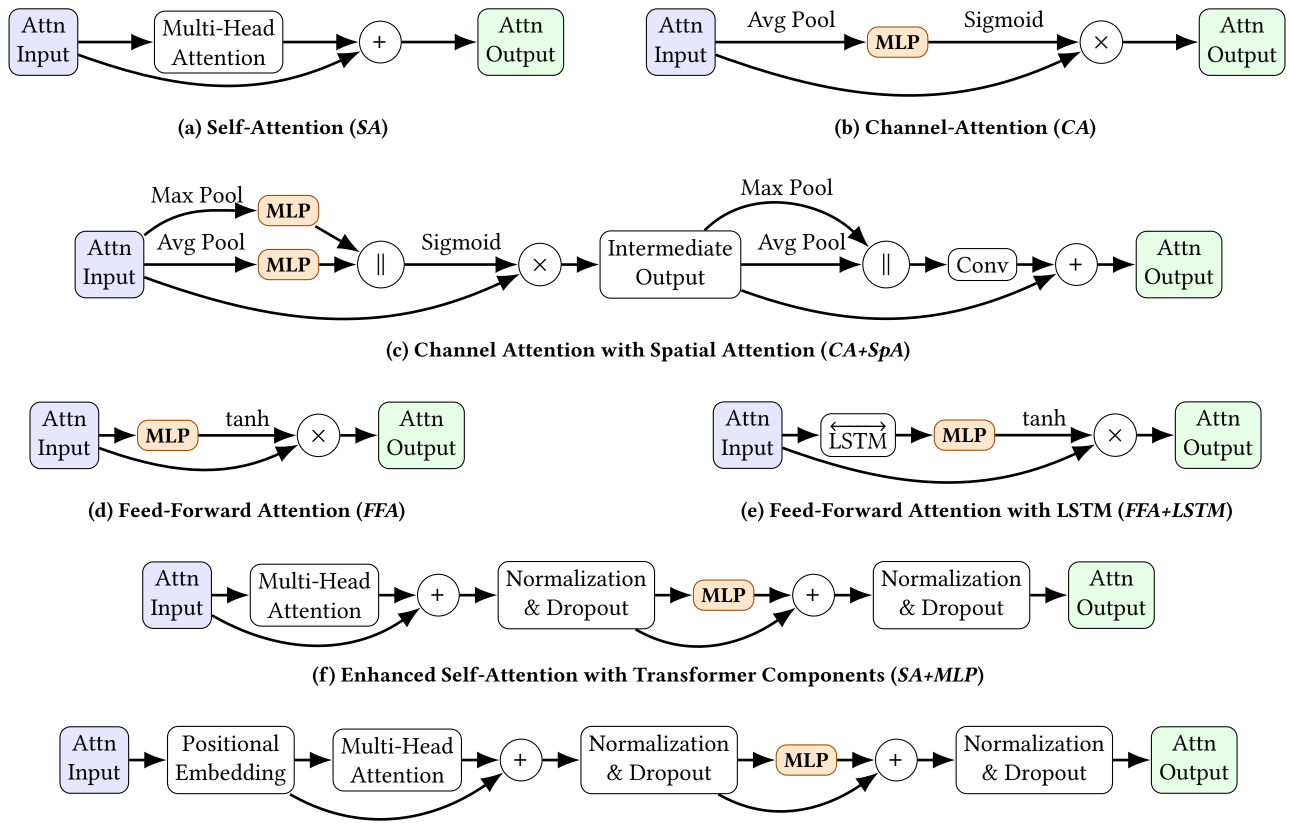
In this work, we explore three distinct attention mechanisms and systematically integrate and evaluate them on the S2P baseline model by integrating selected types of attention blocks between the last convolutional layer and the first dense layer, as shown in Fig. 1. This design follows the principle of *local precedes global*: Convolutional layers first extract local features, followed by attention mechanisms that emphasize relevant patterns through contextual weighting. The baseline S2P model consists of five 1D convolutional layers for feature extraction, followed by dense layers for prediction. Our extensions insert attention modules after the convolutional stack to enhance feature processing.

Our selection of the attention mechanisms is driven by two complementary criteria: architectural alignment with the convolutional backbone of the S2P model and empirical success in prior NILM and sequence-modeling research. Channel Attention (CA) dynamically reweights convolutional feature maps, which directly complements the spatial filtering already present in S2P [15, 39]. FFA offers a lightweight temporal focus without recurrent state to preserve long-term dependencies [2, 28, 29]. And finally SA aggregates global context through multi-head interactions and has recently advanced state-of-the-art performance in NILM tasks [37, 40]. Together, the trio provides a balanced comparison that covers channel-wise, window-wise, and sequence-wise perspectives, allowing us to study how each inductive bias affects load-disaggregation accuracy and efficiency.

3.1 Channel Attention (CA)

The core of CNNs lies in the convolutional operation, which extracts useful features by combining spatial and channel information. However, relying solely on convolutional operations may not fully exploit the information available in each channel. To address this issue, CA has been introduced, which dynamically adjusts the importance of each channel, allowing the network to better utilize the information across different channels [15]. This makes it a natural choice for improving the feature extraction capabilities of our S2P baseline, since it heavily relies on convolutional layers for feature extraction.

CA is based on a Squeeze and Excitation (SE) block, which *squeezes* the global spatial information using a pooling operation and *excites* the channel information by learning features along the channel dimension and establishing interdependency relationships between channels. An extended version also includes spatial attention [39], by allocating attention weights to specific positions within the input, capturing both channel-wise and spatial dependencies. Based on channel attention, the hybrid Convolutional Block Attention Module (CBAM) has been proposed in [39]. Given an intermediate feature map, the CBAM can sequentially generate attention maps along two distinct dimensions (channel and spatial), before subsequently multiplying the feature maps with the input feature map for adaptive feature refinement. Compared to the SE block, the spatial attention map is intended to extract useful information not only from the channel perspective, but also from the spatial perspective. Through adapting these vision-focused mechanisms to 1D energy sequences by treating convolutional filters as



(a) Self-Attention (*SA*)

(b) Channel-Attention (*CA*)

