BIO-INSPIRED EVENTS AND WIFI FOR PRIVACY-PRESERVING HAR: A LARGE-SCALE DATASET AND NOVEL APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Human Activity Recognition (HAR) supports applications ranging from healthcare monitoring to human-robot interaction. Existing HAR methods predominantly fuse RGB cameras and WiFi signals. However, RGB cameras raise privacy concerns, introduce redundancy, and degrade in low-light or occluded scenes, while WiFi lacks the spatial resolution to resolve fine-grained motion. Bio-inspired event cameras capture sparse brightness changes with high temporal resolution and reduced appearance detail, motivating a central research question: *How can asynchronous event streams be effectively integrated with heterogeneous wireless signals to achieve robust and privacy-preserving HAR under real-world conditions?* To investigate this question, we introduce **ExWACT**, the first large-scale event-WiFi HAR dataset, comprising **22.5K** paired samples across **6** scenes in challenging simulated and real-world scenarios. By pairing complementary event streams and WiFi CSI, ExWACT enables systematic evaluation of multimodal fusion under diverse sensing conditions and investigation of domain shifts between simulated and real event data. Based on the dataset, we propose a novel temporally consistent fusion framework. **Our key idea** is *to bridge the modality gap between dense wireless signals and sparse, asynchronous event streams via event-WiFi interaction and temporal alignment*. The framework comprises two components. A **Sparse Dual-branch Modality Encoder** extracts motion-sensitive representations tailored to WiFi CSI and event streams. A **Temporally Consistent Bi-directional Fusion (TCBF)** module enables dynamic interaction between WiFi and events to capture latent spatial-temporal cues while enforcing temporal alignment. This design maintains high HAR accuracy and robustness under challenging visual conditions where traditional RGB-WiFi systems fail. Extensive experiments demonstrate that our method significantly outperforms pure WiFi baselines by **+45.48%** and existing fusion strategies by **+13.65%**. ExWACT establishes a **new direction** for privacy-preserving HAR, bridging wireless WiFi and bio-inspired sensing and laying a solid foundation for further research on real-world HAR.

1 INTRODUCTION

Human Activity Recognition (HAR) is a fundamental sensing problem that plays a critical role with a wide range of applications from healthcare monitoring to human-robot interaction. In this case, accurate and reliable understanding of human actions is essential for safe and intelligent system deployment (Sun et al., 2022; Shin et al., 2025). To achieve HAR, a dominant line of research leverages RGB cameras to capture detailed spatial and temporal patterns of human motion (Carreira & Zisserman, 2017; Simonyan & Zisserman, 2014) or wireless sensors, such as WiFi Channel State Information (CSI) (Liu et al., 2020; 2019; Wang et al., 2018; Ding & Wang, 2019) for unobtrusive sensing in indoor environments. Another line of research, *e.g.*, (Feichtenhofer, 2020; Liu et al., 2022; Feichtenhofer et al., 2019) focuses on fusing RGB cameras and WiFi to exploit complementary visual and wireless cues. The intuition is that RGB cameras provide rich spatial and appearance information, enabling fine-grained activity understanding while WiFi CSI captures motion-induced signal variations independent of lighting conditions. As such, improved robustness and performance in indoor environments can be achieved (Guo et al., 2018; Zou et al., 2019; Zhang & Zhang, 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

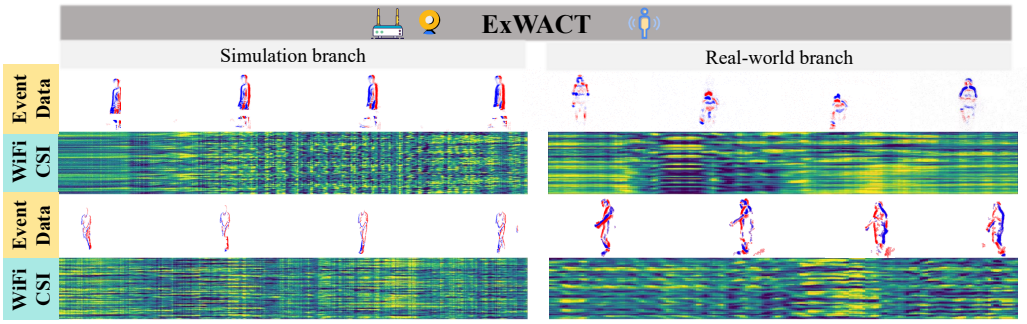

Figure 1: Synthetic and real-world samples from our ExWACT dataset under low-light condition.

However, these kinds of fusion paradigms inherit critical limitations. Specifically, RGB cameras **1)** raise privacy concerns particularly in sensitive indoor settings such as healthcare facilities (Iravantchi et al., 2024; Liu et al., 2025), **2)** introduce substantial data redundancy as consecutive frames often contain overlapping information (Lin et al., 2019; Zheng et al., 2023), and **3)** suffer from performance degradation in low-light, motion blur, or occluded scenes (Tu et al., 2023; Schiappa et al., 2023). By contrast, WiFi signals lack explicit **spatial structure**, resulting in limited spatial resolution and an inability to directly capture fine-grained body configurations (Wang et al., 2018; Ahmad et al., 2024; Li et al., 2025). This often leads to **ambiguity in distinguishing visually similar activities**, especially those involving subtle motion differences or similar temporal patterns. Moreover, WiFi measurements are highly sensitive to environmental dynamics such as multipath propagation, interference, and changes in surrounding objects, which can further degrade recognition performance and generalization (Abuhoureyah et al., 2023; Ahmad et al., 2024).

Event cameras are novel bio-inspired sensors that depart from traditional frame-based imaging by capturing only per-pixel illumination changes asynchronously (*i.e.,* irregularly) and output event streams (Zheng et al., 2023; Gallego et al., 2020). This paradigm offers significant advantages, including microsecond-level temporal resolution, high dynamic range, and less motion blur. This, therefore, motivates us to explore a timely and **novel research question**: *How to effectively integrate event streams with heterogeneous WiFi to achieve robust and privacy-preserving HAR under real-world conditions?* However, addressing this question is non-trivial because WiFi and event cameras exhibit fundamentally different data characteristics: compared with WiFi, event streams are sparse and only encode the edge information (see Fig. 1). This gap introduces two key challenges in representation learning and temporal alignment. **First**, event data is inherently asynchronous in time, making it difficult to directly synchronize with uniformly sampled WiFi signals. Thus, directly applying existing encoding methods, *e.g.,* (Hanchate et al., 2024; Gehrig et al., 2021; Zhu et al., 2025; Zhang et al., 2023b), leads to temporal misalignment (see Tab. 4). **Second**, simply concatenating or aggregating features using existing methods, *e.g.,* (AlRemeithi et al., 2023; Zhou et al., 2023; Xu et al., 2024; Xie et al., 2024) may degrade their complementary nature (see Tab. 3). Collectively, these challenges suggest that simply applying existing RGB-WiFi encoding and fusion strategies is insufficient to fully exploit the potential of event cameras and WiFi.

To this end, we introduce **ExWACT**, the **first event-WiFi HAR dataset**, comprising **22.5K** paired samples across **6** scenes spanning simulated and real-world scenarios (Fig. 1). The subsets include a meeting room, an open laboratory area, and an empty room, covering varied illumination and background human movement, alongside challenging settings involving darkness and partial occlusion. This diversity enables evaluation of cross-scene generalization and robustness to visual degradation and background interference. ExWACT thus establishes a benchmark for exploring complementary event-based vision and WiFi CSI through multimodal fusion under controlled and practical conditions. Based on the dataset, we propose a *temporally-consistent fusion framework* that aligns dense wireless representations with sparse asynchronous events through cross-modal interaction and temporal consistency learning. Addressing the first challenge, the **Sparse Dual-branch Modality Encoder** (Sec. 3.2.1) extracts motion-sensitive features from each modality, using *spike-driven voxelization and temporal aggregation* for events while preserving modality-specific characteristics. Addressing the second, the **Temporally Consistent Bi-directional Fusion** module (Sec. 3.2.2) combines attention-based feature exchange with temporal alignment to capture complementary spatiotemporal cues and mitigate inconsistencies from irregular event timing and WiFi variations.

In summary, **our main contributions** are: (**I**) We introduce **ExWACT**, to the best of our knowledge, the **first event-WiFi HAR dataset**, with **22.5K** paired samples across **6** scenes, supporting systematic evaluation under simulated and real-world conditions (Sec. 3.1). (**II**) We establish, to the best of our knowledge, the **first** *event-WiFi multimodal HAR paradigm*, introducing event-based vision as a privacy-preserving and temporally efficient alternative to RGB in conventional RGB+WiFi sensing. (**III**) We propose the **framework**, bridging dense WiFi CSI and sparse asynchronous events through sparse dual-branch encoding (Sec. 3.2.1) and temporally consistent cross-modal attention fusion (Sec. 3.2.2). (**IV**) Extensive experiments demonstrate improvements of **45.48%** over *pure WiFi* baselines and **4.96%** over *pure event* methods, alongside enhanced **privacy preservation** and **temporal efficiency**. They also reveal the impact of simulated-to-real event **domain shift**, informing future research on robust HAR.

2 Related Work

Visual motion sensing for HAR. RGB videos have long dominated vision-based HAR due to their rich cues on posture, appearance, motion, and context (Dang et al., 2020; Pareek & Thakkar, 2021). Methods have evolved from hand-crafted features to deep architectures, including 2D CNNs with temporal modules (Lin et al., 2019), motion-aware designs such as two-stream and recurrent models (Simonyan & Zisserman, 2014; Donahue et al., 2015), 3D spatiotemporal convolutions (Feichtenhofer et al., 2019), and Transformer-based video models (Bertasius et al., 2021; Liu et al., 2022). Despite strong performance, these approaches rely on dense RGB observations, raising privacy concerns and remaining sensitive to illumination and occlusion, limiting their applicability in privacy-critical indoor settings (Yang et al., 2024; Grover et al., 2023). Event cameras offer an alternative by asynchronously capturing brightness changes instead of full frames, providing high temporal resolution, low redundancy, and inherent suppression of static appearance (Gallego et al., 2020). Existing methods typically represent events as frames, voxel grids, or time surfaces for compatibility with CNNs and Transformers (Lagorce et al., 2016; Zhu et al., 2019; Gehrig et al., 2019; De Blegiers et al., 2023), while others design architectures aligned with event sparsity, such as spiking neural networks and graph-based models (Amir et al., 2017; Bi et al., 2020; Liu et al., 2021). These properties make event cameras particularly suitable for privacy-aware and dynamic HAR, especially under degraded RGB conditions, although they remain limited in capturing static semantic details (Wang et al., 2024).

