HARMONY: A Privacy-preserving and Sensor-agnostic Tele-monitoring system

Qipeng Xie[1] , **Hao Guo**[2] , **Weizheng Wang**[3] , **Yongzhi Huang**[1] , **Linshan Jiang**[5] , **Jiafei Wu**[4] , **Shuxin Zhong**[1] *__Lu Wang__ [6] and **Kaishun Wu**[1]
[1]Hong Kong University of Science and Technology, HKUST (Guangzhou)
[2]Zhejiang Lab
[3]City University of Hong Kong
[4]The University of Hong Kong
[5]National University of Singapore
[6]ShenZhen University
qxieaf@connect.ust.hk, guoh@zhejianglab.org, linshan@nus.edu.sg, jcjiafeiwu@gmail.com
{shuxinzhong, wuks}@hkust-gz.edu.cn, {huangyongzhi, wanglu}@email.szu.edu.cn

Abstract

Global aging necessitates tele-monitoring systems to provide real-time tracking and timely assistance for older adults living independently. While pervasive wireless devices (e.g., CSI, IMU, UWB) enable cost-effective, non-intrusive monitoring, existing systems lack flexibility, limiting their adaptability to different environments. In this work, we posit that the motion dynamics of human movement are invariant across sensing modalities, inspiring the design of HARMONY—a privacy-preserving, sensor-agnostic system that supports multi-modal inputs and diverse tele-monitoring tasks. HARMONY incorporates *Modality-agnostic Data Processing* to uniformly encrypt multi-modal signals and *Task-specific Activity Recognition* for seamless tasks adaptation. A novel *Encrypted-processing Engine* then significantly accelerates computations on encrypted data by optimizing matrix and convolution operations. Evaluations across five different sensing modalities show that HARMONY consistently achieves high accuracy while delivering 3.5 × to 130 × speedups over state-of-the-art baselines. Our results demonstrate that HARMONY is a **practical**, **scalable**, and **privacy-centric** prototype for next-generation remote healthcare.

1 Introduction

Global aging is accelerating, with the over-65 population expected to double by 2050 [Bloom and Luca, 2016; Yang *et al.*, 2024]. This surge poses critical challenges for healthcare systems, particularly in supporting older adults who live alone and are at elevated risk of falls and emergencies [Schütz *et al.*, 2022]. Although traditional caregiver-dependent practices offer personalized support, they do not scale well to meet growing demands. Consequently, researchers are exploring pervasive wireless sensing (e.g., CSI, IMU, UWB) as a cost-effective, unobtrusive means of continuously monitoring elder behavior [Wang *et al.*, 2016]. These sensors promise real-time tracking, preserving independence and enhancing patient-centered care, while offering a scalable solution toward the next generation of remote healthcare [Liu *et al.*, 2016].

Recent research has made significant strides in telemonitoring [Fernandes *et al.*, 2024]. For example, Li et al. combine smartphone accelerometers and microphones to track daily activities and assess medication efficacy [Li *et al.*, 2023]. Zhang et al. introduce RF signals for long-term heart rate variability assessment [Zhang *et al.*, 2024]. Ouyang et al. fuse camera, mmWave radar, and microphone data to predict Alzheimer's progression [Ouyang *et al.*, 2024]. However, they fail to treat multi-modal signals as flexible, interchangeable inputs and support seamless integration of different tasks, which reduces their adaptability to various settings.

To address these limitations, we posit that human movement can be universally characterized by velocity, displacement, and trajectory—independent of the sensing modality. Building on this principle, we propose a unified multi-modal system that encodes these fundamental motion dynamics, thereby enabling consistent and scalable solutions for tasks such as fall detection and behavior recognition [Xu *et al.*, 2024]. Nevertheless, collecting sensitive motion data (e.g., facial features and gait patterns) raises substantial privacy concerns, including risks of identity theft and unauthorized surveillance [Ouyang *et al.*, 2021; Jiang *et al.*, 2024].

To this end, we introduce HARMONY, a privacy-preserving and sensor-agnostic tele-MONItoring system designed to support diverse beHAhavior Recognition tasks. Building upon Homomorphic Encryption (HE) [Zhang *et al.*, 2021], HARMONY executes all computations on encrypted data, ensuring robust privacy. To alleviate HE's high computational overhead, we introduce a novel *Encrypted-processing Engine* that optimizes matrix and convolution operations via structured computations and alignment-aggregation strategies, significantly reducing intermediate costs. Seamless inte-

*Corresponding Author: Shuxin Zhong.

gration of this engine with *Modality-agnostic Data Processing* and *Task-specific Activity Recognition* unifies varied sensing modalities and tasks, paving the way for secure, scalable deployment in real-world healthcare environments. The key contributions are summarized as follows:

- To the best of our knowledge, we introduce HARMONY, the first tele-monitoring system that unifies accessibility, adaptability, and robust privacy guarantees—offering a prototype for next-generation elder care.
- Technically, HARMONY contains: i) a *Modality-aware Data Filtering* component that ensures data reliability through heterogeneous filtering mechanisms; ii) a *Modality-agnostic Data Processing* component that transforms multi-modal sensor inputs into a consistent polynomial representation for uniform, secure processing; iii) a *Task-specific Activity Recognition* component that adapts seamlessly across various tasks (e.g., behavior recognition or fall detection); and iv) a *Encrypted-processing Engine* serves as *the core acceleration component*, enabling efficient computations directly on encrypted data.
- We implement and evaluate HARMONY using five sensing modalities (e.g., CSI, IMU, UWB) in realistic tele-monitoring scenarios. Results show that HARMONY maintains high accuracy while achieving 3.5 × to 130 × speedups over state-of-the-art privacy-preserving baselines, demonstrating its feasibility for practical, large-scale tele-monitoring deployments.

2 Related Work

The section reviews existing research on tele-monitoring systems and privacy-preserving techniques.

