Pexformer: Robust Indoor Human Localisation via Patch-level Tokenisation and Semi-Permeable Attention

Baobing Zhang[1], Sehrish Rafique[1], Mohamad Reza Shahabian Alashti[1], Shadiya Alingal Meethal[1], Vignesh Velmurugan[1], Patrick Holthaus[1], Gabriella Lakatos[1], Angela Dickinson[1] and Farshid Amirabdollahian[1]

Robotics Research Group, University of Hertfordshire, Hatfield, United Kingdom
(b.zhang6, s.rafique, m.r.shahabian, s.alingal-meethal4, v.velmurugan, p.holthaus, g.lakatos, a.m.dickinson, f.amirabdollahian2)@herts.ac.uk

Abstract. Social robots are increasingly explored within Ambient Assisted Living (AAL) settings, to support people living alone in ways that preserve autonomy and dignity. To facilitate effective interactions and prompt assistance, such as proactive check-ins, safe navigation, or escalation when something appears wrong, knowing the person's location in the home is invaluable. Non-intrusive and exteroceptive sensors are often used to estimate a person's location in a house. However, existing data-driven methods often struggle with extreme class imbalance characterised by long-tail distributions of room occupancy and the inherent noise of sparse sensor triggers emerging from AAL settings. To address these challenges, this paper introduces *Patch-Excel-Transformer (pexformer)*, a novel architecture that adapts efficient computational paradigms from the tabular domain to time-series localisation tasks, to maintain accuracy with such sparse data. pexformer leverages *Patch-level Tokenisation* to effectively capture local temporal dynamics and integrates a *Semi-Permeable Attention (SPA)* mechanism to construct hierarchical feature interactions, reducing computational complexity while preserving critical information. Notably, we further observe that a simple *Random Permutation* of tokens acts as an effective regulariser and performs comparably to or better than Mutual Information (MI)-based ordering, avoiding the need for costly statistical pre-computation. Experiments on real-world smart home datasets confirm that pexformer achieves state-of-the-art localisation accuracy and *strong balanced performance, particularly for under-represented room categories,* without relying on complex oversampling techniques. From a social robotics perspective, this perceptual reliability is a prerequisite for proactive human-robot interaction (HRI). The results highlight the model's capability to substantially improve recognition in sparsely visited areas and validate its robustness in handling highly imbalanced heterogeneous sensor data. The code is available to the public at `https://github.com/baobingzhang/pexformer`.

Keywords: Indoor Localisation, Ambient Assisted Living, Heterogeneous Sensing, Class Imbalance, Semi-Permeable Attention, Social Robots

1 Introduction

Socially assistive robots (SARs) are increasingly being explored as part of care strategies that support older adults living alone, particularly where social isolation, chronic conditions, and frailty increase daily support needs [11]. They can be integrated in an environment-based Ambient Assisted Living (AAL) ecosystem, where the home contains a network of sensors and connected devices to support people living alone in ways that preserve autonomy and dignity [10]. This allows combining the embodied interaction of the robots with the contextual awareness provided by the sensors in the house. To be both socially competent and practically useful in the home, knowledge of the person's location augments the robot's social awareness helping to (i) navigate safely while respecting personal space, (ii) facilitate context-aware interaction [9], and (iii) make sense of daily routines and unusual patterns (e.g., prolonged inactivity in the bathroom versus the living room). Such locations are used in AAL activity monitoring systems, which have in turn been shown to help identify evolving trends in a person's daily activities, thereby supporting healthcare professionals in planning timely interventions and helping family members monitor wellbeing remotely [1, 2].

In contrast to privacy-intrusive cameras or compliance-dependent interactive wearables, non-intrusive ambient sensors (e.g., Passive Infrared (PIR) sensors, magnetic door contacts) have emerged as the preferred solution in social robotics and smart homes due to their low cost and privacy-preserving nature [18].

However, achieving precise and robust localisation using these sparse sensor data presents significant challenges. The primary obstacle is *extreme class imbalance*, characterised by the long-tail distribution of room occupancy in real-world scenarios (e.g., vastly more time spent in the Bedroom than in the Hallway or Storage room). Traditional Deep Learning (DL) approaches, such as Long Short-Term Memorys (LSTMs) and Convolutional Neural Networks (CNNs), often exhibit bias towards majority classes, resulting in poor performance in sparsely visited areas [14]. Furthermore, the *sparsity and noise* inherent in binary sensor triggers make it difficult for standard time-series models to capture long-range spatiotemporal dependencies. While Transformers have revolutionised Natural Language Processing and Computer Vision [17], directly applying heavy attention mechanisms to such low-dimensional binary signals often leads to overfitting or computational inefficiency.

Recently, in the domain of Tabular Data learning, novel Transformer architectures designed to surpass Gradient Boosted Decision Trees (GBDTs), such as *TabTransformer* [7] and *SAINT* [15], have demonstrated remarkable efficiency in feature interaction and robustness. Inspired by these advances, we propose a novel hypothesis: *Can the efficient computation paradigms of tabular attention be adapted to time-series localisation by treating temporal sensor windows as specialised "patches"?*

To this end, we introduce *Patch-Excel-Transformer (pexformer)*, a Transformer architecture tailored for non-intrusive indoor localisation. pexformer leverages *Patch-level Tokenization* to aggregate local temporal context and employs

a *Semi-Permeable Attention (SPA)* mechanism to construct hierarchical feature interactions. Departing from the conventional reliance on Mutual Information (MI) for feature prioritisation, we empirically observe that a simple *Random Permutation* strategy acts as an effective regulariser for time-series data, performing comparably to or better than MI-based ordering while avoiding costly statistical pre-computation. Experiments on real-world smart home datasets demonstrate that pexformer not only achieves higher results in overall accuracy but also delivers substantial improvements in *Macro-F1* performance, validating its strong capability in handling long-tail occupancy distributions.