Wireless sensing for HAR. WiFi CSI-based sensing provides a device-free and privacy-preserving alternative to camera-based HAR by inferring human motion from wireless propagation dynamics (Yousefi et al., 2017; Ding & Wang, 2019; Yan et al., 2024). Early approaches relied on handcrafted features and classical classifiers, with systems such as E-eyes demonstrating activity recognition from fine-grained CSI signatures (Wang et al., 2014). Subsequent works introduced physically grounded models, e.g., Fresnel-zone-based formulations, to improve interpretability and capture geometry-dependent signal variations (Zhang et al., 2023a; Abuhoureyah et al., 2024). With deep learning, recurrent, convolutional, and attention-based models became dominant for learning representations directly from CSI (Yousefi et al., 2017; Chen et al., 2018; Ding & Wang, 2019). Recent methods further exploit CSI structure for enhanced spatiotemporal modeling, as in THAT and AutoDLAR (Li et al., 2021; Lu et al., 2024), while advancing toward robust and human-centric perception, including domain-invariant representations (e.g., Widar3.0) and structured outputs such as pose and 3D reconstruction (e.g., WiPose and Person-in-WiFi 3D) (Zheng et al., 2019; Jiang et al., 2020; Yan et al., 2024). Despite these advances, WiFi sensing remains sensitive to environmental factors and offers limited spatial granularity compared to vision (Zheng et al., 2019; Wang et al., 2019b), motivating its integration with visual modalities for richer spatial-temporal understanding.

Multimodal fusion for privacy-aware HAR. The complementary nature of event-based and wireless sensing has catalyzed a growing body of multimodal HAR research, offering a privacy-preserving alternative where sparse temporal events capture motion dynamics and WiFi signals provide robust sensing beyond visual occlusions. Existing WiFi-video systems show that combining radio and vision can improve robustness under occlusion, user-position variation, or degraded visibility, since WiFi provides privacy-friendlier coarse motion cues while the visual branch contributes stronger spatial understanding (Zou et al., 2019; Chen et al., 2023; Wei et al., 2024; Deng et al., 2022). Beyond direct inference-time fusion, recent works have also explored cross-modal supervision and multimodal pretraining, where vision assists WiFi representation learning during training while preserving more

Table 1: Comparison of wireless sensing datasets with auxiliary visual modalities for human activity understanding. *Note:* * denotes datasets that are not publicly available.

Wireless Sensing Dataset	Auxiliary Visual Modality	Visual Privacy Processing	Multi-view	Illumination Variation	Occlusion	# Scenes	Scale
UT-HAR (Yousefi et al., 2017)	N/A	N/A	✗	✗	✗	1	4,973
NTU-HAR (Yang et al., 2022)	N/A	N/A	✗	✗	✗	1	1,200
ARIL (Wang et al., 2019a)	N/A	N/A	✗	✗	✗	1	1,394
WiAR (Guo et al., 2019)	N/A	N/A	✗	✗	✗	3	4,161
Yang et al.* (Yang et al., 2021)	N/A	N/A	✗	✗	✗	3	1,050
RF-NET (Ding et al., 2020)	N/A	N/A	✗	✗	✗	6	12,000
WiMANS (Huang et al., 2024)	RGB	Anonymization required	✔	✗	✔	3	11,286
OPERAnet (Bocus et al., 2022)	Skeleton	Annotation / extraction required	✗	✗	✗	2	6,235
MMFi (Yang et al., 2023b)	RGB-D / LiDAR	Anonymization required	✗	✗	✗	4	1,080
ExWACT (Ours)	Event	No post anonymization required	✔	✔	✔	6	22,572

flexible deployment options (Lu et al., 2024; Yang et al., 2024). Public multimodal datasets such as OPERAnet, MM-Fi and WiMANS have further enabled systematic study of multimodal sensing and generalization (Bocus et al., 2022; Yang et al., 2023b; Huang et al., 2024).

3 METHODOLOGY

3.1 DATASET CONSTRUCTION

Since event-WiFi sensing is a novel paradigm and no existing dataset provides such paired modalities as shown in Tab. 1, we construct the first unified dataset in this work that comprises both generated and real-world data as shown in Fig. 2 and Appendix A. Specifically, we leverage publicly available RGB-WiFi datasets and convert the RGB modality into simulated event streams using the *v2e* converter (Hu et al., 2021), forming generated event-WiFi pairs. In addition, we also collect real-world scenarios with naturally synchronized event data and WiFi signals. This unified dataset enables comprehensive evaluation of the proposed framework in both simulated and real scenarios.

We use WiMANS (Huang et al., 2024) as the primary source for the simulated multimodal benchmark dataset. Since WiMANS does not provide event data directly, a video-to-event conversion process is applied to generate simulated event streams from the RGB videos. Specifically, *v2e* is implemented, which converts frame-based videos into synthetic DVS events by approximating the event generation process of dynamic vision sensors. The converter first transforms RGB frames into luma frames, optionally upsamples them in time using the Super-SloMo interpolation network, and then maps the intensity values into a logarithmic domain before generating events based on brightness changes. This design makes the converted event data closer to the operating principle of event cameras rather than a frame differencing process (Hu et al., 2021).



Figure 2: Key attributes of our ExWACT dataset. More info in App. A.

In our implementation, the duration of DVS exposure is set at 0.033, the output sensor resolution is set to `DVS346` (346 × 260), and the timestamp resolution is set to 0.001. The Super-SloMo interpolation network is used inside the converter to upsample the input RGB videos temporally, so that intermediate brightness changes can be better approximated before event generation. As a result, the simulated event data preserve the fine-grained motion change information in the original RGB videos while providing an event-based representation compatible with the proposed multimodal framework. Examples of the simulated event data are shown in Fig. 1 and Appendix A.1.

To further evaluate the proposed model under real-world conditions, we construct a real-world event-WiFi subset as illustrated in Appendix A.2. It comprises 10 subjects performing 7 additional daily activities – *standing*, *waving*, *jumping*, *picking up*, *walking*, *sitting down*, and *standing up* – across 3

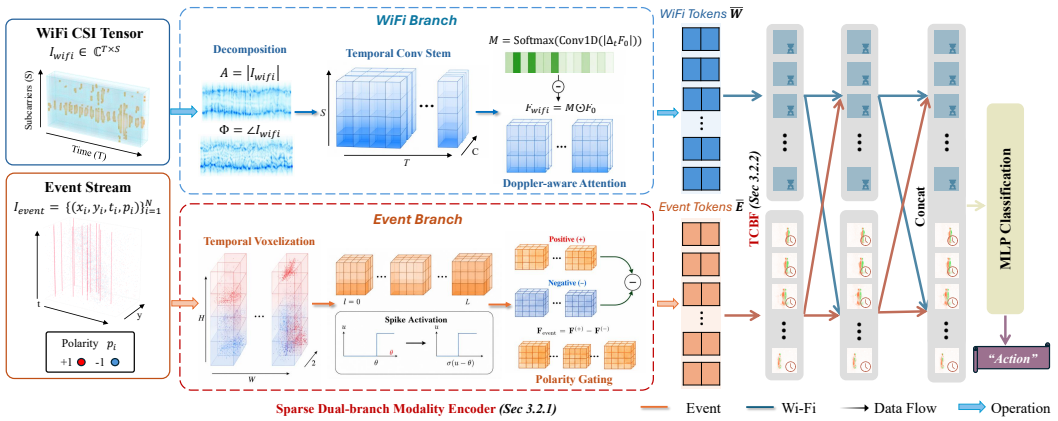

Figure 3: **Overview of the ExWACT framework.** The Sparse Dual-branch Modality Encoder extracts motion-sensitive WiFi and event representations, while TCBF integrates complementary cues through bidirectional attention and temporal consistency alignment.

indoor environments with extreme visual conditions, including low-light and dynamic background scenarios. Each sample is recorded for 5 seconds, with 10 repetitions per activity per subject. To enhance diversity and cross-scene generalization, participant assignments are partially overlapped across scenes, enabling both subject variability and domain continuity. WiFi CSI is collected using modern ***IAX* tool** (Zhang et al., 2025) with a 2×2 MIMO configuration and 114 subcarriers, resulting in data of shape $T \times 2 \times 2 \times 114$. Event data are captured using a *Prophesee Metavision EVK5* camera. A unified Linux-based acquisition pipeline ensures synchronized recording with timestamp alignment at 0.1 ms precision, preserving fine-grained temporal correspondence for cross-learning.

Remarks. Overall, the dataset spans both pure and challenging sensing conditions (*e.g.,* low-light and dynamic-occlusion conditions), enabling comprehensive evaluation of our approach and future methods in both real-world and simulation scenarios as shown in Fig. 14 and Fig. 15. *More details on data collection and statistics are provided in Appendix A.*

3.2 The Proposed Approach

Overview. To support evaluation on the ExWACT dataset, we provide an accompanying framework, also named ExWACT (Fig. 3). It combines a **Sparse Dual-branch Modality Encoder** (Sec. 3.2.1) with **Temporally Consistent Bi-directional Fusion (TCBF)** (Sec. 3.2.2) to learn complementary, temporally aligned representations from WiFi CSI and event streams. Detailed formulations and implementation are provided in Appendix B.

3.2.1 Sparse Dual-branch Modality Encoder

Insights. WiFi CSI captures global motion-induced signal variations but is affected by static reflections and limited spatial resolution. Event streams provide fine-grained motion cues but require processing that accommodates their sparsity and asynchronous timing. The dual-branch encoder addresses these differences through modality-specific processing before projecting features into a shared embedding space.