(c) Channel Attention with Spatial Attention (*CA+SpA*)

(d) Feed-Forward Attention (*FFA*)

(e) Feed-Forward Attention with LSTM (*FFA+LSTM*)

(f) Enhanced Self-Attention with Transformer Components (*SA+MLP*)

(g) Enhanced Self-Attention with Positional Encoding (*SA+MLP+PE*)

Figure 2: Overview of the seven considered attention blocks that were inserted into the S2P model. As shown in Figure 1, the *Attn Input* is the feature tensor produced by the last convolutional layer (i.e., an intermediate representation), and the *Attn Output* is passed to the dense layer. The || symbol indicates tensor concatenation, + and × signify element-wise residual addition and multiplication of two tensors, respectively.

channels, we have implemented two channel-focused variants for their use in NILM:

- *CA* (Fig. 2b): Applies Squeeze and Excitation (SE) blocks [15] to the convolutional outputs of the S2P model. Treats convolutional filter outputs as *channels*, using global average pooling and learned channel weights to emphasize discriminative features like high-frequency appliance signatures.
- *CA+SpA* (Fig. 2c): Extends *CA* with 1D spatial attention from CBAM [39]. Applies max and average pooling along channels followed by 7-width convolution to highlight critical temporal regions (e.g., microwave oven operation).

The motivation of inserting SE blocks directly after the final convolutional layer is illustrated in Fig. 3. As visible in the figure, for a representative fridge trace from the unseen test house, the baseline S2P substantially under-estimates both the edge surges and the steady-state consumption, while it overshoots during the cool-down phase. Adding a lightweight SE unit allows the network to re-weight informative feature maps on-the-fly, which sharpens the detection of the compressor start, reduces the steady-state offset, and shortens the lag when the appliance turns off. Qualitatively, the SE-augmented model tracks the ground-truth signal much more closely and avoids several spurious peaks. These observations confirm that channel-wise recalibration is an effective, low-cost means of improving feature exchange across filters before the dense head, and they motivate the two CA variants.

3.2 Feed-Forward Attention (FFA)

FFA is a simplified version of Bahdanau's Attention [2], which is an attempt to regain lost information during the compression step of variable-length inputs into fixed-length vectors in the RNN-based encoder-decoder framework for machine translation. Specifically, as the length of the input sentence increases, earlier information is more susceptible to being lost than more recent information, leading to a significant drop in translation accuracy. To address this limitation, Bahdanau et al. introduced a set of vectors that represent different parts of the sequence rather than relying on single fixed-length vector to summarize the entire input sequence. This approach enables the model to dynamically select and combine components of these vectors based on their relevance to the current

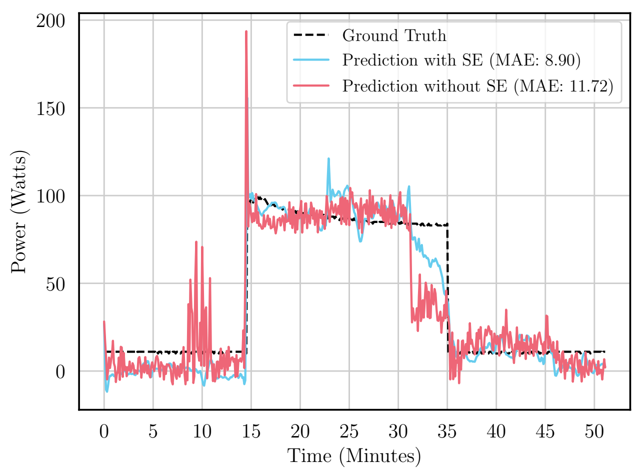


Figure 3: Prediction of S2P with and without the added SE units along with the ground truth for the fridge appliance from house 2 in the UK-DALE dataset. The MAE is reported in Watts.

decoding step. The core idea of the attention mechanism is to calculate attention weights by measuring the similarity of the hidden states between the decoder and the encoder. A higher similarity score results in a greater weight, allowing the model to emphasize specific parts of the input sequence. This enables the decoder to adaptively focus on the most relevant parts of the source sentence at each step.

In the context of solving the load disaggregation problem, the sequence length is fixed. The aggregated power consumption is aligned with the load of the corresponding appliance, and the alignment is known ahead of time. Following the regression branch of Load Disaggregation with Attention (LDwA) [28], we insert a lightweight temporal attention block over the convolutional features. The block produces a context vector by a weighted aggregation of features, emphasizing informative windows. In *FFA+LSTM*, the same mechanism is applied to bidirectional LSTM features instead of purely convolutional features. It was demonstrated that incorporating this attention mechanism into the encoder-decoder framework improved the model's generalization ability, which confirms its effectiveness in improving the performance of NILM models [28]. Thus, we integrate FFA into the S2P model through two variants:

- *FFA* (Fig. 2d): Integrating FFA [2, 29] into S2P's convolutional outputs rather than RNN states. Uses position-wise two-layer Multilayer Perceptron (MLP) with *tanh* activation to weight critical windows (e.g., kettle boiling phases), enabling focused temporal attention without sequence recursion.
- *FFA+LSTM* (Fig. 2e): Hybrid architecture adding bidirectional LSTM cells before FFA. Combines CNN's local feature extraction with LSTM's temporal memory to model appliance state transitions, inspired by [28].

3.3 Self-Attention (SA)

SA was originally proposed as part of the *transformer* architecture for machine translation [37] and has since been adopted in numerous application domains, including NILM [5, 40]. Its key idea is to learn dependencies within the same input sequence by computing attention weights over all positions, allowing each element to incorporate contextual information.

Concretely, this mechanism derives three distinct matrices: Query, Key, and Value. The term *self*-attention stems from the fact that all matrices are obtained by applying learned projections to the same input sequence. Each input attends to all other elements in the input sequence, including itself, when computing its attention weights. This allows the model to incorporate the context of the entire sequence. Additionally, the concept of *Multi-Head Attention* replicates this process in parallel over multiple subspaces, allowing the model to capture various relationships within the input. In practice, positional encodings are often added to the input to preserve ordering information [37]. By combining contextual focus, multi-head diversity, and positional awareness, SA proves highly effective for sequence applications like load disaggregation.

