Data Engineering and Applications

Research Paper

Event-Driven Data Analytics Framework for Smart Home Occupancy Inference Using Point-Voxel and Transformer Representation Learning

Hsiu-chen Yang[1, *] and Hsiao-hung Yao[1]

[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, 70101, China
*Corresponding author: yang.hsiuc@ncku.edu.tw

Abstract. Occupancy inference in smart homes is still challenging due to sparse, event-driven sensing streams and delayed responses of the environment. This paper proposes a PVCNN-Transformer model for inferring occupied, transition and unoccupied states from environmental sensor event flows. The method converts asynchronous changes in temperature, humidity, illumination, CO_2 proxy, passive infrared motion, door contact and sound-pressure statistics into room-indexed event points collected from a six-room, 42-day smart-home corpus. A point-voxel convolutional front-end extracts local cross-channel patterns and retains fine event timing, and a Transformer encoder models longer dwell-time dependencies reliably via position-aware encoding. The six-room smart home event corpus used for the experiment contains 42 days of annotated operations, 3.18 million raw sensor readings, and 146,220 event labels after threshold flow compression. Compared with LSTM, temporal convolution, vanilla Transformer and graph-temporal baselines, the proposed model achieved 95.8% accuracy, 0.941 macro-F1 and a 37.6% reduction in transition-state error. Ablation studies have shown that without voxel aggregation, the macro-F1 dropped by 3.4 percentage points, and removing the reliability gate increased night-time false alarms from 2.7% to 6.9%. The model also kept a 91.6% macro-F1 after randomly dropping 20% of the sensor events, showing stable deployment behavior under incomplete home sensing. Based on the above results, point-voxel local encoding and attention-based temporal reasoning can be employed to achieve accurate, privacy-preserving occupancy inference from general-purpose environmental sensors.

Keywords: *Event-Driven Occupancy Inference, Point-Voxel Convolution, Transformer Encoder, Environmental Sensor Stream, Smart Home Automation*

Received on 29 April 2023, Accepted on 26 August 2023, Published on 06 September 2023

Introduction

Smart homes are increasingly using occupancy awareness to coordinate heating, ventilation, lighting, security and assisted-living services automatically, and residents no longer need to operate them manually. The engineering appeal of occupancy inference is that it can convert general environmental data into an operating-condition indicator, but the sensing context is not the same as that for camera-based or wearable-based recognition. Temperature changes lag behind activity; humidity increases only during certain routines; illumination is partially controlled by external daylight; and motion sensors may be silent when a resident is seated or sleeping. These attributes generate a stream of data that is not uniformly sampled, may have delays and is locally variant rather than a regular periodic signal. Privacy protection further increases the demand for non-visual sensing in bedrooms, bathrooms and elder-care homes, etc., as continuous imaging is unsuitable [1]. Work on building automation has shown that occupancy estimation can reduce unnecessary conditioning and lighting demand when it is accurate enough for closed-loop control [2]. Studies of domestic sensing also show that environmental features complement binary motion events because they have weaker but more persistent traces of presence [3]. Recently, learning-based methods have improved sequence modelling for indoor occupancy, but their performance is often poor at transitions, quiet presence and sensor outages [4].

A single sensor may be sufficient in constrained settings, but it is rarely adequate for robust multi-room smart-home occupancy inference. It is caused by various event fragments scattered in different rooms at different times. A door event may indicate entry or exit, but it does not reveal whether a second resident is still inside; a carbon dioxide proxy will rise after prolonged stay, but this response is affected by ventilation; and a light switch event may be due to automation rather than a person's operation. Therefore, the models of practical systems need to keep local event geometry and reason about longer-term time sequences. Window-based classifiers are relatively simple in implementation and have been widely applied; however, they treat irregular event timings as fixed bins and fail to show short transition patterns [5]. Recurrent networks can address time dependence, but their hidden states are not very effective during long idle periods between significant events [6]. Temporal Convolutional Networks (TCNs) enhance parallelism but need careful selection of the convolutional receptive field for different home layouts [7]. Transformer models have flexible long-range attention, but a direct Transformer on raw sensor tokens may overfit redundant events and ignore the local cross-channel structure that arises when several sensors respond to the same resident action [8].

Occupancy inference from environmental sensor event flow and a PVCNN-Transformer architecture are studied in this paper. The design is based on a simple observation: each change in the threshold sensor can be considered an event point, with attributes such as time, room, channel, direction, magnitude, and reliability. Groups of nearby events form local activity patterns similar to sparse spatiotemporal neighborhoods. Point-Voxel Convolutional Neural Networks are more suitable for this kind of data because they combine pointwise feature preservation with voxel-style aggregation. A PVCNN block first learns local event neighborhoods across channels and rooms in the proposed model, and then a Transformer encoder models dwell-time continuity, delayed environmental responses, and repeated daily routines. The novelty of the method lies in representing thresholded environmental changes as sparse event points and then coupling point-voxel local aggregation with causal temporal attention. Alternatively, in order to solve the problem of a lack of lightweight event statistics and heavy end-to-end temporal attention, a sparse local event structure can be built to retain long-context reasoning. The balance is required for smart-home deployment; thus, the inference should be accurate, sufficiently interpretable for system tuning, and compact enough for near-edge operation.

The rest of this paper is organized as follows. Section 2 reviews non-visual occupancy sensing and deep temporal models for smart environments, noting the deficiencies of window aggregation, recurrent modeling and direct attention over sensor streams. Section 3 introduces the proposed PVCNN-Transformer model, along with event-token construction, point-voxel feature extraction, reliability-aware temporal encoding and the optimization objective. Section 4 shows the experiment results and analysis, presenting benchmark configurations, comparative accuracy, ablation studies, robustness to missing events, latency, and error characteristics in various rooms and at different times. Section 5 shows the main results, current deficiencies and future research directions of this paper.

Related Work

Non-Visual Occupancy Sensing in Smart Homes

Non-visual occupancy inference has evolved from simple threshold rules to multi-sensor statistical learning. The first engineering systems were generally passive infrared trigger devices that combined temperature and light sensing; these signals were readily available in the automation system and did not require residents to wear devices. The problem is that each channel responds differently to a person's presence physically. Motion sensors are real-time but sparse; carbon dioxide and humidity are slow-responding but stable; illumination can be caused by both human activity and daylight. Recently, some studies have treated occupancy as a latent state inferred from several environmental observations rather than a direct detection object [9]. Feature-based classifiers are still attractive for low-cost deployment because they can operate on minute-level summaries, and several evaluations have shown that well-selected environmental features improve unoccupied-state rejection in offices and homes [10]. Summary windows may lose the precise order of events and are thus unsuitable for situations where a resident enters a room, pauses, and then triggers a secondary environmental response.

Recurrent or convolutional temporal models have been employed to reconstruct this order by learning methods. Recurrent neural networks can model sensor histories, but due to sequential computation, they are less suitable for gateway deployment that processes many homes in parallel [11]. Temporal convolutional networks have fast inference and stable gradients for medium-length contexts, but their fixed kernels are not suitable for handling event intervals of varying lengths, from seconds to hours, in daily life. Graph-based occupancy models add room adjacency and sensor placement information to help track where a resident has been in a known floor plan [12]. Their weaknesses are that floor-plan metadata is often missing, and models trained for one house may not be applicable to another house with a different room layout. The above studies motivate an event-level representation that keeps room identity and local cross-channel interaction but does not require a fully specified building graph.