WiFi Branch: Motion-sensitive Temporal Encoding. Given CSI measurements $\mathcal{I}_{\text{wifi}} \in \mathbb{C}^{T \times S}$, with T temporal samples and S subcarriers, we compute amplitude $\mathbf{A} = |\mathcal{I}_{\text{wifi}}|$ and phase $\mathbf{\Phi} = \angle \mathcal{I}_{\text{wifi}}$. Temporal differencing emphasizes dynamic variations while suppressing static components:

$$\mathbf{X}_{\text{wifi}} = \text{Concat}(\Delta_t \mathbf{A}, \Delta_t \mathbf{\Phi}) \in \mathbb{R}^{T \times S \times 2}. \tag{1}$$

A lightweight temporal convolutional stem produces $\mathbf{F}_0 = \text{Conv1D}(\mathbf{X}_{\text{wifi}})$, capturing local temporal dependencies. A dynamic Doppler-aware attention module then computes motion saliency and refines the features:

$$\mathbf{M} = \text{Softmax}(\text{Conv1D}(|\Delta_t \mathbf{F}_0|)), \qquad \mathbf{F}_{\text{wifi}} = \mathbf{M} \odot \mathbf{F}_0. \tag{2}$$

Here, $\odot$ denotes element-wise multiplication. This weighting emphasizes motion-sensitive responses while reducing static interference. The refined features are projected into WiFi tokens $\mathbf{W} \in \mathbb{R}^{L_w \times D}$.

Event Branch: Spike-driven Spatio-temporal Encoding. An event stream $\mathcal{I}_{\text{event}} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$ records spatial coordinates, timestamps, and polarities $p_i \in \{-1, +1\}$. We accumulate events into B temporal bins to obtain a voxel grid $\mathbf{V} \in \mathbb{R}^{B \times H \times W \times 2}$, retaining separate positive and negative polarity channels. A polarity-aware spiking convolutional encoder processes the grid:

$$\mathbf{F}_{l+1} = \sigma(\text{Conv3D}(\mathbf{F}_l) - \theta), \tag{3}$$

where σ is a thresholding function and θ is a learnable firing threshold. Sparse activations capture motion edges and temporal discontinuities. Polarity gating computes $\mathbf{F}_{\text{event}} = \mathbf{F}^{(+)} - \mathbf{F}^{(-)}$, retaining the distinction between opposite brightness changes. The resulting features are flattened and projected into event tokens $\mathbf{E} \in \mathbb{R}^{L_e \times D}$, complementing the global motion cues encoded by the WiFi branch.

3.2.2 Temporally Consistent Bi-directional Fusion (TCBF)

Insights. TCBF combines bidirectional feature interaction with explicit temporal alignment. Within an action-level window Δt, event and WiFi tokens receive positional encodings that preserve their temporal structure. Their sequence lengths L_e and L_w may differ, allowing interaction without strict index-wise correspondence.

Bidirectional Cross-modal Attention. Event tokens attend to WiFi tokens, and WiFi tokens symmetrically attend to event tokens:

$$\begin{aligned} \text{Attn}_{e \leftarrow w} &= \text{Softmax}\left(\frac{\mathbf{Q}_e \mathbf{K}_w^\top}{\sqrt{d}}\right)\mathbf{V}_w, \\ \text{Attn}_{w \leftarrow e} &= \text{Softmax}\left(\frac{\mathbf{Q}_w \mathbf{K}_e^\top}{\sqrt{d}}\right)\mathbf{V}_e, \end{aligned} \tag{4}$$

where queries, keys, and values are learned linear projections of the corresponding tokens, and d is the query/key dimension. This bidirectional exchange allows each modality to incorporate complementary motion cues from the other.

Temporal Consistency Alignment. Cross-attention supports feature interaction but does not explicitly enforce temporal coherence. We therefore aggregate tokens within corresponding temporal segments:

$$\bar{\mathbf{e}}_t = \frac{1}{|\mathcal{S}_e(t)|} \sum_{i \in \mathcal{S}_e(t)} \mathbf{E}_i, \qquad \bar{\mathbf{w}}_t = \frac{1}{|\mathcal{S}_w(t)|} \sum_{j \in \mathcal{S}_w(t)} \mathbf{W}_j, \tag{5}$$

where $\mathcal{S}_e(t)$ and $\mathcal{S}_w(t)$ contain tokens associated with segment t. Temporal consistency is encouraged:

$$\mathcal{L}_{\text{TC}} = \sum_t \|\phi(\bar{\mathbf{e}}_t) - \psi(\bar{\mathbf{w}}_t)\|_2^2, \tag{6}$$

where ϕ and ψ are learnable projections into a shared latent space. The constraint aligns corresponding temporal summaries despite differing token counts and modality-specific representations. **Fusion and Optimization.** The refined tokens $\mathbf{E}'$ and $\mathbf{W}'$ are concatenated and pooled into a global representation for activity prediction: $\hat{\mathbf{y}} = f_{\text{cls}}(f_{\text{pool}}([\mathbf{E}'; \mathbf{W}']))$. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{TC}}$, where λ balances classification and temporal alignment. Together, feature exchange and temporal consistency support recognition under heterogeneous and asynchronous sensing conditions.

4 Experiments and Evaluation

4.1 Experimental Settings and Implementation Details

Datasets. We evaluate our method on the proposed **ExWACT** dataset, consisting of synchronized WiFi CSI and event-based vision data collected in real-world indoor environments. The dataset contains 22,572 CSI samples from dual WiFi bands with aligned event clips, covering diverse scenes including classroom, meeting room, and empty room settings.

Table 2: Validation of the proposed ExWACT dataset. Existing WiFi- and event-based baseline models are applied to validate across different scenes.

Modality	Model	Classroom	Meeting Room	Empty Room	Average
WiFi	ARIL (Wang et al., 2019a)	23.46	16.76	22.91	21.04
	STRF (Yousefi et al., 2017)	28.61	28.58	28.51	28.57
	SenseFi (Yang et al., 2023a)	28.98	29.09	28.71	28.93
	LSTM (Yousefi et al., 2017)	29.77	29.68	29.06	29.50
	ABLSTM (Chen et al., 2018)	30.73	26.82	27.93	28.49
	CLSTM (Mo & Kim, 2021)	**31.12**	**30.65**	**30.70**	30.82
	THAT (Li et al., 2021)	30.73	27.93	29.61	29.42
Event	EvT (Sabater et al., 2022)	**73.74**	**77.65**	**78.77**	76.72
	DECOLLE (Kaiser et al., 2020)	58.66	58.10	45.81	54.19
	Slayer (Shrestha & Orchard, 2018)	58.66	46.93	54.19	53.26

Baselines. We compare against representative WiFi-based methods (ARIL (Wang et al., 2019a), ABLSTM (Chen et al., 2018), THAT (Li et al., 2021)) and event-based models (EvT (Sabater et al., 2022), DECOLLE (Kaiser et al., 2020), Slayer (Shrestha & Orchard, 2018)), covering both conventional and neuromorphic approaches. *Detailed baseline constructions are provided in Appendix C.1.*

Implementation Details. We implement **ExWACT** in PyTorch and train it end-to-end using AdamW with cosine learning rate scheduling and a 5% warmup. The objective combines cross-entropy loss with the proposed temporal consistency alignment (Sec. 3.2.2), weighted by λ. For preprocessing, WiFi CSI is transformed into time-frequency representations using amplitude and phase decomposition and temporal differencing, while event streams are voxelized into $B{=}5$ temporal bins with polarity channels. The WiFi encoder uses a temporal convolutional stem with Doppler-aware attention, and the event encoder adopts a polarity-aware sparse convolutional design. Both branches project features into a shared embedding space ($D{=}256$), followed by a multi-head (8 heads) bidirectional cross-modal attention module. We use a base learning rate of 3×10^{-4}, weight decay 0.05, batch size 64, and gradient clipping at 1.0. Label smoothing (0.1) and a warm-in strategy for λ are applied for stable optimization. The model is trained for 300 epochs on 4 GPUs, and the best checkpoint is selected based on validation accuracy. *Additional implementation details are provided in Appendix C.2.*

4.2 Quantitative Comparison and Validation

Validating the ExWACT Dataset. We conduct quantitative comparisons on the proposed ExWACT benchmark to validate the effectiveness of event-based sensing, WiFi sensing, and their multimodal fusion for human activity recognition. As shown in Tab. 2, single-modality baselines exhibit clear modality-dependent behavior across different scenes. WiFi-based methods achieve relatively limited performance, with average accuracies ranging from **21.04%** to **30.82%**. This indicates that although CSI signals can capture global human-induced channel variations, they remain insufficient for distinguishing fine-grained activities under complex environmental changes. In contrast, event-based models obtain substantially higher accuracy, with EvT achieving the best average performance among single-modality methods at **76.72%**. This suggests that event streams provide more motion-sensitive cues than WiFi alone, especially in visually dynamic scenarios.

Evaluating ExWACT. Table 3 further compares different methods on the real-world and simulation branches of ExWACT. Among unimodal baselines, event-based methods generally outperform WiFi-based methods in the real-world branch, while their performance drops significantly in the simulation branch for several models. For instance, DECOLLE achieves 97.46% on the real-world branch but only 46.27% on the simulation branch, indicating limited robustness under domain variation. Similarly, WiFi-only methods obtain much lower simulation accuracy, with the best WiFi baseline reaching only 33.58%. These results demonstrate that the simulation branch introduces a more challenging evaluation protocol and is necessary for testing the generalization ability of HAR models beyond standard real-world recordings.

Compared with unimodal and simple fusion baselines, the proposed ExWACT achieves the best overall performance on both branches. Specifically, ExWACT obtains **99.18%** accuracy in the real-world branch and 91.02% accuracy in the simulation branch, outperforming the strongest single-modality baseline by **1.72%** and **2.77%**, respectively. To further verify the effectiveness of the proposed fusion design, we additionally implement several basic multimodal fusion strategies by combining strong single-modality backbones from the WiFi and event branches. Notably, even

Table 3: Validation of the proposed method. Performance is tested on ExWACT dataset. *Note:* * denotes our implementation of basic multimodal fusion strategies.

Method	Modality	Real-world Accuracy (%)	Simulation Accuracy (%)	Overall Accuracy (%)
ARIL (Wang et al., 2019a)	WiFi	69.84	25.00	39.95
ABLSTM (Chen et al., 2018)	WiFi	72.70	33.58	46.62
THAT (Li et al., 2021)	WiFi	78.73	33.02	48.26
EvT (Sabater et al., 2022)	Event	89.84	88.25	88.78
DECOLLE (Kaiser et al., 2020)	Event	97.46	46.27	63.33
Slayer (Shrestha & Orchard, 2018)	Event	88.89	45.34	59.86
Vanilla Cross Attention*	WiFi+Event	95.32	72.47	80.09
Late Feature Concatenation*	WiFi+Event	96.84	67.93	77.57
AFE + THAT*	WiFi+Event	98.66	70.61	79.96
ExWACT (Ours)	**WiFi+Event**	**99.18**	**91.02**	**93.74**

when compared with manually designed fusion baseline, AFE + THAT, which pairs competitive event-based and WiFi-based models, ExWACT still achieves superior performance. Specifically, ExWACT outperforms the best fusion method by **13.65%** on the overall dataset, demonstrating that our framework is not a simple aggregation of strong unimodal models, but instead learns a more effective and generalizable cross-modal representation. The improvement is particularly significant in the simulation branch, showing that our method **does not merely fit the real-world data distribution but learns more transferable cross-domain representations**. This gain can be attributed to the proposed event-WiFi collaborative modeling, which jointly leverages the fine-grained asynchronous motion cues from event streams and the environment-aware wireless dynamics from WiFi CSI.