We frame this work as an architecture-and-benchmarking study, rather than as the proposal of a fundamentally new attention mechanism. Our contributions are:

1. A cross-domain adaptation that transposes patch-level tokenisation and Semi-Permeable Attention from the tabular-data domain to sparse, heterogeneous ambient-sensor localisation by treating each sensor's temporal window as an atomic patch token.
2. An integration of the SPA mask into the ambient-sensing setting, which in our experiments reduces cross-sensor interference and improves recognition of under-represented room classes, without claiming a new attention primitive.
3. An empirical observation that, under our evaluation protocol, a simple Random Permutation of sensor tokens performs comparably to or better than MI-based ordering, avoiding the cost of statistical pre-computation; we interpret this effect only tentatively and leave a formal theoretical characterisation to future work.
4. A comprehensive benchmarking study against 11 cross-paradigm baselines and, on efficiency grounds, against four attention-based deep tabular baselines in terms of Params, FLOPs, latency, throughput, and memory, showing consistent improvements in Accuracy and Macro-F1 for imbalanced classes at a deployment-feasible inference cost.

The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 details the proposed model; Section 4 presents the experimental setup, results, and analysis; Section 5 discusses the implications for social robotics; and Section 6 concludes the study with future directions.

2 Related Work

This section reviews existing literature across three dimensions: DL-based indoor localisation, efficiency optimisation in time-series Transformers, and recent advances in tabular DL. We highlight how pexformer bridges the gaps in current methodologies through cross-domain adaptation.

2.1 Sensor-based Deep Indoor Localisation

Indoor localisation based on non-intrusive ambient sensors (e.g., PIR, door contacts) is a fundamental task in smart home environments. Early approaches primarily relied on probabilistic models (e.g., Hidden Markov Model (HMM)) or classical machine learning algorithms (e.g., Random Forest, Support Vector Machine (SVM)). While computationally lightweight, these methods suffer from feature engineering bottlenecks and fail to capture complex non-linear spatiotemporal patterns [18]. With the advent of DL, LSTMs and 1D-Convolutional Neural Networks (1D-CNNs) have become dominant, significantly improving localisation accuracy [14]. However, these models often falter when facing *extreme class imbalance*, tending to bias predictions towards frequently visited areas while neglecting critical low-frequency zones. Furthermore, extracting robust semantic features from sparse and noisy sensor streams remains an unaddressed challenge.

2.2 Time-Series Transformers and Computational Bottlenecks

Transformers have demonstrated superior performance in long-sequence modelling due to their global attention mechanism and are widely applied in time-series forecasting and classification [17]. Notable Transformer-based time-series models include *Informer* [22], which introduces ProbSparse attention for efficient long-horizon forecasting; *TimesNet* [20], which transforms 1D sequences into 2D tensors to capture multi-periodicity; and *PatchTST* [12], which segments time series into subseries-level patches for channel-independent processing. However, the quadratic computational complexity $O(L^2)$ of standard self-attention impedes their deployment on resource-constrained AAL edge devices. To mitigate this, various efficiency-oriented variants have been proposed: Kermani et al. [8] investigated structured pruning and quantisation for energy-efficient inference, while Götz et al. [6] utilised Token Merging strategies to reduce sequence length by combining adjacent tokens. Although these methods alleviate computational burden, they are typically designed for continuous numerical time series and fail to exploit the inherent *sparsity* and *event-driven* nature of binary sensor data in AAL settings. In contrast, pexformer employs a more targeted *Patch-level Tokenisation* strategy that treats each sensor's temporal window as an atomic token, effectively compressing temporal windows into semantic units for dimensionality reduction.

2.3 Deep Tabular Learning: A Cross-Domain Inspiration

Recent advances in *Tabular DL* offer a novel perspective for time-series analysis. To surpass traditional GBDTs (e.g., XGBoost), innovative architectures such as *TabR* [5] and *VisTabNet* [21] have emerged, prioritising high-order feature interaction over sequential dependency. While prior research has leveraged Semi-Permeable Attention and MI sorting to mitigate feature interference [3], pexformer builds upon the insight that short-term sensor trigger patterns effectively mirror the structural characteristics of tabular rows. By adapting this

