CSI-Bench: A Large-Scale In-the-Wild Dataset for Multi-task WiFi Sensing

Guozhen Zhu, Yuqian Hu, Weihang Gao, Wei-Hsiang Wang, Beibei Wang, K. J. Ray Liu
Origin Research
{guozhen.zhu,yuqian.hu,weihang.gao,weihsiang.wang,beibei.wang,ray.liu}
@originwirelessai.com

Abstract

WiFi sensing has emerged as a compelling contactless modality for human activity monitoring by capturing fine-grained variations in Channel State Information (CSI). Its ability to operate continuously and non-intrusively while preserving user privacy makes it particularly suitable for health monitoring. However, existing WiFi sensing systems struggle to generalize in real-world settings, largely due to datasets collected in controlled environments with homogeneous hardware and fragmented, session-based recordings that fail to reflect continuous daily activity.

We present CSI-Bench, a large-scale, in-the-wild benchmark dataset collected using commercial WiFi edge devices across 26 diverse indoor environments with 35 real users. Spanning over 461 hours of effective data, CSI-Bench captures realistic signal variability under natural conditions. It includes task-specific datasets for fall detection, breathing monitoring, localization, and motion source recognition, as well as a co-labeled multitask dataset with joint annotations for user identity, activity, and proximity. To support the development of robust and generalizable models, CSI-Bench provides standardized evaluation splits and baseline results for both single-task and multi-task learning. CSI-Bench offers a foundation for scalable, privacy-preserving WiFi sensing systems in health and broader human-centric applications. *Links: CSI-Bench Dataset*; *CSI-Bench Code*; *Project Page*

1 Introduction

Today's smart IoT devices, such as smart speakers, smart bulbs, and various smart display devices, are commonly connected to home routers or mesh network hubs via WiFi. Beyond their primary role in communication, the WiFi signals between these devices inherently capture rich information about the surrounding environment through their propagation paths [25, 44, 26]. This has positioned WiFi sensing as a compelling alternative to vision- or wearable-based systems for human monitoring in smart environments. By capturing fine-grained temporal and spatial variations in Channel State Information (CSI), commodity WiFi devices can infer a wide range of human-centric phenomena—from gross motor events such as falls to subtle physiological signals like breathing. These properties make WiFi sensing especially attractive for health-related applications in smart homes, where privacy, continuous operation, and ease of deployment are critical. Moreover, because these signals are already being transmitted by existing infrastructure, WiFi-based sensing enables non-intrusive, cost-effective, and passive monitoring without requiring additional sensors or user instrumentation.

Despite increasing research interest, existing WiFi sensing studies suffer from a fundamental limitation: a lack of large-scale, diverse, and real-world datasets. Most current datasets are collected in controlled laboratory settings, often using limited types of homogeneous hardware configurations and a narrow range of tasks. As a result, models trained on these datasets struggle to generalize to new users, devices, or environments, limiting their practical utility.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

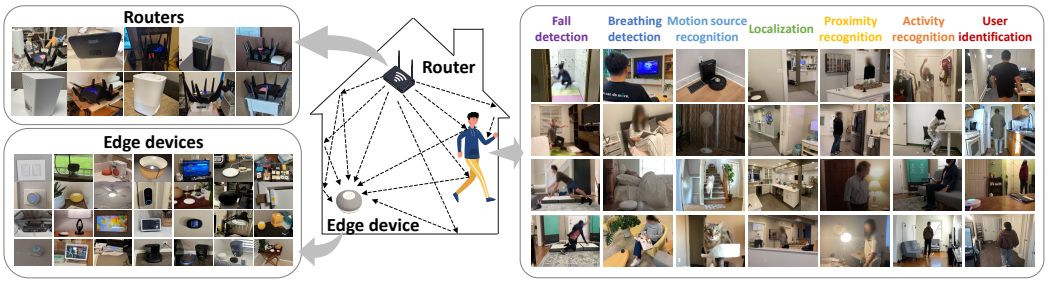


Figure 1: CSI-Bench overview. The benchmark features multiple commercial routers and IoT devices deployed in real homes and offices to collect CSI data. It supports a wide range of human-centric sensing tasks, enabling robust model development across diverse hardware setups and real-world scenarios.

To address these gaps, we introduce CSI-Bench, the first large-scale, in-the-wild benchmark dataset supporting multi-task WiFi sensing as illustrated in Figure 1. Using commercial edge devices, CSI-Bench captures real-world signal variability across diverse environments, including apartments, multi-room houses, offices, and public indoor spaces. Data is recorded continuously from a broad spectrum of WiFi chipsets (Qualcomm, Broadcom, Espressif, MediaTek, and NXP), under both line-of-sight (LoS) and non-line-of-sight (NLoS) conditions, and during natural human activities with minimal intervention.

CSI-Bench advances the field in three key ways:

Large-scale, real-world coverage. The dataset spans over **461** hours of CSI data from **35** users, **26** distinct environments, and **16** device configurations. It reflects realistic deployment conditions with background interference, user mobility, and ambient network traffic.

Multi-task and co-labeled annotations. We provide both single-task specialist datasets (e.g., fall detection, breathing monitoring, localization, and motion source recognition) and a multi-task dataset with joint labels for user identification, activity recognition, and proximity estimation. The co-labeled samples enable efficient multi-task learning and low-latency inference on resource-constrained edge devices.

Standardized benchmarking protocols. We establish strong baselines under supervised learning and multi-task learning. Our findings highlight generalization gaps and the promise of parameter-efficient multi-task learning.

CSI-Bench aims to catalyze robust model development for passive, privacy-preserving WiFi sensing. By offering a unified platform for realistic, diverse, and reproducible evaluation, it provides a foundation for scalable AI applications in smart health, home monitoring, and beyond.

2 Related Work

2.1 WiFi Sensing

Compared to vision-, audio-, or wearable-based systems, WiFi sensing offers a scalable, privacy-preserving, and non-intrusive alternative or complementary solution for continuous monitoring in smart environments and healthcare applications [23, 12, 35, 50]. WiFi sensing has demonstrated substantial potential in tasks such as activity recognition [24, 29], gesture detection [32, 47], indoor tracking and localization [42, 40, 51, 52], fall detection [33], proximity detection [20], and vital sign monitoring [41, 13]. However, most existing studies rely on data collected in constrained settings, which limits generalization to diverse users, hardware platforms, and real-world deployment scenarios.

2.2 WiFi Sensing Dataset

A number of WiFi sensing datasets have contributed valuable resources to the community. Widar3.0 [49] offers large-scale CSI data for gesture recognition using Intel 5300 NICs [15].

Table 1: Comparison of CSI-Bench with published datasets.

Dataset (Year)	Platform	#Edge Device Type	#Samples	#Tasks	#Envs	#Users	In-the-Wild
WiAR [14] (2019)	Intel 5300 NIC	1	4.8k	1	3	10	✗
ARIL[37] (2019)	USRP	1	1.4k	2	1	1	✗
Widar3.0 [49] (2021)	Intel 5300 NIC	1	271.1k	1	3	16	✗
XRF55 [38] (2024)	Intel 5300 NIC	1	42.9k	1	4	39	✗
SignFi [28] (2018)	Intel 5300 NIC	1	14.3k	1	2	5	✗
WiMANs [22] (2024)	Intel 5300 NIC	1	11.3k	3	3	5	✗
CSIDA [19] (2021)	Intel 5300 NIC	1	3k	1	2	5	✗
MM-Fi [46] (2023)	Atheros CSI Tool	1	1.1k	1	4	40	✗
CSI-Bench	Broadcom Qualcomm, MediaTek Espressif, NXP	16	231.6k	7	26	35	✓

SignFi [28] focuses on sign language recognition, capturing fine-grained hand gestures. MM-Fi [46] enables cross-modal analysis by combining WiFi CSI with synchronized video and depth data. XRF55 [38] introduces a large corpus of RF-based activity data for action recognition. Additional datasets such as ARIL [37], CSIDA [19] support tasks like activity recognition and localization.

While these datasets have advanced the field, they share several limitations as illustrated in Table 1. First, most are confined to controlled laboratory settings, offering limited variability in user behavior, device types, and environmental complexity. Second, they primarily support single-task scenarios, lacking the multi-task supervision needed for training general-purpose models. Third, nearly all rely on the Intel 5300 chipset, which does not support continuous CSI recording. As a result, data is collected in fragmented, pre-scripted sessions using manual triggers, which limits dataset scale and fails to capture users' natural daily activities. There remains a growing demand for a unified benchmark that reflects the complexity of real-world deployments, supports multiple sensing tasks, and enables evaluation across diverse users, environments, and hardware platforms. To address this need, we introduce CSI-Bench, a large-scale in-the-wild benchmark for passive WiFi sensing.