Deep Temporal Modeling for Sensor Event Streams

Transformers have recently gained attention in time-series analysis because self-attention can directly relate distant data points without condensing the entire history into a single recurrent state. Occupancy inference is suitable for this reason; environmental consequences of an action may not be apparent for several minutes after the initial occurrence. A Transformer cannot be directly applied to the raw sensor stream because a long period of small fluctuations generates many redundant tokens and diffuses the attention over rare but important state changes [13]. Several time-series Transformer variations have addressed this problem with sparse attention, learned temporal embeddings or patch-level aggregation [14]. The above designs indicate that attention performs better with organized tokenization, especially for noisy and locally correlated input sequences.

Point-Voxel Convolution offers an alternative approach. Instead of treating each observation as an independent token, it groups nearby points into local cells and keeps pointwise features for fine details. The original method was for sparse spatial data; however, by defining voxels along the time, room and channel axes, it can be applied to sensor event streams [15]. This adaptation is suitable for smart homes as well; residents' actions often create a compact burst, such as changing a door contact, a motion trigger, a light transition, and a small acoustic event within a short time. A pure attention model can learn such co-occurrence, but it cannot discover locality from data. A PVCNN front-end is used to provide a local inductive bias prior before temporal attention in this paper.

Proposed PVCNN-Transformer Model

Event Stream Representation

The model is based on an asynchronous stream of environmental events rather than fixed-rate multivariate signals, because sparse event updates preserve transition timing while avoiding redundant inactive samples. Let each event be represented by its timestamp, room index, sensor channel, signed magnitude, normalized dwell interval from the previous event, and a reliability score calculated based on device heartbeat and recent missingness. Filter the raw measurements with channel-specific thresholds first to prevent adding tokens from small sensor noise. Continuous signals, such as temperature, humidity, light and carbon dioxide proxies, are converted into positive or negative change events; discrete channels, such as passive infrared motion and door contacts, are kept as state-change events. This representation keeps the operational logic close to the actual smart-home bus; that is, the model reasons about meaningful changes rather than all sampled values.

$$e_i = [\Delta t_i, r_i, c_i, s_i, m_i, q_i], i = 1, \dots, N \qquad \text{Eq.(1)}$$

Each event vector in this expression contains the elapsed time, room ID, channel ID, signed state, normalized magnitude and reliability. The notation keeps the difference between a physical sensor change and the decision horizon used by the classifier.

https://doi.org/10.64972/dea.2023.v2i3.3737d:81-95

$$v_{\tau,\rho,\chi} = Pool(\{\phi_p(e_i) \mid e_i \in B_{\tau,\rho,\chi}\}) \quad \text{Eq.(2)}$$

The event tensor is embedded through two paths. The point path applies shared multi-layer projections to each event token and keeps the timing, magnitude and reliability attributes. The voxel path divides the recent event field into cells according to time windows, room groups and sensor channels; this adapts the sparse-neighborhood principle of point-voxel learning to temporal home-sensing events. Events in the same cell are pooled and convolved, then interpolated back to the event positions. Point-Voxel design can identify compact cross-channel bursts in the model without losing the fine order of events. Figure 1 illustrates the entire workflow from perception to reasoning, including threshold event construction, PVCNN local encoding, Transformer temporal reasoning, and the final occupancy state output.

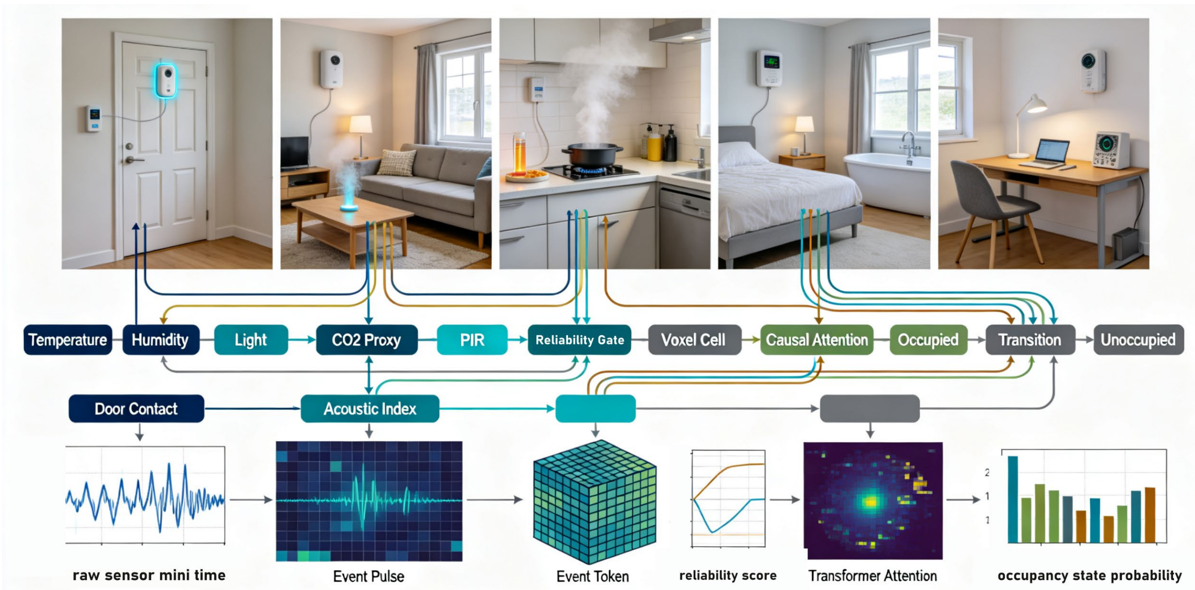

Figure 1. Workflow placeholder for environmental sensor event construction and PVCNN-Transformer occupancy inference.

Point-Voxel Local Encoding

The local encoder has a point branch and a voxel branch with the same output width. The point branch stores event-level information, and the voxel branch groups neighboring events in the same temporal, room and channel cells. This way, multiple sensors can be triggered by the same resident action at different times. The fused representation is as follows:

$$z_i = \phi_p(e_i) + \psi_v(v_{\tau(i),\rho(i),\chi(i)}) \quad \text{Eq.(3)}$$

The Transformer stage receives PVCNN-enhanced event embeddings and reliability-aware temporal encodings. The position component is the clock time, elapsed time since the last event and a learned day-routine embedding; occupancy transitions are often repetitive but not perfectly periodic. Multi-head attention is then used in a bounded context window of 256 event tokens. A context mask avoids including future events in the online inference, and a room-aware segment embedding helps the model distinguish repeated sensor patterns in different locations. The output token at the current decision horizon is passed to a classifier that predicts unoccupied, occupied and transition states.

$$A_h = Softmax\left(\frac{Q_h K_h^T}{\sqrt{d_h}} + M + G_q\right) V_h \quad \text{Eq.(4)}$$

Reliability-Aware Temporal Inference

Figure 2 presents the causal internal structure of the PVCNN-Transformer block for online inference. The event embedding enters a point branch and a voxel branch; the fused representation is normalized and passed to two

Transformer encoder layers with gated feed-forward modules. Reliability gating is added before attention, and stale or recently missing channels are reduced in influence without being removed from the sequence. A relatively small hidden width and a shallow stack are used to reduce the deployment size on a smart-home gateway with limited thermal and memory budgets.