Our transformer-inspired variants include:

- *SA* (Fig. 2a): Standard multi-head self-attention [37] with 2 heads, capturing global temporal relationships.
- *SA+MLP* (Fig. 2f): Enhances *SA* with transformer components including layer normalization, skip connections, and an MLP (200 hidden units) for non-linear feature transformation.
- *SA+MLP+PE* (Fig. 2g): Adds sinusoidal positional encoding to *SA+MLP*, providing explicit sequence order information absent in standard SA.

3.4 Summary

The seven considered attention variants enable the systematic comparison of the impact of attention in the context of NILM: channel vs. spatial vs. temporal focus, additive vs. transformative processing, and explicit vs. implicit position handling. All attention outputs are combined with skip connections to preserve original features. Hyperparameters follow original proposals to maintain comparability. It must be remarked, though, that none of them has been specifically designed with energy disaggregation as the use case. We seek to address this gap in the research landscape by not only assessing the potential of one attention mechanism in the context of NILM, but rather comparatively evaluate the seven possible realization of attention within the used neural networks. By doing this in conjunction with the NILM performance evaluation for five widely used appliance types, this allows us to make generalizable statements about the situations in which attention is worth the additional computational efforts, and by what degree it improves the overall accuracy.

4 Evaluation Setup

We evaluate our attention variants on two public datasets: UK-DALE and REDD. The United Kingdom Domestic Appliance-Level Electricity (UK-DALE) dataset contains power measurements from five households sampled at 6 s resolution [21]. The Reference Energy Disaggregation Data Set (REDD) contains data from six houses in the US sampled every 3 s [22]. We have resampled all channels to

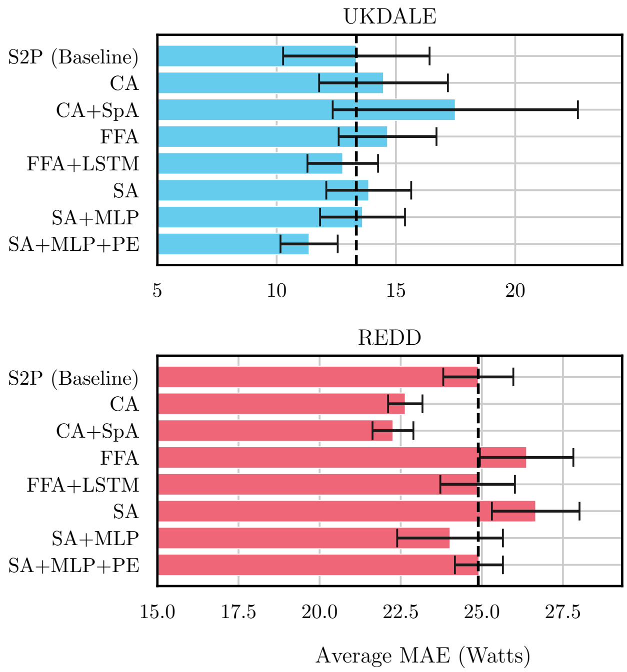


Figure 4: Average Mean Absolute Error (MAE) across the tested appliances available in each dataset (four per dataset). The dashed vertical line marks the S2P baseline average. Bars show the mean Mean Absolute Error (MAE) value across three independent runs with different initial random seeds. Error bars indicate the average standard deviation. Lower is better.

6 s for direct comparability. As of 2020, REDD and UK-DALE are among the top three cited datasets in the context of NILM [18].

The disaggregation performance is known to vary by appliance type [32]. In light of the fact that not all types of electrical appliances are present in all datasets, we focus our analysis on five widely used target appliances. The microwave oven is a device with short operational durations on the order of seconds to minutes, but notable power consumption during its activity. The fridge features activity phases on the order of tens of minutes, which moreover occur periodically, but exhibits a lower power demand than the microwave oven. Washing machines follow operational sequences for durations of several tens of minutes, with starkly varying power needs, but are only being used infrequently. In addition to these three types, we also consider the water kettle (only present in UK-DALE, characteristics similar to the microwave oven) and dishwasher (only present in REDD, characteristics similar to the washing machine) appliances to corroborate the impact of attention on these appliance characteristics. This appliance selection not only covers a representative sample of typical household appliances, but also allows for a direct comparability to related works, such as [28, 40, 41].

For each target appliance we follow the standardized train–test split from prior work [41]: in UK-DALE houses 1, 4, and 5 are used for training and house 2 for testing. House 3 is excluded due to its short recording duration (39 days) and limited appliance coverage (only 4 metered loads). In REDD we train on all houses except house 1, which serves as the test set. Within the training portion of each house we hold out the final 20% of the timeline as a validation set (chronological split). Training uses early stopping on validation Mean Squared Error (MSE) with a patience of 10 epochs and a cap of 25 epochs. The checkpoint with the lowest validation loss is used for testing. Each input window contains 599 samples (≈1 h in real-time), which are processed by five 1D convolutional layers followed by the respective attention module (see Fig. 1). All models are trained with batch size 512 using MSE loss. All experiments use PyTorch on NVIDIA A100 GPUs.

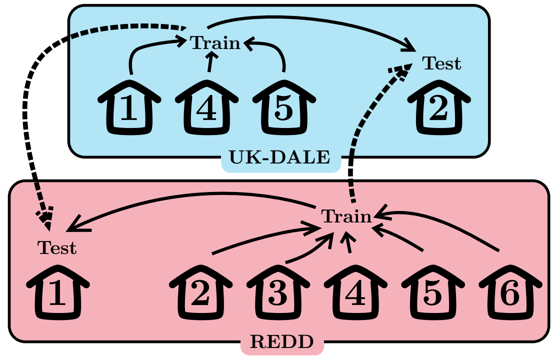


Figure 5: Training and testing settings in both within-dataset and cross-dataset settings. Dashed arrows indicate the cross-dataset scenario.

Each experiment is repeated three times with three different initial seeds. All reported error metrics are the mean and standard deviation across these three runs. Although all runs were executed on the same hardware, wall-clock training times may vary slightly because the GPU server is a shared resource. In addition to accuracy, we report parameter counts, theoretical GFLOPs per forward pass for a 599 sample window, and average training time. These quantities underpin our design guidance with practical cost figures.