2.1 Tele-monitoring systems

Existing tele-monitoring systems have utilized both single- and multi-sensor setups to capture behavior-related signals. Single-modality solutions—for example, RF-based monitoring of heart rate variability [Zhang *et al.*, 2024] or acoustic respiration analysis [Song *et al.*, 2020]—excel at specific tasks but offer limited coverage. In contrast, multi-sensor systems expand capabilities by fusing diverse data streams: Li et al. combine smartphone accelerometers and microphones for gait and medication monitoring [Li *et al.*, 2023], while Ouyang et al. merge camera, mmWave radar, and audio for Alzheimer's staging [Ouyang *et al.*, 2024]. Further, some systems integrate RGB cameras, IMUs, and behavioral logs to detect high-risk activities [Fernandes *et al.*, 2024], or unify multiple sensor streams with electronic health records to mitigate missing modalities [Zhang *et al.*, 2022; Xu *et al.*, 2024]. Despite these advancements, two crucial gaps remain. First, multi-modal sensors are often treated as fixed, limiting flexibility across different living scenarios (e.g., WiFi in bedrooms vs. cameras in living rooms). Second, most solutions are designed to address specific tasks in isolation, rather than supporting a broader range of activities.

2.2 Privacy-Preserving Techniques

As tele-monitoring solutions scale, privacy has become a cornerstone concern [Juvekar *et al.*, 2018; Yang *et al.*, 2023b]. Early strategies mitigated risks by collecting only non-sensitive signals (e.g., PDVocal's focus on breathing sounds [Zhang *et al.*, 2019]) or by injecting noise via differential privacy (DP) to mask sensitive information [**?**]. While federated learning helps by decentralizing data [Ouyang *et al.*, 2024], it remains vulnerable to inference attacks and gradient leakage. HE provides an alternative by enabling computations on encrypted data [Chien *et al.*, 2023], albeit at the cost of substantial computational overhead.

3 Real-time Tele-monitoring in Elderly Care

This section introduces the real-time tracking scenario in elderly care, identifies its challenges and social impacts, and subsequently provides a formal problem formulation.

3.1 Challenges and Social Impact

The global population is aging rapidly, with the percentage of individuals over 65 expected to double by 2050 [Bloom and Luca, 2016]. This demographic shift exerts substantial pressure on healthcare systems, particularly for older adults living independently who require around-the-clock monitoring due to elevated risks of falls or sudden health deteriorations [Schütz *et al.*, 2022]. Traditional care practices that rely on professional caregivers or costly infrastructures struggle to meet these demands at scale. In response, pervasive wireless sensing (e.g., CSI, IMU) has emerged as a cost-effective, continuous, and non-intrusive tracking solution [Enshaeifar *et al.*, 2020; Wu *et al.*, 2024]. However, existing systems often suffer from rigid sensor-input configurations, narrow task specificity, and insufficient privacy safeguards—factors that limit their applicability in diverse real-world scenarios. Our work addresses these limitations by proposing a **privacy-preserving, sensor-agnostic** system capable of supporting multiple behavior-recognition tasks in tele-monitoring contexts. This design introduces a **scalable, user-centric** prototype for next-generation remote healthcare, enhancing both autonomy and safety for older adults.

3.2 Problem Formulation

We formalize the sensor-agnostic tele-monitoring task as learning a unified classification function, $\mathcal{F}_\theta$, capable of processing signals from various sensor modalities to identify behaviors across multiple tasks. Let $x_{c(i)}$ represent the input signal from a specific modality $c(i) \in \mathcal{C}$, where $\mathcal{C}$ is the set of all possible sensor types (e.g., WiFi, IMU, Camera). Similarly, let $y_{t(i)}$ denote the label associated with the recognition task $t(i) \in \mathcal{T}$, where $\mathcal{T}$ encompasses various tasks (e.g., gesture recognition, activity monitoring). Formally:

$$y_{t(i)} = \mathcal{F}_\theta(x_{c(i)}; \theta), \tag{1}$$

where θ denotes the learnable parameters of $\mathcal{F}$.

4 Methodology

We present HARMONY, a privacy-preserving, sensor-agnostic tele-monitoring system for diverse behavior recognition tasks, designed to enhance safety and independence in elderly care (see Figure 1). HARMONY comprises four components:

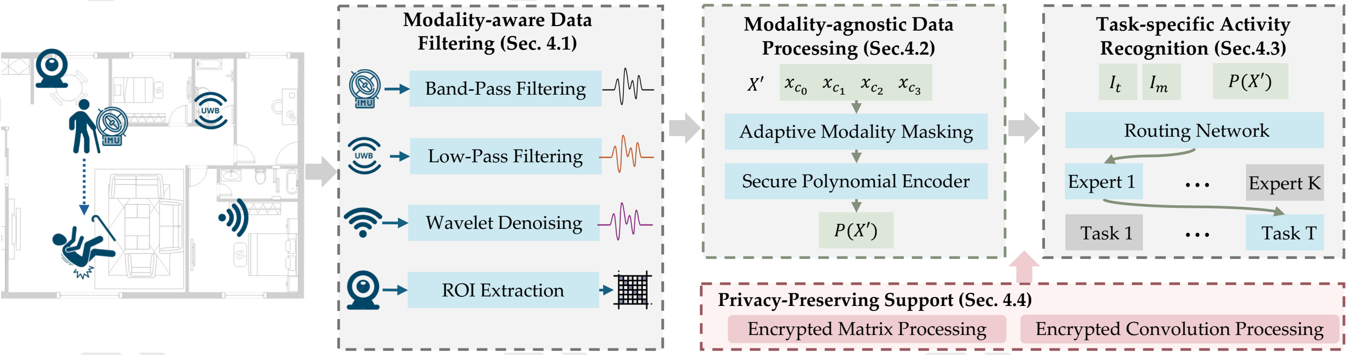
Modality-aware Data Filtering (Sec. 4.1)
IMU
Band-Pass Filtering
UWB
Low-Pass Filtering
Wavelet Denoising
ROI Extraction
Modality-agnostic Data Processing (Sec.4.2)
X'
x_{c_0} x_{c_1} x_{c_2} x_{c_3}
Adaptive Modality Masking
Secure Polynomial Encoder
$P(X')$
Task-specific Activity Recognition (Sec.4.3)
I_t I_m $P(X')$
Routing Network
Expert 1
…
Expert K
Task 1
…
Task T
Privacy-Preserving Support (Sec. 4.4)
Encrypted Matrix Processing
Encrypted Convolution Processing

Figure 1: Overview of the HARMONY system. It consists of four main components: (1) *Modality-aware Data Filtering*, (2) *Modality-agnostic Data Processing*, and (3) *Task-specific Activity Recognition*. Additionally, *Encrypted-processing Engine* serves as the core acceleration component that enables efficient computations directly on encrypted data.