3 Dataset Collection

3.1 Overview

To support robust and generalizable WiFi sensing research, we build a diverse collection of datasets captured in real-world environments using commercial WiFi devices. CSI-Bench spans over **461 hours** of CSI recordings across **35 unique users**, **26 environments**, and **16 device types**, covering both routers and edge devices operating under varied network conditions. Data is collected in homes, offices, and public indoor areas with minimal control over ambient interference or user behavior. Each dataset is designed to support one or more sensing tasks, including fall detection (Fall), breathing monitoring (Breath), localization (Loc.), human activity recognition (HAR), user identification (UID), and proximity estimation (Prox.). Representative CSI samples illustrating task-specific signal patterns are visualized in Figure 2. The following section details the hardware, environments, and collection protocols used to capture the datasets.

3.2 Devices and Hardware Setup

Hardware. To emulate the heterogeneity of real-world WiFi sensing deployments, we employ a diverse set of commercial WiFi routers and edge IoT devices commonly found in residential and commercial environments. The selected devices span five major chipset vendors, including Qualcomm, MediaTek, Broadcom, Espressif, and NXP [7, 6, 2, 3], and cover a broad spectrum of hardware configurations, including 1×1 to 2×2 MIMO and 1×4 antenna setups. All devices support IEEE 802.11 n/ac/ax standards, operating across both 2.4 GHz and 5 GHz bands with channel bandwidths of 20, 40, and 80 MHz. These heterogeneous devices are intentionally chosen to reflect the real IoT ecosystem deployed in homes, offices, and small businesses, where heterogeneous devices with varying wireless capabilities coexist in complex indoor environments. Their detailed specifications, including chipset model, antenna configuration, bandwidth, frequency band, and empirically measured average RSSI, are listed in Appendix A (Table 7).

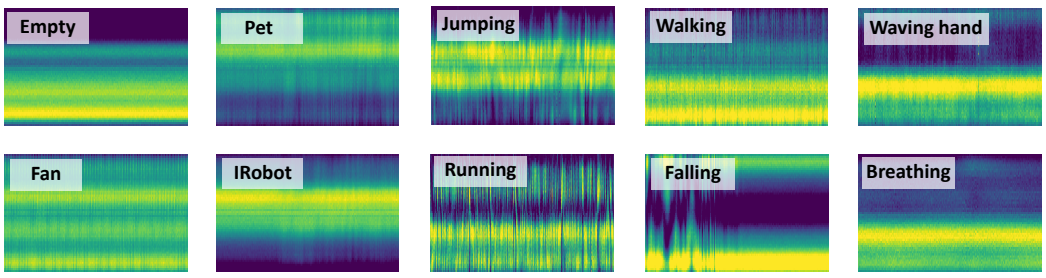


Figure 2: Representative CSI samples are shown for various scenarios, including human actions (jumping, running, walking, hand waving, falling, breathing), non-human motions (pet movement, iRobot, fan), and empty environments. In each sample, the x-axis represents time, and the y-axis represents the subcarrier index.

CSI extraction and synchronization. In our system, IoT client devices periodically transmit CSI packets to routers at two sounding rates: 100 Hz for general sensing tasks and 30 Hz for breathing detection, accommodating different temporal dynamics. Given the distributed nature of these devices, propagation delays and clock drifts cause misalignment in CSI data streams. To address this, the router coordinates data collection by sending batch requests with defined time windows, asking devices to record and upload CSI within the same interval. Each device uses its own system clock to timestamp the data, which allows us to later align the streams in software. Routers handle CSI extraction, buffering, and data upload to cloud servers, running either Linux or FreeRTOS depending on their chipset.

CSI format. Due to hardware diversity, the CSI data in CSI-Bench varies in subcarrier granularity, antenna configurations, and supported bandwidths across different chipset architectures. For example, the NXP 88W8997 provides a 2×2 MIMO configuration with 58 subcarriers at 40 MHz on 5 GHz, while the ESP32-S3, with a 1×1 setup, captures 64 subcarriers at 20 MHz on 2.4 GHz. Qualcomm IPQ4019/IPQ4018 devices offer a 1×2 MIMO configuration, supporting 128 subcarriers at 40 MHz and 256 subcarriers at 80 MHz on 5 GHz. In contrast, the Broadcom BCM4345 employs a 1×4 antenna configuration, providing only 14/28 subcarriers at 20/40 MHz due to proprietary subcarrier grouping. These variations ensure CSI-Bench captures a wide spectrum of signal characteristics, enabling comprehensive evaluation of model generalization across heterogeneous hardware platforms.

3.3 Continuous Data Recording

To overcome the limitations of prior works that typically rely on controlled environments or predefined protocols, we develop an integrated pipeline enabling scalable, in-the-wild CSI data collection across diverse residential settings. Leveraging commercial routers with developer-accessible CSI extraction, cloud infrastructure, and user-friendly annotation tools, our system unobtrusively captures large-scale CSI data from everyday WiFi usage without device-side modifications.

We collaborate with multiple router chipset vendors, who provided firmware and drivers with CSI extraction capabilities enabled, along with proprietary CSI capture utilities for CSI extraction. Building on this, we develop our own tools to programmatically capture and manage CSI data. Specifically, we design separate tools for Linux or FreeRTOS [4], each design to send commands from the Linux application layer directly to the WLAN kernel module, enabling continuous collection and buffering of CSI from all registered devices into unified binary files, which are periodically uploaded to cloud storage via AWS S3 APIs [1]. Each file is timestamped using the router's local system time embedded in the filename, ensuring straightforward temporal alignment across deployments. Upload frequency dynamically adjusts based on device count and bandwidth utilization.

We also develop a lightweight user annotation tool integrated into Google Spreadsheet [5], allowing users to optionally log daily activities—such as waking up, sleeping, leaving or returning home, room occupancy, or inactivity—by tapping buttons that record local timestamps. This design minimizes user effort while ensuring accurate temporal alignment between activity logs and CSI data. An illustration of the annotation tool is provided in Appendix A.10. Our system queries and retrieves CSI files matching these events, concatenates the relevant segments, and refines alignment using embedded packet-level timestamps, resulting in precisely labeled CSI data segments. We collect CSI

Table 2: Summary of tasks, dataset statistics, partitions, and evaluation protocols. *ST* = single-task specialist, *MT* = multi-task joint.

Task	#Classes	Dataset	#Samples	#Users	#Envs	#Devices	Split, Setting
Fall Detection	2	ST	6.7k	17	6	2	70/15/15, *easy/med/hard*
Breath Detection	2	ST	100k	3	3	6	70/15/15, *easy/med/hard*
Motion Source Recognition	4	ST	60.9k	35	10	1	70/15/15, *easy/med/hard*
Room-level Localization	6	ST	7.1k	8	6	8	70/15/15, *easy/med/hard*
Proximity Recognition	4	MT	20.3k	6	6	11	70/15/15, *user/env/device*
Human Activity Recognition	5	MT	41.5k	6	6	11	70/15/15, *user/env/device*
User Identification	6	MT	20.3k	6	6	11	70/15/15, *device*

of motion from non-human sources like pets and cleaning robot when users are not home. When possible, time-aligned external information is collected through camera recordings and local sensor logs to annotate non-human motions or highlight environmental changes.

This pipeline enables extensive, accurately labeled CSI data collection reflective of authentic user behaviors and diverse environments, supporting a wide range of large-scale research applications.

3.4 Environments and Contexts

We collect our data across a broad range of environments, including compact apartments, multi-room houses, offices, hallways, and open indoor public spaces, as detailed in Appendix A.2. These settings introduce diverse physical characteristics, including complex layouts, clutter, variable wall materials, and occlusions, that significantly affect signal propagation.

Unlike prior datasets collected under controlled conditions, our data captures CSI under authentic, in-the-wild conditions. Devices were positioned freely by users, and data was recorded continuously during natural daily activities. Consequently, the CSI reflects realistic variability introduced by NLoS links, neighboring motion, background activity from appliances, WiFi traffic, and environmental factors such as wind and even rain drops. This level of interference is critical for benchmarking the robustness of WiFi sensing models, particularly for healthcare applications where reliable and through-the-wall monitoring in uncontrolled home environments is essential.

3.5 Data Collection Protocols

Although participants are free to move naturally and perform tasks as they would in daily life, we implement basic data collection protocols to ensure consistency and repeatability. Each session begin with a brief calibration phase to verify device connectivity, synchronize timestamps, and confirm stable CSI logging. The recorded activities spans a range of motion patterns, including sitting still, walking, waving hands, and running through hallways. All participants signed a consent form prior to participation, with expenses around $20 /hr. Data from non-human motion sources—such as pets, cleaning robots, and electrical appliances like fans—are collected when users are not present. Detailed task-specific data collection procedures are provided in Appendix A.