To assess robustness under domain shift, we conduct cross-dataset (out-of-distribution) evaluations in both directions by evaluating the models trained on the training houses of one dataset on the test house of the other, as illustrated in Fig. 5. No fine-tuning, hyper-parameter re-tuning, or normalization recalibration is performed on the target dataset. For direct comparability, we restrict the computations of the accuracy in the cross-dataset scenario to those appliances that are present in both datasets, i.e., washing machine, fridge, and microwave. We compute the Mean Absolute Error (MAE) per appliance on the target test house and report the mean across all three models for each appliance as a cross-dataset score. This protocol probes generalization beyond cross-house variation and captures real distribution shifts between datasets. It therefore complements the within-dataset results by revealing which attention families are stable under covariate and label-noise shifts.

5 Evaluation Results

We initially report averages of our evaluation runs for the sake of completeness, and subsequently interpret per appliance and per dataset patterns, since the same attention module can lead to improvements in one context but reduce performance in another. We also examine the computational cost, because practical deployments often face budget constraints.

5.1 Within-Dataset Results

The results for both datasets are summarized in Table 1 per appliance and in Fig. 4 as dataset-level averages along with their standard deviations.

5.1.1 UK-DALE. When considering the averages overall (cf. Fig. 4), the enhanced SA variant with positional encoding (*SA+MLP+PE*) attains the best result, reducing the MAE by 15% on average. The gains are concentrated on transient loads, however. Microwave (−45%) and kettle (−20%) experience significant improvements, whereas the error even increases slightly when disaggregating fridge and washing machine signals. The best model for fridge disaggregation is the SA variant without positional encoding (*SA+MLP*) (−4.6% error), and the best attention type for washing machine disaggregation is *FFA+LSTM* (−5%).

Even though the operation patterns of microwave ovens and kettles are similar in principle (short operational duration at high power consumption), plain *FFA* is very effective on microwaves (43% lower MAE), but fails to improve the disaggregation of the kettle (80% higher MAE). This underscores its sensitivity to load dynamics.

At long last, the results show that channel-attention variants are generally weaker on UK-DALE: The use of plain *CA* results in MAE values that are +9% higher than the baseline, and *CA+SpA* even increases these errors to +31%.

5.1.2 REDD. The disaggregation error distribution is almost reversed when applying attention mechanisms to the REDD dataset. Channel-attention generalizes best on average. *CA+SpA* yields the lowest mean MAE (−11% compared to the baseline on average), with notable gains for the fridge (−23%) and washing machine (−11%), and a modest improvement on the microwave (−4%). *CA* in its plain variant can achieve even better results for the washing machine (−14%). Its usage does, however, not lead to improvements at all when disaggregating the (operationally similar) dishwasher appliance, where *SA+MLP* performs best (−4%).

It is also noteworthy that plain *FFA* and plain *SA* increase the dataset-average error by around 6% and 7%, respectively, indicating that their inductive biases are less robust to REDD's distribution characteristics. An example of a microwave window with a short, high-amplitude spike is shown in Figure 6. *SA* and *CA* suppress off-periods from S2P and focus on the active segment, yielding the lowest per-window error, whereas plain *FFA* under- or over-shoots the transient, despite giving less attention to the surrounding background signal.

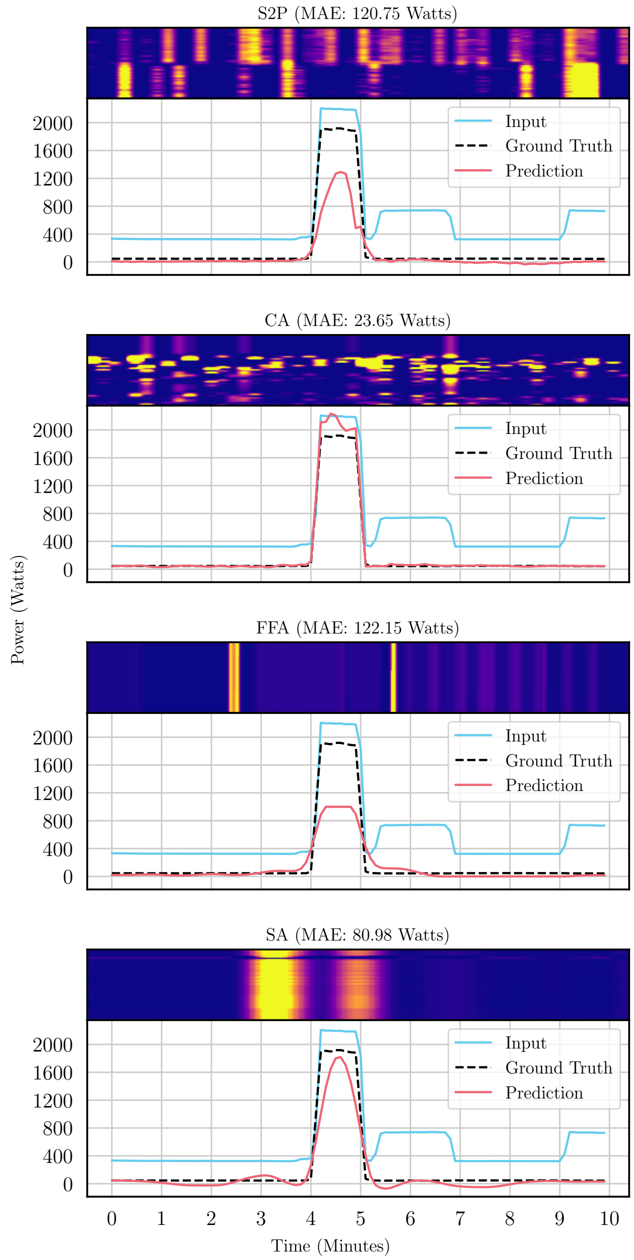


Figure 6: Qualitative attention analysis on a sample window for a microwave from REDD between the baseline S2P and plain attention variants *CA*, *FFA*, and *SA*. The heatmap over each panel indicates per-time attention/activation. The MAE printed in each title refers to this window only (dataset-level averages appear in Table 1).