- **Modality-aware Data Filtering** (Sec. 4.1) applies heterogeneous techniques to filter environmental and device-induced noise, thereby improving data reliability.
- **Modality-agnostic Data Processing** (Sec. 4.2) uses a flexible masking mechanism to preprocess multi-modal signals, adapting dynamically to varying inputs. It then encodes plain-text signals into structured polynomial representations via HE, ensuring secure computation.
- **Task-specific Activity Recognition** (Sec. 4.3) implements a MoE paradigm to dynamically route inputs to specialized expert modules for task-specific analysis.
- **Encrypted-processing Engine** (Sec. 4.4) forms HARMONY's foundation, introducing efficient matrix and convolution operations on encrypted data to facilitate pracical deployment.

4.1 Modality-aware Data Filtering

Modality-aware Data Filtering customizes de-noising strategies based on the unique properties of each signal, effectively mitigating noise caused by the environment and devices. Specifically, band-pass filtering confines Inertial Measurement Unit (IMU) signals to a designated frequency band, preserving meaningful motion data [Cesareo *et al.*, 2018]; low-pass filtering removes high-frequency noise from Ultra-Wideband (UWB) signals [Ma and Yeo, 2010]; wavelet denoising decomposes Channel State Information (CSI) signals into distinct frequency components to isolate relevant patterns [Wang *et al.*, 2014]; and Region of Interest (ROI) detection filters out non-essential background elements, retaining only critical regions for further analysis [Li *et al.*, 2017]. Formally, this process is represented as:

$$x^f_{c(i)} = \mathcal{F}_{c(i)}(x_{c(i)}, \phi_{c(i)}), c(i) \in \mathcal{C} \tag{2}$$

where $x_{c(i)}$ is the i-th signals collected from sensor modality $c(i)$, $x^f_{c(i)}$ is the filtered output, $\mathcal{F}_{c(i)}$ is the sensor-specific filtering function, and $\phi_{c(i)}$ are its parameters.

4.2 Modality-agnostic Data Processing

Modality-agnostic Data Processing incorporates an *Adaptive Modality Masking* mechanism for flexible processing of multi-modal signals and a *Secure Polynomial Encoder* to encode plain-text signals to structured polynomial representations, ensuring secure transportation and computation.

Adaptive Modality Masking

To enable adaptive processing of multi-modal data, we first unify the filtered modality signals $x^f_{c(i)}$ from sensors $\mathcal{C}$ into a single representation, denoted as $X = [x^f_{c(i)}]$, where $c(i) \in \mathcal{C}$. For task-specific selection of modalities, we introduce a masking mechanism $M = [m_{c(i)}]$, where each $m_{c(i)}$ corresponds to the signal $x_{c(i)}$. Here, $m_{c(i)}$ is a scalar that determines the inclusion of $x^f_{c(i)}$ in the computation: $m_{c(i)} = 1$ means the modality is active, while $m_{c(i)} = 0$ excludes it. For instance, if we use CSI for fall detection [Wang *et al.*, 2016], the mask $m_{c(i)}$ for CSI is set to 1, ensuring its contribution. Formally, the masked signals are calculated as:

$$X' = X \odot M, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This mechanism allows HARMONY to flexibly adapt to diverse real-world scenarios with varying device constraints.

Secure Polynomial Encoder

To preserve data confidentiality throughout processing, we adopt HE [Zhang *et al.*, 2021; Yang *et al.*, 2023b], which encodes plain-text signals into a structured polynomial representation. For masked time-series signals $X' = [x'_1, \ldots, x'_n]$, it generates:

$$P(X') = x'_1 + x'_2 z + \ldots + x'_n z^{n-1}, \tag{4}$$

with n as the signal length and z is the polynomial base. This representation is directly fed into *Task-specific Activity Recognition* for encrypted feature extraction.

4.3 Task-specific Activity Recognition

To adaptively handle various tasks, we adopt a Mixture of Experts (MoE) paradigm [Xu *et al.*, 2024]. With this, a routing network $\mathcal{R}(\cdot)$ directs input signals to specialized expert modules designed for task-specific analysis. Concretely, for a given signal i from sensor modality $c(i)$ associated with task

$t(i)$, $\mathcal{R}(\cdot)$ computes assignment weights $\tau_{(t(i),k)}$ over K experts using task-specific indicators $I_{t(i)}$ and modality-specific indicators $I_{c(i)}$:

$$\tau_{(t(i),k)} = \mathcal{R}(P(X'), I_{t(i)}, I_{c(i)}), \text{where} \sum_{k=1}^{K} \tau_{(t(i),k)} = 1. \tag{5}$$

Each expert k then processes the feature representation $P(X')$ through its designated network to produce an output $\mathcal{F}_k(P(X'))$. The detailed computation processes and network design are discussed in Sec. 4.4. Finally, the inference for i-th signal is obtained by aggregating the expert outputs weighted by their respective assignment scores:

$$y_i = \sum_{k=1}^{K} \tau_{(t(i),k)} \cdot \mathcal{F}_k(P(X')). \tag{6}$$

Model Optimization. We adopt a classification setup and optimize HARMONY using the Cross Entropy (CE) loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|M_{t(i)}|} y_{(i,j)} log(\hat{y}_{(i,j)}), \tag{7}$$

where $|M_{t(i)}|$ is the number of categories for task t.

