RIM-NET: LEARNING FROM ROTATION-INVARIANT MOTION STATES FOR WEARABLE HUMAN ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Even when the human activity remains unchanged, uncontrolled variations in sensor orientation can substantially alter the inertial measurements along the sensor's three local axes, posing a major challenge to robust wearable human activity recognition. We explicitly transform these axis-dependent readings into an intrinsic motion state jointly characterized by the magnitudes of dynamic acceleration and angular velocity and their geometric alignment. This motion state preserves motion intensity and the geometric relationship between the two inertial modalities while remaining unchanged under rotations of the sensor coordinate frame, thereby mapping the same physical motion observed under different sensor orientations to a consistent and stable input representation. Building on this motion state, we propose RIM-Net with two core components: Adaptive Quadrature Harmonic Encoding (AQHE) and Cross-Part Phase-Relation Learning (CPRL). To address the limited ability of standard convolutions to explicitly capture motion rhythms at different temporal frequencies, AQHE analyzes the motion-state sequence across multiple frequency bands and adaptively combines the resulting quadrature responses to extract discriminative local motion patterns. Moreover, local frequency and amplitude features alone do not describe whether movements at different body locations are synchronized or exhibit frequency-specific lead–lag relationships. CPRL explicitly models these cross-part temporal relations and uses them to guide feature aggregation. Experiments on five public human activity recognition benchmarks under subject-independent protocols demonstrate that RIM-Net consistently achieves strong recognition performance while maintaining a compact parameter footprint. These results support our central hypothesis that rotation-invariant motion states constructed from inertial measurements provide a reliable foundation for robust wearable activity recognition.

1 INTRODUCTION

Wearable inertial sensors enable unobtrusive human activity recognition in health monitoring and ubiquitous computing (Chen et al., 2021; Haresamudram et al., 2025). Reliable recognition, however, remains challenging when devices are attached at different orientations: a rotation of the sensor coordinate frame changes the axis-wise measurements even when the underlying motion is unchanged, and each device may have a different coordinate frame in multi-location settings. Although deep HAR has advanced through convolutional, attention-based, and cross-sensor models (Yang et al., 2015; Abedin et al., 2021; Miao et al., 2022; Xie et al., 2025), most methods still learn from axis-aligned signals or generic temporal features, making their representations sensitive to sensor orientation and leaving the rhythmic timing between body locations implicit. Orientation-independent features, rotation augmentation, and $\mathrm{SO}(3)$-equivariant networks improve robustness to sensor-frame variations (Wu & Jafari, 2019; Yu et al., 2023; Baek et al., 2026). Separately, frequency-aware time-series models exploit periodic structures to characterize repeated motion (Wu et al., 2023; Mu et al., 2025; Mohapatra et al., 2025). However, orientation-robust methods do not explicitly adapt local frequency analysis to the current motion state, while periodic models generally do not describe frequency-specific relative timing across multiple sensor locations. This gap is important for activities with similar local motion magnitudes but different patterns of body-part coordination. We illustrate the comparison between RIM-NET with typical methods in Figure 1.

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.

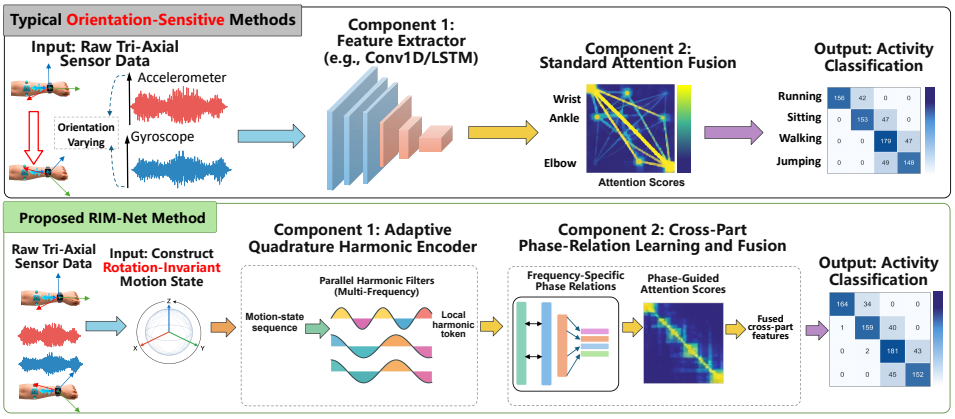

Figure 1: **Comparison between RIM-Net and typical methods.** The top row illustrates typical methods based on orientation-sensitive model's input, fixed filter, and standard attention-based cross-part fusion. The bottom row shows RIM-Net, which constructs a rotation-invariant motion state, adaptively modulates quadrature harmonic filters, and uses phase relations to guide attention.

Our key observation is that ***both local motion rhythms and cross-part timing can be modeled without retaining an arbitrary sensor orientation***. The norms of the acceleration and angular-velocity vectors and their inner product are preserved under a common three-dimensional rotation. Meanwhile, the relative phase indicates whether motion magnitudes at different body locations change simultaneously or with a time lag, rather than describing motion along a particular sensor axis. Based on this observation, we propose RIM-NET, which first constructs a **rotation-invariant motion state** from the magnitudes of dynamic acceleration and angular velocity and their geometric alignment. An Adaptive Quadrature Harmonic Encoding (AQHE) block uses this state to adaptively combine multi-frequency responses and extract local motion rhythms. Cross-Part Phase-Relation Learning (CPRL) then determines, at each frequency, whether motion magnitudes at different body locations change simultaneously or with a time lag, and uses this timing information to guide cross-part feature aggregation. **Our contributions are threefold. (1)** We construct a compact rotation-invariant motion state to modulate multi-frequency quadrature responses. Under a fixed sensor-frame rotation over the filter support, the motion state remains unchanged, while the end-to-end model's rotation robustness is evaluated empirically. **(2)** We introduce CPRL to model frequency-specific phase relations between motion-magnitude signals and use them to guide cross-part attention. **(3)** Across five public HAR benchmarks, RIM-NET achieves an average weighted F1-score of 93.96%, ranks first among the compared methods on four datasets, and requires only 58K–104K trainable parameters.

2 RELATED WORK

2.1 DEEP AND EFFICIENT WEARABLE ACTIVITY RECOGNITION

DeepConvLSTM combines convolutional feature extraction with recurrent temporal modeling (Ordóñez & Roggen, 2016), while DynamicWHAR uses a graph-based architecture to learn input-dependent correlations among wearable sensors (Miao et al., 2022). TinyHAR integrates lightweight convolution, recurrent modeling, and attention (Zhou et al., 2022), whereas Tinier-HAR further reduces the model size using depthwise separable convolutions and gated recurrent units (Bian et al., 2025). Although these methods provide effective temporal, cross-sensor, or lightweight modeling, they mainly operate on axis-dependent features and do not explicitly describe frequency-specific timing relations between body locations.