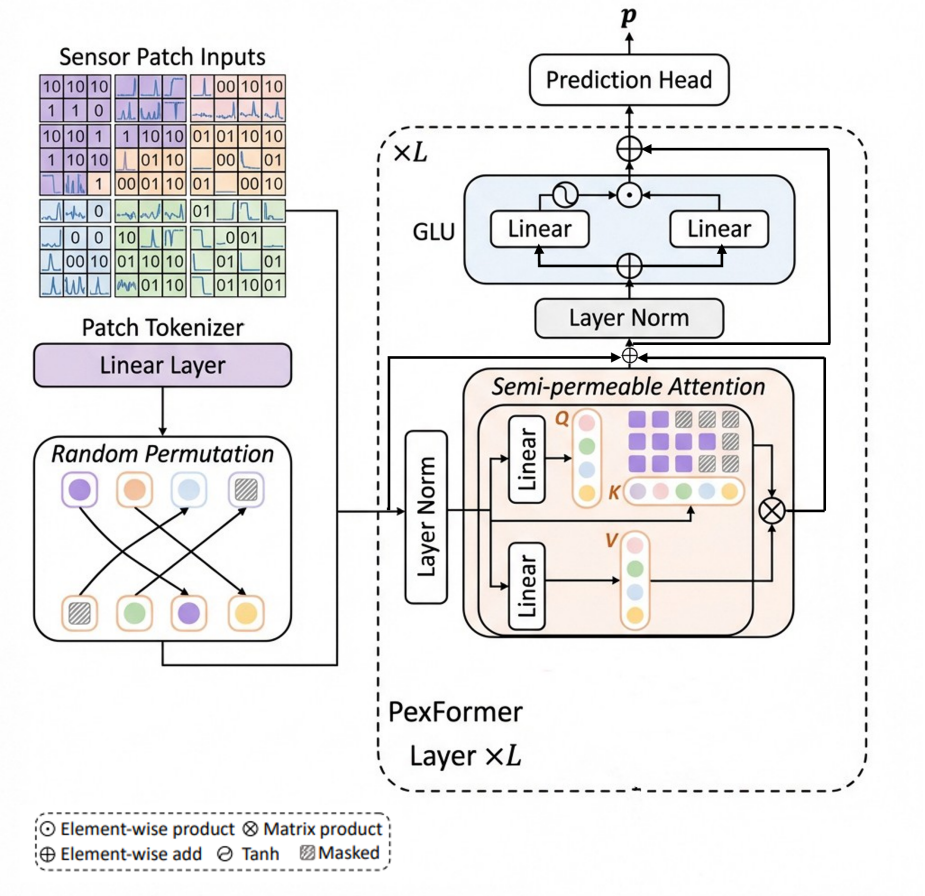

Fig. 1. The architecture of pexformer. The model takes sensor patches as input, tokenised via linear projection. A Random Permutation module shuffles feature order for regularisation. The backbone stacks SPA layers—utilising a mask on permuted tokens—and Gated Linear Units (GLU). A single prediction head outputs final probabilities.

efficient attention paradigm to the spatiotemporal localisation task, we observe that for temporal sensor data, Random Permutation acts as an effective regulariser, performing comparably to or better than information-theoretic sorting while avoiding costly pre-computation. This improves robustness on long-tail classes, yielding a higher Macro-F1 score, while maintaining computational efficiency.

3 Methodology

3.1 Problem Formulation

This study focuses on human localisation within AAL environments using ambient sensor data. Given a dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{X}_i \in \mathbb{R}^{S \times W}$ denotes the raw signals from S sensors collected within a sliding window of length W. $y_i \in \{1, \ldots, C\}$ represents the room location label. Our goal is to develop a mapping function $f : \mathbf{X} \to y$ that achieves high-accuracy classification for non-stationary tabular spatio-temporal features.

3.2 Overview of pexformer Architecture

We describe **pexformer** (Patch-Excel-Transformer) as a cross-domain adaptation rather than a new attention mechanism: it combines temporal patching techniques with the asymmetric interaction mechanism of ExcelFormer [**?**] and transposes them from numeric tabular inputs to heterogeneous ambient-sensor windows. To process longer contexts and capture complex sensor trajectories, pexformer uses a scaled-up backbone with an embedding dimension of $d = 256$ and $L = 6$ layers, and employs a Pre-Norm residual structure to enhance training stability in deeper configurations.

3.3 PatchVectorTokenizer

Unlike traditional point-wise embedding methods that process independent time steps, pexformer treats the entire window vector of each sensor as an atomic semantic unit through the **PatchVectorTokenizer**. For the j-th sensor, the token embedding $\mathbf{z}_j$ is computed as a direct linear projection:

$$\mathbf{z}_j = \mathbf{x}_j \mathbf{W}_p + \mathbf{b}_p, \quad \mathbf{z}_j \in \mathbb{R}^d \tag{1}$$

where $\mathbf{x}_j \in \mathbb{R}^W$ is the raw signal slice for sensor j, and $\mathbf{W}_p \in \mathbb{R}^{W \times d}$ is the learnable projection matrix. This approach effectively captures local dynamic patterns and trajectories within the sensor window.

3.4 Semi-Permeable Attention (SPA)

To break rotational invariance and suppress noise from uninformative features, the SPA module constrains information flow using a global lower-triangular mask matrix $\mathbf{M}$ applied over all S sensor tokens. Concretely, tokens attend only to tokens with equal or higher priority in the permutation order:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{d_{head}}}\right)\mathbf{V} \tag{2}$$

The elements M_{ij} are defined based on a feature priority function $\pi(\cdot)$:

$$M_{ij} = \begin{cases} 0, & \text{if } \pi(i) \geq \pi(j) \\ -\infty, & \text{if } \pi(i) < \pi(j) \end{cases} \tag{3}$$

In pexformer, the priority order π is determined via Random Permutation or MI (MI) sorting. When Random Permutation is used, a single permutation is sampled once before training under a fixed random seed and remains constant throughout training and inference, ensuring full reproducibility. Empirically, the Random-Permutation variant attains comparable or slightly better generalisation (93.30% accuracy) than the MI-based variant under our evaluation protocol; one possible interpretation is that it acts as an implicit regulariser in the spirit of random projection, though we do not claim a rigorous mechanistic account and a formal theoretical characterisation is left for future work.

3.5 Interaction Attenuated Initialisation (IAI)

To prevent blind coupling between features during early training stages, we apply the Interaction Attenuated Initialization (IAI) strategy to the weights $\mathbf{W}$ of all attention modules:

$$\mathrm{Var}(\mathbf{W}) = \gamma \cdot \mathrm{Var}_{\mathrm{Kaiming}}(\mathbf{W}) \tag{4}$$

By setting the scaling factor $\gamma = 10^{-4}$, the model initially operates with minimal feature interaction, subsequently allowing dependencies to evolve driven by the data distribution.