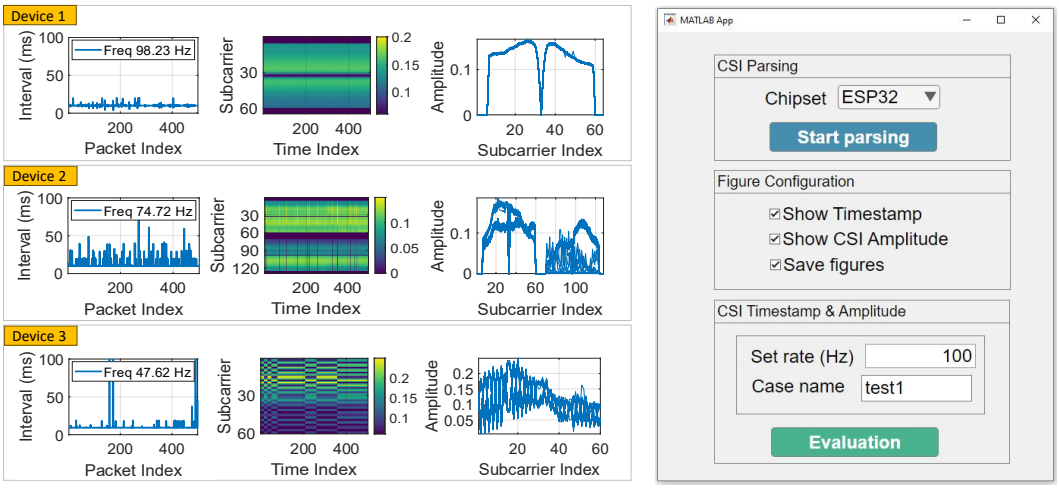
3.6 Dataset Statistics

CSI-Bench spans seven classification tasks with varied sensing objectives. Table 2 summarizes dataset scale and coverage, including the number of samples, recording duration, users, environments, and device types. This diversity reflects real-world deployment conditions and supports robust generalization benchmarking.

4 Data Quality and Preprocessing

4.1 CSI Quality Verification

Motivation. CSI quality checking is critical for ensuring data reliability, as raw measurements often suffer from signal dropouts, high noise levels, or inconsistent timestamps. These issues can arise due to differences in chipset design, CSI extraction algorithms, hardware configurations (e.g., antenna



(a) (b)

Figure 3: MATLAB-based CSI verification tool. (a) Visualization of CSI quality from three devices, showing variations in sampling interval, time-subcarrier heatmap, and amplitude response. (b) User interface for parsing and evaluating CSI data, supporting timestamp checks, amplitude analysis, and figure export to ensure data reliability in CSI-Bench.

layout, RF circuitry), and deployment conditions. As illustrated in Figure 3a, the CSI quality varies from device to device. Device 1 exhibits the best CSI quality, with consistent temporal patterns and a stable sampling rate near the nominal 30 and 100 Hz. Device 2 shows moderate quality with occasional outliers and a lower sampling rate, while Device 3 suffers from the poorest quality, marked by irregular sampling intervals and temporal clustering of CSI frames. Given the diverse hardware platforms and settings in CSI-Bench, these quality variations must be systematically addressed to enable meaningful benchmarking.

Verification tool. To systematically assess and ensure CSI data quality, we adopt a structured evaluation framework introduced in an existing work [21], which models CSI verification as a multilayered pipeline. Each layer of this pipeline targets a specific aspect of data integrity using customized metrics, covering timestamp consistency, CSI amplitude stability, and other modality-specific characteristics. This design allows us to characterize various perspectives of CSI quality and adapt the evaluation to different sensing tasks. In the context of CSI-Bench, we apply this framework to filter out samples with timestamp irregularities, unstable or flat CSI amplitude, and signal dropout, ensuring that only reliable traces are included in the benchmark. The CSI verification tool is implemented in MATLAB, as shown in Figure 3b, to facilitate systematic quality control before incorporating data into CSI-Bench.

4.2 CSI Preprocessing Pipeline

Amplitude extraction. In real-world measurements, CSI is often corrupted by phase noise caused by timing and frequency synchronization offsets, as well as additive thermal noise. In the literature, two main approaches are used to handle phase distortions: phase cleaning [10, 31, 45] and phase elimination [39, 43, 48]. Phase cleaning aims to correct the distorted phase but cannot fully eliminate initial phase offsets, making it less reliable for consistent processing across diverse devices. Therefore, in our benchmark, we adopt the phase elimination approach. Specifically, if the extracted CSI at time t and subcarrier frequency f is represented as $H(f, t)$, we use the amplitude $|H(f, t)|$ as input, eliminating the unreliable phase component.

Data segmentation. To facilitate supervised model training, the collected CSI data is segmented into fixed-duration, non-overlapping samples. For tasks including Fall Detection, Localization, Motion Source Recognition, and the Multi-Task dataset, we segment CSI data into 5-second intervals. For the Breathing Detection dataset, considering the slower temporal variations inherent to respiration

signals, we segment the CSI data into 10-second intervals. We also provide the unsegmented CSI recordings of part of our dataset to support

Amplitude normalization. To mitigate the effects of varying signal strengths, we normalize each CSI sample across all subcarriers and time indices by removing the mean and scaling by the standard deviation. This ensures consistent scaling across samples while preserving the relative temporal–spectral dynamics within each sample. The normalized CSI is computed as $\hat{H}(f_k, t) = \frac{H(f_k,t)-\mu_H}{\sigma_H+\epsilon}$ where μ_H and σ_H denote the mean and standard deviation of $H(f_k, t)$ over the entire CSI matrix, and ϵ is a small constant added for numerical stability.

Subcarrier standardization. Due to hardware differences, the number of subcarriers in CSI samples can vary across different platforms, leading to inconsistent input shapes along the frequency dimension. To standardize the data, we select a fixed number of subcarriers and apply zero-padding or clipping in the frequency dimension as needed. This ensures all samples have consistent input shapes across the dataset.

5 Benchmark Design

5.1 Task Suite and Metrics

CSI-Bench supports a suite of supervised classification tasks for WiFi sensing, covering key applications in health monitoring and ambient intelligence. Each task operates on a fixed-length CSI tensor $\mathbf{X} \in \mathbb{R}^{C\times K\times T}$, where C is the channel count, K is the standardized subcarrier dimension over antenna arrays, and T is the temporal length of samples (5 seconds for most tasks, and 10 seconds for breathing detection).

Single-task specialized dataset. The benchmark includes four single-task datasets: *Fall Detection* (binary classification of fall vs. non-fall), *Breathing Detection* (binary detection during sleep, sampled at 30 Hz), *Motion Source Recognition* (four-class classification of human, pet, robot, and fan motion), and *Room-Level Localization* (six-way classification of the user location). These are evaluated independently using dedicated datasets.

Multi-task joint dataset. A multi-task dataset contains co-labeled samples for three tasks: *Human Activity Recognition* (five-class classification), *User Identification* (multi-class over 6 users), and *Proximity Recognition* (four-class distance estimation). This enables parameter-efficient multi-task training with a shared backbone and task-specific heads.

All tasks are evaluated using overall accuracy and weighted F1-score. Accuracy provides a global measure of classification correctness, while the weighted F1-score accounts for class imbalance by averaging per-class F1-scores weighted by class frequency. This is especially relevant for tasks with skewed distributions such as fall detection or proximity recognition.

5.2 Evaluation Protocols

CSI-Bench provides standardized train/validation/test splits for all tasks to ensure fair comparison and reproducibility. For each dataset, 70% of samples are used for training, 15% for validation, and the remaining 15% for testing, with class balance and environment distribution preserved. Evaluation protocols and statistics for each task are summarized in Table 2.

To evaluate real-world robustness, each test sample is annotated with a difficulty level—Easy, Medium, or Hard—based on signal quality, environment, and subject complexity. For the multi-task dataset, we define three out-of-distribution (OOD) splits—cross-user, cross-environment, and cross-device—reflecting domain shifts in deployment. These settings enable systematic robustness and generalization evaluation. Full details are provided in Appendix A.

5.3 Baseline Models

To establish reference performance and benchmark learning effectiveness on CSI-Bench, we implement a suite of baseline models across single-task supervised and multi-task learning settings.