TRI-HAR addresses independent orientation shifts in multi-IMU systems using a shared $\mathrm{SO}(3)$-equivariant backbone and an invariant projection at each sensor location (Baek et al., 2026). This design provides structural invariance to fixed, location-specific rotations before the resulting features are fused for classification. However, TRI-HAR primarily focuses on eliminating orientation dependence and does not explicitly model local motion rhythms or frequency-specific lead–lag relations across body locations. In contrast, RIM-NET constructs a compact rotation-invariant motion state

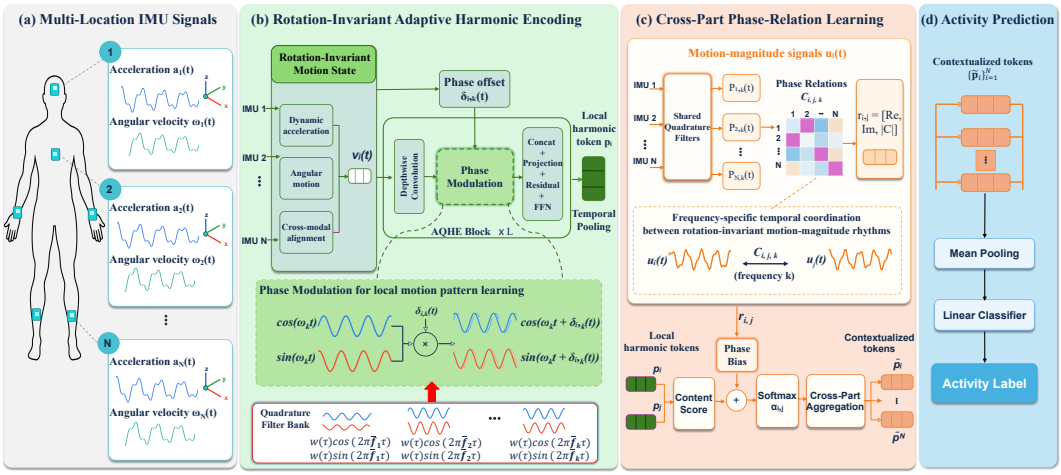

Figure 2: **Overview of RIM-NET.** Rotation-Invariant Adaptive Harmonic Encoding captures local motion rhythms, while Cross-Part Phase-Relation Learning models frequency-specific coordination across body locations.

to adapt multi-frequency harmonic responses and uses cross-part phase relations to guide feature aggregation.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

For wearable human activity recognition, inertial sensors are always placed at multiple body locations. An input window is denoted by

$$\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^{N}, \qquad \mathbf{X}_i = [\mathbf{a}_i; \boldsymbol{\omega}_i] \in \mathbb{R}^{6\times T}, \tag{1}$$

where $\mathbf{a}_i(t) \in \mathbb{R}^3$ and $\boldsymbol{\omega}_i(t) \in \mathbb{R}^3$ denote the tri-axial acceleration and angular velocity measured at body location i, respectively. Given $\mathcal{X}$, the objective is to predict an activity label $y \in \{1, \ldots, C\}$. However, directly using $\mathcal{X}$ as input to a deep learning model can lead to predictions that are sensitive to changes in sensor orientation. We aim to learn representations that are robust to sensor-coordinate variations while preserving local motion rhythms and cross-part temporal coordination. As shown in Figure 2, our network consists of two components. First, the ***Rotation-Invariant Adaptive Harmonic Encoding*** constructs a **rotation-invariant motion state** for each sensor and uses these states to adaptively encode local harmonic responses. Second, the ***Cross-Part Phase-Relation Learning*** module models frequency-specific relative timing between rotation-invariant motion-magnitude signals at different body locations. Rotation-Invariant Adaptive Harmonic Encoding describes how each body part moves, whereas Cross-Part Phase-Relation Learning captures how different body parts coordinate.

3.2 ROTATION-INVARIANT ADAPTIVE HARMONIC ENCODING

Rotation-invariant Motion State. We seek a local motion state that remains unchanged under these inevitable sensor rotations. To this end, we construct a three-component state from two motion magnitudes and one geometric alignment term shown in Eq. (2). All three components depend only on vector norms or inner products, which are preserved under three-dimensional rotations. For each body location and time step, we then define the motion state as

$$\mathbf{v}_i(t) = \Big[\log\big(|\mathbf{a}_i^d(t)|_2 + \epsilon\big),\ \log\big(\gamma|\boldsymbol{\omega}_i(t)|_2 + \epsilon\big),\ \widetilde{h}_i(t)\Big]^\top, \tag{2}$$

where $\epsilon > 0$ provides numerical stability and $\gamma > 0$ balances the numerical scales of the two modalities. The first two components quantify the magnitudes of dynamic acceleration and angular motion,

respectively. The third component measures their instantaneous geometric alignment. Because the raw acceleration contains a slowly varying component, we first extract a dynamic acceleration residual:

$$\mathbf{g}_i(t) = \mathcal{L}(\mathbf{a}_i)(t), \qquad \mathbf{a}_i^d(t) = \mathbf{a}_i(t) - \mathbf{g}_i(t), \tag{3}$$

where $\mathcal{L}$ is a fixed linear low-pass filter applied identically to all three axes. We implement this filter using a normalized symmetric Hann kernel. The resulting $\mathbf{a}_i^d(t)$ is a high-frequency residual relative to the estimated low-frequency component. It should not be interpreted as an exact estimate of inertial-frame linear acceleration. Next, we compute the geometric alignment term as $\widetilde{h}_i(t) = c_i(t)h_i(t)$. Thus, $h_i(t)$ is a numerically stabilized cosine similarity between the dynamic-acceleration and angular-velocity vectors. Although $h_i(t)$ remains numerically well-defined at small magnitudes, its directional information can then be dominated by sensor noise. We therefore attenuate the geometric term using the deterministic reliability factor

$$h_i(t) = \frac{\mathbf{a}_i^d(t)^\top \boldsymbol{\omega}_i(t)}{\sqrt{\left(|\mathbf{a}_i^d(t)|_2^2 + \epsilon_a\right)\left(\|\boldsymbol{\omega}_i(t)\|_2^2 + \epsilon_\omega\right)}}, \qquad c_i(t) = \frac{|\mathbf{a}_i^d(t)|_2}{|\mathbf{a}_i^d(t)|_2 + \tau_a}\frac{|\boldsymbol{\omega}_i(t)|_2}{|\boldsymbol{\omega}_i(t)|2 + \tau_\omega}, \tag{4}$$

where $\epsilon_a > 0$ and $\epsilon_\omega > 0$ are stabilization constants, τ_a and τ_ω characterize the noise scales of the accelerometer and gyroscope. This multiplicative weighting suppresses the alignment term whenever either motion magnitude approaches its corresponding noise floor.

Adaptive Quadrature Harmonic Encoder. We propose an Adaptive Quadrature Harmonic Encoding (AQHE) block that uses the rotation-invariant motion state to adaptively modulate a bank of multi-frequency quadrature filters, producing orientation-robust local motion representations that capture discriminative rhythmic patterns. We project motion states into an H-dimensional feature sequence:

$$\mathbf{F}_i^{(0)} = \mathbf{W}_{\text{in}} *_{1\times1} \mathbf{v}_i \in \mathbb{R}^{H\times T}, \tag{5}$$

where the projection and subsequent AQHE parameters are shared across body locations. To capture multi-scale motion rhythms, we construct K frequency-specific quadrature filter pairs:

$$\kappa_k^{\cos}(\tau) = w(\tau)\cos(2\pi\bar{f}_k\tau), \qquad \kappa_k^{\sin}(\tau) = w(\tau)\sin(2\pi\bar{f}_k\tau). \tag{6}$$

where $w(\tau)$ is a fixed Hann window (Harris, 1978), f_k is the physical center frequency of band k, and $\bar{f}_k = f_k/f_s$ is its normalized frequency in cycles per sample at sampling rate f_s. Each filter is normalized to unit ℓ_2 norm, yielding $\overline{\kappa}_k^{\cos}$ and $\overline{\kappa}_k^{\sin}$.