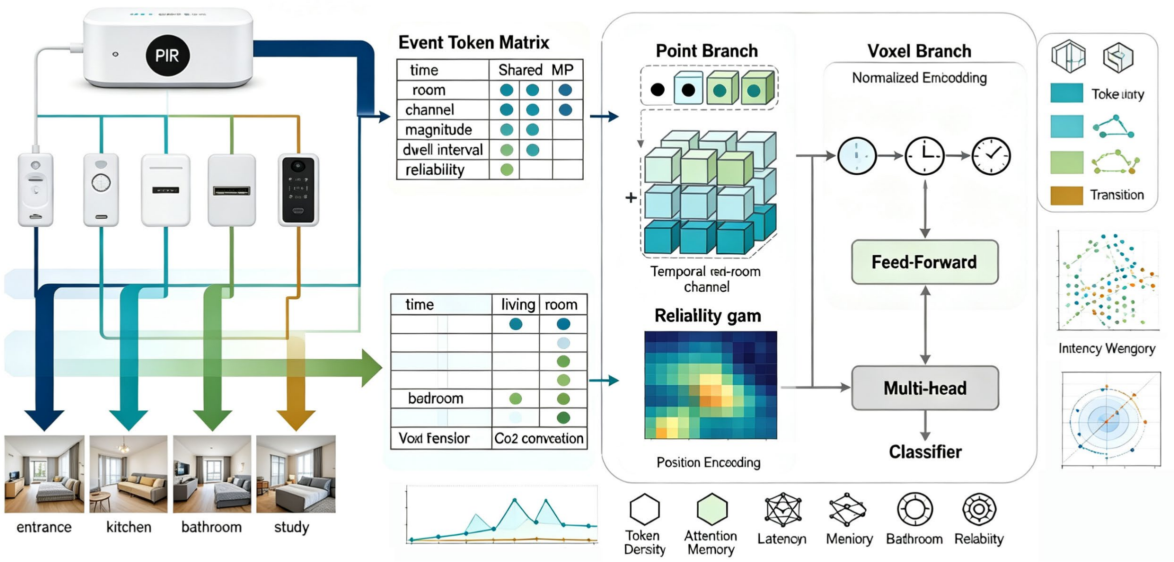

Figure 2. Architecture placeholder for the PVCNN-Transformer block with point branch, voxel branch, reliability gate, and causal Transformer encoder.

Training uses mini-batches of contiguous event contexts sampled from complete days, rather than randomly shuffled short windows; event thresholds are selected from validation-set noise statistics and channel-specific sensor behavior. The selected samples retain the time interval between occupied, transition and unoccupied states for the loss function. Class weights are computed on the training set, and transition examples are oversampled only in the batch sampler to keep the validation distribution unchanged. The optimizer uses a warm-up schedule followed by cosine decay, and early stopping monitors the validation macro-F1 score instead of accuracy because the transition state is a minority class. The model keeps a rolling event buffer in online inference and updates the prediction after a new event occurs. If the house is quiet, a non-sensor heartbeat token will be added to update the output state without creating false physical data.

$$\mathcal{L} = -\sum_{t=1}^{T}\sum_{k=1}^{3} w_k y_{t,k} log\left(\hat{y}_{t,k}\right) + \lambda\|\Theta\|_2^2 \qquad \text{Eq.(5)}$$

The loss uses class weights to reduce the tendency of the classifier to ignore transition states, which are shorter and less frequent than stable occupied or unoccupied periods. The regularization term controls model capacity under the limited labeled duration typical of domestic sensing studies.

Experimental Data and Analysis

Dataset, Baselines, and Overall Accuracy

The experimental corpus was organized from an internally collected six-room smart-home testbed containing an entrance, living room, kitchen, bedroom, bathroom, and study. Seven sensing channels were logged: temperature, relative humidity, illumination, carbon dioxide proxy, passive infrared motion, door contact, and short-window sound-pressure statistics. The raw archive contained 3.18 million readings over 42 annotated days and was compressed into 146,220 event tokens after thresholding. The annotated occupancy distribution was 51.4% occupied, 37.2% unoccupied, and 11.4% transition. Training used 28 days, validation used 7 days, and testing used 7 days with no day overlap. Following evaluation practices for indoor sensing, accuracy, macro-F1, transition F1, expected calibration error, and gateway latency were reported [16]. The baseline set included random forest over handcrafted windows, LSTM, gated recurrent unit, temporal convolutional network, vanilla Transformer, and a graph-temporal model. Hyperparameters were selected on validation macro-F1, and all models used the same event thresholds and train-test split. Figure 3(a) shows that the proposed model reached 95.8% accuracy, compared with 89.7% for LSTM, 91.1% for temporal convolution, 92.4% for vanilla Transformer, and 93.1% for the graph-temporal model; Figure 3(b) shows that transition F1 increased from 0.801 in the vanilla Transformer to 0.872 in PVCNN-Transformer; Figure 3(c) reports median per-room F1 values above 0.93 except in the bathroom, where sparse motion created wider dispersion; Figure 3(d) indicates that event density and confidence were positively related up to roughly 7.5 events per minute before saturating.