Table 3: Performance comparison of supervised models across four core WiFi sensing tasks. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Fall Detection		Breathing Detection		Room-Level Localization		Motion Source Recognition	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MLP [34]	92.16 ±0.91	92.17 ±0.92	97.59 ±0.08	97.59 ±0.08	87.14 ±0.80	86.90 ±0.83	98.86 ±0.07	98.86 ±0.07
ResNet-18 [16]	94.88 ±0.26	94.89 ±0.26	98.58 ±0.17	98.58 ±0.17	100.00 ±0.00	100.00 ±0.00	99.56 ±0.07	99.56 ±0.07
LSTM [17]	94.93 ±0.51	94.92 ±0.50	98.62 ±0.17	98.62 ±0.17	99.12 ±0.27	99.12 ±0.26	98.42 ±0.19	98.42 ±0.19
Transformer [36]	94.28 ±0.72	94.26 ±0.72	98.64 ±0.19	98.64 ±0.19	99.27 ±0.22	99.27 ±0.22	98.61 ±0.27	98.61 ±0.27
ViT [11]	93.58 ±0.71	93.59 ±0.70	98.63 ±0.17	98.63 ±0.17	99.94 ±0.11	99.94 ±0.11	98.74 ±0.10	98.74 ±0.10
PatchTST [30]	94.03 ±0.74	94.03 ±0.73	98.84 ±0.13	98.84 ±0.13	99.91 ±0.10	99.91 ±0.10	98.86 ±0.19	98.86 ±0.19
TimeSformer-1D [8]	93.86 ±1.16	93.87 ±1.13	98.68 ±0.21	98.68 ±0.21	100.00 ±0.00	100.00 ±0.00	98.38 ±0.17	98.39 ±0.17

Table 4: Comparison of task-specific and multi-task training for the Transformer model across shared-data tasks. The improvements (Δ) are reported as mean ± std (%) over three runs.

Task	Task-Specific Training		Multi-Task Joint Training		Improvement	
	Acc	F1	Acc	F1	ΔAcc	ΔF1
Human Activity Recognition	75.40 ±0.93	75.49 ±0.73	88.06 ±0.76	86.00 ±2.05	**+12.66**	**+10.51**
User Identification	99.51 ±0.32	99.51 ±0.32	99.55 ±0.06	99.70 ±0.27	+0.04	+0.19
Proximity Recognition	77.52 ±3.13	77.35 ±3.24	86.41 ±1.97	87.09 ±1.46	**+8.89**	**+9.74**

Supervised learning. We evaluate representative architectures spanning fully connected networks (MLP) [34], recurrent models (LSTM) [17], convolutional backbones (ResNet-18) [16], and transformer-based sequence learners, including Vision Transformer (ViT) [11], PatchTST [30], and TimeSformer-1D [8]. All models are trained independently on each task using the corresponding specialist dataset. Input CSI tensors are amplitude-only with hyperparameters tuned using validation performance.

Multi-task learning. To explore parameter efficiency and cross-task knowledge sharing, we also implement multi-task learning using a shared backbone with lightweight task-specific adapters [9]. We adopt the same backbones as in the supervised setting and attach low-rank (LoRA) adapters [18] and separate classification heads for each task. During training, task-labeled samples are drawn from the joint multi-task dataset, and optimization proceeds with shared backbone updates and task-specific losses.

All models are trained using the AdamW optimizer [27] with a cosine learning rate schedule and early stopping. Detailed architecture configurations and training hyperparameters are provided in Appendix B.

5.4 Baseline Evaluation

We report performance on all tasks using both standard supervised learning baselines. Table 3 summarizes accuracy and weighted F1-score for supervised models trained on the specialist datasets. Among the models, transformer-based architectures—particularly TimeSformer-1D and PatchTST—consistently achieve strong performance, highlighting their effectiveness in capturing temporal dynamics in high-dimensional CSI data. Simpler models such as MLP and LSTM perform adequately on some tasks but show clear limitations in harder cases.

Multi-task learning results are presented in Table 4. Compared to task-specific training, our multi-task models with a shared Transformer backbone and lightweight adapter-based heads achieve improved performance across multiple tasks. These findings highlight the effectiveness of joint training in capturing shared representations while preserving task-specific specialization through adapters. They also suggest that multi-task learning can improve generalization in real-world settings where sensing tasks are naturally co-located and co-labeled.

In addition to strong performance, our multi-task framework significantly reduces model complexity and training cost. By consolidating three single-task Transformers into a single backbone with task-specific adapters, we reduce the total parameter count by over 60%. This compression is achieved without degrading task performance. Moreover, because all tasks are trained jointly in a single pass,

Table 5: Cross-domain performance of Transformer model on three tasks. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Task	Cross-Device		Cross-Env		Cross-User	
	Acc	F1	Acc	F1	Acc	F1
Human Activity Recognition	61.82 ±0.95	57.80 ±0.78	54.92 ±0.98	47.17 ±1.12	54.72 ±0.84	46.67 ±1.00
Human Identification	59.94 ±0.77	59.81 ±0.96	/	/	/	/
Proximity Recognition	30.68 ±3.11	28.76 ±3.51	29.67 ±1.63	27.12 ±1.79	30.26 ±1.94	25.97 ±2.26

the wall-clock training time is reduced by nearly $3\times$ compared to training separate models for each task. These gains in model size and training efficiency make our approach especially suitable for deployment on resource-constrained edge devices, where memory and compute budgets are limited.

We also report task-wise performance stratified by difficulty levels (Easy, Medium, Hard) for the single-task datasets in Appendix C.1. Performance drops on hard samples for tasks like fall detection due to signal degradation, cluttered environments, and hardware diversity, reinforcing the need for deployment-aware evaluation.

5.5 Evaluation on OOD Splits

We evaluate the Transformer-based multi-task model under three OOD conditions: cross-device, cross-environment, and cross-user (Table 5). Compared with the in-distribution results in Table 4, all OOD accuracies and F1-scores drop substantially, indicating that models trained on seen domains fail to generalize effectively to unseen users, environments, or hardware. This degradation reflects WiFi CSI's strong sensitivity to device, environmental, and user variations, revealing a significant generalization gap between in-distribution and OOD domains. The results highlight the need for domain-adaptive, calibration-free learning frameworks to improve real-world robustness in WiFi sensing. More detailed results and analysis can be found in Appendix C.2.

5.6 Discussion and Takeaways

CSI-Bench enables scalable research on high-dimensional CSI-based sensing under real-world conditions. Its large scale, diverse hardware coverage, and co-labeled tasks support the development of unified multi-task models for on-device health monitoring. Multi-task learning yields competitive performance while significantly reducing model size and inference cost, making it well-suited for resource-constrained edge deployment. However, performance drops notably under OOD settings, particularly in cross-device scenarios, exposing persistent generalization challenges. Failure cases often arise from hardware heterogeneity, cluttered environments, or degraded signal quality. Overall, CSI-Bench offers a realistic and comprehensive testbed for developing robust, efficient, and generalizable WiFi sensing systems in unconstrained environments.

6 Limitations

The dataset uses amplitude-only CSI features due to phase instability across platforms. While this is practical, it limits exploration of techniques that exploit calibrated phase or angle-of-arrival information. CSI-Bench is designed around classification tasks. Extensions to regression (e.g., continuous sign estimation) and more temporally structured tasks (e.g., long-term activity tracking) are promising but not yet included. We release all data, tools, and splits to support community-driven extensions and improvements.

7 Conclusion

We introduce CSI-Bench, a large-scale, in-the-wild benchmark dataset designed to advance research in WiFi-based sensing for health and human-centric applications. Collected using commercial WiFi edge devices deployed in real residential settings, CSI-Bench captures natural signal variability across users, devices, and environments—providing a realistic foundation for developing deployable, privacy-preserving WiFi sensing systems. The dataset includes single-task datasets for fall detection,

breathing monitoring, localization, and motion source recognition, as well as a co-labeled multi-task dataset supporting user identification, activity recognition and proximity recognition. This enables the development of multi-task models that support efficient joint inference while allowing rigorous evaluation under both in-distribution and out-of-distribution conditions.

To the best of our knowledge, CSI-Bench is the largest available dataset of its kind and can enable learning pipelines that benefit from high-dimensional CSI signals, diverse commercial edge devices, and real-world data ("in the wild"). Beyond the dataset, CSI-Bench includes a suite of baseline models and training protocols under supervised and multi-task settings. Our results show that multi-task learning can reduce model size and inference cost while maintaining competitive accuracy, making it suitable for health monitoring on resource-constrained devices. At the same time, performance drops under domain shifts highlight the need for future research on adaptive and generalizable sensing models. CSI-Bench provides a comprehensive testbed to support this work and offers a scalable, practical resource for advancing WiFi sensing systems in healthcare and beyond. We release the full dataset and benchmark code to facilitate reproducibility and further innovation in this space.