Adaptive Quadrature Harmonic Encoding Block. For the ℓ-th AQHE block, depthwise convolution of the pre-normalized input $\mathbf{Z}_i^{(\ell)} = \mathrm{LN}(\mathbf{F}_i^{(\ell)})$ produces the paired responses

$$\mathbf{Y}_{i,k}^{\cos,(\ell)} = \mathbf{Z}_i^{(\ell)} * \overline{\kappa}_k^{\cos}, \qquad \mathbf{Y}_{i,k}^{\sin,(\ell)} = \mathbf{Z}_i^{(\ell)} * \overline{\kappa}_k^{\sin}. \tag{7}$$

Together, the paired responses capture frequency-specific temporal structure and local phase information. To adaptively combine the cosine and sine responses in each frequency band according to the current motion state, we use a lightweight conditioning network to generate bounded, band-specific phase offsets $\boldsymbol{\delta}_i^{(\ell)}(t)$ from the temporally aligned motion-state sequence $\mathbf{v}_i^{(\ell)}$:

$$\boldsymbol{\delta}_i^{(\ell)}(t) = \delta_{\max}\tanh\left(g_\theta^{(\ell)}\left(\mathbf{v}_i^{(\ell)}\right)(t)\right) \in \mathbb{R}^K, \tag{8}$$

where $g_\theta^{(\ell)}$ is implemented using an MLP followed by a short temporal convolution. Then, the offset dynamically combines the quadrature responses:

$$\widehat{\mathbf{Y}}_{i,k}^{(\ell)}(t) = \mathbf{Y}_{i,k}^{\cos,(\ell)}(t)\cos\delta_{i,k}^{(\ell)}(t) - \mathbf{Y}_{i,k}^{\sin,(\ell)}(t)\sin\delta_{i,k}^{(\ell)}(t). \tag{9}$$

The modulated responses from all bands are concatenated, projected back to H channels, and incorporated through a residual AQHE block:

$$\begin{aligned}\mathbf{F}_i^{(\ell+\frac{1}{2})} &= \mathbf{F}_i^{(\ell)} + \mathbf{W}_o^{(\ell)} *_{1\times1} \mathrm{Concat}_{k=1}^{K}\,\widehat{\mathbf{Y}}_{i,k}^{(\ell)},\\ \mathbf{F}_i^{(\ell+1)} &= \mathbf{F}_i^{(\ell+\frac{1}{2})} + \mathrm{FFN}^{(\ell)}\left(\mathrm{LN}\left(\mathbf{F}_i^{(\ell+\frac{1}{2})}\right)\right).\end{aligned} \tag{10}$$

After stacking L AQHE blocks, temporal average pooling produces the local harmonic token of body location i:

$$\mathbf{p}_i = \frac{1}{T_L} \sum_{t=1}^{T_L} \mathbf{F}_i^{(L)}(t) \in \mathbb{R}^H. \tag{11}$$

The resulting token $\mathbf{p}_i$ encodes the state-adaptive, multi-frequency motion pattern of the i-th body location.

3.3 Cross-Part Phase-Relation Learning

Although content-based attention (Essa & Abdelmaksoud, 2023) enables feature exchange between body locations, its scores do not explicitly capture their frequency-specific relative timing, such as whether one motion-magnitude rhythm consistently leads or lags another. CPRL addresses this limitation in two steps. It first estimates phase relations between the rotation-invariant motion-magnitude signals and then incorporates these relations as biases into cross-part attention to update the local harmonic tokens.

Rotation-Invariant Motion-Magnitude Phase Relations. For each body location, we combine the dynamic-acceleration and angular-velocity magnitudes into a scalar motion-magnitude signal:

$$u_i(t) = \sqrt{\|\mathbf{a}_i^d(t)\|_2^2 + \gamma^2 \|\boldsymbol{\omega}_i(t)\|_2^2}. \tag{12}$$

Here, γ balances the numerical scales of the two sensing modalities. Because $u_i(t)$ depends only on vector norms, it is invariant to rotations of the local sensor coordinate frame. Importantly, $u_i(t)$ represents a scalar motion-magnitude trajectory rather than the directional motion of the original tri-axial signals. Consequently, the phase relations defined below characterize the temporal alignment of motion-magnitude rhythms, not the directional phase relation between raw motion vectors.

We apply the normalized quadrature filters from Eq. (6) to this scalar signal. For frequency band k, the cosine and sine responses form the complex band-limited signal

$$P_{i,k}(t) = \big(u_i * \overline{\kappa}_k^{\cos}\big)(t) + \mathrm{j}\big(u_i * \overline{\kappa}_k^{\sin}\big)(t), \tag{13}$$

where $\mathrm{j} = \sqrt{-1}$. The real and imaginary components are the cosine- and sine-filter responses, respectively. Reusing the AQHE filter bank aligns the local encoding and cross-part relation analysis at the same temporal frequencies.

For body locations i and j, we summarize their phase relation in band k as

$$C_{ij,k} = \frac{1}{T'} \sum_{t=1}^{T'} \frac{P_{i,k}(t) P_{j,k}^*(t)}{|P_{i,k}(t)||P_{j,k}(t)| + \epsilon_{\mathrm{ph}}}. \tag{14}$$

Here, $(\cdot)^*$ denotes complex conjugation, T' is the temporal length of the filtered response, and $\epsilon_{\mathrm{ph}} > 0$ stabilizes low-energy samples. If $P_{i,k}(t) = A_i(t) e^{\mathrm{j}\varphi_i(t)}$, the normalized cross-product approximately represents $e^{\mathrm{j}(\varphi_i(t) - \varphi_j(t))}$. Therefore, $C_{ij,k}$ averages the instantaneous phase differences across the input window rather than comparing response amplitudes. The magnitude $|C_{ij,k}|$ measures the temporal consistency of the phase difference. A value near one indicates a stable offset, whereas a value near zero indicates a varying or unreliable relation. The real and imaginary components retain the offset without introducing the discontinuity of a wrapped phase angle. We concatenate these quantities across frequency bands:

$$\mathbf{r}_{ij} = \mathrm{Concat}_{k=1}^{K} \left[\mathrm{Re}(C_{ij,k}), \mathrm{Im}(C_{ij,k}), |C_{ij,k}|\right] \in \mathbb{R}^{3K}. \tag{15}$$

The explicit magnitude provides the relation encoder with a direct cue for phase consistency. Moreover, $C_{ji,k} = C_{ij,k}^*$, so we explicitly impose this conjugate relation when constructing the reverse-direction descriptor.