4.4 Encrypted-processing Engine

Real-time systems demand highly efficient processing. We employ fully connected networks (FCNs) for time-series data and convolutional neural networks (CNNs) for images. However, performing fundamental operations—such as matrix-vector multiplication in FCNs and convolution in CNNs—directly on encrypted signals $P(X')$ is computationally expensive. To address this challenge, we propose *Encrypted Matrix Processing* and *Encrypted Convolution Processing*. These components leverage structured computation and alignment-aggregation strategies to optimize matrix and convolution operations on encrypted data, significantly reducing computational overhead.

Encrypted Matrix Processing

We first introduce an efficient matrix-vector multiplication between a plain-text weight matrix W and an encrypted input vector $\widetilde{X}$ to reduce computational overhead. The process is illustrated in Figure 2 and consists of four key steps:

- *Matrix Decomposition and Encoding (Step 1).* The plain-text weight matrix $W \in \mathcal{R}^{(f_o, f_i)}$ is decomposed row-wise into f_o vectors w_0, w_1, w_{f_0-1} using diagonal encoding. This encoding rearranges each row into a structured format that enables element-wise multiplication with the encrypted input vector $\widetilde{X}_i$, avoiding intermediate data shifts and computational bottlenecks.
- *Element-wise Multiplication (Step 2).* Each encoded row w_j is multiplied element-wise with $\widetilde{X}_i$, resulting in intermediate encrypted vectors $\widetilde{Y}_i$. These intermediate results represent partial dot products for each row.
- *Rotation for Alignment (Step 3).* The intermediate vectors $\widetilde{Y}_i$ undergo a series of rotations to align corresponding elements across ciphertexts, ensuring that terms contributing to the same dot product are properly positioned. This step enables efficient summation in the encrypted domain.
- *Summation (Step 4).* Finally, the rotated terms within each ciphertext are summed to compute the dot product for each row of W, producing the output vector $\widetilde{U}_i = [u_i^0, u_i^1, \dots, u_i^{f_o}]^T$. For example:

$$\begin{aligned} u_0 &= w_{00}x_0 + w_{11}x_1 + w_{22}x_2 + w_{33}x_3, \\ u_1 &= w_{10}x_0 + w_{01}x_1 + w_{12}x_2 + w_{03}x_3. \end{aligned} \tag{8}$$

Strengths. By leveraging structured computations and rotation operations, our methods significantly minimize reliance on expensive permutation operations, enabling scalable and efficient matrix-vector multiplication for encrypted data.

Encrypted Convolution Processing

We then introduce an efficient method for convolution computation between a plain-text convolutional kernel $K = [k_{ij}]$ and an encrypted input image $\widetilde{X}$, which is represented as a polynomial using Eq. 4. The encrypted image is represented as $P(X') \in \mathcal{R}^{(l_1, l_2)}$ and the kernel matrix $K \in \mathcal{R}^{(l_3, l_4)}$ encodes learnable parameters, capturing spatial-temporal dependencies. The procedure involves four steps:

- *Sliding Window Transformation.* The encrypted input $P(X')$ is divided into overlapping patches based on the kernel dimensions. Each patch corresponds to a sub-matrix of the input, and these patches are rearranged into rows of a new transformed matrix, denoted as $\widetilde{X}$.
- *Kernel Expansion.* The kernel K is broadcasted to match the size of $\widetilde{X}$, ensuring each kernel element k_{ij} aligns with the corresponding features in $\widetilde{X}$.
- *Element-wise Multiplication.* The transformed matrix $\widetilde{X}$ is element-wise multiplied with the expanded kernel K, capturing features interactions. The resulting matrix is denoted as $M = \widetilde{X} \cdot K$.
- *Aggregation.* The output is computed as $Y = M + b$, where b is the bias term added to the aggregated result.

To meet the time constraints of real-world deployments, the output is not passed to a second convolutional layer, as suggested in [Brutzkus *et al.*, 2019]. Instead, it is directly fed into *Encrypted Matrix Processing* for further feature extraction.

Complexity Analysis

Table 1 compares the computational complexity of the *Naive*,*Diagonal*, and HARMONY (Ours) methods, focusing on the number of Rot, SCMult, and Add operations. Among these, Rot operations are the most computationally expensive due to their exponential growth in the Naive method with respect to the output dimension m [Zhang *et al.*, 2021]. The Diagonal method mitigates this issue by strategically arranging matrix elements, but its efficiency degrades when the input dimension n increases linearly. To overcome these limitations, HARMONY employs Encrypted-processing