5.2 Cross-Dataset Transfer

We next test models trained on UK-DALE against the REDD test house and vice versa, as shown in Fig. 5. Results are visualized in Fig. 7. They show that when applying UK-DALE trained models to REDD, *FFA* yields the lowest MAE, closely followed by *FFA+LSTM*. When training models on REDD and applying them to UK-DALE, *CA+SpA* generalizes best, and plain *CA* comes second. Pure *SA* does not lead to improvements in either direction, and while adding positional encoding (*SA+MLP+PE*) helps in-distribution, it did not improve the cross-dataset results in a statistically significant way.

Table 1: Performance comparison of the S2P baseline and attention-augmented variants on the UK-DALE and REDD test houses. MAE values are given as mean ± standard deviation over three independent runs with different random seeds. Lower values are better. Highlighted values indicate the best method per column.

Model	UK-DALE				REDD			
	Washing Machine	Fridge	Microwave	Kettle	Washing Machine	Fridge	Microwave	Dishwasher
S2P (Baseline)	13.21 ± 1.25	16.71 ± 0.4	14.27 ± 10.13	9.18 ± 0.5	28.28 ± 1.0	30.77 ± 2.97	18.39 ± 0.12	22.13 ± 0.22
CA	16.17 ± 2.62	17.19 ± 0.63	13.26 ± 6.64	11.32 ± 0.89	**24.39 ± 0.44**	25.71 ± 1.47	18.01 ± 0.16	22.45 ± 0.07
CA+SpA	13.5 ± 1.16	16.97 ± 0.81	29.94 ± 17.52	9.57 ± 1.05	25.18 ± 1.66	**23.58 ± 0.38**	**17.7 ± 0.15**	22.56 ± 0.34
FFA	16.62 ± 1.94	17.28 ± 1.4	8.18 ± 3.32	16.53 ± 1.53	28.89 ± 1.37	30.57 ± 3.12	22.27 ± 0.72	23.78 ± 0.53
FFA+LSTM	**12.55 ± 1.06**	16.62 ± 0.77	8.43 ± 3.28	13.48 ± 0.81	31.2 ± 3.28	24.52 ± 0.69	19.61 ± 0.18	24.16 ± 0.45
SA	14.67 ± 1.56	16.72 ± 0.39	15.03 ± 5.04	9.04 ± 0.11	33.88 ± 1.66	31.55 ± 1.67	19.76 ± 1.24	21.43 ± 0.84
SA+MLP	15.12 ± 2.73	**15.94 ± 0.28**	11.73 ± 2.12	11.59 ± 1.98	25.1 ± 2.66	30.12 ± 2.52	19.64 ± 0.61	**21.21 ± 0.74**
SA+MLP+PE	13.51 ± 0.82	16.8 ± 0.63	**7.78 ± 2.66**	**7.37 ± 0.7**	30.59 ± 0.73	28.72 ± 1.78	17.76 ± 0.19	22.57 ± 0.27

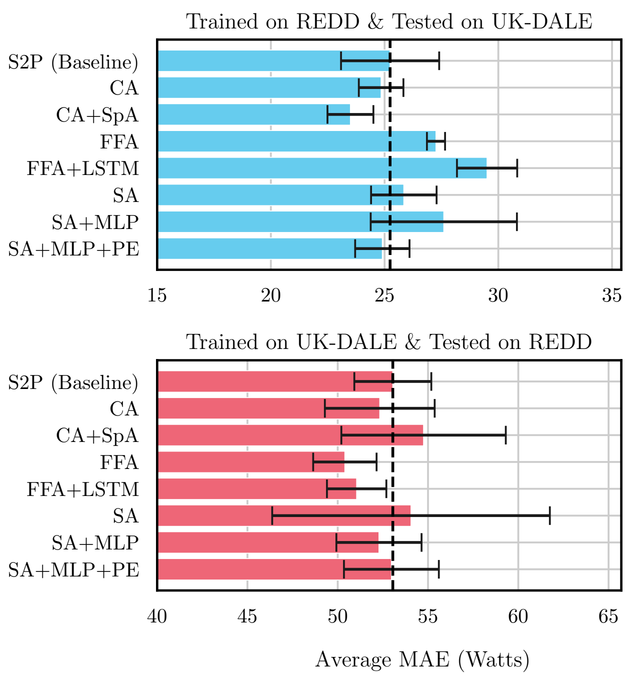


Figure 7: Cross-dataset transfer performance (out-of-distribution) as measured by the MAE in Watts over the three appliances common to both datasets: washing machine, fridge, and microwave.

The practical takeaway is to treat attention selection as a domain-sensitive hyperparameter: Choose the family according to the expected target distribution and validate the transfer explicitly.

When comparing the within-dataset and cross-dataset outcomes in Figures 4 and 7, it appears that *FFA* is beneficial when transferring from the more heterogeneous UK-DALE to REDD, whereas *CA+SpA* is advantageous in the opposite direction. Despite these relative rankings, absolute cross-dataset errors remain markedly higher than within-dataset results, especially when transferring from UK-DALE to REDD. We attribute this to the differences in appliance mix, state-change frequency, background loads, and noise characteristics between datasets. Closing this gap likely requires architectures and training techniques explicitly optimized for domain transfer. For example, transformer-based NILM models coupled with transfer learning using sinusoidal positional encoding, principled source-task selection, and targeted fine-tuning have been reported to improve performance and cross-domain generalization while reducing labeled-data needs [33]. These techniques are, however, orthogonal to our goal of isolating attention effects within the S2P model.

5.3 Enhancements of Within-Family Attention Variants

To isolate the effect of each enhancement inside an attention family, we report the change in MAE relative to the plain variants of that family (i.e., the S2P model augmented only with the CA, FFA, or SA block) in Table 2. As before, negative values indicate improvements. In general, all extensions show on average improvement compared to the plain versions. Distinct patterns stand out for each dataset, however.

5.3.1 UK-DALE. Adding positional encoding to the self-attention method (*SA+MLP+PE*) gives the largest within-family gain on average and is the underlying reason for the significant reduction of the disaggregation error for the microwave, while *SA+MLP* alone caters to the better disaggregation of the fridge, yet simultaneously increases the MAE for the kettle.