4.3 Special Scene Qualitative Analysis

To further examine the robustness of ExWACT under practical deployment conditions, we conduct a qualitative comparison in two challenging scenes: low-light environments and occluded environments, as shown in Fig. 4. These cases are representative failure scenarios for RGB-based human activity recognition, where visual appearance cues are either severely degraded by insufficient illumination or partially missing due to object-level occlusion. As a result, RGB-based methods tend to rely on incomplete or ambiguous spatial observations, leading to incorrect predictions such as confusing "sitting down" with "picking up", or "standing up" with visually similar actions.

In contrast, ExWACT maintains reliable recognition in both scenarios by exploiting the complementary properties of event streams and WiFi CSI. Under the dark scene, although the RGB frames contain limited texture and body-pose information, event signals still capture asynchronous motion changes caused by human movements, while WiFi responses provide additional environment-aware motion dynamics. This enables ExWACT to preserve discriminative temporal patterns and correctly identify actions such as "sitting down", "walking", and "standing up". Similarly, under occlusion, when the human body is partially blocked in RGB frames, the event modality remains sensitive to visible motion boundaries, and WiFi signals can sense body-induced channel variations beyond the direct line of sight. The fused representation therefore remains informative even when the visual observation is incomplete.

4.4 Ablation Studies and Analysis

Modalities. As shown in Tab. 4, both single-modality variants perform worse than multimodal fusion, indicating the complementary nature of WiFi and event-based sensing. The WiFi-only branch captures global motion-induced signal variations but lacks sufficient spatial resolution, while the event-only branch provides fine-grained motion cues but is sensitive to scene-dependent event sparsity. Naive fusion improves over single-modality baselines, but remains limited because it simply combines features without explicitly modeling cross-modal interaction.

Modules. Removing the Doppler-aware attention module leads to a performance drop, demonstrating its importance in emphasizing dynamic human-induced perturbations while suppressing static back-

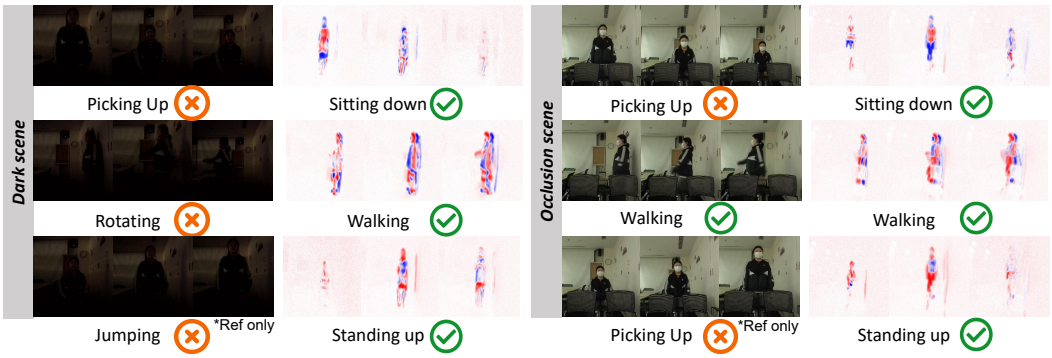

Figure 4: **Comparison between RGB-based methods and the proposed ExWACT under challenging scenes.** RGB images are not part of the dataset and are shown for visual reference only.

Table 4: Ablation studies on core modules.

Variant	WiFi	Event	DA	PA	Bi-CA	TC	Acc. (%)
WiFi only	✔	–	✔	–	–	–	70.37
Event only	–	✔	–	✔	–	–	86.35
Simple Fusion	✔	✔	–	–	–	–	87.92
w/o DA	✔	✔	–	✔	✔	✔	91.63
w/o PA	✔	✔	✔	–	✔	✔	92.08
w/o Bi-CA	✔	✔	✔	✔	–	✔	90.94
w/o TC	✔	✔	✔	✔	✔	–	92.31
Full ExWACT	✔	✔	✔	✔	✔	✔	**93.74**

ground reflections. Similarly, removing the polarity-aware event encoder weakens the model's ability to distinguish motion direction and temporal discontinuities in event streams. Replacing bidirectional cross-modal attention with late fusion further degrades performance, confirming that mutual feature exchange between WiFi and event tokens is necessary for effective multimodal reasoning. Finally, the removal of temporal consistency alignment reduces recognition accuracy, suggesting that explicitly regularizing asynchronous modality correspondence is beneficial for robust event-WiFi fusion.

Cross-scene/domain performance. ExWACT generalizes well on ExWACT across unseen scenes (75.43%-89.70%), and retains strong cross-domain performance (85.36%/87.12%), indicating only moderate degradation under domain shift. Joint training further yields high overall accuracy (93.74%), demonstrating the effectiveness of the proposed multimodal fusion for robust generalization. Detailed results can be found in Tab. 5 and Tab. 6 of Appendix D.

Overall, the full ExWACT model achieves the best performance, validating that sparse dual-branch encoding, bidirectional cross-modal attention, and temporal consistency learning jointly contribute to robust and privacy-preserving HAR.

5 Conclusion and Future Work

In this paper, we introduced **ExWACT**, the first large-scale event-WiFi HAR dataset, comprising **22.5K** paired samples across **6** scenes spanning simulated and real-world scenarios. Its diverse environments and challenging sensing conditions establish a benchmark for studying complementary event-based vision and wireless sensing for privacy-aware HAR. We also proposed the **framework**, which integrates dense WiFi CSI and sparse asynchronous event streams through modality-specific motion encoding and temporally consistent bidirectional fusion. Extensive experiments demonstrate improvements of **+45.48%** over pure WiFi baselines and **+13.65%** over existing fusion strategies, while identifying simulation-to-real domain shift as a key challenge. Future work will investigate domain adaptation and cross-scene generalization to improve robustness under unseen environments and challenging sensing conditions. We will publicly release the dataset, source code, and documentation to support reproducible evaluation and further research on privacy-preserving HAR.

AI USE STATEMENT

We used generative AI tools to assist with refining the interpretation and presentation of author-provided methodological descriptions and experimental results. Additional uses included reorganizing and polishing manuscript text, troubleshooting LaTeX formatting, and developing the ExWACT logo. Generative AI is not a component of the ExWACT recognition model. The simulated event data are produced through the video-to-event conversion pipeline described in Section 3.1, while the real-world subset comprises recordings acquired using event cameras and WiFi hardware. Formal theorem development and proof generation are not applicable to this empirical study. The authors take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ExWACT combines simulated event streams derived from an existing RGB-WiFi dataset with real-world event-WiFi recordings of volunteers performing predefined daily activities. Real-world data collection was conducted with administrative approval and the volunteers' informed consent. The source dataset and external software are acknowledged in the manuscript. Public release will respect participant permissions and the terms governing source data and software, with identifying metadata excluded and any necessary redistribution restrictions documented. Also, event-based vision and WiFi measurements reduces the exposure of dense appearance information compared with RGB recordings.

REPRODUCIBILITY STATEMENT

The construction of ExWACT, including simulated event generation, real-world acquisition, and modality synchronization, is described in Section 3.1. Appendix A provides supplementary dataset statistics, recording environments, and representative samples. The accompanying framework is specified in Sections 3.2.1 and 3.2.2, with additional methodological details in Appendix B. Experimental settings and evaluation results are presented in Section 4, while Appendix C documents preprocessing, baseline implementations, model configurations, and training procedures. Additional cross-scene and cross-domain evaluations are reported in Appendix D. We will publicly release the dataset, source code, and documentation once accepted, subject to the applicable participant permissions and source-data terms, to facilitate reproduction of the reported results and consistent evaluation of future event-WiFi HAR methods.

REFERENCES

Fahd Abuhoureyah, Yan Chiew Wong, Ahmad Sadhiqin Bin Mohd Isira, and Mohammed Nasser Al-Andoli. Csi-based location independent human activity recognition using deep learning. *Human-Centric Intelligent Systems*, 3(4):537–557, 2023.

Fahd Saad Abuhoureyah, Yan Chiew Wong, and Ahmad Sadhiqin Bin Mohd Isira. Wifi-based human activity recognition through wall using deep learning. *Engineering Applications of Artificial Intelligence*, 127:107171, 2024.

Iftikhar Ahmad, Arif Ullah, and Wooyeol Choi. Wifi-based human sensing with deep learning: Recent advances, challenges, and opportunities. *IEEE Open Journal of the Communications Society*, 5: 3595–3623, 2024.

Hamad AlRemeithi, Issac Niwas Swamidoss, Abdulla Al Mansoori, Abdulrahman AlMarzooqi, Slim Sayadi, and Tarek Bouamer. Analysis of modern learning-based multimodal fusion algorithms for neuromorphic vision sensors. In *SPIE Future Sensing Technologies 2023*, volume 12327, pp. 380–391. SPIE, 2023.

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7243–7252, 2017.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, pp. 4, 2021.

Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020.

Mohammud J. Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, and Robert Piechocki. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific Data*, 9(1):474, 2022.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jiajing Chen, Kun Yang, Xiaolong Zheng, Shengbo Dong, Liang Liu, and Huadong Ma. Wimix: A lightweight multimodal human activity recognition system based on wifi and vision. In *Proceedings of the IEEE International Conference on Mobile Ad Hoc and Smart Systems*, pp. 406–414, 2023.

Zhenghua Chen, Le Zhang, Chaoyang Jiang, Zhiguang Cao, and Wei Cui. Wifi csi based passive human activity recognition using attention based blstm. *IEEE Transactions on Mobile Computing*, 18(11):2714–2724, 2018.

L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

Tristan De Blegiers, Ishan Rajendrakumar Dave, Adeel Yousaf, and Mubarak Shah. Eventtransact: A video transformer-based framework for event-camera based action recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–7. IEEE, 2023.

Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Gaitfi: Robust device-free human identification via wifi and vision multimodal learning. *IEEE Internet of Things Journal*, 10(1):625–636, 2022.