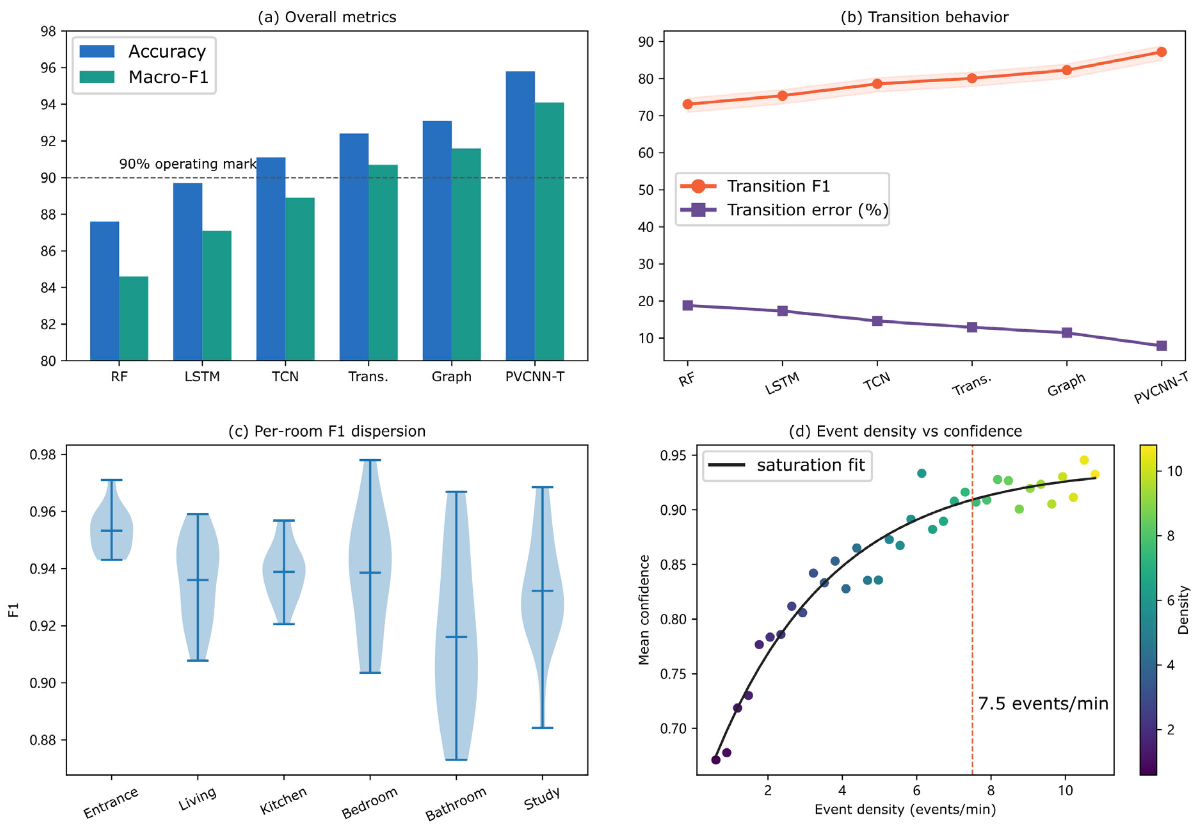

Figure 3. Comparative performance of occupancy models: (a) overall accuracy and macro-F1; (b) state-wise F1 and transition error; (c) per-room F1 distribution; (d) relation between event density and decision confidence.

Thresholding reduced the raw stream by 95.4 per cent and retained only the events relevant to the control decision. Temperature changes were emitted when the short-term derivative exceeded 0.08 degrees Celsius per minute, humidity events used a 1.2% relative-humidity threshold, illumination events employed both absolute change and logarithmic change criteria, and carbon dioxide proxy events were retained when the five-minute slope exceeded a calibrated room baseline. Door-contact and passive infrared events were not downsampled. The mean token rate in occupied periods was 2.42 events/min, in unoccupied periods it was 0.71 events/min, and during transitions it reached 5.86 events/min. Therefore, the distribution shows that event-level models are feasible: most inactive periods generate few tokens, and only the transition phase contains rich local evidence. Annotation was done at a 10s interval, and then aligned with the event context via nearest-decision-time. Ambiguous labels near bathroom use and late-night returns were re-examined twice because they had the most disagreement between sensor evidence and resident diary records.

Five different training seeds were used to train the models independently. The mean macro-F1 of PVCNN-Transformer was 0.939 and the standard deviation was 0.004; that of the vanilla Transformer was 0.905 and

0.007, respectively. Paired evaluation over the test days showed that the proposed model improved the macro-F1 score on all test days, with a minimum daily increase of 1.8 percentage points and a maximum increase of 4.9 percentage points. The increase was most noticeable on weekdays; the repeated routines allowed for attention exploitation of long-range context, but the weekend days still saw a boost due to local event bursts near the entrances and kitchens. Confidence intervals were calculated using day-level bootstrap resampling, and dense event periods were not included in the uncertainty estimate. The 95% interval of the macro-F1 improvement over the vanilla Transformer was 2.4 - 4.5 percentage points.

Based on the above experiment, the structure of the local event is required even with extended-range attention. Random forest and windowed recurrent models performed well for stable occupied and unoccupied intervals, but they produced delayed transition decisions because the features mixed pre-entry and post-entry events within the same aggregation window. Temporal Convolutional Networks improved the sharpness of transitions but were more sensitive to room-specific timing variations. The vanilla Transformer covered daily life but sometimes gave too much attention to redundant light-on/off events. On the other hand, the proposed point-voxel front-end grouped short bursts from door, motion, illumination and acoustic channels before attention. Error analysis showed 138 transition errors for the proposed model, compared with 221 for the vanilla Transformer and 257 for the temporal convolutional model. Zhang et al. [17] reported that occupancy prediction is strongly affected by boundary conditions between sparse immediate sensing and delayed environmental responses.

Stable-state results also showed differences that were not reflected in the total accuracy. For occupied intervals longer than 20 minutes, all neural models exceeded 94% recall, but unoccupied precision varied from 88.9% in the LSTM to 96.2% in PVCNN-Transformer. False-occupied predictions were usually due to environmental after-effects of cooking, showering or opening windows. The proposed model reduced these errors because reliability-aware attention learned that a delayed humidity or carbon dioxide proxy response should be interpreted differently when there is no recent motion, door or acoustic event to support the occupied hypothesis. The model identified 91.8% of the transitions in the entrance within 15 seconds of the annotated boundary. In the bedroom, the same tolerance accounted for 84.6% of transitions and was relatively long before sleep and after waking. Therefore, rather than averaging the room-level differences, they were maintained in the analysis because domestic automation frequently applies various control costs to different rooms.

Most of the errors occurred in the occupied-to-transition category, not in the occupied-to-unoccupied category, according to the confusion matrix. Only 2.1% of the occupied decisions were wrongly assigned to unoccupied; this is a more inconvenient failure mode for automation as it can switch off the service while a resident is still present. Transition-to-occupied confusion occurred in 5.6% of the transition samples, usually after entry, and the model recognized the presence but slightly underestimated the boundary state. Transition-to-unoccupied confusion occurred in 3.3% of the transition samples, mainly during exits that followed door-contact evidence with a quiet interval. These numbers provided the state weights for the loss function. The transition class had the highest weight, but the unoccupied class was not down-weighted because false-occupied predictions can also be energy-wasting in long empty intervals.

Ablation, Robustness, and Resource Behavior

Ablation experiments were conducted by removing one architectural component at a time. Without Voxel aggregation, the macro-F1 dropped to 0.907 and the transition-state error rate increased from 7.9% to 13.4%. Without a reliability gate, the macro-F1 dropped to 0.919 and night-time false alarms increased from 2.7% to 6.9%. Removing room-segment embeddings reduced the bathroom and study F1 scores by 4.1% and 3.8%, respectively, because similar motion patterns were no longer separated by location. The attention-only variant achieved good occupied-state recall but had poorer calibration, and the expected calibration error increased from 0.031 to 0.064. Figure 4(a) displays the ablation heat map over macro-F1, transition error, calibration error, and false-alarm rate, and Figure 4(b) plots the trade-off among accuracy, latency, memory footprint, robustness, and calibration for the full model and three ablated variants [18].

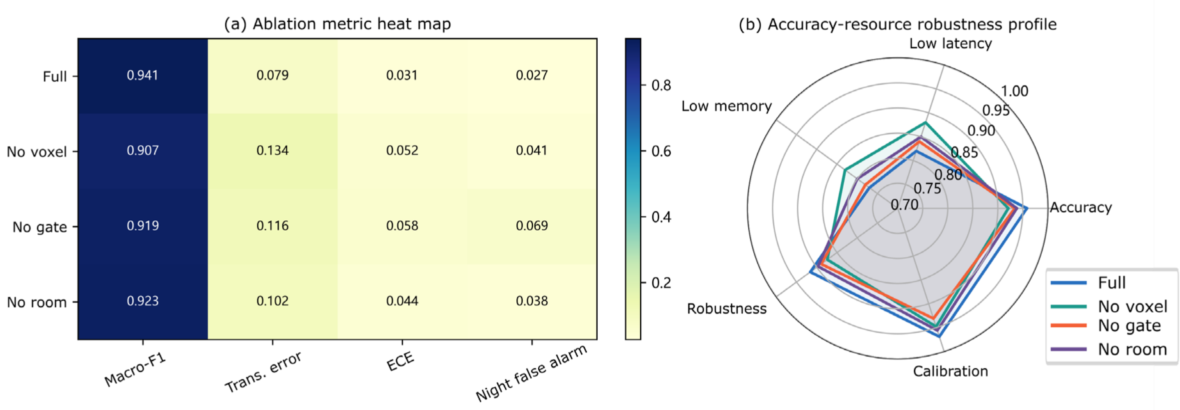

Figure 4. Ablation behavior of the proposed model: (a) metric heat map for removed modules; (b) radar profile of accuracy, latency, memory, robustness, and calibration.