Phase-guided Cross-part Aggregation. We combine content similarity with the estimated phase relation to determine how information is exchanged between body locations. For target location i and source location j, a_{ij}^{cont} measures the content compatibility between their local harmonic tokens. The phase-relation descriptor $\mathbf{r}_{ij}$ is mapped to an additive phase bias a_{ij}^{phase}, whose maximum

magnitude is controlled by λ_{ph}. The two scores are added and normalized over all source locations, producing the attention weight α_{ij}. This weight determines how strongly location j contributes when updating location i. Formally,

$$a_{ij}^{\text{cont}} = \frac{(\mathbf{W}_q \mathbf{p}_i)^\top (\mathbf{W}_k \mathbf{p}_j)}{\sqrt{d}}, \qquad a_{ij}^{\text{phase}} = \lambda_{\text{ph}} \tanh\big(\mathbf{w}_r^\top \phi(\mathbf{W}_r \mathbf{r}_{ij})\big), \tag{16}$$

$$\alpha_{ij} = \text{softmax}_j \left(a_{ij}^{\text{cont}} + a_{ij}^{\text{phase}} \right). \tag{17}$$

Here, d denotes the query-key dimension, and ϕ is the activation function in the relation encoder. Because the hyperbolic tangent is bounded, the phase bias satisfies $a_{ij}^{\text{phase}} \in [-\lambda_{\text{ph}}, \lambda_{\text{ph}}]$. Thus, λ_{ph} limits how strongly the phase relation can modify the content-based attention score. We set $a_{ii}^{\text{phase}} = 0$ because a location's phase relation with itself contains no cross-part information.

Finally, the attention weights aggregate location-dependent context through a residual update:

$$\widetilde{\mathbf{p}}_i = \mathbf{p}_i + \sum_{j=1}^{N} \alpha_{ij} \mathbf{W}_v \mathbf{p}_j. \tag{18}$$

The residual connection preserves the original local token, while the weighted sum introduces information selected jointly by content similarity and phase coordination. Because the summation includes $j = i$, self-information can still enter through content attention, although its phase bias is disabled.

3.4 Prediction and Training Objective

Given the cross-part contextualized tokens $\{\widetilde{\mathbf{p}}_i\}_{i=1}^N$, we obtain a global activity representation by mean pooling over all sensor locations:

$$\mathbf{z} = \frac{1}{N} \sum_{i=1}^{N} \widetilde{\mathbf{p}}_i \in \mathbb{R}^H. \tag{19}$$

A linear classifier then maps $\mathbf{z}$ to the class logits:

$$\mathbf{o} = \mathbf{W}_c \mathbf{z} + \mathbf{b}_c \in \mathbb{R}^C, \tag{20}$$

where $\mathbf{W}_c \in \mathbb{R}^{C \times H}$ and $\mathbf{b}_c \in \mathbb{R}^C$ are learnable classifier parameters. We optimize all model parameters jointly by minimizing the mini-batch cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp(o_{b,y_b})}{\sum_{c=1}^{C} \exp(o_{b,c})}, \tag{21}$$

where B is the mini-batch size, y_b is the ground-truth label of sample b, and $o_{b,c}$ denotes its logit for class c. The model is trained end-to-end using the cross-entropy loss $\mathcal{L}_{\text{cls}}$.

4 Experiments

4.1 Datasets, Protocol, and Baselines

Datasets. We evaluate RIM-Net on five public benchmarks covering both single-device and multi-location wearable human activity recognition: MotionSense (Malekzadeh et al., 2019), UCI-HAR (Anguita et al., 2013), PAMAP2 (Reiss & Stricker, 2012), MHEALTH (Baños et al., 2015), and Opportunity (Chavarriaga et al., 2013). MotionSense and UCI-HAR contain smartphone-based inertial measurements with 12 and 9 input channels, respectively. PAMAP2 and MHEALTH provide 27 and 23 channels collected from multiple body-worn sensors, while Opportunity represents a more densely instrumented multi-sensor setting. These datasets allow us to evaluate both local motion encoding and cross-part interaction under diverse sensing configurations. For the single-location datasets, only the rotation-invariant adaptive harmonic encoder is used, and the cross-part phase-relation module is omitted. For the multi-location datasets, input channels are grouped according to their physical sensor locations before being processed by RIM-Net. All signals are converted to consistent physical units, and normalization statistics are estimated exclusively from the training subjects. More details of sensor channels used in the experiments is shown in the Appendix.

Table 1: **Weighted F1-score on five HAR benchmarks over five random seeds (mean ± standard deviation, %).** Parameters are reported in thousands (K); ranges reflect dataset-specific model instantiations. Bold and underlined values indicate the best and second-best recognition results, respectively.

Model	Params (K)	MotionSense	UCI-HAR	PAMAP2	MHEALTH	Opportunity
TRI-HAR	35.1–98.1	89.50 ± 3.02	90.26 ± 5.03	$\underline{85.18 \pm 3.33}$	88.89 ± 7.77	91.82 ± 0.67
DeepConvLSTM	654.7–1639.1	84.73 ± 2.51	89.94 ± 1.01	72.59 ± 3.93	88.48 ± 3.11	88.20 ± 1.10
DynamicWHAR	730.7–1911.8	86.40 ± 2.04	88.33 ± 1.70	70.97 ± 3.54	84.55 ± 3.27	$\underline{92.79 \pm 0.71}$
TinyHAR	4.3–12.0	$\underline{92.67 \pm 0.67}$	$\mathbf{96.53 \pm 0.41}$	73.22 ± 3.58	$\underline{93.72 \pm 3.15}$	89.80 ± 0.90
TinierHAR	0.7–2.2	91.99 ± 0.60	$\underline{96.37 \pm 0.59}$	74.07 ± 1.16	92.50 ± 2.78	88.20 ± 1.20
RIM-NET (ours)	58–104	$\mathbf{95.09 \pm 1.18}$	93.47 ± 0.32	$\mathbf{89.27 \pm 1.42}$	$\mathbf{96.91 \pm 1.21}$	$\mathbf{95.05 \pm 0.48}$

Evaluation protocol. We adopt a subject-independent evaluation protocol in which the training, validation, and test sets contain mutually exclusive subjects. All windows from the same subject are assigned to a single partition, thereby preventing subject-specific information from leaking into the test set. Model selection and early stopping are performed using the validation weighted F1-score, and the held-out test set is evaluated only after training is completed. Within each dataset, all methods use identical window configurations, data partitions, and preprocessing procedures. Each experiment is repeated using five random seeds, and we report the mean and standard deviation of the test weighted F1-score. For a dataset with C activity classes, the weighted F1-score is defined as

$$F_1^{\mathrm{w}} = \sum_{c=1}^{C} \frac{n_c}{\sum_{j=1}^{C} n_j} F_{1,c}, \tag{22}$$

where n_c and $F_{1,c}$ denote the number of test samples and the weighted F1-score of class c, respectively. This metric accounts for class imbalance while retaining the contribution of every activity class.

Baselines. We compare RIM-NET with representative recurrent, convolution-based, and lightweight HAR architectures. DeepConvLSTM (Ordóñez & Roggen, 2016) combines convolutional feature extraction with recurrent temporal modeling. TRI-HAR (Baek et al., 2026) uses edge convolution followed by MLP layers. DynamicWHAR (Miao et al., 2022) explicitly learns input-dependent interactions among wearable sensors. TinyHAR (Zhou et al., 2022) and TinierHAR (Bian et al., 2025) are included as parameter-efficient baselines designed for resource-constrained HAR. Unless otherwise stated, all baselines are trained and evaluated locally under the same protocol as RIM-NET rather than using scores copied from their original publications.