Jianyang Ding and Yong Wang. Wifi csi-based human activity recognition using deep recurrent neural network. *IEEE Access*, 7:174257–174269, 2019.

Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pp. 517–530, 2020.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5633–5643, 2019.

Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021.

Shresth Grover, Vibhav Vineet, and Yogesh Rawat. Revealing the unseen: Benchmarking video action recognition under occlusion. *Advances in Neural Information Processing Systems*, 36: 65642–65664, 2023.

Linlin Guo, Lei Wang, Jialin Liu, Wei Zhou, and Bingxian Lu. Huac: Human activity recognition using crowdsourced wifi signals and skeleton data. *Wireless Communications and Mobile Computing*, 2018(1):6163475, 2018.

Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. WiAR: A public dataset for WiFi-based activity recognition. *IEEE Access*, 7:154935–154945, 2019. doi: 10.1109/ACCESS.2019.2948103.

Abhishek Hanchate, Himanshu Balhara, Vishal S Chindepalli, and Satish TS Bukkapatnam. Process signature-driven high spatio-temporal resolution alignment of multimodal data. *arXiv preprint arXiv:2403.06888*, 2024.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1312–1321, 2021.

Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A. McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. In *Proceedings of the European Conference on Computer Vision*, pp. 72–91, 2024.

Yasha Iravantchi, Thomas Krolikowski, William Wang, Kang G Shin, and Alanson P Sample. Privacylens: On-device pii removal from rgb images using thermally-enhanced sensing. *Proc. Priv. Enhancing Technol.*, 2024(4):872–891, 2024.

Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, pp. 1–14, 2020.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (DECOLLE). *Frontiers in Neuroscience*, 14:424, 2020. ISSN 1662-453X. doi: 10.3389/fnins.2020.00424. URL https://www.frontiersin.org/article/10.3389/fnins.2020.00424.

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

Bin Li, Xin Jiang, Yirui Du, Yanzuo Yu, and Ruonan Zhang. Wi-ssr: Wi-fi-based lightweight high-resolution model for human activity recognition. *IEEE Sensors Journal*, 25(4):6556–6571, 2025.

Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 286–293, 2021.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

Jian Liu, Hongbo Liu, Yingying Chen, Yan Wang, and Chen Wang. Wireless sensing for human activity: A survey. *IEEE Communications Surveys & Tutorials*, 22(3):1629–1645, 2019.

Jiao Liu, Guanlong Teng, and Feng Hong. Human activity sensing with wireless signals: A survey. *Sensors*, 20(4):1210, 2020.

Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *Ijcai*, pp. 1743–1749, 2021.

Yang Liu, Siao Liu, Xiaoguang Zhu, Hao Yang, Jielin Li, Juncen Guo, Liangyu Teng, Dingkang Yang, Yan Wang, and Jing Liu. Privacy-preserving video anomaly detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Xinxin Lu, Lei Wang, Chi Lin, Xin Fan, Bin Han, Xin Han, and Zhenquan Qin. Autodlar: A semi-supervised cross-modal contact-free human activity recognition system. *ACM Transactions on Sensor Networks*, 20(4):1–20, 2024.

Hyunggeun Mo and Seungku Kim. A deep learning-based human identification system with wi-fi csi data augmentation. *IEEE Access*, 9:91913–91920, 2021.

Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3), 2021.

Alberto Sabater, Luis Montesano, and Ana C. Murillo. Event transformer: A sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2677–2686, 2022.

Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14698–14708, 2023.

Jungpil Shin, Najmul Hassan, Abu Saleh Musa Miah, and Satoshi Nishimura. A comprehensive methodological survey of human activity recognition across diverse data modalities. *Sensors*, 25 (13):4028, 2025.

Sumit Bam Shrestha and Garrick Orchard. SLAYER: Spike layer error reassignment in time. In *Advances in Neural Information Processing Systems*, volume 31, pp. 1419–1428. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7415-slayer-spike-layer-error-reassignment-in-time.pdf.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

Zhigang Tu, Yuanzhong Liu, Yan Zhang, Qizi Mu, and Junsong Yuan. Dtcm: Joint optimization of dark enhancement and action recognition in videos. *IEEE Transactions on Image Processing*, 32: 3507–3520, 2023.

Fangxin Wang, Wei Gong, and Jiangchuan Liu. On spatial diversity in wifi-based human activity recognition: A deep learning-based approach. *IEEE Internet of Things Journal*, 6(2):2035–2047, 2018.

Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. Joint activity recognition and indoor localization with wifi fingerprints. *IEEE Access*, 7:80058–80068, 2019a. doi: 10.1109/ACCESS.2019.2923743.

Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-wifi: Fine-grained person perception using wifi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5452–5461, 2019b.

Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5615–5623, 2024.

Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, pp. 617–628, 2014.

Mengli Wei, Daguo Zhao, Lei Zhang, Cheng Wang, Yonggang Zhang, Qi Wang, Xiaochen Fan, Yaping Zhong, and Shiwen Mao. Wi-fitness: Improving wi-fi sensing with video perception for smart fitness. *IEEE Internet of Things Journal*, 12(3):2876–2889, 2024.

Bochen Xie, Yongjian Deng, Zhanpeng Shao, and Youfu Li. Eisnet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Transactions on Multimedia*, 26:8639–8650, 2024.

Haoran Xu, Peixi Peng, Guang Tan, Yuan Li, Xinhai Xu, and Yonghong Tian. Dmr: Decomposed multi-modality representations for frames and events fusion in visual reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26508–26518, 2024.

Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 969–978, 2024.

Jianfei Yang, Xinyan Chen, Han Zou, Dazhuo Wang, Qianwen Xu, and Lihua Xie. EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression. *IEEE Internet of Things Journal*, 9(15):13086–13095, 2022. doi: 10.1109/JIOT.2021.3139954.

Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Patterns*, 4 (3), 2023a.

Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18756–18768, 2023b.

Jieming Yang, Yanming Liu, Zhiying Liu, Yun Wu, Tianyang Li, and Yuehua Yang. A framework for human activity recognition based on WiFi CSI signal enhancement. *International Journal of Antennas and Propagation*, 2021(1):6654752, 2021. doi: 10.1155/2021/6654752.

Yanni Yang, Pengfei Hu, Jiaxing Shen, Haiming Cheng, Zhenlin An, and Xiulong Liu. Privacy-preserving human activity sensing: A survey. *High-Confidence Computing*, 4(1):100204, 2024.

Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. A survey on behavior recognition using WiFi channel state information. *IEEE Communications Magazine*, 55 (10):98–104, 2017. doi: 10.1109/MCOM.2017.1700082.

Dongheng Zhang, Liquan Fang, Xuecheng Xie, Shuai Yang, Jie Chen, and Yan Chen. Enhancing csi-based wireless sensing with open source linux 802.11 ax csi tool. *IEEE Sensors Journal*, 2025.

Hualei Zhang, Zhu Wang, Zhuo Sun, Wenchao Song, Zhihui Ren, Zhiwen Yu, and Bin Guo. Understanding the mechanism of through-wall wireless sensing: A model-based perspective. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–28, 2023a.

Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9781–9790, 2023b.

Zheng Zhang and Junxing Zhang. Actionfi: Human action recognition via multimodal feature fusion of restructured csi and optical flow. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 6824–6829. IEEE, 2025.

Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services*, pp. 313–325, 2019.

Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7808–7815. IEEE, 2023.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 989–997, 2019.

Yabin Zhu, Xiao Wang, Chenglong Li, Bo Jiang, Lin Zhu, Zhixiang Huang, Yonghong Tian, and Jin Tang. Crsot: Cross-resolution object tracking using unaligned frame and event cameras. *IEEE Transactions on Multimedia*, 2025.

Han Zou, Jianfei Yang, Hari Prasanna Das, Huihan Liu, Yuxun Zhou, and Costas J. Spanos. Wifi and vision multimodal learning for accurate and robust device-free human activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

A SUPPLEMENTARY CHARACTERIZATION OF EXWACT

A.1 EXAMPLES

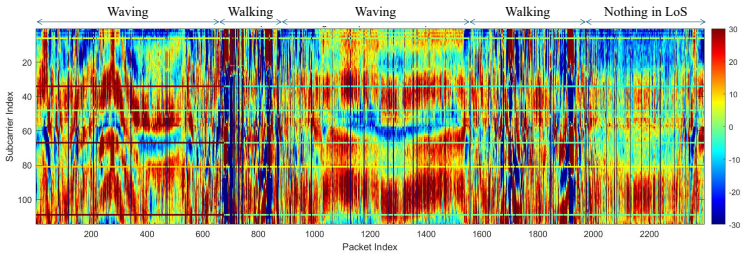

Figure 5: Raw WiFi CSI extracted using IAX tool (Zhang et al., 2025) when performing different actions.

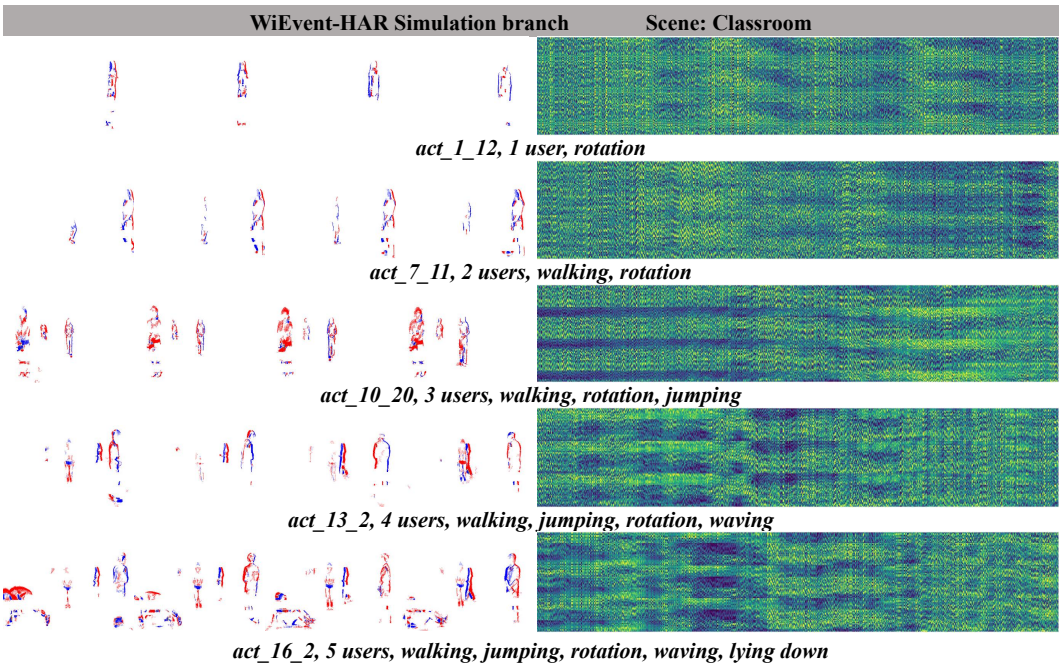

Figure 6: Paired Event data and WiFi CSI examples from ExWACT simulation branch. Event-frame visualization accumulated over a short temporal window.