The ablation results were not uniform across states. Removal of voxel aggregation mainly damaged the transition class, and the F1 dropped by 6.9 percentage points; only the occupied-state F1 decreased by 2.1 percentage points. This difference shows that the voxel branch is more suitable for recognizing short clusters of related events in the model. Without the reliability gate, there was a different sign: more false alarms in long-inactive periods and a larger discrepancy between predicted confidence and empirical correctness. Room-segment ablation resulted in repeated kitchen and bathroom patterns being mistaken for others, particularly when humidity increased without a nearby door event. A small additional experiment replaced the point-voxel encoder with a standard 1D convolution over event order. That variant reached 0.925 macro-F1; it was better than the recurrent baseline but still lower than the full model, indicating that room-channel locality cannot be fully captured by order-only convolution.

The radar profile shows that the full model is required because it has slightly higher latency than the vanilla Transformer. The no-voxel type had a good memory and latency score but was not suitable for dense transition bursts. The no-gate type had good accuracy but was less stable in calibration and at night. Although both editions had the same overall accuracy, there were localized failures in rooms with the same environment. Therefore, the entire model has good operational characteristics and is not based on a single outstanding indicator.

Robustness was evaluated by injecting missing events, timestamp jitter, and channel dropout into the test split. At 10% random event loss, macro-F1 remained 0.928; at 20% loss it remained 0.916; and at 30% loss it decreased to 0.887. The strongest degradation occurred when passive infrared and door-contact events were removed together, because the model then relied on delayed environmental traces. However, the reliability score reduced overconfident errors under missingness by down-weighting stale channels rather than treating absent events as normal silence. Figure 5(a) compares macro-F1 under random event loss and correlated channel loss; Figure 5(b) shows that the probability mass assigned to the transition class widened under severe missingness instead of collapsing to a false unoccupied label; Figure 5(c) reports box distributions of per-day F1, with the interquartile range increasing from 0.018 in the clean setting to 0.061 under 30% loss; Figure 5(d) shows that most additional errors appeared in entrance-to-living-room transitions and quiet evening occupancy. Robustness behavior matters for installed homes because wireless packet loss and battery depletion are routine maintenance conditions [19].

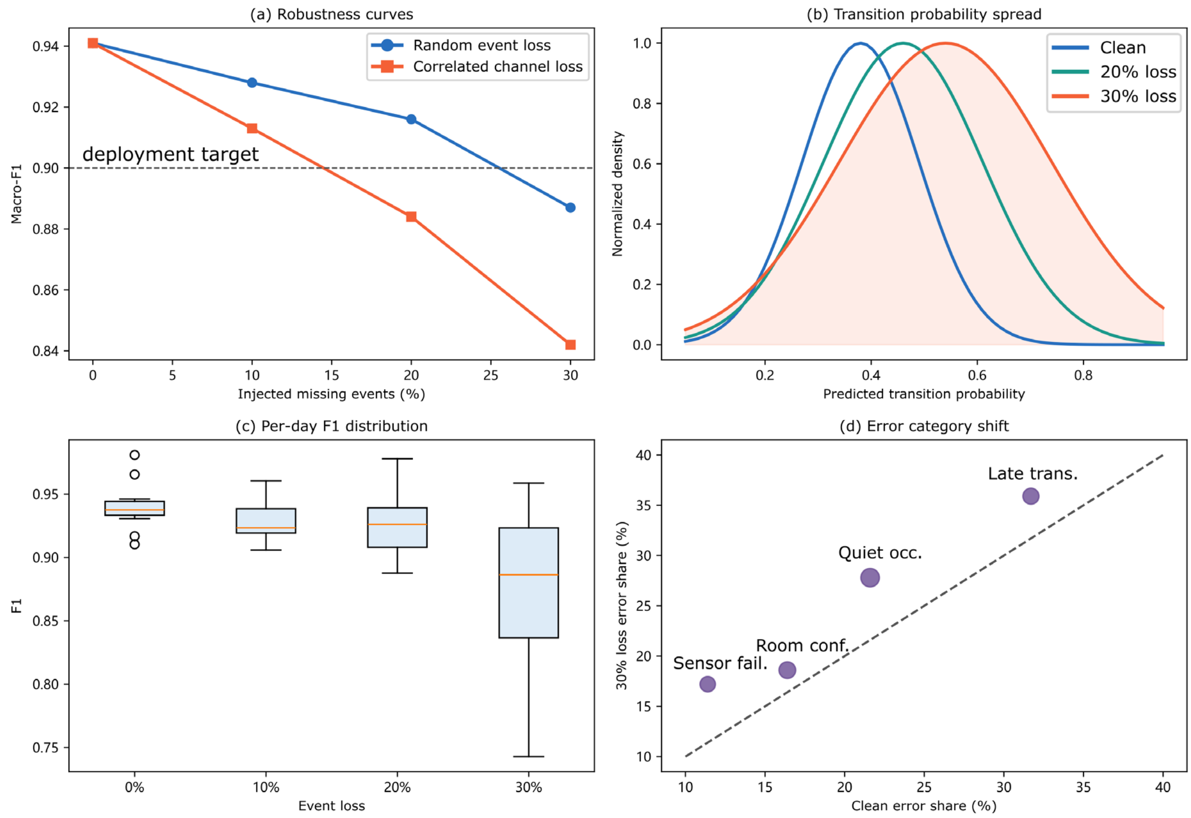

Figure 5. Robustness under missing or corrupted events: (a) macro-F1 under random and correlated loss; (b) transition probability spread; (c) per-day F1 distribution; (d) error-category changes under degradation.

Timestamp jitter was less severe than event deletion. When Gaussian timestamp noise with a 5 s standard deviation was added, macro-F1 decreased by 0.7 percentage points; when the standard deviation increased to 20 s, macro-F1 decreased by 2.8 percentage points. The model was still feasible because the temporal voxel branch pooled events in a short neighborhood and the elapsed-time embedding represented intervals coarsely enough to tolerate small bus delays. Correlated channel dropouts were relatively large. Removing all illumination events reduced the macro-F1 score by 1.9 per cent, removing acoustic events reduced it by 2.4 per cent, and removing both passive infrared and door contact reduced it by 7.6 per cent. The strongest connection is at the entrance, and the door channel will show an earlier change in occupancy. Based on the above tests, robustness is not a single quantity but a channel-dependent feature determined by sensor physics and the purpose of the room.

Missingness simulation also investigated recovery after a sensor was brought back online. The channel was restored two hours after being removed, and the reliability score gradually increased over the next 15 minutes instead of immediately returning to 1. This prevented the classifier from overreacting to the first restored event, as it may be that accumulated physical changes rather than a new resident action. In the restored-channel test, immediate trust recovery resulted in 14 more false transition alarms, and gradual recovery produced five. Therefore, the most noticeable results were in the carbon dioxide proxy and humidity channels, as they have older room data. Door contact recovered more rapidly due to discrete events and reduced impact of environmental inertia. This behavior is a deployment detail, but it directly affected the false-alarm stability of the testbed.