Broader Impact

CSI-Bench establishes a foundational step toward scalable, privacy-preserving, and contactless healthcare enabled by commodity WiFi infrastructure. By leveraging signals already emitted in everyday environments, it supports continuous health and behavioral monitoring without requiring wearables or cameras, thereby reducing barriers to adoption and ensuring dignity, inclusivity, and accessibility in care.

The benchmark supports multiple healthcare-relevant sensing capabilities: fall detection and breathing monitoring for safety and wellness tracking; activity recognition and localization for rehabilitation and behavioral analysis; user identification and proximity estimation for personalized assistance and social-interaction assessment; and motion source recognition to reduce false alarms in automated home-care systems. Collectively, these tasks lay the groundwork for comprehensive, unobtrusive smart health monitoring in homes, hospitals, and assisted living facilities.

Beyond practical healthcare applications, CSI-Bench contributes to the broader machine learning and sensing communities by providing a large-scale, in-the-wild dataset that captures the true complexity of human environments. Its diversity in users, hardware, and contexts fosters reproducible research and robust model development for real-world deployment. The high-dimensional CSI signals pose unique challenges in time-series modeling, representation learning, and domain generalization, motivating advances in trustworthy and adaptive AI systems that extend beyond healthcare.

All data are collected with consent, anonymized, and ethically managed. CSI-Bench serves both as a foundation for practical healthcare solutions and as a benchmark for high-dimensional, human-centered AI systems.

References

[1] Amazon s3 api reference. `https://docs.aws.amazon.com/AmazonS3/latest/API/Welcome.html`, 2024.

[2] Broadcom inc. `https://www.broadcom.com/`, 2024.

[3] Espressif systems. `https://www.espressif.com/`, 2024.

[4] Freertos: Real-time operating system for microcontrollers. `https://www.freertos.org/`, 2024.

[5] Google sheets. `https://www.google.com/sheets/about/`, 2024.

[6] Nxp semiconductors. `https://www.nxp.com/`, 2024.

[7] Qualcomm technologies inc. `https://www.qualcomm.com/`, 2024.

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824. PMLR, 2021.

[9] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[10] Chen Chen, Yan Chen, Yi Han, Hung-Quoc Lai, and K. J. Ray Liu. Achieving centimeter-accuracy indoor localization on WiFi platforms: A frequency hopping approach. *IEEE Internet of Things Journal*, 4(1):111–121, 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Pengsong Duan, Xianguang Diao, Yangjie Cao, Dalong Zhang, Bo Zhang, and Jinsheng Kong. A comprehensive survey on wi-fi sensing for human identity recognition. *Electronics*, 12(23), 2023.

[13] Qinghua Gao, Jingyu Tong, Jie Wang, Zhouhua Ran, and Miao Pan. Device-free multi-person respiration monitoring using WiFi. *IEEE Transactions on Vehicular Technology*, 69(11):14083–14087, 2020.

[14] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access*, 7:154935–154945, 2019.

[15] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11n traces with channel state information (csi). Technical report, University of Washington, 2011.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[19] Pengli Hu, Chengpei Tang, Kang Yin, and Xie Zhang. Wigr: A practical wi-fi-based gesture recognition system with a lightweight few-shot network. *Applied Sciences*, 11(8):3329, 2021.

[20] Yuqian Hu, Guozhen Zhu, Beibei Wang, and K. J. Ray Liu. Robust proximity detection using on-device gait monitoring. In *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pages 01–06, 2023.

[21] Yuqian Hu, Guozhen Zhu, Wei-Hsiang Wang, Beibei Wang, and K. J. Ray Liu. What you need is a good csi. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1608–1610. ACM, 2024.

[22] Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. *arXiv preprint arXiv:2402.09430*, 2024.

[23] Abdullah Khalili, Abdel-Hamid Soliman, Md Asaduzzaman, and Alison Griffiths. Wi-fi sensing: applications and challenges. *The Journal of Engineering*, 2020:87–97, 2020.

[24] Bing Li, Wei Cui, Wei Wang, Le Zhang, Zhenghua Chen, and Min Wu. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[25] K. J. Ray Liu and Beibei Wang. *Wireless AI: Wireless sensing, positioning, IoT, and communications*. Cambridge University Press, 2019.

[26] K. J. Ray Liu and Beibei Wang. Statistical principles of time reversal [perspectives]. *IEEE Signal Processing Magazine*, 41(1):31–37, 2024.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[28] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. Signfi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1), 2018.

[29] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi. Sharp: Environment and person independent activity recognition with commodity ieee 802.11 access points. *IEEE Transactions on Mobile Computing*, 22(10):6160–6175, 2023.

[30] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, page 6. ACM, 2017.

[32] Sai Deepika Regani, Beibei Wang, Yuqian Hu, and K. J. Ray Liu. Gwrite: Enabling through-the-wall gesture writing recognition using WiFi. *IEEE Internet of Things Journal*, 10(7):5977–5991, 2023.

[33] Sai Deepika Regani, Beibei Wang, Yuqian Hu, Guozhen Zhu, and K. J. Ray Liu. Fallaware: An explainable learning approach to robust fall detection with WiFi. *IEEE Journal of Selected Areas in Sensors*, 2:71–83, 2025.

[34] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[35] Julio C.H. Soto, Iandra Galdino, Egberto Caballero, Vinicius Ferreira, Débora Muchaluat-Saade, and Célio Albuquerque. A survey on vital signs monitoring based on wi-fi csi data. *Computer Communications*, 195:99–110, 2022.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc., 2017.

[37] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. Joint activity recognition and indoor localization with WiFi fingerprints. *IEEE Access*, 7:80058–80068, 2019.

[38] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1), 2024.

[39] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of WiFi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 65–76. ACM, 2015.

[40] Wei-Hsiang Wang, Beibei Wang, Yuqian Hu, Guozhen Zhu, and K. J. Ray Liu. Device-free room-level localization with WiFi utilizing spatial-frequency–time diversity. *IEEE Internet of Things Journal*, 11(21):35689–35698, 2024.

[41] Xuanzhi Wang, Anlan Yu, Kai Niu, Weiyan Shi, Junzhe Wang, Zhiyun Yao, Rahul C. Shah, Hong Lu, and Daqing Zhang. Understanding the diffraction model in static multipath-rich environments for WiFi sensing system design. *IEEE Transactions on Mobile Computing*, 23(11):10393–10410, 2024.

[42] Xuyu Wang, Lingjun Gao, Shiwen Mao, and Santosh Pandey. Csi-based fingerprinting for indoor localization: A deep learning approach. *IEEE Transactions on Vehicular Technology*, 66(1):763–776, 2017.

[43] 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, 2017.

[44] Chenshu Wu, Beibei Wang, Oscar C. Au, and K.J. Ray Liu. Wi-Fi can do more: Toward ubiquitous wireless sensing. *IEEE Communications Standards Magazine*, 6(2):42–49, 2022.

[45] Qinyi Xu, Yan Chen, BeiBei Wang, and K. J. Ray Liu. Radio biometrics: Human recognition through a wall. *IEEE Transactions on Information Forensics and Security*, 12(5):1141–1155, 2017.

[46] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[47] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, 6(6):10763–10772, 2019.

[48] Feng Zhang, Chenshu Wu, Beibei Wang, Hung-Quoc Lai, Yi Han, and K. J. Ray Liu. Widetect: Robust motion detection with a statistical electromagnetic model. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3), 2019.

[49] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8671–8688, 2022.

[50] Guozhen Zhu, Beibei Wang, Weihang Gao, Yuqian Hu, Chenshu Wu, and K. J. Ray Liu. SrcSense: Robust WiFi-based motion source recognition via signal-informed deep learning. *IEEE Journal of Selected Areas in Sensors*, 2:40–53, 2025.

[51] Guozhen Zhu, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. EZMap: Boosting automatic floor plan construction with high-precision robotic tracking. *IEEE Internet of Things Journal*, 10(8):6988–6998, 2023.

[52] Guozhen Zhu, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. Floor plan reconstruction with high-precision rf-based tracking. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5073–5077, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 6.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully discLoSe all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data is provided or not)?

   Answer: [Yes]

   Justification: The experiments are reproducible. The code, dataset and detailed instructions are provided.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please find the links to our code and dataset attached in the abstract. We have detailed instructions on how to use our data and benchmark code included.