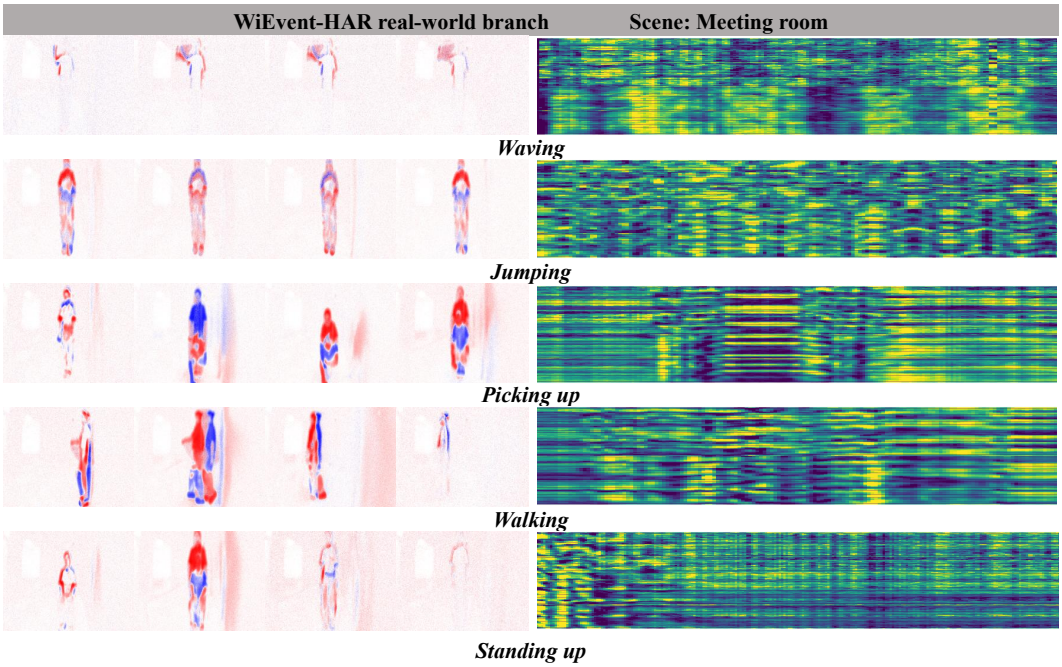

Figure 7: Paired Event data and WiFi CSI examples from ExWACT real-world branch. Event-frame visualization accumulated over a long temporal window.

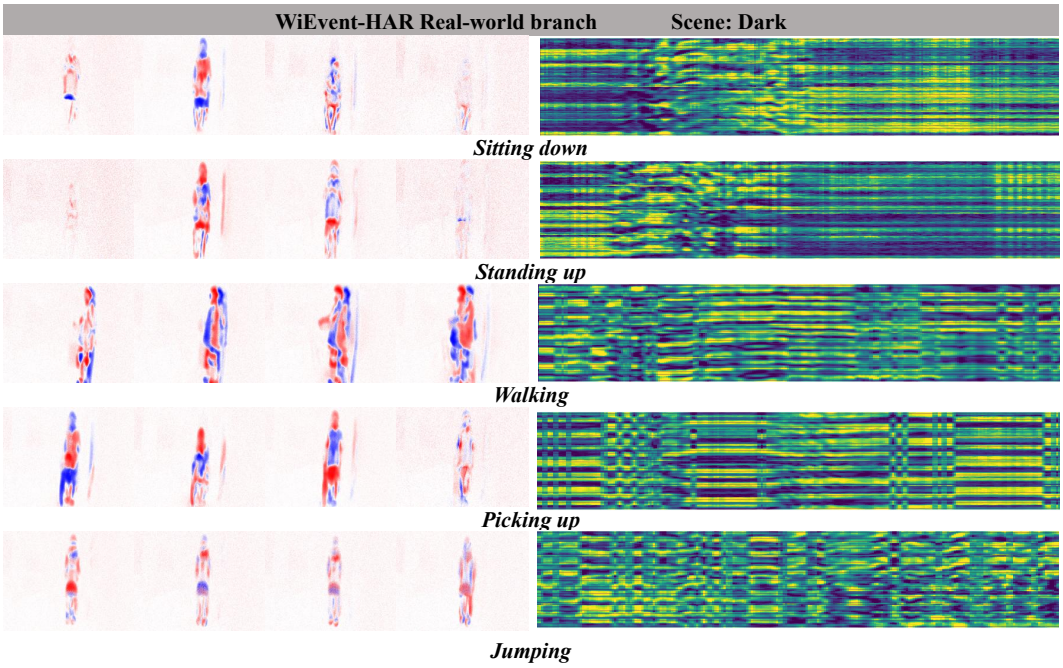

Figure 8: Paired Event data and WiFi CSI examples from ExWACT real-world branch in dark scene. Event data can record recognizable movements. Event-frame visualization accumulated over a long temporal window.

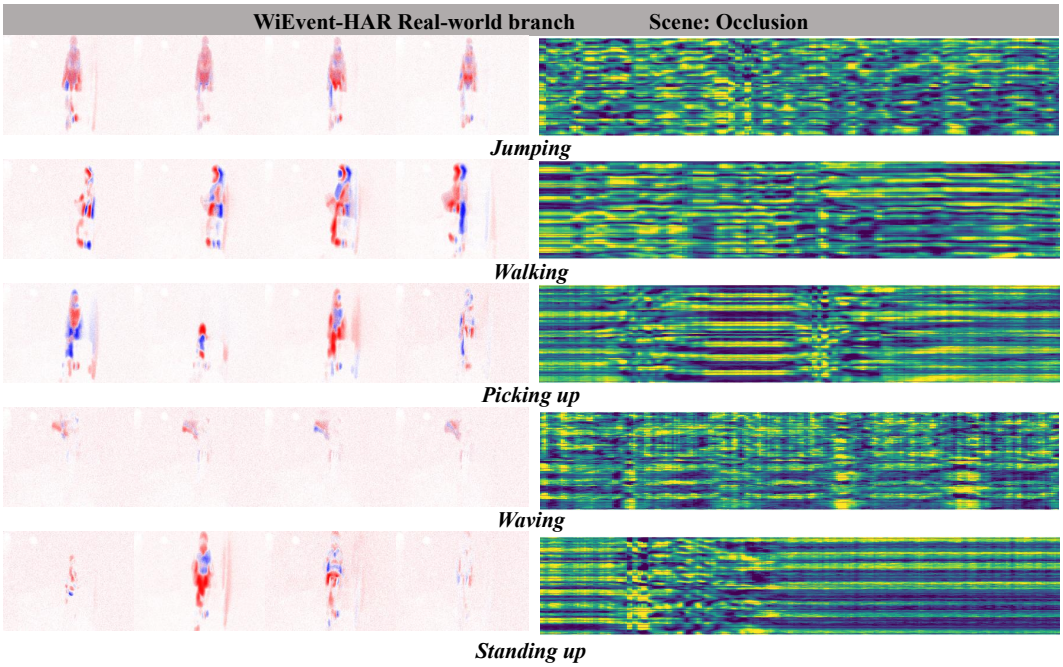

Figure 9: Paired Event data and WiFi CSI examples from ExWACT real-world branch in occlusion scene. WiFi CSI captures motion occurrences when obstacles are present between the transmitter and receiver. Event-frame visualization accumulated over a long temporal window.

A.2 Real-World Data Collection Settings

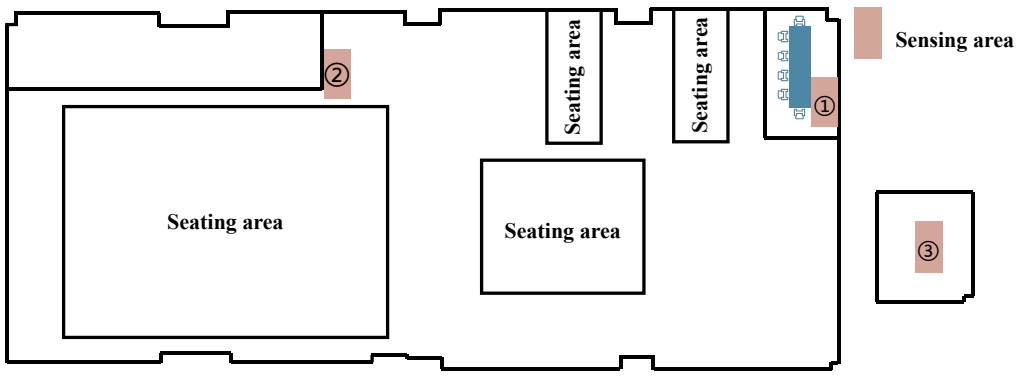

Figure 10: Environmental settings for ExWACT real-world data collection: Scene①, meeting room; Scene②, open laboratory area near the seating section; Scene③, empty room.

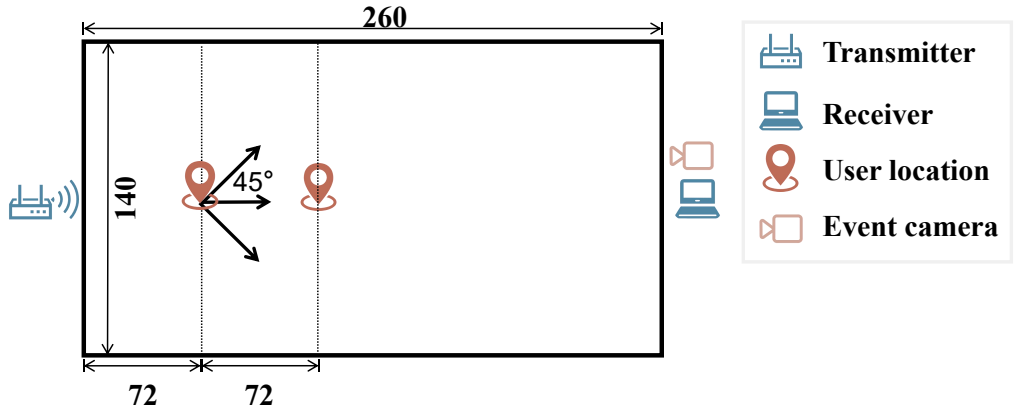

Figure 11: Sensing-area configuration for ExWACT real-world data collection. Participants perform actions at two locations and under three orientations.