5.3.2 REDD. Here, *SA+MLP* is the most consistent enhancement inside its family, and *FFA+LSTM* improves the disaggregation performance for fridge and microwave, but reduces accuracy when looking at the washing machine. Spatial attention on top of channel attention (*CA+SpA*) is unstable on UK-DALE (due to its impact on the microwave), yet improves REDD fridge disaggregation performance.

Across families, the direction of the within-family MAE changes largely matches what each enhancement is intended to contribute. Making SA order sensitive through positional encodings (*SA+MLP+PE*) sharpens time localization around the prediction point, which fits short high-power bursts such as microwaves and kettles, while giving little or mixed change on fridges and washing machines where long plateaus dominate. Adding the transformer MLP and normalization (*SA+MLP*) enriches feature mixing and can stabilize context, which helps signals with more regular segments such as fridges, yet it may distract the model on very sharp on-off events like many kettle activations. Extending FFA with recurrent memory (*FFA+LSTM*) can exploit dependencies that span multiple samples and can smooth noisy activations, which explains the gains on appliances with sustained phases or predictable transitions such as fridges and washing machines, but it can also spread attention over too many steps and reduce focus on a single spike, which yields mixed outcomes on microwaves and dishwashers. Appending a spatial attention to CA (*CA+SpA*) is designed to highlight the active region within the window and can aid loads with structured backgrounds such as fridges, yet for extremely brief or highly variable spikes, as seen in many microwave windows, the spatial map can conflict with channel weights and either suppress the event or amplify noise. In summary, although no enhancement dominates across all appliances, positional terms tend to help short transients and recurrent memory tends to help extended or phase-driven behavior.

5.4 Resource Footprint and Computational Overhead

We have tabulated the resource footprints of the considered methods in Table 3. *SA* variants increase the required GFLOPs by 2× to 2.2× and training time by around 4× with only about 5% model-size growth. This is due to the large number of operations required by the multi-head attention block to compute the three attention matrices, which account for more than 45% of the additional computation. *FFA+LSTM* doubles the number of trainable parameters from 34.34 M to 71.67 M, and also nearly doubles the amount of computation required due to the added LSTM layer, which increases the post-convolution parameter count (from 12.6 k to 40.8 k). *CA* and *CA+SpA* add only around 4% parameters and modest runtime overheads. In short, the most accurate UK-DALE choice (*SA+MLP+PE*) is compute-hungry, yet REDD's best average (*CA+SpA*) is comparatively lightweight.

5.5 Summary and Limitations

Our results point to a clear interaction between appliance dynamics and the inductive bias of each attention family. Transient, high-power bursts (microwaves, kettle) benefit most from mechanisms that encode temporal focus with positional context. Hence, *SA+MLP+PE* and *FFA+LSTM* were observed to perform best on UK-DALE. In particular, *SA+MLP+PE* provided the most consistent gains across transients (cf. Table 1). In contrast to this, loads with repetitive or quasi-steady patterns (fridge) respond well to either modest global context (*SA+MLP* on UK-DALE) or channel re-calibration augmented by spatial weighting (*CA+SpA* on REDD). For complex, multi-state cycles (washing machine, dishwasher), improvements are moderate in all considered cases: the best UK-DALE washing-machine result (*FFA+LSTM*) comes at a greatly increased memory and computational demand, while on REDD the dishwasher benefits only marginally from *SA+MLP*. Taken together, this evidence supports our main claim: Attention helps in many cases, but its benefits are context dependent and must be balanced against cost.

Keeping in mind that our results are applied with 6 s sampling and a 599 sample window, as is the default configuration for the S2P implementation [41], different sampling rates or window sizes may shift the tradeoffs among attention types. We also kept most hyper-parameters fixed to isolate attention effects. Thus, even modest tuning could further improve some variants (e.g., head counts in SA or MLP width in the various variants). In summary, however, our results have been able to effectively demonstrate that many attention types come at a modest computational overhead, and lead to measurable improvements of NILM accuracy in many cases.

Table 2: Within-family MAE deltas in Watts relative to the plain attention variant. Negative values indicate lower error. WM = Washing Machine, Fr = Fridge, Mw = Microwave, Ke = Kettle, Dw = Dishwasher, Avg = Average.

(a) UK-DALE

Variant vs. Plain	WM	Fr	Mw	Ke	Avg
CA+SpA − CA	−2.67	−0.22	+16.68	−1.75	+3.01
FFA+LSTM − FFA	−4.07	−0.66	+0.25	−3.05	−1.88
SA+MLP − SA	+0.45	−0.78	−3.30	+2.55	−0.27
SA+MLP+PE − SA	−1.16	+0.08	−7.25	−1.67	−2.50

(b) REDD

Variant vs. Plain	WM	Fr	Mw	Dw	Avg
CA+SpA − CA	+0.79	−2.13	−0.31	+0.11	−0.38
FFA+LSTM − FFA	+2.31	−6.05	−2.66	+0.38	−1.51
SA+MLP − SA	−8.78	−1.43	−0.12	−0.22	−2.64
SA+MLP+PE − SA	−3.29	−2.83	−2.00	+1.14	−1.75

Table 3: Resource footprint of the baseline S2P model and attention-augmented variants. Reported are model size in MB, theoretical compute overhead per forward pass in GFLOPs for a single window, parameter count in millions, and average training time in hours. Training time is averaged across the three runs for each appliance-specific model. Lower values indicate better efficiency.