4.2 RECOGNITION RESULTS

Table 1 shows that RIM-NET achieves the highest weighted F1-score on four of the five benchmarks. It obtains $95.09 \pm 1.18\%$ on MotionSense, $89.27 \pm 1.42\%$ on PAMAP2, $96.91 \pm 1.21\%$ on MHEALTH, and $95.05 \pm 0.48\%$ on Opportunity. Relative to the strongest baseline on each dataset, the corresponding gains are 2.42, 4.09, 3.19, and 2.26 percentage points, respectively. The largest margin occurs on PAMAP2, where RIM-NET improves over TinierHAR from 85.18% to 89.27%. The method also consistently outperforms both lightweight baselines on the other multi-location benchmarks, MHEALTH and Opportunity. These results are consistent with the intended use of local harmonic encoding and cross-part relation learning in multi-location sensing. On UCI-HAR, RIM-NET reaches $93.47 \pm 0.32\%$. This result exceeds DeepConvLSTM, DynamicWHAR, and TRI-HAR, but is 3.06 percentage points below TinyHAR and 2.90 points below TinierHAR. UCI-HAR is therefore the only benchmark on which RIM-NET does not attain the best result. Nevertheless, its standard deviation of 0.32 is the lowest among the compared methods on this dataset.

4.3 ABLATION STUDY

Sensor-Coordinate Rotation Robustness. We evaluate the robustness of different input representations and training strategies under controlled sensor-coordinate rotations. Two conventional raw-axis baselines are constructed using DeepConvLSTM. DeepConvLSTM-Raw is trained on the orig-

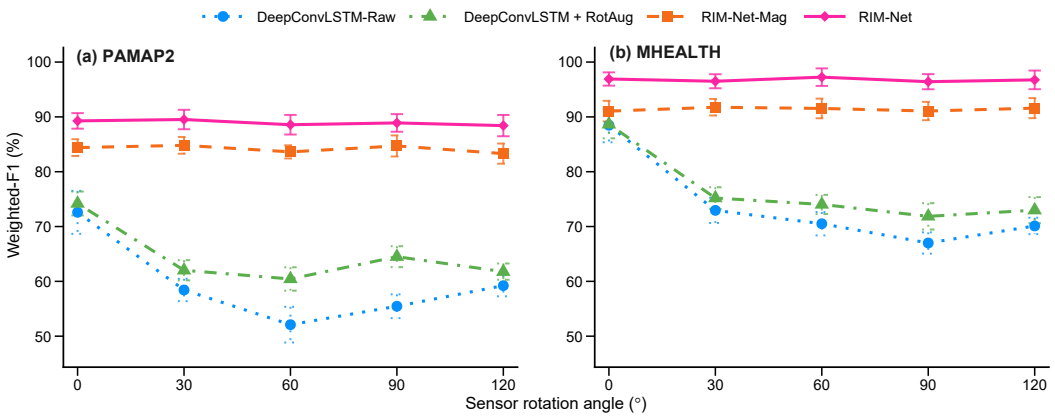

Figure 3: **Testing the rotation robustness under different rotation angles on PAMAP2 and MHEALTH.** Error bars denote the standard deviation over five runs. DeepConvLSTM-Raw is trained on the original signals. DeepConvLSTM-Raw + RotAug applies randomly sampled rotation augmentation during training. RIM-Net-Mag derives from RIM-Net, but it replaces the $\widetilde{h}_i(t)$ with 0 in the motion state.

Table 2: **Component-Level ablation of AQHE and CPRL.** Weighted F1-score (%) is reported as mean $\pm$ standard deviation over five runs. The motion-state representation is kept unchanged in all variants.

Variant	Local encoder	Cross-part fusion	PAMAP2	MHEALTH	Opportunity	Params (K)↓
Base (w/o both)	Matched temporal Conv	Mean pooling + matched MLP	71.23 $\pm$ 1.72	80.65 $\pm$ 1.49	81.65 $\pm$ 1.35	81.88
w/o AQHE	Matched temporal Conv	CPRL	81.30 $\pm$ 1.31	88.92 $\pm$ 1.43	90.88 $\pm$ 1.20	81.91
w/o CPRL	AQHE	Mean pooling + matched MLP	84.66 $\pm$ 1.19	93.02 $\pm$ 1.24	92.16 $\pm$ 1.08	81.85
Full model	AQHE	CPRL	**89.27 $\pm$ 1.42**	**96.91 $\pm$ 1.21**	**95.05 $\pm$ 0.48**	81.88

inal tri-axial accelerometer and gyroscope signals without rotation augmentation. DeepConvLSTM-Raw + RotAug uses the same architecture and raw-axis input, but applies randomly sampled rotation augmentation during training. The comparison between these two baselines evaluates whether rotation augmentation can improve the robustness of a conventional axis-dependent model. To isolate the contribution of the geometric alignment component $\widetilde{h}_i(t)$ in our motion state, we additionally construct RIM-Net-Mag. Specifically, its input state is defined as

$$\mathbf{v}_i^{\mathrm{mag}}(t) = \left[\log\left(\|\mathbf{a}_i^d(t)\|_2 + \epsilon\right), \log\left(\gamma\|\boldsymbol{\omega}_i(t)\|_2 + \epsilon\right), 0\right]^\top. \tag{23}$$

Both RIM-Net-Mag and RIM-Net are trained using only the original, unrotated training data. During evaluation, rotations are applied to the raw sensor signals before any preprocessing operation. For each body location i, the accelerometer and gyroscope share the same rotation matrix $\mathbf{R}_i$, whereas different locations are rotated independently. The rotation matrix remains fixed within each input window, consistent with the assumption used in our invariance analysis. We evaluate fixed rotation angles of $0°$, $30°$, $60°$, $90°$, and $120°$. These experimental results are shown in Figure 3. DeepConvLSTM models exhibit substantial performance degradation under sensor rotations, whereas RIM-Net-Mag and RIM-Net remain stable. Adding rotation augmentation makes the DeepConvLSTM perform better, but the improvement is limited. RIM-Net consistently outperforms RIM-Net-Mag, demonstrating the discriminative contribution of the acceleration–angular-velocity alignment component. More theoretical proof of rotation-invariance can be seen in the Appendix.

Table 3: **Mechanism-Level ablation of AQHE and CPRL.** When evaluating AQHE variants, the full CPRL is retained. When evaluating CPRL variants, the full AQHE is retained. Weighted F1-score (%) is reported as mean ± standard deviation over five runs.

Module	Variant	Controlled change	PAMAP2	MHEALTH	Opportunity
AQHE	Matched Conv	Harmonic encoding ⇒ temporal conv	81.30 ± 1.31	88.92 ± 1.43	90.88 ± 1.20
	Static-QHE	State-conditioned offsets ⇒ static offsets	87.80 ± 1.52	94.47 ± 1.27	93.81 ± 1.09
	Full AQHE		**89.27 ± 1.42**	**96.91 ± 1.21**	**95.05 ± 0.48**
CPRL	Mean fusion	Remove cross-part interaction	84.33 ± 2.09	89.09 ± 2.31	91.70 ± 1.50
	Content-only attention	Set $a_{ij}^{\text{phase}} = 0$	87.40 ± 1.72	94.21 ± 1.46	94.21 ± 1.19
	Frequency-averaged CPRL	Remove frequency -specific relations	85.09 ± 2.11	93.81 ± 1.83	90.50 ± 2.26
	CPRL w/o reliability	Set $\lvert C_{ij,k}\rvert = 0$	83.46 ± 1.73	87.27 ± 1.66	89.05 ± 1.81
	Full CPRL		**89.27 ± 1.42**	**96.91 ± 1.21**	**95.05 ± 0.48**

Component-Level Ablation. Table 2 evaluates AQHE and CPRL using a 2×2 design. When AQHE is removed, it is replaced by a depthwise-separable temporal convolutional encoder matched in depth, channel width, receptive field, and parameter count; the fixed quadrature filters are retained only for CPRL to extract phase relations. When CPRL is removed, its cross-part attention is replaced by mean pooling followed by a parameter-matched MLP. All variants use the same motion state and training protocol. Either module alone improves over the matched base model, while their combination achieves the best results on all three datasets, indicating that AQHE and CPRL provide complementary benefits.