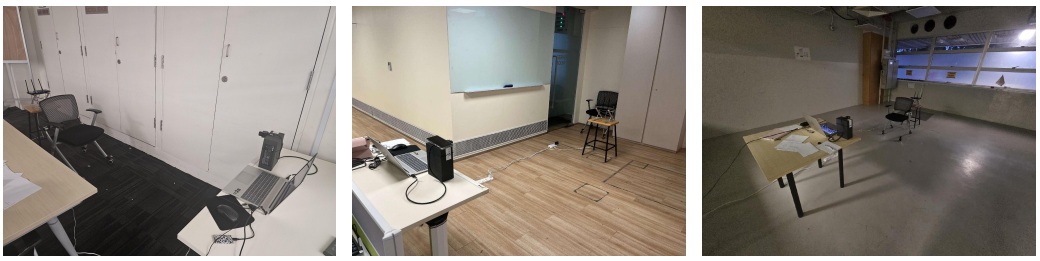

Figure 12: The three environments of ExWACT real-world data cover different levels of scene complexity: (a) Scene①, a meeting room with tables and chairs under medium illumination without main lighting; (b) Scene②, an empty area in a laboratory under normal illumination with random background human movement; and (c) Scene③, an empty room without background human interference but with varying background illumination.

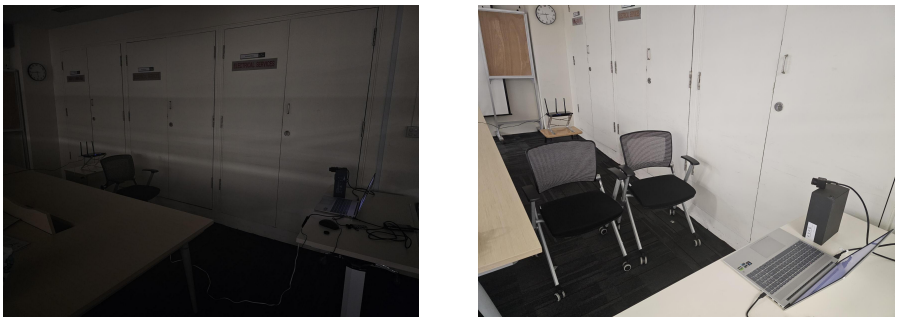

Figure 13: Special scene settings for evaluating model robustness under darkness and occlusion. The dark scene (a) contains no indoor light source, while the occlusion scene (b) includes obstacles that partially block the human body.

A.3 Dataset Validation

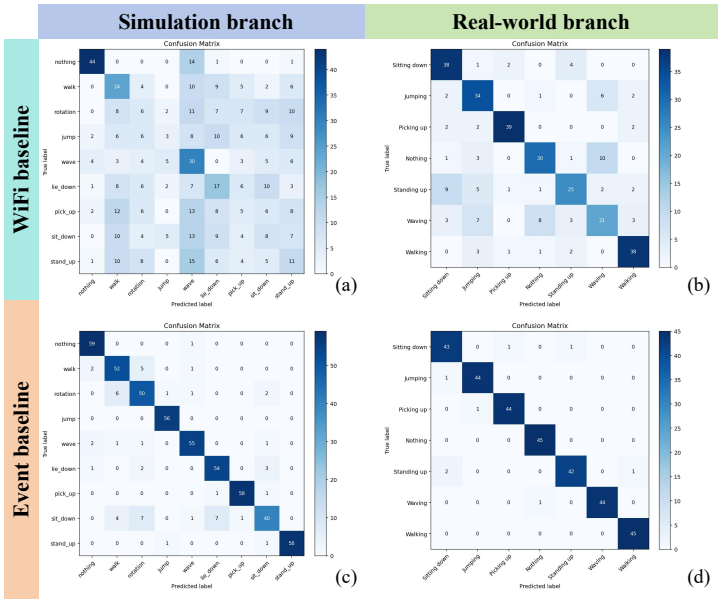

Figure 14: Validation of the proposed ExWACT dataset. Confusion matrices of the best-performing baselines on the simulation and real-world branches of ExWACT. Multiple WiFi- and event-based baseline methods are evaluated on each branch, and the representative results with the highest performance are visualized. The selected baseline models are (a) ABLSTM, (b) THAT, (c) EvT, (d) DECOLLE.

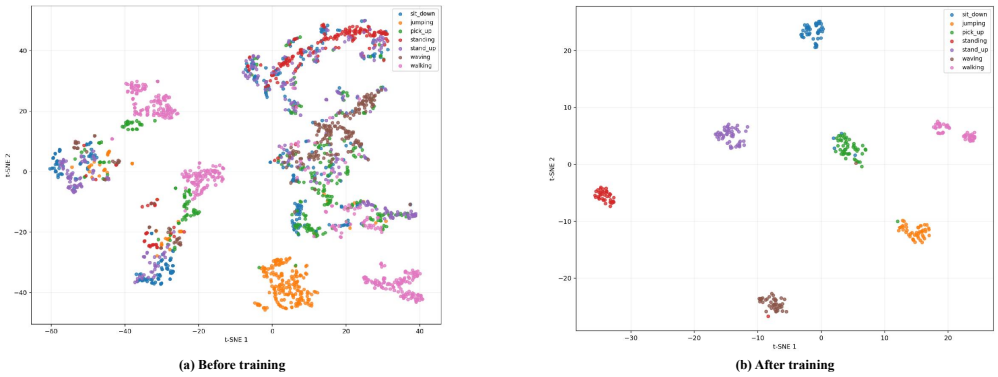

Figure 15: Validation of the proposed ExWACT dataset. t-SNE visualisation of feature distributions before and after training on ExWACT real-world branch.

B Methodology Details

B.1 The Proposed Approach

Overview. An overview of ExWACT is illustrated in Fig. 3. The key idea is to *learn temporally consistent and complementary representations from heterogeneous event and WiFi modalities through a unified cross-modal encoding and fusion framework*. At its core, the framework consists of two sequential components: (i) a **Sparse Dual-branch Modality Encoder** (Sec. 3.2.1), which performs modality-specific temporal modeling and intermediate representation alignment, and (ii)

a **Temporally Consistent Bi-directional Fusion (TCBF)** module (Sec. 3.2.2), which facilitates token-level interaction and enforces temporal consistency across modalities for robust HAR.

B.1.1 Sparse Dual-branch Modality Encoder

Insights. The design of the **Sparse Dual-branch Modality Encoder** is motivated by the intrinsic limitations of existing approaches that either rely on RGB-based representation or apply generic fusion strategies without modeling event-WiFi heterogeneity. In practice, WiFi CSI lacks spatial resolution and is easily dominated by static reflections, while event streams are sparse, asynchronous, and difficult to align with conventional frame-based processing. These mismatches make simple feature extraction and fusion suboptimal. The proposed module is designed to explicitly capture the complementary motion characteristics of WiFi and event modalities under heterogeneous and asynchronous sensing. The WiFi branch focuses on *global motion dynamics* by transforming CSI into time-frequency representations with Doppler-aware attention to suppress static interference, while the event branch encodes *fine-grained spatio-temporal structure* through polarity-aware, spike-driven processing that preserves motion discontinuities and directionality.

WiFi Branch: Time-frequency Motion Encoding. Given raw CSI measurements $\mathcal{I}_{\text{wifi}} \in \mathbb{C}^{T\times S}$, where T denotes the number of temporal samples and S the number of subcarriers, we first decompose the signal into amplitude and phase components: $\mathbf{A} = |\mathcal{I}_{\text{wifi}}|$, $\mathbf{\Phi} = \angle\mathcal{I}_{\text{wifi}}$. To capture motion-sensitive variations, we construct a **time-frequency response tensor**:

$$\mathbf{X}_{\text{wifi}} = \text{Concat}\left(\Delta_t\mathbf{A},\ \Delta_t\mathbf{\Phi}\right) \in \mathbb{R}^{T\times S\times 2}, \tag{7}$$

where Δ_t denotes temporal differencing to emphasize dynamic changes while suppressing static components. The tensor is then processed by a lightweight temporal convolutional stem $\mathbf{F}_0 = \text{Conv1D}(\mathbf{X}_{\text{wifi}})$, which captures local temporal dependencies.

To further enhance motion discrimination, we introduce a **dynamic Doppler-aware attention module**. Specifically, we compute a motion saliency map:

$$\mathbf{M} = \text{Softmax}\left(\text{Conv1D}(|\Delta_t\mathbf{F}_0|)\right), \tag{8}$$

which highlights high-frequency temporal variations corresponding to human motion. The refined WiFi features are obtained via $\mathbf{F}_{\text{wifi}} = \mathbf{M}\odot\mathbf{F}_0$, where $\odot$ denotes element-wise multiplication. This mechanism adaptively emphasizes dynamic components while suppressing static background reflections. Finally, the features are projected into token embeddings $\mathbf{W} = \text{Proj}(\mathbf{F}_{\text{wifi}}) \in \mathbb{R}^{L_w\times D}$.

Event Branch: Spike-driven Spatio-temporal Encoding. Given an event stream $\mathcal{I}_{\text{event}} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, where (x_i, y_i) denotes spatial location, t_i timestamp, and $p_i \in \{-1, +1\}$ polarity, we first perform **temporal voxelization** by partitioning events into B temporal bins $\mathbf{V} \in \mathbb{R}^{B\times H\times W\times 2}$, where the last dimension corresponds to positive and negative polarities accumulated within each bin.