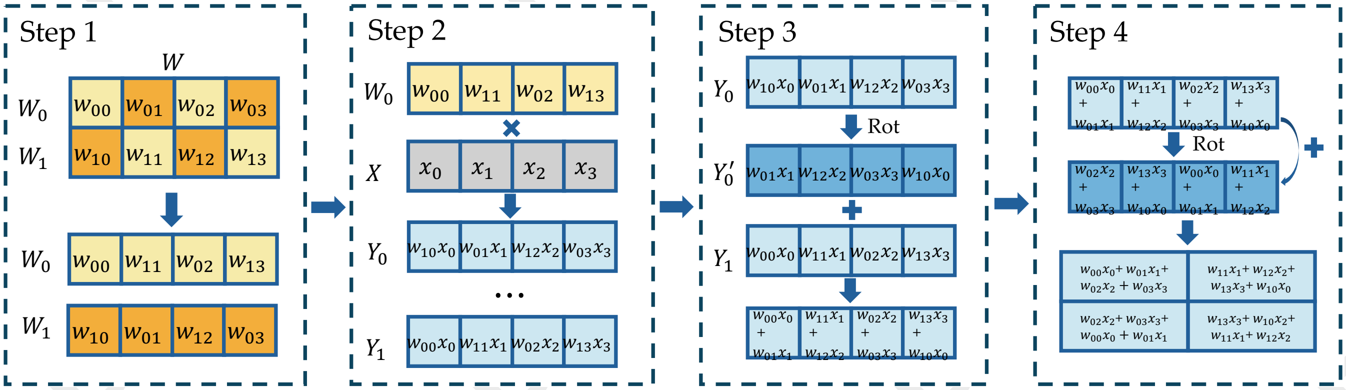
Step 1
W
W_0
w_{00} w_{01} w_{02} w_{03}
W_1
w_{10} w_{11} w_{12} w_{13}
W_0
w_{00} w_{11} w_{02} w_{13}
W_1
w_{10} w_{01} w_{12} w_{03}
Step 2
W_0
w_{00} w_{11} w_{02} w_{13}
×
X
x_0 x_1 x_2 x_3
Y_0
$w_{10}x_0$ $w_{01}x_1$ $w_{12}x_2$ $w_{03}x_3$
…
Y_1
$w_{00}x_0$ $w_{11}x_1$ $w_{02}x_2$ $w_{13}x_3$
Step 3
Y_0
$w_{10}x_0$ $w_{01}x_1$ $w_{12}x_2$ $w_{03}x_3$
Rot
Y_0'
$w_{01}x_1$ $w_{12}x_2$ $w_{03}x_3$ $w_{10}x_0$
+
Y_1
$w_{00}x_0$ $w_{11}x_1$ $w_{02}x_2$ $w_{13}x_3$
$w_{00}x_0 + w_{01}x_1$ $w_{11}x_1 + w_{12}x_2$ $w_{02}x_2 + w_{03}x_3$ $w_{13}x_3 + w_{10}x_0$
Step 4
$w_{00}x_0 + w_{01}x_1$ $w_{11}x_1 + w_{12}x_2$ $w_{02}x_2 + w_{03}x_3$ $w_{13}x_3 + w_{10}x_0$
Rot
+
$w_{02}x_2 + w_{03}x_3$ $w_{13}x_3 + w_{10}x_0$ $w_{00}x_0 + w_{01}x_1$ $w_{11}x_1 + w_{12}x_2$
$w_{00}x_0+ w_{01}x_1+ w_{02}x_2 + w_{03}x_3$
$w_{11}x_1+ w_{12}x_2+ w_{13}x_3+ w_{10}x_0$
$w_{02}x_2+ w_{03}x_3+ w_{00}x_0 + w_{01}x_1$
$w_{13}x_3+ w_{10}x_2+ w_{11}x_1+ w_{12}x_2$

Figure 2: Illustration of encrypted matrix processing with structured computation and rotation operations.

	`Rot`	`SCMult`	`Add`
Naive	$f_o log_2 f_i$	f_o	$f_o log_2 f_i$
Diagonal	$f_i - 1$	f_i	1
Ours (Input)	$\frac{2f_o f_i'}{N} - 1 + log_2 \frac{N}{2f_o}$	$\frac{2f_o f_i'}{N}$	$\frac{2f_o f_i'}{N} - 1 + \log_2 \frac{N}{2m}$
Ours (Output)	$\frac{2f_o' f_i}{N} - 1 + \log_2 \frac{N}{2f_o'}$	$\frac{2f_o' f_i}{N}$	$\frac{2f_o' f_i}{N} - 1 + \log_2 \frac{N}{2f_o'}$

Table 1: Complexity comparison of matrix processing

Engine that minimizes the most time-consuming `Rot` operations. To adapt to arbitrary input and output, we consider the matrix-vector computation for both the input and output stages. Furthermore, as N grows sufficiently large, the inequality $\frac{2mn'}{N} - 1 + \log_2 \frac{N}{2m} < n - 1$ is more likely to hold, ensuring that HARMONY achieves significantly lower computational costs compared to baseline methods.

5 Experiment

This section evaluates the effectiveness of HARMONY by addressing four key research questions:

- **RQ1**: How accurately and cost-effectively does HARMONY recognize various activities? (Sec. 5.2)
- **RQ2**: How efficient are the HE operations within HARMONY? (Sec. 5.3)
- **RQ3**: How do cryptographic parameter settings affect the performance of HARMONY? (Sec. 5.4)
- **R4**: How effective is HARMONY when deployed in real-world scenarios? (Sec. 5.5)

5.1 Evaluation Settings

Datasets

We evaluated HARMONY using 5 datasets collected from different sensor modalities (UWB, IMU, Depth Camera, WiFi, and Camera). These datasets cover diverse human activity recognition tasks, including fall detection [Ouyang *et al.*, 2021; Yang *et al.*, 2023a]. Below is a brief description:

- **UWB** signals was collected in the parking lot, corridor, and room using two Decawave DWM1000 UWB nodes placed 3 meters apart and sampled at 5 Hz. Eight participants contributed 663 data records, with scenarios both involving and not involving a person walking between the nodes.
- **IMU** captured three walking-related activities (corridor walking, upstairs, downstairs) in two buildings. Seven participants contributed to 1,369 data records. Each frame, sampled at 50,Hz, includes 9-axis data. Using a 2-second window, every recording is a 900-dimensional vector.
- **Depth Camera** recorded five hand gestures (good, ok, victory, stop, fist) using a PicoZense DCAM710 depth-sensing camera under outdoor, dark, and indoor conditions. Nine participants contributed to 7,422 data records.
- **WiFi (CSI)** detected six human gestures (box, circle, clean, fall, run, walk) using the Atheros CSI tool. Twenty participants contributed 1,200 data records.
- **Camera (RGB)** captured three human activities (stand, walk, fall) from 20 participants, totaling 1,000 data records.

Metrics

We focus on designing a unified system capable of simultaneously processing signals from different sensor modalities, prioritizing the efficiency of privacy-preserving computation over task-specific performance. To evaluate the performance of HARMONY, we use two metrics: **Accuracy** and **Latency (ms)**. Accuracy measures the proportion of correct predictions, providing a general assessment of HARMONY's effectiveness in performing the intended tasks. Latency quantifies the time required to complete specific computational processes, reflecting HARMONY's efficiency in privacy-preserving operations.