3.6 Gated Linear Feedforward Network (GLU)

To fit the irregular decision boundaries characteristic of tabular sensor states, pexformer utilises a **GLU**-based feedforward network instead of a vanilla MLP:

$$\mathrm{GLU}(\mathbf{z}) = (\mathbf{z}\mathbf{W}_1 + \mathbf{b}_1) \odot \tanh(\mathbf{z}\mathbf{W}_2 + \mathbf{b}_2) \tag{5}$$

where $\odot$ denotes element-wise multiplication. The gating mechanism, combined with the tanh activation, facilitates the modelling of sharp transitions in sensor activities.

3.7 Optional Post-hoc Logit Adjustment

To investigate whether explicit long-tail compensation can further benefit pexformer, we study a post-hoc Logit Adjustment strategy. During inference, the output logits may be shifted relative to the class prior probabilities π_c:

$$f'_c(\mathbf{x}) = f_c(\mathbf{x}) - \tau \log \pi_c \tag{6}$$

where τ is a tuneable factor that controls the strength of the adjustment. When $\tau = 0$, no adjustment is applied. As we show in Section 4, pexformer achieves its best overall performance at $\tau = 0$, suggesting that its architecture already provides intrinsic robustness to class imbalance.

As illustrated in Fig. 1, pexformer adapts a specialised tabular Transformer architecture for time-series data. Specifically, we introduce a Patch Tokeniser to capture local temporal dynamics and replace the mutual-information-based sorting with a Random Permutation mechanism, which we find acts as a robust regularizer for sensor patch tokens.

Fig. 2. The experimental environment of the **Robot House**. Left: The floor plan depicting the spatial layout of functional zones and sensor anchor distributions. Right: Real-world photographic views of the interior scenes and interactive robots.

Table 1. Systematic profile of the indoor localisation dataset highlighting spatiotemporal density and sensory infrastructure.

Core Metric	**Dataset Specifications**
Spatiotemporal Scale	8,067 samples
Sampling Density	1.6 Hz average state-update frequency
Spatial Anchors	21 Heterogeneous nodes (Motion, Door, Seat, etc.)
Feature Space	67-D raw vector (including spatial coordinates)
Semantic Labels	8 Refined localization categories
Primary Hotspots	Sofa Area (36.1%), Bedroom (24.0%), Kitchen (12.1%)
Data Fidelity	100% Validity (0.00% missing rate for labels)

4 Experiments and Result Analysis

4.1 Experimental Setup

Experimental environment and Dataset The experiments are conducted using data collected from the Robot House [16], a specialised AAL environment equipped with over 60 heterogeneous sensors, including PIR motion sensors, door/cupboard contact sensors, pressure mats, and appliance power monitors. As illustrated in Fig. 2, the testbed comprises multiple functional zones (e.g., Kitchen, Bedroom, Bathroom) where domestic service robots and ambient sensors are integrated to capture high-fidelity human activity and localisation data. This setup provides a rich, multi-modal context for evaluating the proposed indoor localisation framework. For more detailed information about the data,

please refer to [16]. The dataset captures fine-grained human trajectories with high-precision room-level annotations. Detailed information about the data is shown in Table 1.

Preprocessing and Protocol The dataset comprises 8,067 state-update samples collected over multiple days of continuous recording at an average frequency of 1.6 Hz. No data augmentation is applied; the raw sensor triggers are used directly. We utilise a sliding window approach with a window size of $W = 10$ to capture temporal dependencies. Non-stationary ambient features are normalised using a Quantile Transformer to ensure numerical stability during training. We adopt a randomised train-test split (80/20) to evaluate the model's generalisation ability. To mitigate overfitting on the relatively small sample size, the model employs dropout ($p = 0.1$), weight decay (1×10^{-4}), and cosine annealing learning rate scheduling.

To ensure a fair comparison, all baselines receive identical input features and sliding windows. Hyperparameters for each baseline are tuned via grid search over their respective parameter spaces. All experiments use a fixed random seed for reproducibility.

Evaluation Metrics To provide a comprehensive assessment under class imbalance, we report Accuracy, Macro-averaged F1-score, Weighted F1-score, Balanced Accuracy, and the Matthews Correlation Coefficient (MCC).

4.2 Implementation Details

pexformer is implemented using the PyTorch framework and contains approximately 3.6M trainable parameters. We employ a scaled-up architecture consisting of an embedding dimension $d = 256$, a depth of $L = 6$ layers, and 16 attention heads. The model is optimised using the AdamW optimiser with a weight decay of 1×10^{-4} and a learning rate of 1×10^{-3}. A Cosine Annealing scheduler is used for learning rate decay across 50 epochs. Training is accelerated via Automatic Mixed Precision (AMP). All experiments are conducted on a single NVIDIA GeForce RTX 3070 Laptop GPU (8 GB). Training pexformer for 50 epochs completes in approximately 3 minutes; inference on the full test set takes less than 1 second.

4.3 Performance Benchmarking Against Conventional ML and DL Baselines

Comprehensive Comparative Analysis. Table 2 presents an exhaustive comparison between pexformer and 11 cross-paradigm baselines, demonstrating substantial improvements in indoor localisation performance. In a comprehensive evaluation spanning discriminative, ensemble, and standard DL models, pexformer's accuracy (93.30%) and MCC (90.80%) not only vastly exceed

Table 2. Comprehensive performance comparison of pexformer against 11 baselines in indoor localization. Best results are in **bold**, second best (best baseline) are underlined. Type denotes the model category (DL: Deep Learning, ML: Traditional Machine Learning). Δ denotes the relative improvement over the best-performing baseline.