   Guidelines:

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training details are discussed in Section B.3.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results in the paper are reported with error bars representing the standard deviation across three random seeds, please see Table 3 and Table 4 in Section 5.3. The primary sources of variability include random initialization, data shuffling, and train/validation/test splits. These are consistently applied to all baseline and proposed models.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computing requirements are discussed in Appendix B.3.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: Data were passively collected via commercial WiFi devices in participants' homes, with informed consent and no direct researcher interaction. An internal ethics review addressed privacy and risk, and all data were anonymized. Participants received fair compensation.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] Please see Section 7.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve the release of models or data that carry a high risk for misuse or dual-use concerns. We do not release any generative models or scraped data from public sources, and the released dataset does not pose foreseeable risks that require additional safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original papers and repositories that produced the code packages and models used in our work. For each asset, we have included the appropriate references in the paper, along with the license information and relevant URLs where applicable. All assets used are open-source and have been used in compliance with their respective licenses

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new large-scale dataset and benchmark code suite for WiFi-based sensing tasks. Please find the links under abstract.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our dataset includes data from human subjects in their homes, see 3.5, with written informed consent and fair compensation provided. Consent procedures and compensation details are summarized in the main paper, with full instructions and sample screenshots available in the supplementary material for transparency.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were discLoSed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Our research involved passive data collection in participants' homes using commercial WiFi devices, with no direct interaction or intervention, see 3.5.

A Dataset Description

This appendix details the composition and collection protocols of CSI-Bench.

A.1 Subjects and Scenarios

CSI-Bench includes CSI data from 35 individual users (U01–U35), comprising 26 males and 9 females aged between 23 and 42 years, with heights ranging from 155 to 185 cm and body weights from 45 to 90 kg. In addition, six two-user sessions (UM01–UM06) are recorded to support multi-person interaction analysis. To further diversify subject types, the dataset includes 20 pets (P01–P20), with body weights from 6 to 40 kg, and a dedicated two-pet scenario (PM01). Finally, four distinct fan-based motion scenes (F01–F04) capture ambient signal patterns caused by oscillating and ceiling fans.

A.2 Environments

Data is collected across 26 distinct real-world environments (E01–E26), including studio apartments, multi-bedroom apartments, townhouses, and multi-floor single-family houses. These environments vary in layout complexity, room geometry, wall materials, and furniture density, introducing rich multipath and occlusion effects. A summary of all environments is provided in Table 8.

A.3 Devices and Hardware Diversity

To ensure broad coverage of real-world IoT infrastructure, we select 16 types of commercial WiFi-enabled edge devices operating across both 2.4 GHz and 5 GHz bands, with bandwidths of 20, 40, and 80 MHz. Devices span major chipset vendors such as Qualcomm, Broadcom, Espressif, and NXP, covering configurations from low-cost smart plugs to high-performance routers and smart speakers. Prior to data collection, each device is evaluated using our in-house CSI verification tool to assess signal consistency, sampling stability, and amplitude dynamics. Devices that pass quality thresholds are used in deployment. Figure 4 presents the CSI quality scores across candidate devices; Table 7 lists their specifications, together with average RSSI values (in dBm) measured over one-hour static indoor sessions for each device. These RSSI readings characterize real-world signal strength and serve as a practical proxy for transmit-power variability across hardware families.

A.4 Task-Specific Dataset Statistics

CSI-Bench supports both single-task specialist datasets and a co-labeled multi-task dataset. Task-wise breakdowns include the number of samples, users, environments, and devices, as detailed in Table 6. Each task is annotated with appropriate labels to support supervised learning, multi-task training, and cross-domain evaluation.

Note on Evaluation Splits. For rigorous benchmarking, CSI-Bench defines task-specific evaluation splits based on difficulty levels (Easy, Medium, Hard) and out-of-distribution (OOD) axes (cross-user, cross-environment, cross-device). These splits are introduced in Sections A.5–A.9 for each task and are used to generate the experimental results reported in Appendix C.

A.5 Fall Detection

The Fall Detection dataset is designed to evaluate human fall recognition in real residential settings using commodity WiFi hardware. Data is collected with synchronized video ground-truth under varied hardware and environmental conditions.

Subjects and Scenarios. The dataset includes 17 participants across 6 indoor environments. Activities include casual walking, sitting, lying down, and falling. Scenarios include both LoS and NLoS layouts, with added noise from ambient sources such as ceiling fans to simulate realistic deployments.

Hardware Setup. WiFi CSI data is primarily collected using NXP88W8997 2×2 802.11ac chipsets operating at 5.18 GHz with a 40 MHz bandwidth. Each transmitter-receiver pair forms 4 spatial links and records 58 subcarriers at a sampling rate of 100 Hz. Additionally, a smaller portion of the data is

Table 6: Summary of tasks and dataset statistics.

Task	Users	Envs	Gender	Age (yrs)	Height (cm)	Weight (kg)
Fall	U06 - U22	E21 - E26	14M / 3F	23 - 42	156 - 182	46 - 90
Breath	U06, U23, U24	E05, E09, E10	1M / 2F	27 - 32	160 - 173	60 - 88
Loc.	U01, U05, U06 UM01 - UM05	E01, E03, E04 E06 - E08	4M / 4F	27 - 41	155 - 175	48 - 90
Prox.	U01 - U06	E01 - E06	2M / 4F	26 - 41	163 - 173	45 - 90
HAR	U01 - U06	E01 - E06	2M / 4F	26 - 41	163 - 173	45 - 90
UID	U01 - U06	E01 - E06	2M / 4F	26 - 41	163 - 173	45 - 90
MSR	P01 - P20, PM01	E11 - E13, E15 - E20	-	-	-	6 - 40
	U03 - U04, U08 - U10, U13, U18, U20, U25 - U27, U29 - U35, UM06	E11 - E20	15M / 3F	23 - 35	155 - 185	50 - 90
	F01 - F04	E11 - E13	-	-	-	-

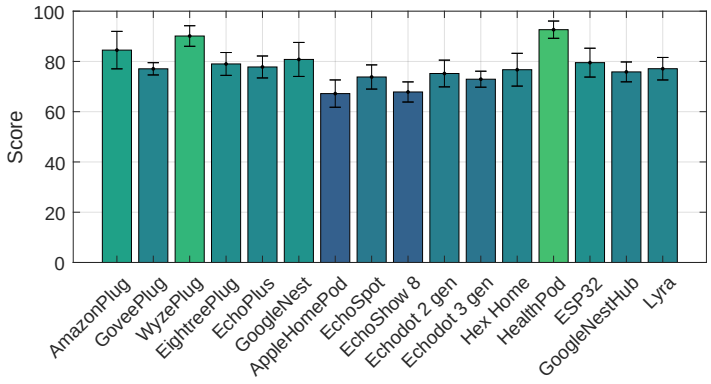


Figure 4: Average CSI quality scores of 16 widely used IoT devices evaluated using our CSI verification tool. Each bar represents the mean score across five measurement trials, with error bars indicating the standard deviation.

collected using ESP32-S3 devices, which operate at 2.4 GHz with a 1×1 antenna setup and capture 64 subcarriers.

Data Collection Protocol. Each session lasts 1–5 minutes, capturing both routine and fall-related activities. Fall events are annotated using synchronized video recordings.

Scale and Composition: 6 environments (homes and offices); 17 participants; 2,770 fall events; 3,930 non-fall activities.

Difficulty-Level Evaluation. Test samples are stratified into *Easy*, *Medium*, and *Hard* tiers based on environmental complexity and device quality. Medium includes fan-induced interference; Hard includes ESP32-based low-quality CSI.

A.6 Breathing Detection

This dataset captures subtle respiration signals under natural sleep conditions using diverse IoT hardware in real homes.

Table 7: Summary of edge devices, WiFi chipsets, and specifications.

Device	Chipset	Model	Antenna	Bandwidth	Band	RSSI
AmazonPlug	MediaTek	MT7697N	1x1	20MHz	2.4G	-49.95 dBm
GoveePlug	Espressif	ESP8266/ESP8285	1x1	20MHz	2.4G	-50.26 dBm
WyzePlug	Espressif	ESP8266/ESP8285	1x1	20MHz	2.4G	-55.57 dBm
EightreePlug	Espressif	ESP8266/ESP8285	1x1	20MHz	2.4G	-54.01 dBm
EchoPlus	MediaTek	MT8516	1x1	20/40/80MHz	2.4G & 5G	-42.24 dBm
GoogleNest	Qualcomm	IPQ4019	1x1	20/40MHz	2.4G & 5G	-49.41 dBm
AppleHomePod	–	–	1x1	20/40MHz	2.4G & 5G	-52.56 dBm
EchoSpot	MediaTek	MT6625L	1x1	20/40MHz	2.4G & 5G	-51.20 dBm
EchoShow 8	MediaTek	MT8183	1x1	20/40/80MHz	2.4G & 5G	-38.14 dBm
Echodot 2 gen	MediaTek	MT6625LN	1x1	20/40MHz	2.4G & 5G	-76.83 dBm
Echodot 3 gen	MediaTek	MT7658CSN	1x1	20/40/80MHz	2.4G & 5G	-57.75 dBm
Hex Home	Qualcomm	–	1x2	20/40MHz	5G	–
HealthPod	NXP	88W889	2x2	20/40/80MHz	5G	–
ESP32	Espressif	S3	1x1	20/40MHz	2.4G	-62.86 dBm
GoogleNestHub	Broadcom	BCM4345	1x1	20/40/80MHz	2.4G & 5G	-60.39 dBm
Lyra	Qualcomm	–	2x2	20/40/80MHz	2.4G & 5G	–

Subjects and Scenarios. Breathing data is collected from 3 participants across 3 residential environments. Deployment setups range from same-room (LoS) to cross-room (NLoS), with and without fan interference.