Latency and resource allocation were measured on an edge gateway equipped with a quad-core low-power processor and 4GB of memory. Batch size was fixed to one to reflect online inference. The full PVCNN-Transformer used 1.84 million parameters, required 13.6 ms median inference time per decision, and consumed 68 MB of peak memory. The vanilla Transformer required 11.8 ms but had lower transition F1, while the graph-temporal model required 18.9 ms because room adjacency operations were less efficient in the runtime used for deployment. Figure 6 reports latency, memory, parameter count, and energy per 1000 decisions for all models; the proposed model used 0.41 Wh per 1000 decisions, compared with 0.37 Wh for the vanilla Transformer and 0.56 Wh for the graph-temporal model. The additional cost over direct attention is modest relative to the accuracy gain, which is important because smart-home controllers often share compute resources with automation, logging, and local privacy services [20].

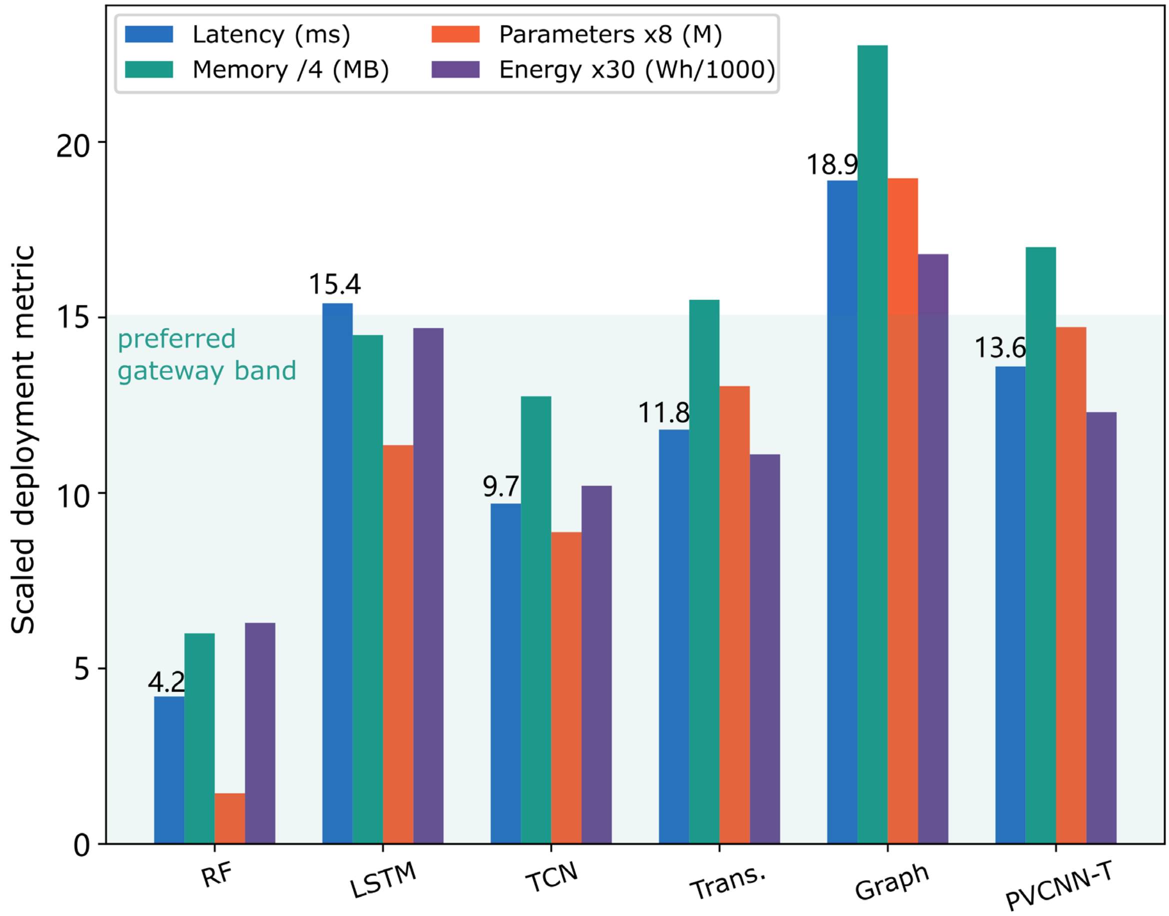

Figure 6. Edge deployment profile comparing latency, memory, parameter count, and energy consumption across occupancy models.

Many home gateways run multiple services simultaneously, so we checked the model size. Quantizing the linear layer to 8-bit weights reduced the model size from 68 MB to 39 MB and only slightly increased the median latency due to reduced memory traffic and a small compute overhead from dequantization. Macro-F1 after post-training quantization was 0.936, a drop of 0.005 from the floating-point model. Pruning 20% of the feed-forward weights achieved a similar reduction in memory but increased the transition-F1 loss; thus, compression should preserve attention and point-voxel fusion parameters before removing temporal capacity. The last deployment profile chose quantization over pruning. The rolling-buffer inference mechanism also limited memory growth by not storing the raw stream history after event conversion and feature extraction.

Calibration, Temporal Stability, and Error Analysis

Energy interpretation depends on decision frequency. Approximately 9.4 event arrivals were triggered per minute during the active evening period, and only one heartbeat decision per minute occurred in the unoccupied daytime. Given the above mixed schedule, the estimated daily inference energy is 0.038 Wh and significantly lower than that of lighting or climate control. Therefore, the relevant deployment cost is not average energy consumption but latency under bursts, as several sensors may trigger almost simultaneously when a resident

https://doi.org/10.64972/dea.2023.v2i3.3737d:81-95

enters the home. Stress testing with 20 events arriving in five seconds produced a 95th-percentile decision latency of 21.7 ms, which was still below the 100 ms internal limit of the automation controller. The queue length never exceeded four decisions because event conversion and model inference were both lightweight.

Calibration and temporal stability were examined because an occupancy classifier used for control must not merely be accurate; it must also express uncertainty in ambiguous intervals. Reliability diagrams showed that the proposed model was close to the diagonal for confidence bins between 0.4 and 0.9, while the vanilla Transformer was overconfident in the 0.7-0.9 range. The expected calibration error of PVCNN-Transformer was 0.031, compared with 0.047 for the graph-temporal model and 0.064 for the attention-only ablation. Figure 7(a) plots predicted confidence against empirical correctness, Figure 7(b) shows the signed delay distribution for transition detection with a median of 8.2 s, Figure 7(c) presents the cumulative error duration by room where the bathroom contributed 23.5% of total error time, and Figure 7(d) compares false-alarm density by hour with peaks at 06:00 and 22:00. Similar calibration concerns have been discussed in sensor-driven building control because overconfident false occupancy can waste energy and overconfident false vacancy can reduce comfort [21].