Model	Type	Acc. (%)	F1-W (%)	F1-M (%)	Bal. Acc (%)	MCC (%)
BiLSTM	DL	85.79	84.91	46.46	59.67	55.46
GRU	DL	84.36	82.37	39.60	56.46	51.10
ResNet-1D	DL	70.27	70.54	16.27	19.26	20.61
Random Forest	Traditional Machine Learning (ML)	84.17	81.04	37.47	52.81	47.75
LightGBM	ML	84.11	80.32	33.03	49.98	46.82
Gradient Boosting	ML	84.11	80.69	38.51	51.47	45.79
XGBoost	ML	83.80	80.18	33.87	49.66	44.76
SVM	ML	83.74	80.96	39.90	45.15	45.71
AdaBoost	ML	80.38	77.45	25.38	29.67	39.18
GaussianNB	ML	10.99	8.63	20.16	37.92	32.65
k-Nearest Neighbors (KNN)	ML	10.86	17.14	25.44	38.75	18.09
pexformer (Ours)	DL	**93.30**	**93.15**	**83.52**	**83.84**	**90.80**
Δ vs. Best Baseline (%)	-	**+7.51**	**+8.24**	**+37.06**	**+24.17**	**+35.34**

the best ensemble learner, Random Forest (84.17%), but also display a significant performance gap over *ResNet-1D* (70.27%), a model specifically designed for temporal sequences. This suggests that standard DL architectures often experience representational collapse when processing ambient sensor data with tabular characteristics, due to their inability to model asymmetric feature interactions and irregular decision boundaries. Furthermore, the poor performance of GaussianNB and KNN underscores the high non-linearity and complex spatial overlapping inherent in Robot House sensory data. By synergising the PatchVectorTokenizer with the SPA mechanism, pexformer achieves a notable 37.06% improvement in F1-Macro relative to the strongest baseline, *BiLSTM*. This signifies that pexformer significantly mitigates the representational collapse of tail-class recognition in AAL, providing a highly robust localisation representation for complex, non-stationary smart home environments.

Error Analysis. An examination of the per-class predictions of pexformer reveals three distinct error regimes, each traceable to a specific sensor-level failure mode rather than to a generic classification difficulty. *(i) Transition zones* (e.g., hallways and doorways) account for the largest share of residual errors. Occupancy of these zones is typically signalled by a short burst of overlapping PIR triggers from two adjacent rooms, without a door-contact or pressure-mat anchor to disambiguate the label; within a $W{=}10$ window, such episodes are often indistinguishable from the tail of the subsequent room entry. *(ii) Open-plan boundary confusions* between the *Kitchen* and the *Living/Dining* area persist where two PIR fields-of-view physically overlap, producing nearly symmetric activations for a single occupancy event. *(iii) Rooms with distinctive sensor signatures*, such as the *Bathroom* (unique appliance/water-related triggers) and the *Bedroom* (pressure-mat-dominated activations), are recognised with near-perfect accuracy, indicating that the residual error budget is concentrated in sensor-ambiguity-driven cases rather than in generic model under-fitting.

Table 3. Main Results on Robot House Dataset. Comparison of pexformer against recent state-of-the-art deep tabular models. **Bold** indicates the best performance. All metrics are reported in percentages (%). Note that pexformer achieves a substantial gain in Macro-F1, indicating superior handling of imbalanced classes.

Model	Type	Acc. (%)	F1-Macro (%)	F1-Weighted (%)	Bal. Acc. (%)
NODE [13]	Tree-NN	83.67	33.33	79.63	49.44
TabTransformer [7]	Transformer	80.88	33.69	80.81	50.01
DCN V2 [19]	Cross-Net	83.49	39.50	81.25	54.43
SAINT [15]	Attention	83.80	37.10	83.15	52.29
Trompt [4]	Prompt-Based	86.34	42.03	83.23	59.52
TabR [5]	Retrieval	83.30	36.44	80.01	51.35
pexformer (Ours)	**Patch-Trans**	**93.30**	**83.52**	**93.15**	**83.84**
Δ vs. Best Baseline (%)	-	**+6.96**	**+41.49**	**+ 9.92**	**+24.32**

Comparing this breakdown with the tabular baselines in Table 3 highlights a qualitative shift in the nature of the errors. Baseline Macro-F1 scores between 33.69% and 42.03% imply that most baselines collapse on minority classes irrespective of sensor signature: underrepresented rooms such as *Office*-like or *Storage*-like zones are essentially missed even when their sensor evidence is unambiguous, which is characteristic of a *long-tail* failure mode driven by class prior bias rather than by genuine sensor-level ambiguity. pexformer, by contrast, recovers these minority rooms close to the majority-class accuracy (reflected in its Macro-F1 of 83.52%), so that the remaining errors are overwhelmingly boundary cases where the input signal itself does not carry sufficient information to separate two candidate rooms.

This pattern is consistent with the view that Patch-level Tokenisation and SPA together mitigate the long-tail bias but cannot resolve cases that are already ambiguous at the sensor level. Further improvements would therefore likely need to come from richer physical context, e.g., sensor-topology priors, boundary-aware attention, or short-range fusion with door-contact events, rather than from additional modelling capacity. We note that the above breakdown is qualitative and rests on aggregate per-class statistics from a single evaluation run with a fixed random seed; a formal confusion matrix across multiple seeds and a dedicated boundary-zone analysis are identified as future work in Section 5.

4.4 Benchmarking against Deep Tabular Baselines

Table 3 benchmarks pexformer against prominent deep tabular learning models published between 2020 and 2024. The results highlight three key observations: First, *Efficacy on Ambient Sensing Tasks.* pexformer achieves an accuracy of *93.30%*, surpassing the second-best model, Trompt [4] (86.34%). This suggests that integrating the patching mechanism into the ExcelFormer [**?**] architecture allows for more effective capture of spatio-temporal features in ambient sensor data, leading to improved performance on non-stationary distributions. Second, *Robustness to Class Imbalance.* This is notably evident in the *Macro-F1* metric.