Mechanism-Level Ablation. Table 3 examines the mechanisms underlying AQHE and CPRL while retaining the other in full. For AQHE, Matched Conv replaces the entire harmonic encoder with the matched temporal convolution, whereas Static-QHE retains quadrature filtering but replaces the state-conditioned offsets $\delta_{i,k}^{(\ell)}(t)$ with input-independent learnable offsets $\bar{\delta}_k^{(\ell)}$. For CPRL, Mean fusion directly averages location tokens without attention; content-only attention sets $a_{ij}^{\text{phase}} = 0$ while retaining the content-attention and residual-update operations; frequency-averaged CPRL averages the phase relations across frequency bands; and CPRL w/o reliability sets the explicit consistency channel $|C_{ij,k}|$ to zero while retaining its real and imaginary components. Full AQHE and CPRL consistently perform best, supporting state-conditioned harmonic adaptation and frequency-specific, reliability-aware cross-part modeling.

5 Conclusion

To address sensor-coordinate variations in wearable HAR, we model local motion rhythms and cross-part coordination using rotation-invariant quantities. Based on this insight, we proposed RIM-Net, which constructs a compact motion state from acceleration and angular-velocity magnitudes and their reliability-weighted geometric alignment. AQHE uses this state to adapt multi-frequency quadrature responses, while CPRL incorporates frequency-specific phase relations between motion-magnitude signals into cross-part attention. Under subject-independent evaluation on five public HAR benchmarks, RIM-Net ranks first among the compared methods on four datasets, and requires only 58K–104K trainable parameters. Controlled rotation experiments show that its performance remains stable across sensor-coordinate changes without rotation augmentation, whereas raw-axis models degrade substantially. Component- and mechanism-level ablations further support the contributions of the invariant motion state, state-conditioned harmonic encoding, and frequency-specific phase-relation learning. These findings demonstrate that physically structured, rotation-invariant representations provide an effective basis for robust wearable activity recognition.

AI USE STATEMENT

Generative AI tools were used solely to assist with writing the experimental code. All other aspects of this work were completed by the authors without generative AI assistance. The authors reviewed and tested the AI-assisted code to verify its correctness. The authors take full responsibility for the final content of this work, including the AI-assisted code.

ETHICS STATEMENT

This work studies wearable human activity recognition using existing public datasets and does not involve the collection of new human-subject data. The recorded activities and participant populations may not fully represent real-world users, and differences in sensor placement, physical ability, demographics, and data-collection protocols may introduce dataset-specific biases. Our conclusions are therefore limited to the evaluated datasets, activity classes, sensor configurations, and experimental settings. Wearable sensor data may also contain privacy-sensitive behavioral information. Any practical deployment should obtain informed consent, protect collected data, and evaluate performance across the intended user population. Finally, the reported results do not establish the suitability of the proposed method for clinical, safety-critical, or surveillance applications.

REPRODUCIBILITY STATEMENT

Section 3 specifies the rotation-invariant motion state, AQHE, CPRL, and the training objective. Section 4 describes the five public datasets, subject-independent evaluation protocol, baseline implementations, evaluation metric, and five-run uncertainty reporting. Appendix A.5 details the retained sensor channels, physical-unit conversion, resampling, and window segmentation, while Appendix A.6 provides the model hyperparameters, optimization settings, checkpoint selection, early stopping, and frequency-bank configuration. Appendix A.1 states the assumptions underlying rotation invariance, and Appendix A.4 documents the controlled sensor-rotation evaluation. Together, these details specify the model, preprocessing, training, and evaluation settings needed to reproduce the reported experiments.

REFERENCES

Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1:1–1:22, 2021. doi: 10.1145/3448083.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the European Symposium on Artificial Neural Networks*, pp. 437–442, 2013.

Seungyeol Baek, Yoonbyung Chai, Yonghyeon Lee, Sungjoon Choi, and Sungho Suh. Rotation-invariant multi-IMU activity recognition under independent per-location orientation shifts. In *Proceedings of the 2026 ACM International Symposium on Wearable Computers*, 2026. doi: 10.1145/3830727.3834824.

Oresti Baños, Claudia Villalonga, Rafael Garcia, Alejandro Saez, Miguel Damas, Juan A. Holgado-Terriza, Sungyong Lee, Hector Pomares, and Ignacio Rojas. Design, implementation and validation of a novel open framework for agile development of mobile health applications. *BioMedical Engineering OnLine*, 14:S6, 2015. doi: 10.1186/1475-925X-14-S2-S6.

Sizhen Bian, Mengxi Liu, Vítor Fortes Rey, Daniel Geißler, and Paul Lukowicz. TinierHAR: Towards ultra-lightweight deep learning models for efficient human activity recognition on edge devices. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers*, pp. 163–169, 2025. doi: 10.1145/3715071.3750410.

Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15): 2033–2042, 2013. doi: 10.1016/j.patrec.2012.12.014.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):77:1–77:40, 2021. doi: 10.1145/3447744.

Ehab Essa and Islam R Abdelmaksoud. Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors. *Knowledge-Based Systems*, 278:110867, 2023.

Harish Haresamudram, Chi Ian Tang, Sungho Suh, Paul Lukowicz, and Thomas Plötz. Past, present, and future of sensor-based human activity recognition using wearables: A surveying tutorial on a still challenging task. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):34:1–34:44, 2025. doi: 10.1145/3729467.

Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Shuangjian Li, Tao Zhu, Furong Duan, Liming Chen, Huansheng Ning, Christopher Nugent, and Yaping Wan. HARMamba: Efficient and lightweight wearable sensor human activity recognition based on bidirectional Mamba. *IEEE Internet of Things Journal*, 12(3):2373–2384, 2025. doi: 10.1109/JIOT.2024.3463405.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the ACM/IEEE International Conference on Internet-of-Things Design and Implementation*, pp. 49–58, 2019. doi: 10.1145/3302505.3310068.

Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. Towards a dynamic inter-sensor correlations learning framework for multi-sensor-based wearable human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):130:1–130:25, 2022. doi: 10.1145/3550331.

Payal Mohapatra, Lixu Wang, and Qi Zhu. Phase-driven generalizable representation learning for nonstationary time series classification. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=cb3nwoqLdd.

Yang Mu, Muhammad Shahzad, and Xiao Xiang Zhu. MPTSNet: Integrating multiscale periodic local patterns and global dependencies for multivariate time series classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19572–19580, 2025. doi: 10.1609/aaai.v39i18.34155.