Baseline Methods

To evaluate HARMONY, we compare it with two categories of baselines: federated learning-based remote monitoring systems, and HE-based privacy-preserving computation methods, which ensure data privacy through encryption.

- **ClusterFL** [Ouyang *et al.*, 2021] is designed to monitor different activities with different sensors, addressing privacy concerns through distributed learning techniques.
- **Plain-text** computes directly on signals, achieving high efficiency but remaining vulnerable to attacks.
- **Naive** [Brutzkus *et al.*, 2019] computes on ciphertexts by applying basic additions and multiplication relying on frequent bootstrapping to refresh ciphertexts.

	UWB	**IMU**	**Depth**	**CSI**	**RGB**
ClusterFL	89.1%	90.5%	71.8%	99.7%	90.0%
Plain-text	94.7%	97.6%	95.9%	98.8%	90.0%
Naive	94.6%	97.6%	95.8%	98.8%	90.0%
Diagonal	94.6%	97.6%	95.8%	98.7%	90.0%
HARMONY	94.7%	97.6%	95.9%	98.8%	90.0%

Table 2: Performance comparison of accuracy across different datasets. The best results are underlined.

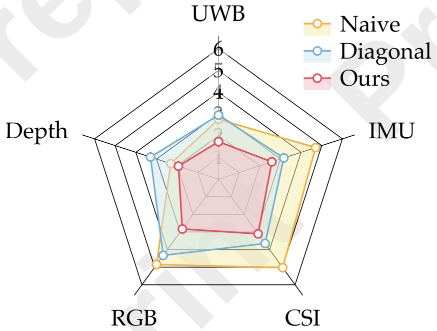

Figure 3: Latency comparison across different methods (log_{10}-scaled). Lower values indicate higher computational efficiency.

- **Diagonal** [Zhang *et al.*, 2021] limits computations to the diagonal, which reduces interactions and simplifies the complexity of homomorphic operations.

Implementation Details

We configured HARMONY with the following parameters: the model was set with 5 tasks (T) and 5 experts (K), using a learning rate of 5×10^{-5} optimized by Adam. For FHE, the operations were implemented using the SEAL library, with cryptographic parameters configured to a multiplicative depth of 10, a scaling factor bit of 40, an HE slot number of 4096, and a security level of 128 bits. The hardware setup included an Intel i7-7700 CPU (32 GB memory) with an NVIDIA Tesla 4096 GPU for server-side computations and a Jetson Orin for edge-side operations.

5.2 Overall Performance

Table 2 compares the recognition accuracy of HARMONY with four baseline methods across all datasets. Specifically, we evaluate HARMONY against *ClusterFL* [Ouyang *et al.*, 2021], a federated learning-based system for remote monitoring that ensures privacy by leveraging edge computing. Although *ClusterFL* effectively addresses privacy, its performance is slightly lower than HARMONY due to the separate processing of each task. We also benchmark HARMONY against *Plain-text* (no encryption), *Naive*[Brutzkus *et al.*, 2019], and *Diagonal* [Zhang *et al.*, 2021], which are two homomorphic encryption (HE)-based techniques, using the same architecture. The results show that HARMONY achieves competitive accuracy, with values of 95.9% for Depth, 94.7% for UWB, 97.6% for IMU, 98.8% for CSI, and 90.0% for RGB. Notably, the small error (0.01) reflects the precision loss due to HE. In terms of latency, as shown in Figure 3, *Plain-text* incurs minimal latency (ranging from 0.3 ms for UWB to 7.5 ms for CSI), so it is excluded from the plot. HARMONY demonstrates significant reductions in computational time, achieving speedups of 3.5× to 130× for more computation-heavy tasks. The variation in improvements is due to different data complexity and dimensionality for each signal type. These results underline HARMONY superior efficiency without sacrificing accuracy.

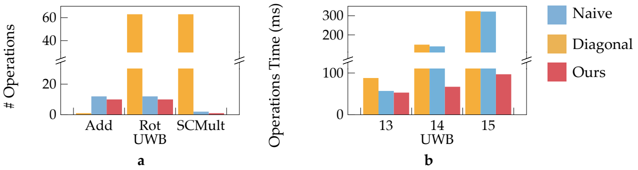

Figure 4: Efficiency analysis for UWB modalities.

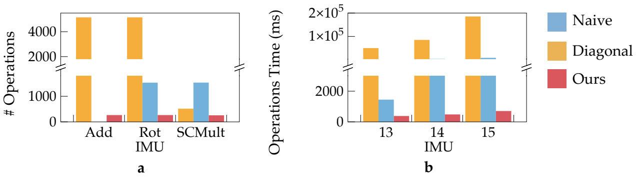

Figure 5: Efficiency analysis for IMU modalities.

5.3 Effectiveness of Encrypted-processing Engine

To evaluate the effectiveness of Encrypted-processing Engine, the core component of HARMONY, we compare the number of operations—Addition (Add), Rotation (Rot), and Scalar Multiplication (SCMulti)-with *Naive* and *Diagonal* across five sensing modalities. The number of operations directly impacts efficiency, with rotation being particularly resource-intensive due to its high complexity and significant effect on performance. To evaluate efficiency, we compare the number of operations for two methods (*Naive* and *Diagonal*) across five sensing modalities, as shown in from Figures 5 (a) to 9(a). The results demonstrate that while *Diagonal* reduces operations compared to *Naive*, HARMONY achieves a substantial reduction in operations compared to *Diagonal*, demonstrating its effectiveness in optimizing the most resource-intensive tasks while maintaining strong overall performance. Specifically, the number of Rot operations is reduced by 1.2× to 7.7× compared to *Naive*. Although our method involves slightly more Add operations than the *Diagonal* method, the overall impact on computational cost is negligible due to the relatively low complexity of addition operations.