Table 4. Efficiency Comparison with Attention-based Deep Tabular Baselines. Measurements are taken on a single NVIDIA GeForce RTX 3070 Laptop GPU (PyTorch 2.9, CUDA 12.8). Latency is measured at batch size 1; throughput and peak GPU memory at batch size 256. Each value is the mean of three independent runs of 100 iterations after 20 warmup iterations. All models consume the same underlying 10×21 sensor window; reshape to each model's native input layout is a free view operation.

Model	Type	Params (M)	FLOPs (M)	Latency (ms)	Throughput (s/s)	Mem (MB)
TabTransformer [7]	Transformer	0.214	0.15	0.947	243,765	10.7
SAINT [15]	Attention	0.181	0.21	0.988	244,568	11.7
Trompt [4]	Prompt-Based	0.562	0.30	1.858	127,315	12.5
ExcelFormer [?]	Scalar-Trans	0.227	27.97	9.655	5,354	131.1
pexformer (Ours)	**Patch-Trans**	3.564	49.88	3.710	13,805	91.8

Key baselines, such as SAINT [15] and DCN V2 [19], yield Macro-F1 scores between 33% and 42%, indicating challenges in detecting under-represented sparse classes and a potential bias towards high-frequency classes. In comparison, pexformer improves the Macro-F1 to *83.52%*. This indicates that the synergy of *Patch Tokenisation* and *Semi-Permeable Attention* assists in preserving the feature integrity of sparse events, thereby mitigating feature oversmoothing often observed in deep networks. Third, *Architectural Adaptability*. While pure Transformer architectures often face convergence challenges on small-scale tabular data, the results demonstrate that by introducing appropriate inductive biases (specifically Patch Tokenisation), pexformer can outperform both hybrid architectures (e.g., NODE [13]) and retrieval-augmented frameworks (e.g., TabR [5]) on unstructured sensor datasets.

4.5 Efficiency Comparison with Deep Tabular Baselines

To complement the accuracy analysis in Table 3, Table 4 reports a direct efficiency comparison of pexformer against the four attention-based deep tabular baselines on five deployment-oriented metrics: parameter count, FLOPs per sample, single-sample latency (batch = 1), batched throughput (batch = 256), and peak GPU memory. Three observations stand out. First, *pexformer substantially reduces the computational overhead of its architectural ancestor,* ExcelFormer. Per-sample latency drops from 9.655 ms to 3.710 ms (a 2.6× speed-up), peak GPU memory decreases by 30% (131.1 → 91.8 MB), and batched throughput improves by 2.6× (5,354 → 13,805 samples/s). This improvement directly reflects the impact of our *Patch-level Tokenisation*: whereas scalar tokenisation produces $10 \times 21 = 210$ tokens per window, pexformer compresses each sensor's temporal window into a single patch token, shrinking the dominant attention term from $O(210^2)$ to $O(21^2)$, a two-order-of-magnitude reduction in the quadratic cost of self-attention. Second, *the additional parameters of pexformer are effectively utilised rather than wasted*. TabTransformer, SAINT, and Trompt are more parameter-efficient because they adopt shallower depth and smaller

Table 5. Ablation Study on Key Components. We evaluate the contribution of the Permutation Strategy and the Attention Mechanism. **Bold** denotes the best performance. The significant drop in "w/o SPA" confirms the critical role of the semi-permeable mechanism in preventing feature oversmoothing.

Variant	Sorting	Attention	Acc. (%)	F1-Macro (%)	F1-Weighted (%)	Bal. Acc. (%)
pexformer (Ours)	**Random**	**Semi-Permeable**	**93.17**	**83.34**	**93.04**	**83.87**
w/ MI Sorting	Mutual Info	Semi-Permeable	92.99	81.67	92.73	81.92
w/o SPA	Random	Standard (Full)	68.16	48.25	65.42	52.71

embeddings, but this capacity reduction comes at a steep accuracy cost: their Macro-F1 scores on the same task lie between 33.69% and 42.03% (Table 3), roughly half of the 83.52% achieved by pexformer. The larger capacity of pexformer therefore yields a favourable size–performance trade-off on sparse, long-tail ambient sensing data rather than a wasteful over-parameterisation. Third, *absolute inference cost remains well within edge-deployment budgets.* The single-sample latency of 3.710 ms corresponds to an inference rate of approximately 270 Hz, two orders of magnitude higher than the 1.6 Hz sensor update frequency of the Robot House. At batch size 256, peak memory stays below 100 MB. These figures confirm that the accuracy improvements reported in Table 3 are not purchased at a computational cost that would preclude deployment on modest AAL edge hardware.

4.6 Ablation Studies

Table 5 details the ablation study on the key components of pexformer, validating the rationale behind our architectural choices. The ablation experiments are conducted in a separate controlled pipeline with a shared RNG reset to ensure fair comparison among variants; the minor numerical difference from the main results (e.g., 93.17% vs. 93.30%) is attributable to the inherent variance across independent training runs. First, *The Role of SPA.* Replacing SPA with standard attention (Variant "w/o SPA") results in a substantial performance degradation, with accuracy dropping from *93.17%* to *68.16%* and Macro-F1 dropping to *48.25%*. This contrast indicates that, under our evaluation protocol, applying full attention directly to sensor spatio-temporal data is associated with severe degradation. A plausible interpretation is that the masked SPA variant limits direct interaction between sensor tokens and thereby preserves more of the locally discriminative structure needed for minority-class recognition, whereas full attention may permit noise propagation and over-smoothing across tokens. We note this interpretation as a hypothesis consistent with the observed behaviour rather than as a causal mechanism formally established by these experiments. Second, *Random Permutation vs. MI Sorting.* A notable observation is that a simple *Random Permutation* performs comparably to or slightly better than the statistically grounded *MI (MI)* sorting (Accuracy: 93.17% vs. 92.99%; Macro-F1: 83.34% vs. 81.67%), while avoiding the computational cost of MI estimation. We emphasise that the observed gap is small and obtained under a single fixed-seed