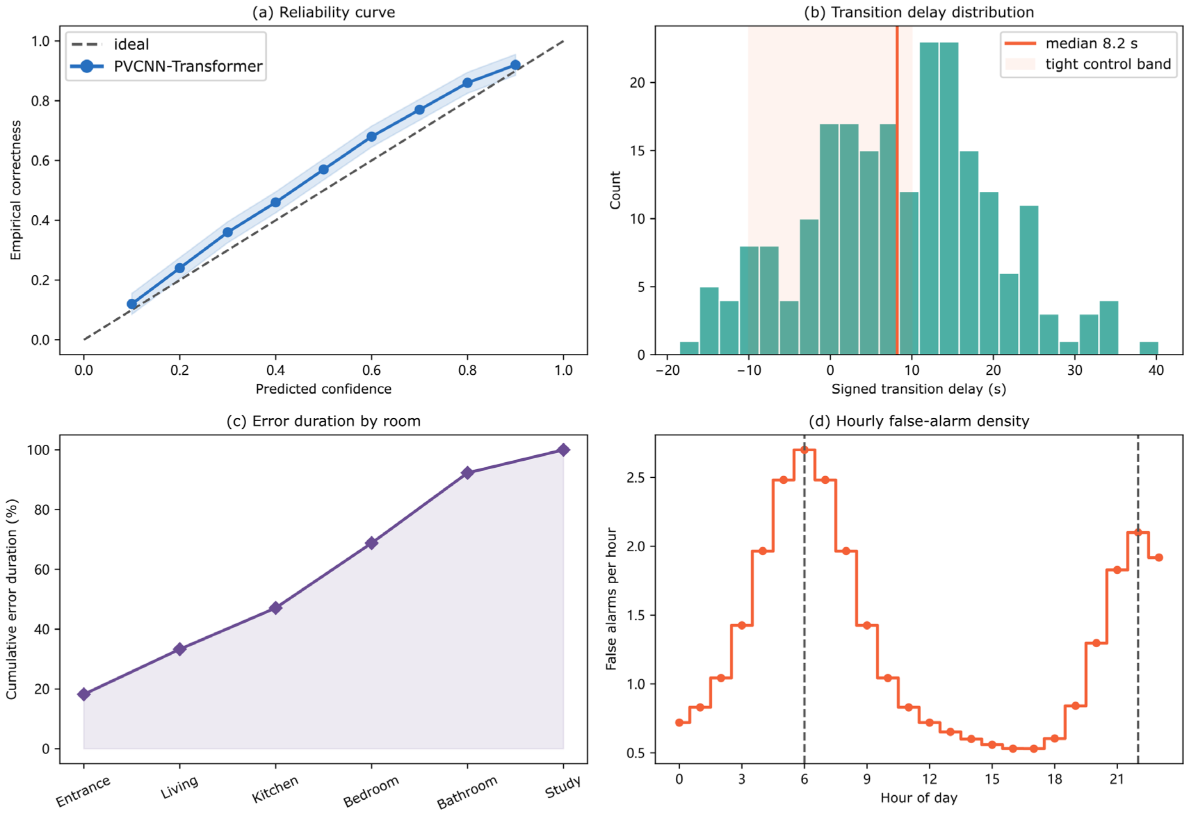

Figure 7. Calibration and temporal error characteristics: (a) reliability curve; (b) transition delay distribution; (c) cumulative error duration by room; (d) hourly false-alarm density.

State oscillations shorter than two minutes were treated as an operational indicator of unstable switching. Oscillation will cause the lighting and HVAC controllers to switch too often; thus, it is not desired. The LSTM produced 43 short oscillations in the test split, the vanilla Transformer produced 37, and PVCNN-Transformer produced 19. Most of the remaining oscillations occurred near the bathroom and entrance, and these involved brief visits or door openings. A hysteresis layer was tested outside the classifier, and two consecutive occupied predictions were required to switch from unoccupied to occupied. The oscillations were reduced to 11 at this level, but the median transition delay rose to 16.7 s. Therefore, the reported results use the raw classifier output and do not perform post-processing to conceal timing errors in the model's behavior.

Finally, the errors in the audit were categorized into late transitions, early transitions, room confusion, quiet occupancy, and sensor faults. Late transitions account for 31.7% of all errors, early transitions for 18.9%, room confusion for 16.4%, quiet occupancy for 21.6%, and sensor faults for 11.4%. Quiet occupied errors were concentrated in the study and bedroom, and passive infrared events can be absent for extended periods there. Sensor-failure errors were less frequent but operationally significant, as a missing door-contact heartbeat could make the entry-and-exit hypothesis ambiguous. Literature on long-term home sensing indicates that label uncertainty and resident routine drift can affect the measured performance [22]. To reduce this effect, the present evaluation used day-level splits and manually reviewed transition labels within a 30 s tolerance. The tolerance did not hide major failures: when the tolerance was reduced to 10 s, transition F1 decreased from 0.872 to 0.839, and when it was expanded to 60 s, transition F1 increased to 0.895. These data show that the model is sensitive to practical control horizons rather than only to coarse daily occupancy states.

A secondary analysis compared event-stream modeling with fixed 60 s and 300 s window aggregation. The 60 s window model produced 93.2% accuracy and 0.902 macro-F1, while the 300 s window model produced 92.5% accuracy and 0.884 macro-F1. The event-stream model produced 95.8% accuracy and 0.941 macro-F1 while using fewer input tokens than the raw 1 Hz stream. The difference was most visible during short kitchen and entrance visits, where window aggregation mixed a brief occupied state with neighboring unoccupied intervals. Prior work has reported comparable boundary effects when occupancy labels are evaluated at coarse temporal resolution [23]. In the present data, 17.8% of all transition intervals lasted less than two minutes, which explains why aggressive aggregation penalized transition F1. The event representation also made the system easier to inspect because each token could be traced to a concrete sensor change.

Then, it was tested on a cross-week regular shift. The first part of the corpus was a regular weekday schedule, and the second part included two remote-work days and one late-night return. Models that heavily rely on clock-time priors showed noticeable degradation for these shifted days. The proposed model reduced this effect by using clock time as one embedding among several, rather than as a primary handcrafted feature. On shifted days, the macro-F1 scores for PVCNN-Transformer, vanilla Transformer, temporal convolution, and LSTM were 0.918, 0.887, 0.861, and 0.842, respectively. Studies of residential automation have shown that the behavior drift of residents is a problem for long-term occupancy learning [24]. Based on the above results, local event neighborhoods help the model adapt to schedule changes because the entry, movement and environment-response patterns are still informative even after altering the hour of occurrence.

Several implementation choices influenced the measured performance. A context length of 256 events provided the best validation macro-F1; reducing it to 128 events lowered macro-F1 by 1.6 percentage points, while increasing it to 512 events added 42% latency with only 0.3 percentage-point improvement. The voxel size was also important: overly coarse temporal cells merged separate actions, whereas very fine cells weakened pooling benefits. The selected 90 s temporal voxel with room-channel grouping balanced transition sensitivity and noise suppression. Comparable sensitivity to temporal scale has been observed in environmental time-series learning for building applications [25]. During deployment simulation, the model updated the occupancy state every time a new event arrived and also emitted a heartbeat decision every 60 s during inactivity. This design avoided stale states after long quiet periods without forcing the gateway to run full inference at high fixed frequency.

The experimental results should be interpreted within the sensing conditions used here and require validation across additional homes, sensor brands and multi-resident routines. The corpus was obtained in a single instrumented apartment, but day-level divisions, multiple rooms and induced missingness were employed to reduce optimistic bias. The model does not infer resident identity, and the labels distinguish occupancy state rather than activity class. This scope is by design, and many automation decisions only need presence, transition and vacancy without fine-grained activity recognition. Multi-resident homes can still produce bursts of local events simultaneously and are thus difficult to attribute to a single-occupancy state. Recent work on privacy-preserving smart-home analytics has emphasized the tension between richer sensing and acceptable domestic monitoring [26]. The proposed Design remains on the environmental-sensor side of that boundary and extracts more value from sensors that are already common in building automation through event structure and temporal attention.