5.4 Sensitivity Analysis

To thoroughly evaluate the performance and applicability of HARMONY under different encryption parameters, we analyze the impact of different HE slot numbers (N) (from 2^{13} to 2^{15}) on three methods (Naive, Diagonal, and HARMONY) across five sensing modalities. The analysis considers two key aspects: communication overhead and execution time. Using $N < 2^{13}$ is not feasible as it fails to meet modern cryptographic security standards. For communication over-

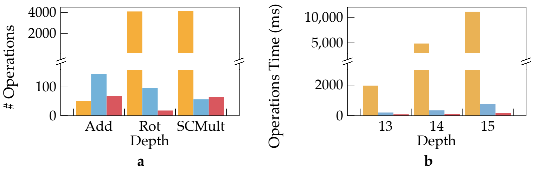

Figure 6: Efficiency analysis for Depth modalities.

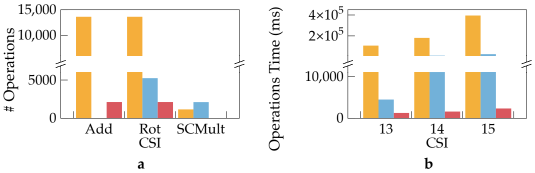

Figure 7: Efficiency analysis for CSI modalities.

head, it increases significantly as N grows, ranging from 0.75 × 1024 to 15 × 1024 and 3 × 1024 kilobytes. Regarding execution time, from Figures 5 (b) to 9(b) shows that the *Naive* and *Diagonal* methods experience dramatic increases with larger N, while our method maintains relatively stable performance, demonstrating the efficiency of the optimizations introduced in *Encrypted-processing Engine*. Considering that our scenario does not demand exceptionally stringent security requirements and prioritizes real-time response and low latency, we select $N = 2^{13}$. This choice ensures sufficient capacity for encapsulating encrypted information while maintaining high efficiency.

5.5 Real-world Deployment

To evaluate the feasibility of HARMONY in real-world scenarios, we conducted deployment tests in both indoor and outdoor environments. The evaluation metrics included accuracy and latency across different processes, as summarized in Table 3. For behavior recognition, we deployed the system in an apartment and tested it with 5 participants using WiFi, IMU, and UWB devices. HARMONY achieved an accuracy of 90.0% with a latency of 1.9 seconds. For gesture recognition, the system was deployed in both indoor and outdoor settings and evaluated with another 5 participants. HARMONY achieved an accuracy of 91.0% with a latency of 1.8 seconds. These results demonstrate that HARMONY performs behavior and gesture recognition effectively, achieving high accuracy and low latency across diverse real-world conditions.

	Encrypt.	Infer.	Decrypt.	Delay	Accuracy
Behavior	958 ms	678 ms	6 ms	1.9 s	90.0%
Gesture	778 ms	87 ms	6 ms	1.8 s	91.0%

Table 3: Latency breakdown (encryption, inference, decryption, and overall delay) and accuracy for two tasks—fall detection and gesture recognition—during real-world deployment.

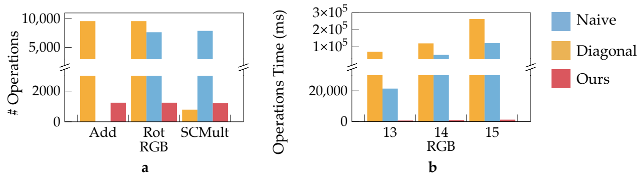

Figure 8: Efficiency analysis for RGB modalities.

6 Discussion

6.1 Insights and Lessons Learned

Through the design, implementation, and real-world deployment of HARMONY, we derived the following key insights:

- **Universal semantic representations empower cross-modal recognition.** Real-world environments often involve dynamically changing sensor availability, making adaptability a critical requirement. Leveraging fundamental motion dynamics—velocity, displacement, and trajectory—as universal semantic representations, HARMONY enables a sensor-agnostic system capable of cross-modal recognition. As demonstrated in Table 2, HARMONY outperform better than *ClusterFL*, which process each modality independently.
- **Structured computation and data reuse optimize encrypted data processing.** Performing computations on encrypted data is typically resource-intensive due to high computational overhead. To address it, we designed a structured computation strategy that combines matrix or vector partitioning with rotation-based data reuse. As illustrated in from Figures 5 to 9, this approach reduces redundant operations and minimizes computational complexity, resulting in speedups ranging from 3.5× to 130×.

6.2 Limitations and Future Work

HARMONY is a privacy-preserving and sensor-agnostic telemonitoring system designed to support diverse recognition tasks in dynamically changing environments, positioning it as a prototype for next-generation remote healthcare systems. However, the current focus is primarily on general monitoring and high-level pattern recognition (e.g., gestures or behaviors), lacking the capability to analyze fine-grained physiological signals essential for healthcare applications, such as heart rate variability or respiratory patterns. Expanding HARMONY to incorporate these detailed physiological features represents a promising direction for future development.

7 Conclusion

We present HARMONY, a privacy-preserving, sensor-agnostic system for remote healthcare monitoring. HARMONYfeatures four components that ensure data reliability, encode multimodal signals uniformly, adapt to various tasks, and optimize encrypted data operations. Our experiments demonstrate high accuracy across five sensing modalities while achieving 3.5× to 130× speedups over state-of-the-art privacy-preserving baselines. These results highlight HARMONY's potential to deliver continuous, unobtrusive monitoring—advancing accessible and patient-centered remote healthcare.

Ethical Statement

This study adheres to high ethical standards. All data collection follows the principle of data minimization, ensuring that only behavior-relevant data is processed. Sensitive data is carefully encrypted in *Encrypted-processing Engine*. Users are fully informed about the types of data collected and their rights through a transparent privacy policy, and explicit informed consent is obtained before data collection.

Acknowledgments

The work of this paper is supported by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things(No.2023B1212010007), China NSFC Grant (No.62472366), 111 Center(No.D25008), the Project of DEGP (No.2024GCZX003, 2023KCXTD042), Shenzhen Science and Technology Foundation (ZDSYS20190902092853047).