protocol; it should therefore be read as evidence that the two ordering strategies are *not statistically distinguishable* within our setting rather than as a definitive claim of superiority. One plausible interpretation is that, in the temporal patch modality, MI captures primarily static pairwise correlations and may not reflect complex nonlinear dependencies between sensors, so that Random Permutation operates as an implicit regulariser analogous to random projection. This interpretation is tentative: the effect may also be driven by optimisation noise or by interactions with the fixed-seed protocol, and a mechanistic justification—ideally supported by multi-seed significance tests and targeted controlled experiments—is left for future work (see Section 5). Third, *The Contribution of Patch Tokenisation.* The importance of the PatchVectorTokenizer can be inferred by comparing pexformer with standard tabular Transformer baselines in Table 3. Models such as TabTransformer and SAINT, which employ point-wise or per-feature embedding without temporal patching, achieve Macro-F1 scores below 38%, whereas pexformer reaches 83.52%. This substantial gap indicates that aggregating each sensor's temporal window into a single semantic token is critical for capturing local activation dynamics, without which the model cannot distinguish between transient noise and meaningful occupancy signals.

Table 6. Sensitivity Analysis of Logit Adjustment Factor τ. We evaluate how the long-tail adjustment strength affects performance. While increasing τ marginally improves Balanced Accuracy (peaking at $\tau = 1.2$), the default setting ($\tau = 0$) yields the best overall Accuracy and Macro-F1, suggesting pexformer's intrinsic robustness to imbalance.

Factor τ	Acc. (%)	F1-Macro (%)	F1-Weighted (%)	Bal. Acc. (%)
0.0	**93.30**	**83.52**	**93.15**	83.84
0.5	92.55	82.19	92.70	83.42
0.8	92.49	82.43	92.76	84.01
1.0	91.87	82.33	92.44	84.63
1.2	91.06	81.91	91.95	**84.83**
1.5	89.01	80.12	90.66	83.67
2.0	79.70	75.80	84.92	81.11

4.7 Parameter Sensitivity Analysis

Table 6 investigates the impact of the adjustment factor τ in the Balanced Softmax Loss on pexformer's performance. This factor is designed to explicitly shift decision boundaries to favour long-tailed classes. We observe a classic *Accuracy-Fairness Trade-off.* As τ increases from 0.0 to 1.2, the *Balanced Accuracy* exhibits a slight upward trend, peaking at **84.83%** with $\tau = 1.2$. This confirms that moderate logit adjustment can indeed force the model to pay more attention to underrepresented samples. However, this gain comes at the cost of

overall Accuracy. Crucially, we find that the default setting ($\tau = 0$) yields the best holistic performance (Accuracy 93.30%, F1-Macro 83.52%). This suggests that *pexformer's architectural design provides intrinsic resilience to class imbalance.* Unlike traditional models that heavily rely on loss engineering to handle long-tail distributions, pexformer's inherent inductive bias allows it to converge to an optimal solution under standard supervision, eliminating the need for extensive hyperparameter tuning.

5 Discussion

The PexFormer architecture proposed in this study achieves a semantic leap from discrete, noise-prone sampling points to continuous "motion trajectories" via patch-level tokenisation, effectively capturing intricate activity semantics while significantly reducing the computational overhead of the self-attention mechanism. This temporal abstraction is further bolstered by the Semi-Permeable Attention (SPA) mechanism, whose structured masking serves as a vital filter to suppress cross-sensor interference and preserve the distinct characteristics of heterogeneous signal sources. Crucially, PexFormer exhibits remarkable robustness to input feature ordering, maintaining state-of-the-art performance even with random permutations. This inherent order-invariance eliminates the need for expensive pre-computed statistical priors, ensuring that the system remains highly reliable and versatile for deployment in dynamic real-world environments where sensor configurations may shift or intermittently fail. Crucially, this technical robustness translates directly into improved HRI outcomes. In the context of SARs, location awareness is not merely a navigation requirement but a foundation for social context. By accurately identifying user presence in rarely visited zones (e.g., hallways or storage areas), robots can avoid intrusive behaviours and better adhere to social norms like proxemics. Furthermore, the ability to handle long-tail distributions ensures that the robot remains "socially present" even when the user is engaged in infrequent but critical routines. This reliable localisation allows the robot to transition from a reactive tool to a proactive companion capable of context-aware check-ins and safer, more dignified support for people living alone.

Limitations

Despite the strong performance of pexformer in indoor human localisation tasks, this study acknowledges several limitations that motivate further work. First, we empirically observed that Random Permutation within the Semi-Permeable Attention mechanism performs comparably to or better than traditional MI-based sorting. While this suggests that randomness may serve as a form of regularisation similar to random projection, effectively preventing overfitting, this phenomenon is primarily established on empirical observations and currently lacks a comprehensive theoretical foundation to fully explain its underlying mechanism. Second, the validation of the proposed model is currently confined to a single

smart home dataset and evaluated under a random 80/20 train-test split. This protocol mirrors the setup used by the deep tabular baselines in our comparison and ensures a like-for-like accuracy benchmark, but it does not directly assess the model's robustness to *temporal distribution shift* (e.g., training on earlier recording sessions and testing on later ones) nor to *cross-home transfer*, where sensor layout, topology, and occupancy behaviour may differ substantially. Extending the evaluation to temporally separated splits and additional external smart-home corpora therefore constitutes an important line of follow-up work to more stringently characterise generalisation. Third, all metrics reported in this paper are point estimates obtained from a single deterministic run with a fixed random seed, primarily to guarantee exact reproducibility of the architecture, initialisation, and training pipeline. While the fixed-seed protocol is consistent across pexformer and every baseline in Tables 2, 3 and 5, so that the ranking is directly comparable, we do not report confidence intervals, standard deviations across repeated runs, or formal significance tests against the strongest baselines. A full multi-seed analysis with such statistical characterisation would provide stronger empirical guarantees on the observed gains and is left as future work.