Dominique Nshimyimana, Vítor Fortes Rey, Mengxi Liu, Bo Zhou, and Paul Lukowicz. LITEWAY: Lightweight HAR via temporal efficient highway. In *Proceedings of the 2026 ACM International Symposium on Wearable Computers*, 2026. URL https://arxiv.org/abs/2608.09421.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the International Symposium on Wearable Computers*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Chongyang Wang, Yuan Gao, Akhil Mathur, Amanda C. de C. Williams, Nicholas D. Lane, and Nadia Bianchi-Berthouze. Leveraging activity recognition to enable protective behavior detection in continuous data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):81:1–81:27, 2021. doi: 10.1145/3463508.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Jian Wu and Roozbeh Jafari. Orientation independent activity/gesture recognition using wearable motion sensors. *IEEE Internet of Things Journal*, 6(2):1427–1437, 2019. doi: 10.1109/JIOT.2018.2856119.

Haoyu Xie, Haoxuan Li, Chunyuan Zheng, Haonan Yuan, Guorui Liao, Jun Liao, and Li Liu. Decomposing and fusing intra- and inter-sensor spatio-temporal signal for multi-sensor wearable human activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (13):14441–14449, 2025. doi: 10.1609/aaai.v39i13.33582.

Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3995–4001, 2015.

Xiaoqun Yu, Tiejun Ma, Jaehyuk Jang, and Shuping Xiong. Data augmentation to address various rotation errors of wearable sensors for robust pre-impact fall detection. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2197–2207, 2023. doi: 10.1109/JBHI.2022.3228598.

Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. TinyHAR: A lightweight deep learning model designed for human activity recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, pp. 89–93, 2022. doi: 10.1145/3544794.3558467.

Yexu Zhou, Tobias King, Haibin Zhao, Yiran Huang, Till Riedel, and Michael Beigl. MLP-HAR: Boosting performance and efficiency of HAR models on edge devices with purely fully connected layers. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pp. 133–139, 2024. doi: 10.1145/3675095.3676624.

A APPENDIX

A.1 PROOF OF ROTATION INVARIANCE.

Scope of the invariance result. The following result concerns the analytically constructed motion state under an ideal sensor-coordinate transformation. We assume that the same rotation is applied to the co-located accelerometer and gyroscope, that the rotation remains constant over the support of the low-pass filter, and that any preprocessing performed before state construction is rotation-equivariant. Under these assumptions, the state is pointwise invariant to the coordinate rotation.

Here, we aim to theoretically demonstrate the rotational invariance of the constructed motion state. For a sensor at body location i, a fixed frame rotation transforms the acceleration and angular velocity as

$$\mathbf{a}_i'(t) = \mathbf{R}_i \mathbf{a}_i(t), \qquad \boldsymbol{\omega}_i'(t) = \mathbf{R}_i \boldsymbol{\omega}_i(t), \qquad \mathbf{R}_i \in \mathrm{SO}(3), \tag{24}$$

where $\mathbf{R}_i$ satisfies $\mathbf{R}_i^\top \mathbf{R}_i = \mathbf{I}$ and $\det(\mathbf{R}_i) = 1$. The accelerometer and gyroscope at the same body location share a coordinate frame and therefore undergo the same rotation. Sensors at different body locations may undergo different rotations. We now establish that the resulting state is invariant to the frame rotation in Eq. (24). Suppose that $\mathbf{R}_i$ remains constant over the support of the low-pass filter. The linearity of $\mathcal{L}$, together with identical filtering along all three axes, gives

$$\mathbf{a}_i^{d\prime}(t) = \mathbf{a}_i'(t) - \mathcal{L}(\mathbf{a}_i')(t) \quad = \mathbf{R}_i \mathbf{a}_i(t) - \mathbf{R}_i \mathcal{L}(\mathbf{a}_i)(t) = \mathbf{R}_i \mathbf{a}_i^d(t). \tag{25}$$

Because $\mathbf{R}_i^\top \mathbf{R}_i = \mathbf{I}$, the rotation preserves both motion magnitudes:

$$|\mathbf{R}_i \mathbf{a}_i^d(t)|_2 = |\mathbf{a}_i^d(t)|_2, \qquad |\mathbf{R}_i \boldsymbol{\omega}_i(t)|_2 = |\boldsymbol{\omega}_i(t)|_2. \tag{26}$$

It also preserves their inner product:

$$\big(\mathbf{R}_i \mathbf{a}_i^d(t)\big)^\top \big(\mathbf{R}_i \boldsymbol{\omega}_i(t)\big) = \mathbf{a}_i^d(t)^\top \boldsymbol{\omega}_i(t). \tag{27}$$

The first two state components are therefore unchanged. Moreover, $h_i(t)$ is unchanged because it depends only on norms and an inner product, while $c_i(t)$ is unchanged because it depends only on norms. Consequently, the motion state $\mathbf{v}_i'(t)$ computed after the sensor-frame rotation is identical to the motion state $\mathbf{v}_i(t)$ computed before the rotation.

This result establishes state-level invariance under the assumptions above. It does not use aggregate recognition scores as evidence of exact invariance. The rotation experiments instead assess the end-to-end empirical robustness of the implemented model.

A.2 COMPLEXITY

Let N, K, H, T', and L_h denote the numbers of sensor locations, frequency bands, feature channels, encoded time steps, and harmonic kernel length, respectively. For one AQHE block, depthwise quadrature filtering requires $\mathcal{O}(NKHT'L_h)$ operations and is therefore linear in the sequence length. With L AQHE blocks of the same temporal resolution, the corresponding complexity becomes $\mathcal{O}(LNKHT'L_h)$. In CPRL, computing the phase relations between all pairs of body locations requires $\mathcal{O}(N^2KT')$ operations, while cross-part attention requires $\mathcal{O}(N^2H)$. The quadratic dependence is limited to the number of sensor locations N, which is typically small in multi-location wearable HAR. Since the AQHE parameters and quadrature filters are shared across body locations, the number of encoder parameters is independent of N.

A.3 PARAMETER AND COMPUTATIONAL EFFICIENCY

RIM-NET contains between 58K and 104K trainable parameters across the evaluated datasets. This parameter count is substantially lower than those of DeepConvLSTM and DynamicWHAR, which contain 654.7K and 927.3K parameters in the reported configurations, respectively. Combined with the recognition results, this comparison shows that RIM-NET achieves stronger performance on most datasets without relying on a large recurrent or graph-based backbone.

RIM-NET is not the smallest model in the comparison. TRI-HAR, TinyHAR, and TinierHAR use fewer parameters, with reported sizes of 35.1K–98.1K, 4.3K–12.0K, and 0.7K–2.2K, respectively.

Table 4: Accelerometer–gyroscope configurations used for the five datasets. Only co-located tri-axial accelerometer–gyroscope pairs are retained. Each selected location therefore contributes six input channels.

Dataset	Co-located Acc–Gyro pairs	Input channels	Sensor configuration
MotionSense	1	6	One tri-axial acceleration stream and one tri-axial angular-velocity stream from a single smartphone.
UCI-HAR	1	6	Tri-axial total acceleration and tri-axial angular velocity from the waist-mounted smartphone.
PAMAP2	3	18	Co-located accelerometer–gyroscope pairs at the hand, chest, and ankle; each location contributes six channels.
MHEALTH	2	12	Co-located pairs at the left ankle and right lower arm. The chest accelerometer is excluded because no corresponding gyroscope is available.
Opportunity	7	42	Seven body-worn IMUs, each providing tri-axial acceleration and tri-axial angular velocity.