Model	Size (MB)	GFLOPs	Params (M)	Time (h)
S2P (Baseline)	131.01	30.43	34.34	2.90
CA	136.68	31.63	35.82	3.10
CA+SpA	136.68	31.64	35.83	3.40
FFA	136.69	32.55	35.83	3.30
FFA+LSTM	273.40	53.58	71.67	6.90
SA	136.72	60.23	35.84	9.50
SA+MLP	136.80	67.57	35.86	11.90
SA+MLP+PE	137.75	67.57	35.86	12.00

6 Conclusion

The integration of attention into neural networks has shown outstanding results in many application areas of artificial intelligence. Still, direct comparisons of different attention families in the context of electrical load disaggregation remain scarce. We have thus systematically integrated seven attention blocks from three families into the common S2P disaggregation model and evaluated them for five appliance types across two public datasets, including cross-dataset transferability analyses. Our principal contribution is a controlled, structural comparison that isolates the effect of attention modules inside a unified evaluation framework. Three main insights emerged: (i) Attention often helps to achieve a better disaggregation performance, but the efficacy of the considered attention mechanisms is strongly dependent on the load dynamics. On UK-DALE, transformer-style Self-Attention (SA) that also encodes the position of each time step achieved the lowest average error, with particularly large gains for short, high-power devices such as microwaves and kettles, reducing the baseline error by about 15% on average (and by up to 45% for microwaves). (ii) Data matters. On REDD, a lightweight attention that reweights convolutional feature channels with a shallow spatial step achieves the best average error. On UK-DALE, however, this attention variant fails to reduce the disaggregation error and even increases it for appliances like microwave ovens. (iii) Efficiency tradeoffs are substantial. SA variants raise the required amount of computation and training time around 4×, and adding a recurrent layer even roughly doubles the number of trainable parameters. Our evaluation has shown that Channel Attention (CA) is the most economical option in both memory and runtime.

These findings clearly indicate that there is no optimal attention configuration. Rather, attention selection must be context- and cost-aware. Our in-depth analysis has shown where attention helps and provides immediate practical guidance: If the power consumption of transient-heavy appliances shall be disaggregated from a building's mains supply, the SA method with positional information can lead to improved results, even though they come at a higher computation demand. For model designers, our ablations suggest several choices that transfer across architectures: Add positional encodings when fine grained temporal localization within the input window is essential, for example short spikes, and omit them when long steady phases dominate. Use channel attention to recalibrate features cheaply in noisy settings, and consider a light spatial step only when the active region is well localized. Pair feed-forward attention with bidirectional LSTMs only when multi-step memory is necessary and the parameter increase is acceptable. In all cases, validate both accuracy and cost on the intended deployment distribution.

References

[1] Awadelrahman M. A. Ahmed, Yan Zhang, and Frank Eliassen. 2020. Generative Adversarial Networks and Transfer Learning for Non-Intrusive Load Monitoring in Smart Grids. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. 1–7. https://doi.org/10.1109/SmartGridComm47815.2020.9302933

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL] https://arxiv.org/abs/1409.0473

[3] Roberto Bonfigli, Andrea Felicetti, Emanuele Principi, Marco Fagiani, Stefano Squartini, and Francesco Piazza. 2018. Denoising autoencoders for Non-Intrusive Load Monitoring: Improvements and comparative evaluation. *Energy and Buildings* 158 (2018), 1461–1474. https://doi.org/10.1016/j.enbuild.2017.11.054

[4] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2021. An Attentive Survey of Attention Models. *ACM Trans. Intell. Syst. Technol.* 12, 5, Article 53 (Oct. 2021), 32 pages. https://doi.org/10.1145/3465055

[5] Kunjin Chen, Yu Zhang, Qin Wang, Jun Hu, Hang Fan, and Jinliang He. 2020. Scale- and Context-Aware Convolutional Non-Intrusive Load Monitoring. *IEEE Transactions on Power Systems* 35, 3 (2020), 2362–2373. https://doi.org/10.1109/TPWRS.2019.2953225

[6] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. *Trans. Multi.* 17, 11 (Nov. 2015), 1875–1886. https://doi.org/10.1109/TMM.2015.2477044

[7] Shuang Dai, Fanlin Meng, Qian Wang, and Xizhong Chen. 2024. DP2-NILM: A distributed and privacy-preserving framework for non-intrusive load monitoring. *Renewable and Sustainable Energy Reviews* 191 (2024), 114091. https://doi.org/10.1016/j.rser.2023.114091

[8] Suryalok Dash and N.C. Sahoo. 2022. Electric energy disaggregation via non-intrusive load monitoring: A state-of-the-art systematic review. *Electric Power Systems Research* 213 (2022), 108673. https://doi.org/10.1016/j.epsr.2022.108673

[9] Suryalok Dash and N. C. Sahoo. 2023. Attention-Based Multitask Probabilistic Network for Nonintrusive Appliance Load Monitoring. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–12. https://doi.org/10.1109/TIM.2023.3273663

[10] Department for Energy Security and Net Zero. 2024. *Energy Consumption in the UK 2024*. Technical Report. UK Government. https://www.gov.uk/government/statistics/energy-consumption-in-the-uk-2024

[11] Michele D'Incecco, Stefano Squartini, and Mingjun Zhong. 2020. Transfer Learning for Non-Intrusive Load Monitoring. *IEEE Transactions on Smart Grid* 11, 2 (2020), 1419–1429. https://doi.org/10.1109/TSG.2019.2938068

[12] Costinela Fortea, Dragos Sebastian Cristea, Monica Laura Zlati, Valentin Marian Antohi, Mihaela Neculita, Nicoleta Cristache, and Ioana Lazarescu. 2024. Evaluation of the effectiveness of energy sustainability measures through the dynamic energy consumption model. *Frontiers in Energy Research* Volume 12 - 2024 (2024). https://doi.org/10.3389/fenrg.2024.1383314

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

[14] S. Hosseini, N. Henao, S. Kelouwani, K. Agbossou, and A. Cardenas. 2019. A Study on Markovian and Deep Learning Based Architectures for Household Appliance-level Load Modeling and Recognition. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*. 35–40. https://doi.org/10.1109/ISIE.2019.8781186

[15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

[16] Patrick Huber, Alberto Calatroni, Andreas Rumsch, and Andrew Paice. 2021. Review on Deep Neural Networks Applied to Low-Frequency NILM. *Energies* 14, 9 (2021). https://doi.org/10.3390/en14092390

[17] International Energy Agency. 2025. *Electricity 2025*. Technical Report. IEA, Paris. https://www.iea.org/reports/electricity-2025 Licence: CC BY 4.0.