6 Conclusion and Future Work

This paper presents pexformer, a Transformer architecture that adapts patch-level tokenisation and Semi-Permeable Attention from the tabular domain to ambient-sensor-based indoor localisation. On a real-world smart home dataset, pexformer achieves 93.30% accuracy and an 83.52% Macro-F1, substantially outperforming 11 baselines spanning classical ML and deep tabular models, without requiring complex resampling techniques. This robust localisation lays the groundwork for socially assistive robots to provide context-aware, proactive support in AAL settings. Future work will pursue four directions: (i) establishing a theoretical framework for the regularisation effect of random permutation in spatiotemporal feature interaction, (ii) exploring adaptive multi-scale patching mechanisms, (iii) integrating pexformer into a social robot for real-time evaluation in multi-room AAL environments, and (iv) conducting multi-seed evaluations with confidence intervals and significance testing, as well as extending benchmarks to temporally separated splits and additional external smart-home datasets, to further characterise generalisation and statistical robustness.

Acknowledgements

We kindly acknowledge the financial support of the Dinwoodie Charitable Company for the Hospital@Home project. We also acknowledge institutional support from the University of Hertfordshire.

References

1. Blackman, S., Matlo, C., Bobrovitskiy, C., Waldoch, A., Fang, M.L., Jackson, P., Mihailidis, A., Nygård, L., Astell, A., Sixsmith, A.: Ambient assisted living tech-

nologies for aging well: a scoping review. Journal of Intelligent Systems **25**(1), 55–69 (2016)
2. Chan, A., Cai, J., Qian, L., Coutts, B., Phan, S., Gregson, G., Lipsett, M., Rincón, A.M.R., et al.: In-home positioning for remote home health monitoring in older adults: systematic review. JMIR aging **7**(1), e57320 (2024)
3. Chen, J., Yan, J., Chen, Q., Chen, D.Z., Wu, J., Sun, J.: Can a deep learning model be a sure bet for tabular prediction? In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 288–296 (2024)
4. Chen, K.Y., Chiang, P.H., Chou, H.R., Chen, T.W., Chang, D.T.H.: Trompt: towards a better deep neural network for tabular data. In: Proceedings of the 40th International Conference on Machine Learning. pp. 4392–4434 (2023)
5. Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., Babenko, A.: Tabr: Tabular deep learning meets nearest neighbors. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=rhgIgTSSxW
6. Götz, L., Kollovieh, M., Günnemann, S., Schwinn, L.: Efficient time series processing for transformers and state-space models through token merging. arXiv preprint arXiv:2405.17951 (2024)
7. Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 (2020)
8. Kermani, A., Zeraatkar, E., Irani, H.: Energy-efficient transformer inference: Optimization strategies for time series classification. arXiv preprint arXiv:2502.16627 (2025)
9. Koay, K.L., Syrdal, D., Bormann, R., Saunders, J., Walters, M.L., Dautenhahn, K.: Initial design, implementation and technical evaluation of a context-aware proxemics planner for a social robot. In: International conference on social Robotics. pp. 12–22. Springer (2017)
10. Luperto, M., Monroy, J., Renoux, J., Lunardini, F., Basilico, N., Bulgheroni, M., Cangelosi, A., Cesari, M., Cid, M., Ianes, A., et al.: Integrating social assistive robots, iot, virtual communities and smart objects to assist at-home independently living elders: the movecare project. International Journal of Social Robotics **15**(3), 517–545 (2023)
11. Mehrabi, F., Ghezelbash, A.: Wired for companionship: a meta-analysis on social robots filling the void of loneliness in later life. The Gerontologist **65**(12), gnaf219 (2025)
12. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730 (2022)
13. Popov, S., Morozov, S., Babenko, A.: Neural oblivious decision ensembles for deep learning on tabular data. In: International Conference on Learning Representations
14. Rafique, S., Holthaus, P., Fang, G., Amirabdollahian, F.: Interpretable room-level human presence detection using ambient sensors in smart homes. In: 17th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2025) (2025)
15. Somepalli, G., Schwarzschild, A., Goldblum, M., Bruss, C.B., Goldstein, T.: Saint: Improved neural networks for tabular data via row attention and contrastive pretraining. In: NeurIPS 2022 First Table Representation Workshop
16. University of Hertfordshire: Robot house. https://robothouse.herts.ac.uk/ (2025), last accessed: 2025/06/12

17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: A survey. Pattern recognition letters **119**, 3–11 (2019)
19. Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., Chi, E.: Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In: Proceedings of the web conference 2021. pp. 1785–1797 (2021)
20. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2d-variation modeling for general time series analysis. arXiv preprint arXiv:2210.02186 (2022)
21. Wydmański, W., Movsum-zada, U., Tabor, J., Śmieja, M.: Vistabnet: Adapting vision transformers for tabular data. In: Proceedings of the 2025 SIAM International Conference on Data Mining (SDM). pp. 497–506. SIAM (2025)
22. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 11106–11115 (2021)