Hardware Setup. Devices include Amazon Echo Dots, Echo Plus, Google Nest Hub, and Qualcomm-based 5 GHz routers. Sampling is fixed at 30 Hz.

Data Collection Protocol. Overnight sessions are passively recorded during natural sleep without intervention. Participants optionally log activity context.

Scale and Composition: $\sim$55,000 breathing samples; $\sim$45,000 empty-room samples; $\sim$11,400 fan-interfered samples; Diverse device placements and heights (0.47–2.18 m).

Difficulty-Level Evaluation. Difficulty is assigned based on device-user distance, interference level, and deployment complexity. Hard tiers involve distant NLoS setups and overlapping fan motion.

A.7 Room-Level Localization

This dataset supports room-level user localization in typical households with both single- and multi-user presence.

Subjects and Scenarios. Data is collected from 8 users in 6 homes. Three rooms per home are labeled for occupancy. Scenarios include both single and two-user activity.

Hardware Setup. Devices span 8 types (Echo, Google Nest, Apple HomePod, etc.) operating on 2.4/5 GHz at 30 or 100 Hz. Bandwidths vary from 20–80 MHz.

Data Collection Protocol. Users annotate their room presence and co-occupancy manually. Sessions reflect natural daily activities.

Scale and Composition: 3,805 single-user samples; 3,257 multi-user samples; 6 diverse environments; 8 device types.

Difficulty-Level Evaluation. Tiers are defined by user count and hardware quality. Easy cases use high-quality CSI from 5 GHz devices; hard cases include 2.4 GHz plugs and multi-user ambiguity.

A.8 Motion Source Recognition

This dataset captures motion patterns from humans, pets, robots, and fans in diverse indoor settings.

Table 8: Summary of environments (*XBXB* indicates *X bedrooms and X bathrooms*).

Env ID	Type	Area (sqft)	Layout Type	# Floors
E01	Single-family house	2400	Multi-room	3
E02	Apartment	633	Studio	1
E03	Apartment	1077	2B2B	1
E04	Apartment	790	1B1B	1
E05	Apartment	714	1B1B	1
E06	Apartment	1652	3B2B	1
E07	Apartment	1250	2B2B	1
E08	Single-family house	1790	Multi-room	2
E09	Apartment	1200	2B2B	1
E10	Single-family house	1904	Multi-room	2
E11	Single-family house	1352	Multi-room	2
E12	Apartment	830	1B1B	1
E13	Apartment	2242	4B2B	1
E14	Single-family house	1700	Multi-room	2
E15	Single-family house	2000	Multi-room	2
E16	Apartment	960	1B1B	1
E17	Apartment	860	1B1B	1
E18	Single-family house	1680	Multi-room	2
E19	Town house	2600	Multi-room	4
E20	Office	1224	Partitioned rooms	1
E21	Apartment	700	2B1B	1
E22	Single-family house	1300	Multi-room	2
E23	Office	1500	Partitioned rooms	1
E24	Single-family house	1250	Multi-room	2
E25	Single-family house	1400	Multi-room	2
E26	Single-family house	900	Multi-room	3

Subjects and Scenarios. Data include 13 humans (ages 23–34), 11 pets, Roomba robots, and oscillating fans. Activities include walking, sneaking, and simulated intrusion. Environments span homes, townhouses, and offices.

Hardware Setup. CSI is collected via NXP88W8997 2×2 devices at 100 Hz over 58 subcarriers.

Data Collection Protocol. Each session lasts 3–8 minutes. Human data is optionally logged by users; non-human motion is passively captured.

Scale and Composition: $\sim$150K seconds of human motion; $\sim$2,000 minutes of pet activity; $\sim$1,000 minutes of robot activity; $\sim$200 minutes of fan motion.

Difficulty-Level Evaluation. Difficulty is based on motion type, subject diversity, and signal quality. Easy cases include clean human walking or small pets; hard cases include multi-subjects, large pets, or intrusion patterns under NLoS.

A.9 Multi-task Dataset

This dataset enables multi-task learning across activity recognition, user identification, and proximity estimation.

Subjects and Scenarios. Six users perform 5 activities across 6 homes: walking (at 4 distances), running, jumping, seated breathing, and waving. Cross-user and cross-environment samples are included.

Hardware Setup. Each environment uses 5–7 IoT devices across 2.4/5 GHz bands. Devices include Echo, Google Nest, Apple HomePod, ESP32 plugs, and more.

Data Collection Protocol. Each activity lasts 3–6 minutes. Participants use a lightweight UI to annotate activity boundaries and proximity distances.

Scale and Composition: 41,503 total samples; $\sim$5,000–6,000 samples per activity; 4 proximity distances: 0.5, 1.5, 2.5, 3.5 m.

Cross-Domain Evaluation. To evaluate generalization, held-out domains include:

- **Cross-User:** U02
- **Cross-Environment:** E05
- **Cross-Device:** Amazon Plug, Echo Spot

These exclusions are reserved for OOD test sets used in Appendix C.

A.10 Annotation Tool

To facilitate user-friendly and accurate labeling during in-the-wild data collection, we developed a lightweight annotation tool based on Google Spreadsheets for accessibility and cross-platform compatibility.

As illustrated in Figure 5, the tool provides a simple interface where users can log activities through "Start/End" button clicks corresponding to predefined motion types (e.g., walking, breathing, jumping, waving hand, running, localization). Each button click automatically records the timestamp, activity label, tester ID, and session duration, ensuring precise temporal alignment with the collected CSI data.

The design emphasizes ease of use and minimal user burden. Participants simply tap the relevant button when starting and finishing an activity—no manual typing or complex input is required. The captured logs include information such as activity type, approximate user–device distance, location context, and timestamps, which are later aligned with the CSI files through automated scripts described in Section 3.3.

During the collection campaign, participants typically logged around three minutes per activity type per day, covering multiple motion categories. While logging was optional, this structured yet flexible protocol ensured sufficient labeled samples for model training while allowing users to behave naturally in their environments.

B Model Architectures and Training Details

To support rigorous evaluation across in-distribution, cross-domain, and few-shot generalization, we implement and benchmark a suite of neural network models representative of contemporary time-series and vision-inspired architectures. All models are implemented in PyTorch and trained under consistent protocols unless otherwise noted.

B.1 Supervised Learning Architectures

We benchmark the following supervised architectures across all tasks:

Multi-Layer Perceptron (MLP) The MLP model consists of three fully-connected layers with ReLU activations and dropout for regularization. The input to the model is a flattened CSI feature vector, capped at a maximum of 10,000 dimensions to control memory usage. Specifically, the architecture is: [Input $\rightarrow$ Linear(512) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.5) $\rightarrow$ Linear(128) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(Output classes)].

Long Short-Term Memory (LSTM) Our LSTM baseline includes a bidirectional LSTM with two layers, each containing 256 hidden units. A linear classifier follows this, accompanied by dropout for regularization: [Input $\rightarrow$ Bi-LSTM(256, 2 layers, dropout=0.3) $\rightarrow$ Linear(256) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(Output classes)].