To preserve sparsity and temporal causality, we adopt a **polarity-aware spiking convolutional encoder**. The voxel grid is processed using convolutional layers with spike-like activation:

$$\mathbf{F}_{l+1} = \sigma(\text{Conv3D}(\mathbf{F}_l) - \theta), \tag{9}$$

where $\sigma(\cdot)$ is a thresholding function (e.g., surrogate Heaviside), and θ is a learnable firing threshold. This formulation enables sparse activation while maintaining sensitivity to motion edges and temporal discontinuities. To explicitly encode motion directionality, we incorporate polarity gating by subtracting negative-polarity responses from positive-polarity responses. Finally, the spatio-temporal features are flattened and projected into event tokens

$$\mathbf{F}_{\text{event}} = \mathbf{F}^{(+)} - \mathbf{F}^{(-)}, \quad \mathbf{E} = \text{Proj}(\mathbf{F}_{\text{event}}) \in \mathbb{R}^{L_e\times D}, \tag{10}$$

where $\mathbf{F}^{(+)}$ and $\mathbf{F}^{(-)}$ correspond to features extracted from positive and negative polarity channels.

Our proposed dual-branch encoder explicitly models *motion dynamics* from two complementary perspectives: the WiFi branch captures global motion-induced signal perturbations through time-frequency analysis, while the event branch encodes fine-grained local motion with sparse spatio-temporal precision. By preserving modality-specific sparsity and temporal structure, the encoder produces discriminative representations that are well-suited for subsequent cross-modal fusion.

B.1.2 Temporally Consistent Bi-directional Fusion (TCBF)

Insights. The **TCBF** module is motivated by the limitation of conventional fusion methods that ignore temporal misalignment between continuous WiFi signals and asynchronous event streams. Direct fusion or standard cross-attention leads to inconsistent feature correspondence. To address this, we jointly model *cross-modal interaction* via bidirectional attention and *temporal consistency* via explicit alignment, enabling robust and coherent fusion under heterogeneous timing.

Cross-modal Token Representation. Given an input sample within an action-level temporal window Δt, the WiFi CSI and event streams are first encoded into modality-specific token sequences: $\mathbf{E} \in \mathbb{R}^{L_e \times D}, \mathbf{W} \in \mathbb{R}^{L_w \times D}$, where L_e and L_w denote the number of event and WiFi tokens, respectively, and D is the embedding dimension. Due to the asynchronous nature of event streams and continuous sampling of WiFi signals, $L_e \neq L_w$ in general. To preserve temporal structure, both modalities are equipped with positional encodings within Δt, enabling implicit temporal correspondence without enforcing strict index-wise alignment.

Bidirectional Cross-modal Attention. We adopt a bidirectional cross-attention mechanism to model inter-modal dependencies. Specifically, event tokens attend to WiFi tokens and symmetrically WiFi tokens attend to event tokens where $\mathbf{Q}_e = \mathbf{E}\mathbf{W}_e^Q$, $\mathbf{K}_w = \mathbf{W}\mathbf{W}_w^K$, and $\mathbf{V}_w = \mathbf{W}\mathbf{W}_w^V$:

$$\mathrm{Attn}_{e \leftarrow w} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_e \mathbf{K}_w^\top}{\sqrt{d}}\right)\mathbf{V}_w, \quad \mathrm{Attn}_{w \leftarrow e} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_w \mathbf{K}_e^\top}{\sqrt{d}}\right)\mathbf{V}_e. \tag{11}$$

This bidirectional design allows each modality to incorporate complementary cues from the other, enhancing representation richness.

Temporal Consistency Alignment. While cross-attention enables feature interaction, it does not explicitly enforce temporal coherence across modalities. To address this, we introduce a **temporal consistency constraint** that aligns the temporal dynamics of the two modalities at the representation level. We first aggregate token sequences into temporal summaries:

$$\bar{\mathbf{e}}_t = \frac{1}{|\mathcal{S}_e(t)|}\sum_{i \in \mathcal{S}_e(t)} \mathbf{E}_i, \quad \bar{\mathbf{w}}_t = \frac{1}{|\mathcal{S}_w(t)|}\sum_{j \in \mathcal{S}_w(t)} \mathbf{W}_j, \tag{12}$$

where $\mathcal{S}_e(t)$ and $\mathcal{S}_w(t)$ denote sets of tokens corresponding to temporal segment t.

We then enforce consistency between modalities using a smooth alignment loss:

$$\mathcal{L}_{\mathrm{TC}} = \sum_{t=1}^{T} \|\phi(\bar{\mathbf{e}}_t) - \psi(\bar{\mathbf{w}}_t)\|_2^2, \tag{13}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are learnable projection functions that map both modalities into a shared latent space. This objective encourages temporally corresponding representations to be consistent while preserving modality-specific characteristics.

Fusion and Optimization. After cross-modal interaction, the refined token sequences $\mathbf{E}'$ and $\mathbf{W}'$ are concatenated and pooled into a global representation which is fed into a classifier for activity prediction $\mathbf{z} = f_{\mathrm{pool}}([\mathbf{E}'; \mathbf{W}']), \hat{\mathbf{y}} = f_{\mathrm{cls}}(\mathbf{z})$. The overall training objective combines classification and temporal consistency $\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{TC}}$, where λ balances discriminative learning and cross-modal temporal alignment.

Compared to conventional fusion strategies, the proposed TCBF module explicitly models both *cross-modal interaction* and *temporal consistency*, enabling robust learning under asynchronous sensing conditions. This design is particularly effective for event-WiFi fusion, where temporal misalignment and modality heterogeneity are fundamental challenges.

C Implementation Details

C.1 Baseline Implementation

We compare our approach against representative WiFi-based and event-based HAR methods. For WiFi sensing, we adopt ARIL (Wang et al., 2019a), ABLSTM (Chen et al., 2018), and THAT (Li

et al., 2021) as baselines. For event-based recognition, we include EvT (Sabater et al., 2022), DECOLLE (Kaiser et al., 2020), and SLAYER (Shrestha & Orchard, 2018). These models cover both conventional sequence modeling and neuromorphic-inspired approaches, providing a diverse comparison across modalities. All baseline models are evaluated using the same data preprocessing pipeline, train/validation/test splits, and evaluation protocol within each branch.

Simulation Branch Implementation. For the simulation branch, all WiFi-only and event-only baselines are evaluated under a single-label activity classification framework. Only single-user samples are selected in this branch and converted into single-label activity labels for a consistent single-user setting. The model architectures are reproduced following the original works. Same train/validation/test splits are used on the event- and WiFi- based models, and the splits are constructed to maintain balanced activity and scene distributions across the selected samples.

For WiFi-only models, CSI samples are reshaped into temporal feature sequences, normalized along each feature dimension, and padded or truncated to a fixed sequence length determined by the dataset statistics, with an upper cap of 3000 frames. All WiFi-only baselines are trained for 200 epochs with a batch size of 128 using the Adam optimizer. The learning rate is set to 1×10^{-3} for all WiFi baselines.

All event-only baselines are trained end-to-end with a batch size of 4. DECOLLE is optimized using Adamax with a learning rate of 3×10^{-4}, SLAYER is optimized using Adam with a learning rate of 5×10^{-4}, and EvT is optimized using AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. For EvT, gradient clipping with a maximum norm of 1.0 is applied during training to improve stability. The best checkpoints are selected based on validation accuracy, with validation loss used as a tie-breaker.

Real-World Branch Implementation. For the real-world branch, we evaluate the same WiFi-only and event-only baseline models as those used in the simulation branch. All real-world baselines are trained and evaluated using the same train/validation/test split, with balanced activity and scene distributions maintained where possible.

For WiFi-only models, each CSI tensor is reshaped into a temporal feature sequence, normalized along each feature dimension, and padded or truncated to a fixed length of 437 frames. The WiFi-only baselines are trained for 50 epochs with a batch size of 16 using the Adam optimizer, with the learning rate set to 1×10^{-3}.

For event-only models, training is conducted end-to-end using event-stream inputs with a batch size of 4. DECOLLE is optimized using Adamax with a learning rate of 3×10^{-4}, SLAYER is optimized using Adam with a learning rate of 5×10^{-4}, and EvT is optimized using AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. For EvT, gradient clipping with a maximum norm of 1.0 is also applied to ensure training stability. All real-world event-only baselines are trained for 50 epochs. The same checkpoint selection criterion as in the simulation branch is applied, where the best checkpoint is selected based on validation accuracy, with validation loss used as a tie-breaker.

C.2 ExWACT Implementation Details

We implement **ExWACT** in PyTorch and train the model end-to-end using the AdamW optimizer with cosine learning-rate scheduling and a linear warmup over the first 5% of training. The overall objective combines cross-entropy classification loss with the proposed temporal consistency alignment loss (Sec. 3.2.2), weighted by a balancing factor λ.

For the WiFi branch, raw CSI signals are preprocessed into amplitude and phase components, followed by temporal differencing to construct frequency-temporal representations. The resulting sequences are normalized per subcarrier and segmented into fixed-length windows. For the event branch, asynchronous event streams are voxelized into B temporal bins (default B=5) to form sparse spatio-temporal volumes, with separate channels for positive and negative polarities. The WiFi encoder consists of a lightweight temporal convolutional stem (kernel size 3) followed by a Doppler-aware attention module, while the event encoder adopts a polarity-aware convolutional architecture with sparse activation. Both branches project features into a shared embedding space of dimension D=256. The cross-modal fusion module employs multi-head attention with 8 heads and applies bidirectional interaction between modalities.

Unless otherwise specified, we use a base learning rate of 3×10^{-4}, weight decay 0.05, and batch size 64. Gradient clipping is applied with a maximum norm of 1.0 to stabilize training. We further adopt label smoothing with a factor of 0.1 to improve generalization. To balance optimization, the temporal consistency weight λ is gradually increased during the first few epochs (warm-in strategy), preventing early-stage domination of alignment loss. All experiments are conducted on a server equipped with 4 NVIDIA GPUs, and the best checkpoint is selected based on validation accuracy. Data preprocessing and augmentation follow standard practices, including random temporal cropping for WiFi sequences and spatial jittering for event volumes.

D Cross-Scene & Cross-domain Experiments

Table 5: Preliminary cross-scene validation on the real-world branch of WiEvent-HAR. The cross-scene results validate the discriminative quality of the real-world data under unseen-scene settings.

Train + Validation	**Test**	**Accuracy (%)**
Scene 1 + Scene 2	Scene 3	89.70
Scene 1 + Scene 3	Scene 2	75.43
Scene 2 + Scene 3	Scene 1	76.00

Table 6: Preliminary cross-domain evaluation of WiEvent-HAR across simulation and real-world environments. The results demonstrate the generalization capability under domain shift between simulated and real-world data.

Train + Validation	**Test**	**Accuracy (%)**
Simulation	Real-World	85.36
Real-World	Simulation	87.12
Overall	Simulation	91.02
	Real-world	99.18
	Overall	93.74