References

[Bloom and Luca, 2016] David E Bloom and Dara Lee Luca. The global demography of aging: facts, explanations, future. In *Handbook of the economics of population aging*, volume 1, pages 3–56. Elsevier, 2016.

[Brutzkus *et al.*, 2019] Alon Brutzkus, Ran Gilad-Bachrach, and Oren Elisha. Low latency privacy preserving inference. In *International Conference on Machine Learning*, pages 812–821. PMLR, 2019.

[Cesareo *et al.*, 2018] Ambra Cesareo, Ylenia Previtali, Emilia Biffi, and Andrea Aliverti. Assessment of breathing parameters using an inertial measurement unit (imu)-based system. *Sensors*, 19(1):88, 2018.

[Chien *et al.*, 2023] Hao-Jen Chien, Hossein Khalili, Amin Hass, and Nader Sehatbakhsh. Enc2: Privacy-preserving inference for tiny iots via encoding and encryption. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2023.

[Enshaeifar *et al.*, 2020] Shirin Enshaeifar, Payam Barnaghi, Severin Skillman, David Sharp, Ramin Nilforooshan, and Helen Rostill. A digital platform for remote healthcare monitoring. In *Companion Proceedings of the Web Conference 2020*, pages 203–206, 2020.

[Fernandes *et al.*, 2024] Glenn J Fernandes, Jiayi Zheng, Mahdi Pedram, Christopher Romano, Farzad Shahabi, Blaine Rothrock, Thomas Cohen, Helen Zhu, Tanmeet S Butani, Josiah Hester, et al. Habitsense: A privacy-aware, ai-enhanced multimodal wearable platform for mhealth applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–48, 2024.

[Jiang *et al.*, 2024] Siyang Jiang, Xian Shuai, and Guoliang Xing. Artfl: Exploiting data resolution in federated learning for dynamic runtime inference via multi-scale training. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 27–38. IEEE, 2024.

[Juvekar *et al.*, 2018] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. {GAZELLE}: A low latency framework for secure neural network inference. In *27th USENIX security symposium (USENIX security 18)*, pages 1651–1669, 2018.

[Li *et al.*, 2017] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017.

[Li *et al.*, 2023] Huining Li, Xiaoye Qian, Ruokai Ma, Chenhan Xu, Zhengxiong Li, Dongmei Li, Feng Lin, Ming-Chun Huang, and Wenyao Xu. Therapypal: Towards a privacy-preserving companion diagnostic tool based on digital symptomatic phenotyping. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.

[Liu *et al.*, 2016] Lili Liu, Eleni Stroulia, Ioanis Nikolaidis, Antonio Miguel-Cruz, and Adriana Rios Rincon. Smart homes and home health monitoring technologies for older adults: A systematic review. *International journal of medical informatics*, 91:44–59, 2016.

[Ma and Yeo, 2010] Kaixue Ma and Kiat Seng Yeo. New ultra-wide stopband low-pass filter using transformed radial stubs. *IEEE Transactions on Microwave Theory and Techniques*, 59(3):604–611, 2010.

[Ouyang *et al.*, 2021] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 54–66, 2021.

[Ouyang *et al.*, 2024] Xiaomin Ouyang, Xian Shuai, Yang Li, Li Pan, Xifan Zhang, Heming Fu, Sitong Cheng, Xinyan Wang, Shihua Cao, Jiang Xin, et al. Admarker: A multi-modal federated learning system for monitoring digital biomarkers of alzheimer's disease. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 404–419, 2024.

[Schütz *et al.*, 2022] Narayan Schütz, Samuel EJ Knobel, Angela Botros, Michael Single, Bruno Pais, Valérie Santschi, Daniel Gatica-Perez, Philipp Buluschek, Prabitha Urwyler, Stephan M Gerber, et al. A systems approach towards remote health-monitoring in older adults: Introducing a zero-interaction digital exhaust. *NPJ digital medicine*, 5(1):116, 2022.

[Song *et al.*, 2020] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. Spirosonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[Wang *et al.*, 2014] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. We can hear you with wi-fi! In *Proceedings of the 20th annual interna-*

tional conference on Mobile computing and networking, pages 593–604, 2014.

[Wang *et al.*, 2016] Yuxi Wang, Kaishun Wu, and Lionel M Ni. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing*, 16(2):581–594, 2016.

[Wu *et al.*, 2024] Haiyang Wu, Kaiwei Liu, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. Demo abstract: Caringfm: An interactive in-home healthcare system empowered by large foundation models. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 255–256. IEEE, 2024.

[Xu *et al.*, 2024] Muhao Xu, Zhenfeng Zhu, Youru Li, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. Flexcare: Leveraging cross-task synergy for flexible multimodal healthcare prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3610–3620, 2024.

[Yang *et al.*, 2023a] 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), 2023.

[Yang *et al.*, 2023b] Xuanang Yang, Jing Chen, Kun He, Hao Bai, Cong Wu, and Ruiying Du. Efficient privacy-preserving inference outsourcing for convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 18:4815–4829, 2023.

[Yang *et al.*, 2024] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, 2024.

[Zhang *et al.*, 2019] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *The 25th annual international conference on mobile computing and networking*, pages 1–16, 2019.

[Zhang *et al.*, 2021] Qiao Zhang, Chunsheng Xin, and Hongyi Wu. Gala: Greedy computation for linear algebra in privacy-preserved neural networks. *arXiv preprint arXiv:2105.01827*, 2021.

[Zhang *et al.*, 2022] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

[Zhang *et al.*, 2024] Bin-Bin Zhang, Dongheng Zhang, Yadong Li, Zhi Lu, Jinbo Chen, Haoyu Wang, Fang Zhou, Yu Pu, Yang Hu, Li-Kun Ma, et al. Monitoring long-term cardiac activity with contactless radio frequency signals. *Nature Communications*, 15(1):1–11, 2024.