Single-device datasets; multi-location datasets with complete Acc–Gyro pairs; datasets containing an additional location without a paired gyroscope. For PAMAP2, one tri-axial accelerometer range is retained at each location to avoid duplicating measurements of the same physical acceleration.

Although RIM-NET allocates more parameters than the two ultra-lightweight models, it achieves higher weighted F1-score on four of the five benchmarks. The results therefore support a favorable recognition–compactness trade-off rather than a claim of minimum model size.

The computational cost of RIM-NET consists primarily of local harmonic encoding and cross-part relation learning. As analyzed in Section A.2, its overall complexity is

$$\mathcal{O}\big(LNKHT'L_h + N^2KT' + N^2H\big)\,. \tag{28}$$

The harmonic encoding stage scales linearly with the temporal length, while the quadratic terms depend only on the number of sensor locations, which is small for the evaluated wearable configurations. Because measured MACs, latency, and energy consumption are not included in Table 1, we restrict the empirical efficiency claim to parameter compactness. The complexity analysis characterizes computational scaling, but hardware-level efficiency requires direct measurements on the intended deployment platform.

A.4 Dataset and Preprocessing Details

Table 4 summarizes the inertial channels retained for each dataset. We use only co-located tri-axial accelerometer–gyroscope pairs because the proposed motion state requires both vectors to be expressed in the same local sensor coordinate frame. Consequently, each retained sensor location contributes six channels. For MHEALTH, the chest accelerometer is excluded because the chest device does not provide angular-velocity measurements. Other modalities, including magnetometers, ECG signals, heart-rate measurements, orientation estimates, object sensors, and ambient sensors, are not used. All inertial signals are resampled to 50 Hz before window segmentation using appropriate anti-alias filtering. We segment each dataset into 128-sample windows with a stride of 64 samples, corresponding to a window duration of 2.56 s and a stride of 1.28 s. This unified temporal resolution ensures that the frequency bands and temporal kernels have the same physical interpretation across datasets.

A.5 Implementation Details

Training configuration. We train RIM-Net for at most 300 epochs using AdamW with an initial learning rate of 10^{-3}, a weight decay of 10^{-4}, and a batch size of 64. The learning rate remains fixed throughout training. We optimize the standard cross-entropy loss without label smoothing and do not apply gradient clipping. After each epoch, we evaluate the model using the validation weighted

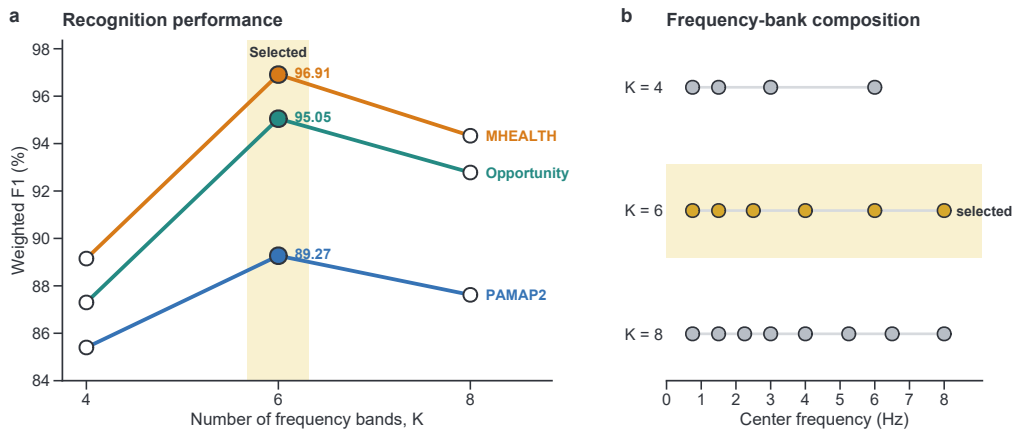

Figure 4: **Sensitivity to the number and placement of frequency bands.** (a) weighted F1-score obtained with different numbers of frequency bands. (b) Corresponding center-frequency configurations. The shaded region denotes the selected six-band configuration.

F1-score and save the checkpoint achieving the best validation result. Training is terminated early if the validation score does not improve for 15 consecutive epochs, after which the best checkpoint is used for test evaluation. Rotation augmentation is disabled for the complete RIM-Net unless explicitly stated otherwise.

Model configuration. Unless otherwise specified, the same model configuration is used across all datasets. For motion-state construction, the low-frequency acceleration component is estimated using a normalized symmetric Hann kernel of length 21. We set the modality scale factor to $\gamma = 1.0$ after converting the measurements to consistent physical units. The logarithmic stabilization constant is set to $\epsilon = 10^{-6}$, while the stabilization constants in the acceleration–angular-velocity alignment term are set to $\epsilon_a = \epsilon_\omega = 10^{-6}$. The reliability thresholds are set to $\tau_a = \tau_\omega = 0.05$.

AQHE configuration. The motion state is projected to a hidden dimension of $H = 64$. The quadrature filter bank contains $K = 6$ bands centered at $\{0.75, 1.5, 2.5, 4.0, 6.0, 8.0\}$ Hz. The filters are distributed more densely over the low- and medium-frequency range, which contains the fundamental rhythms of common human activities, while the higher-frequency bands capture rapid motion variations and motion harmonics. Because all signals are resampled to 50 Hz, these frequencies have the same physical interpretation across datasets. All cosine and sine kernels contain 101 samples, corresponding to approximately 2.02 s, and are normalized to unit ℓ_2 norm.

The local encoder contains two AQHE blocks. In each block, the offset controller first projects the three-dimensional motion state to 32 hidden channels and then applies a temporal convolution with kernel size 5 to generate six frequency-specific offsets. The offsets are bounded using $\delta_{\max} = \pi/2$. The feed-forward subnetwork uses an expansion ratio of 2, corresponding to an intermediate dimension of $2H = 128$. A dropout rate of 0.1 is applied to the AQHE residual projection and feed-forward subnetwork.

CPRL configuration. For the multi-location datasets, the query and key projections have dimension $d = 64$. The relation descriptor for each location pair has dimension $3K = 18$, consisting of the real component, imaginary component, and phase-consistency magnitude from each of the six frequency bands. It is processed by a two-layer relation encoder with a hidden dimension of 32 and a GELU activation. The phase-bias scale is set to $\lambda_{\text{ph}} = 1.0$, bounding the phase-relation bias within $[-1, 1]$. We set $\epsilon_{\text{ph}} = 10^{-6}$ to stabilize the normalized complex cross-product. CPRL is omitted for the single-location MotionSense and UCI-HAR datasets.

A.6 SENSITIVITY TO THE FREQUENCY-BANK CONFIGURATION.

We investigate the influence of the number and placement of quadrature frequency bands by comparing configurations with $K \in \{4, 6, 8\}$ while keeping all other settings fixed. As shown in Fig. 4, increasing the number of bands from four to six consistently improves performance on PAMAP2, MHEALTH, and Opportunity. The six-band configuration achieves an average weighted F1-score of 93.74%, exceeding the four- and eight-band configurations by 6.46 and 2.17 percentage points, respectively. Increasing K further to eight does not provide additional gains, suggesting that the more densely placed bands offer limited complementary information. We therefore adopt $K = 6$ with center frequencies $\{0.75, 1.5, 2.5, 4.0, 6.0, 8.0\}$ Hz as the default configuration.