Conclusion

This paper proposes a PVCNN-Transformer model for smart-home occupancy inference based on environmental sensor event streams. The model represents asynchronous sensor changes as room-indexed event points, extracts local cross-channel neighborhoods via point-voxel convolution, and uses Transformer attention to model extended temporal dependencies. Within the tested sensing conditions, the architecture alleviates key limitations of simple windowed classifiers and direct attention models by combining sparse local structure with long-context reasoning.

The work has deficiencies. The experimental corpus is relatively extensive and covers multiple rooms and various situations, but it still lacks the full scope of multi-resident ambiguity, seasonal HVAC variations, device aging, etc. The model must also set thresholds for event construction, and these thresholds may need to be adjusted according to the sensor brand and installation site. Environmental sensors are used to protect the privacy of sight, but it is unknown whether the place is quiet and occupied or actually empty.

Future work will extend cross-home adaptation, self-supervised pretraining on unlabeled event streams, federated updating and uncertainty-aware control policies that can determine whether an occupancy estimate is sufficiently reliable for automation. Additional work will explore federated updating across homes, resident-controlled privacy settings, and hardware-aware compression to deploy the model on a small gateway while maintaining strong transition detection.

Author Contributions

Hsiu-chen Yang contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Hsiao-hung Yao contributes to analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Rueda, L., Agbossou, K., Cardenas, A., Henao, N., & Kelouwani, S. (2020). A comprehensive review of approaches to building occupancy detection. Building and Environment, 180, 106966. https://doi.org/10.1016/j.buildenv.2020.106966

[2] Trivedi, D., & Badarla, V. (2020). Occupancy detection systems for indoor environments: A survey of approaches and methods. Indoor and Built Environment, 29(8), 1053-1069. https://doi.org/10.1177/1420326X19875621

[3] Sun, K., Zhao, Q., & Zou, J. (2020). A review of building occupancy measurement systems. Energy and Buildings, 216, 109965. https://doi.org/10.1016/j.enbuild.2020.109965

[4] Mena, A. R., Ceballos, H. G., & Alvarado-Uribe, J. (2022). Measuring indoor occupancy through environmental sensors: A systematic review on sensor deployment. Sensors, 22(10), 3770. https://doi.org/10.3390/s22103770

[5] Ye, J., O'Grady, M., & Banos, O. (2020). Sensor technology for smart homes. Sensors, 20(24), 7046. https://doi.org/10.3390/s20247046

[6] Sanchez-Comas, A., Synnes, K., & Hallberg, J. (2020). Hardware for recognition of human activities: A review of smart home and AAL related technologies. Sensors, 20(15), 4227. https://doi.org/10.3390/s20154227

[7] Demrozi, F., Pravadelli, G., Bihorac, A., & Rashidi, P. (2020). Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. IEEE Access, 8, 210816-210836. https://doi.org/10.1109/ACCESS.2020.3037715

[8] Eldib, M., Philips, W., & Aghajan, H. (2020). Discovering human activities from binary data in smart homes. Sensors, 20(9), 2513. https://doi.org/10.3390/s20092513

[9] ElHady, N. E., Jonas, S., Provost, J., & Senner, V. (2020). Sensor failure detection in ambient assisted living using association rule mining. Sensors, 20(23), 6760. https://doi.org/10.3390/s20236760

[10] Vanus, J., Fiedorova, K., Kubicek, J., Gorjani, O. M., & Augustynek, M. (2020). Wavelet-based filtration procedure for denoising the predicted CO2 waveforms in smart home within the Internet of Things. Sensors, 20(3), 620. https://doi.org/10.3390/s20030620

[11] Wang, C., Jiang, J., Roth, T., Nguyen, C., Liu, Y., & Lee, H. (2021). Integrated sensor data processing for occupancy detection in residential buildings. Energy and Buildings, 237, 110810. https://doi.org/10.1016/j.enbuild.2021.110810

[12] Dai, X., Liu, J., & Zhang, X. (2020). A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings. Energy and Buildings, 223, 110159. https://doi.org/10.1016/j.enbuild.2020.110159

[13] Tekler, Z. D., Low, R., Gunay, H. B., Andersen, R. K., & Blessing, L. (2020). A scalable Bluetooth Low Energy approach to identify occupancy patterns and profiles in office spaces. Building and Environment, 171, 106681. https://doi.org/10.1016/j.buildenv.2020.106681

[14] Stopps, H., Huchuk, B., Touchie, M. F., & O'Brien, W. (2021). Is anyone home? A critical review of occupant-centric smart HVAC controls implementations in residential buildings. Building and Environment, 187, 107369. https://doi.org/10.1016/j.buildenv.2020.107369

[15] Esrafilian-Najafabadi, M., & Haghighat, F. (2021). Occupancy-based HVAC control systems in buildings: A state-of-the-art review. Building and Environment, 197, 107810. https://doi.org/10.1016/j.buildenv.2021.107810

[16] Azimi, S., & O'Brien, W. (2022). Fit-for-purpose: Measuring occupancy to support commercial building operations: A review. Building and Environment, 212, 108767. https://doi.org/10.1016/j.buildenv.2022.108767

[17] Zhang, W., Wu, Y., & Calautit, J. K. (2022). A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment. Renewable and Sustainable Energy Reviews, 167, 112704. https://doi.org/10.1016/j.rser.2022.112704

[18] Sayed, A. N., Himeur, Y., & Bensaali, F. (2022). Deep and transfer learning for building occupancy detection: A review and comparative analysis. Engineering Applications of Artificial Intelligence, 115, 105254. https://doi.org/10.1016/j.engappai.2022.105254

[19] Tekler, Z. D., & Chong, A. (2022). Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy. Building and Environment, 226, 109689. https://doi.org/10.1016/j.buildenv.2022.109689

[20] Natarajan, A., Vijayakumar, K., & Singh, M. (2022). Occupancy detection and localization strategies for demand modulated appliance control in Internet of Things enabled home energy management system. Renewable and Sustainable Energy Reviews, 167, 112731. https://doi.org/10.1016/j.rser.2022.112731

[21] Zhang, R., Dong, B., O'Neill, Z., Cheng, H., & Hu, F. (2022). Development of a testing and evaluation protocol for occupancy sensing technologies in building HVAC controls: A case study of representative people counting sensors. Building and Environment, 208, 108610. https://doi.org/10.1016/j.buildenv.2021.108610

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2021). Attention is all you need. Communications of the ACM, 64(6), 108-115. https://doi.org/10.1145/3495724

[23] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[24] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325

[25] Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A Transformer-based framework for multivariate time series representation learning. Proceedings of the 27th ACM SIGKDD

Conference on Knowledge Discovery & Data Mining, 2114-2124. https://doi.org/10.1145/3447548.3467401

[26] Guo, M. H., Cai, J. X., Liu, Z. N., Mu, T. J., Martin, R. R., & Hu, S. M. (2021). Pct: Point cloud transformer. Computational visual media, 7(2), 187-199. https://doi.org/10.1007/s41095-021-0229-5