[18] Hafiz Khurram Iqbal, Farhan Hassan Malik, Aoun Muhammad, Muhammad Ali Qureshi, Muhammad Nawaz Abbasi, and Abdul Rehman Chishti. 2021. A Critical Review of State-of-the-Art Non-Intrusive Load Monitoring Datasets. *Electric Power Systems Research* 192 (March 2021), 106921. https://doi.org/10.1016/j.epsr.2020.106921

[19] Maria Kaselimi, Nikolaos Doulamis, Anastasios Doulamis, Athanasios Voulodimos, and Eftychios Protopapadakis. 2019. Bayesian-optimized Bidirectional LSTM Regression Model for Non-intrusive Load Monitoring. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2747–2751. https://doi.org/10.1109/ICASSP.2019.8683110

[20] 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* (Seoul, South Korea) *(BuildSys '15)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/2821650.2821672

[21] Jack Kelly and William Knottenbelt. 2015. UKDALE: The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes. *Scientific Data* 2, 1 (March 2015), 150007. https://doi.org/10.1038/sdata.2015.7

[22] J Zico Kolter and Matthew J Johnson. 2011. REDD: A Public Data Set for Energy Disaggregation Research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, Vol. 25. 59–62.

[23] Yanjing Lei, Zehui Feng, Xiangqing Lin, Di Cao, and Jiakai Zhang. 2024. NILM-LANN: A Lightweight Attention-based Neural Network in Non-Intrusive Load Monitoring. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 181–186. https://doi.org/10.1109/CSCWD61410.2024.10580201

[24] Nguyen Viet Linh and Pablo Arboleya. 2019. Deep Learning Application to Non-Intrusive Load Monitoring. In *2019 IEEE Milan PowerTech*. 1–5. https://doi.org/10.1109/PTC.2019.8810435

[25] Qi Liu, Jing Zhang, Xiaodong Liu, Yonghong Zhang, Xiaolong Xu, Mohammad Khosravi, and Muhammad Bilal. 2022. Improving wireless indoor non-intrusive load disaggregation using attention-based deep learning networks. *Physical Communication* 51 (2022). https://doi.org/10.1016/j.phycom.2021.101584

[26] Zhuojie Nie, Yongbiao Yang, and Qingshan Xu. 2022. An ensemble-policy non-intrusive load monitoring technique based entirely on deep feature-guided attention mechanism. *Energy and Buildings* 273 (2022). https://doi.org/10.1016/j.enbuild.2022.112356

[27] Guobing Pan, Haipeng Wang, Tao Tian, Yuhan Luo, Songdi Xia, and Qiyu Li. 2024. Research on non-intrusive load decomposition model based on parallel multi-scale attention mechanism and its application in smart grid. *Energy and Buildings* 312 (2024), 114210. https://doi.org/10.1016/j.enbuild.2024.114210

[28] Veronica Piccialli and Antonio M. Sudoso. 2021. Improving Non-Intrusive Load Disaggregation through an Attention-Based Deep Neural Network. *Energies* 14, 4 (2021). https://doi.org/10.3390/en14040847

[29] Colin Raffel and Daniel P. W. Ellis. 2016. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. arXiv:1512.08756 [cs.LG] arXiv preprint arXiv:1512.08756.

[30] Hasan Rafiq, Prajowal Manandhar, Edwin Rodriguez-Ubinas, Omer Ahmed Qureshi, and Themis Palpanas. 2024. A review of current methods and challenges of advanced deep learning-based non-intrusive load monitoring (NILM) in residential context. *Energy and Buildings* 305 (2024). https://doi.org/10.1016/j.enbuild.2024.113890

[31] Hasan Rafiq, Xiaohan Shi, Hengxu Zhang, Huimin Li, and Manesh Kumar Ochani. 2020. A Deep Recurrent Neural Network for Non-Intrusive Load Monitoring Based on Multi-Feature Input Space and Post-Processing. *Energies* 13, 9 (2020). https://doi.org/10.3390/en13092195

[32] Andreas Reinhardt and Christoph Klemenjak. 2020. How does Load Disaggregation Performance Depend on Data Characteristics? Insights from a Benchmarking Study. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (Virtual Event, Australia) *(e-Energy '20)*. Association for Computing Machinery, New York, NY, USA, 167–177. https://doi.org/10.1145/3396851.3397691

[33] Jing Rong, Cong Wang, Qiuzhan Zhou, Yunxue He, and Huinan Wu. 2025. Enhancing Non-Intrusive Load Monitoring through Transfer Learning with Transformer Models. *Energy and Buildings* 330 (March 2025), 115334. https://doi.org/10.1016/j.enbuild.2025.115334

[34] Pascal A. Schirmer and Iosif Mporas. 2023. Non-Intrusive Load Monitoring: A Review. *IEEE Transactions on Smart Grid* 14, 1 (2023), 769–784. https://doi.org/10.1109/TSG.2022.3189598

[35] Zihan Shan, Gangquan Si, Kai Qu, Qianyue Wang, Xiangguang Kong, Yu Tang, and Chen Yang. 2023. Multiscale Self-Attention Architecture in Temporal Neural Network for Nonintrusive Load Monitoring. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–12. https://doi.org/10.1109/TIM.2023.3271009

[36] Changho Shin, Sunghwan Joo, Jaeryun Yim, Hyoseop Lee, Taesup Moon, and Wonjong Rhee. 2019. Subtask gated networks for non-intrusive load monitoring. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 142, 8 pages. https://doi.org/10.1609/aaai.v33i01.33011150

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] Feng Wang and David M. J. Tax. 2016. Survey on the attention based RNN model and its applications in computer vision. arXiv:1601.06823 [cs.CV] https://arxiv.org/abs/1601.06823

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[40] Zhenrui Yue, Camilo Requena Witzig, Daniel Jorde, and Hans-Arno Jacobsen. 2020. BERT4NILM: A Bidirectional Transformer Model for Non-Intrusive Load Monitoring. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring* (Virtual Event, Japan) *(NILM'20)*. Association for Computing Machinery, New York, NY, USA, 89–93. https://doi.org/10.1145/3427771.3429390

[41] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-Point Learning With Neural Networks for Non-Intrusive Load Monitoring. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.11873