Event	Info	Duration (s)	Date	Tester	Start Time	Start Time	End Time	End Time
walking	at 0.5m(0-1m) from the bot	180	4/14/2025	U01	10:43:04	14:43:04	10:46:12	14:46:12
walking	at 0.5m(0-1m) from the bot	180	4/14/2025	U01	10:48:21	14:48:21	10:51:27	14:51:27
walking	1.5m(1-2m) from the bot	180	4/14/2025	U01	10:52:45	14:52:45	10:55:52	14:55:52
walking	1.5m(1-2m) from the bot	180	4/14/2025	U01	10:56:50	14:56:50	10:59:56	14:59:56
walking	2.5m(2-3m) from the bot	180	4/14/2025	U01	11:00:53	15:00:53	11:04:00	15:04:00
walking	2.5m(2-3m) from the bot	180	4/14/2025	U01	11:07:18	15:07:18	11:10:25	15:10:25
walking	3.5m(3-4m) from the bot	180	4/14/2025	U01	11:11:15	15:11:15	11:14:23	15:14:23
walking	3.5m(3-4m) from the bot	180	4/14/2025	U01	11:15:25	15:15:25	11:18:31	15:18:31
breathing	about 1.5m from the bot	180	4/14/2025	U01	11:24:15	15:24:15	11:27:20	15:27:20
breathing	about 1.5m from the bot	180	4/14/2025	U01	11:28:02	15:28:02	11:31:10	15:31:10
jumping	about 1.5m from the bot	180	4/14/2025	U01	11:53:38	15:53:38	11:56:46	15:56:46
jumping	about 1.5m from the bot	180	4/14/2025	U01	11:58:29	15:58:29	12:01:36	16:01:36
wavinghand	about 1.5m from the bot	180	4/14/2025	U01	12:03:37	16:03:37	12:06:43	16:06:43
wavinghand	about 1.5m from the bot	180	4/14/2025	U01	12:06:51	16:06:51	12:09:58	16:09:58
running	about 1.5m from the bot	180	4/14/2025	U01	12:11:42	16:11:42	12:14:47	16:14:47
running	about 1.5m from the bot	180	4/14/2025	U01	12:15:51	16:15:51	12:18:57	16:18:57
localization	Walk in the room near the bot	180	4/14/2025	U01	17:45:07	21:45:07	17:47:50	21:47:50
localization	Walk in the room near the bot	180	4/14/2025	U01	17:48:51	21:48:51	17:51:58	21:51:58
localization	Walk in the room near the bot	180	4/14/2025	U01	17:52:35	21:52:35	17:55:41	21:55:41
localization	Walk in the room near the bot	180	4/14/2025	U01	17:56:00	21:56:00	17:59:04	21:59:04
localization	Walk in the room near the bot	180	4/14/2025	U01	17:59:25	21:59:25	18:02:30	22:02:30
localization	Walk in the room near the bot	180	4/14/2025	U01	18:02:40	22:02:40	18:05:46	22:05:46

Figure 5: Screenshot of the Google Spreadsheet–based annotation tool used by participants to record activities and timestamps during data collection.

ResNet-18 We modify a standard ResNet-18 architecture to accept single-channel input (WiFi CSI data) by adapting the first convolutional layer accordingly. The final fully connected layer is tailored to the task-specific number of classes.

Vision Transformer (ViT) The ViT model converts input CSI data into embedded patches using convolutional patch embeddings, followed by Transformer encoder layers (6 layers, embedding dimension=128, and 4 heads). A class token is prepended for classification tasks. Dropout and layer normalization are employed for stability.

Transformer This architecture employs Transformer encoder layers (4 layers, model dimension=256, 8 attention heads). Inputs are linearly projected into the model dimension, positional encodings are added, and global average pooling is applied before classification. Dropout is set to 0.1 to prevent overfitting.

PatchTST PatchTST utilizes temporal patch embeddings (patch length=16, stride=8) processed through Transformer encoder layers (4 layers, embedding dimension=128, 4 heads). The architecture includes positional encodings, dropout (0.1), and a CLS token or mean pooling strategy for final prediction.

TimeSformer-1D TimeSformer-1D adopts patch embeddings (patch size=4) followed by separate temporal and feature attention within Transformer blocks (4 layers, embedding dimension=128, 8 heads). A class token and positional embeddings are included for classification, with dropout layers added for robustness.

All models use a final linear classifier and are initialized using Xavier uniform initialization unless otherwise specified.

B.2 Multi-Task Learning with Adapters

To enable efficient multi-task learning across diverse WiFi sensing tasks, we implement task-specific adapter modules on top of a shared backbone:

- **LoRA Adapters:** For Transformer backnone model, we apply LoRA to the attention modules. Each task has separate adapter weights (rank=8, α=32, dropout=0.05).
- **Task Adapters:** A residual two-layer bottleneck MLP (down-project, GELU, up-project, followed by LayerNorm) is applied post-backbone for each task.

Table 9: **Fall Detection** performance comparison of supervised models. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Easy		Medium		Hard	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	97.62 ±0.52	97.62 ±0.52	69.12 ±5.63	68.20 ±5.19	67.12 ±2.96	66.05 ±4.03
MLP [34]	94.84 ±0.85	94.84 ±0.85	70.59 ±9.61	70.19 ±9.91	63.70 ±2.37	63.41 ±2.25
PatchTST [30]	97.13 ±0.72	97.13 ±0.72	61.76 ±7.59	56.31 ±12.22	62.67 ±2.82	61.36 ±4.14
ResNet18 [16]	97.27 ±0.32	97.27 ±0.32	77.94 ±5.63	76.96 ±6.46	68.84 ±3.04	68.08 ±3.58
TimeSformer-1D [8]	96.58 ±0.50	96.59 ±0.49	67.65 ±7.59	64.55 ±11.72	65.75 ±9.29	61.19 ±17.16
Transformer [36]	97.08 ±0.54	97.08 ±0.54	69.12 ±5.63	68.10 ±5.91	65.07 ±6.94	63.89 ±7.40
ViT [11]	97.40 ±0.42	97.40 ±0.42	77.94 ±13.04	77.07 ±14.46	65.75 ±3.71	64.06 ±6.66

- **Task-Specific Heads:** Each task has a separate classification head, initialized via Xavier uniform.

During training, we activate one task at a time and update both the shared backbone and the active task's adapter and head.

B.3 Training Protocol

All models are trained with the AdamW optimizer, a batch size of 128, and initial learning rate of $1e-3$. We apply cosine learning rate decay with 5 warm-up epochs and weight decay of $1e-5$. Training lasts up to 100 epochs, with early stopping based on validation loss (patience = 15). We use categorical cross-entropy as the loss function. The hyperparameter are tuned based on models' accuracy on validation dataset. Data is loaded from HDF5 using standardized splits as discussed in Section 5.2 and label mappings. Our experiments utilize NVIDIA GeForce RTX 4090 GPUs and AWS Sagemaker involved training with three random seeds across all datasets. For training in AWS Sagemaker, we use ml.g5.g5.12xlarge, which includes 4 NVIDIA A10G Tensor Core GPUs. The training time for tasks ranges from 0.5 hour to 13 hours.

C Additional Experiments

The results in Appendix C are stratified by the difficulty tiers and OOD evaluation protocols defined in Appendix A.1–A.5. For each task, performance is reported across (i) three difficulty levels (Easy, Medium, Hard) reflecting environmental and signal complexity for single-task datasets (Appendix C.1), and (ii) three out-of-distribution (OOD) axes—cross-user, cross-environment, and cross-device—for multi-task datasets (Appendix C.1). All splits are predefined during data collection and are described per task in Appendix A.

C.1 Evaluation with Difficulty Tiers

Table 9 compares **Fall Detection** performance across three difficulty levels. All models perform well under the Easy setting, with LSTM, PatchTST, ResNet18, and ViT achieving F1-scores above 97%. MLP underperforms due to limited temporal modeling. In the Medium tier, performance drops notably—ResNet18 and ViT remain strong (F1 $\sim$77%), while PatchTST degrades significantly (F1 $\sim$56%). TimeSformer-1D and Transformer show moderate results. In the Hard tier, ResNet18 leads with 68.08% F1, while others degrade further. The larger variance in Medium and Hard tiers is due to smaller dataset sizes, which increase sensitivity to noise and reduce performance stability.

Table 11: **Localization** performance comparison of supervised models. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Easy		Medium		Hard	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	99.72 ±0.32	99.75 ±0.29	100.00 ±0.00	100.00 ±0.00	98.31 ±0.50	98.31 ±0.50
MLP [34]	91.36 ±0.93	92.03 ±0.82	96.11 ±1.31	96.18 ±1.29	80.20 ±1.06	80.03 ±1.19
PatchTST [30]	100.00 ±0.00	100.00 ±0.00	99.90 ±0.19	99.95 ±0.10	99.86 ±0.17	99.86 ±0.18
ResNet18 [16]	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00
TimeSformer-1D [8]	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00
Transformer [36]	99.30 ±0.36	99.40 ±0.24	99.90 ±0.19	99.90 ±0.19	98.95 ±0.66	98.95 ±0.66
ViT [11]	99.79 ±0.42	99.82 ±0.35	99.90 ±0.19	99.90 ±0.19	99.50 ±0.23	99.50 ±0.23

Table 12: **Motion Source Recognition** performance comparison of supervised models. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Easy		Medium		Hard	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	96.65 ±0.96	96.99 ±0.78	98.79 ±0.11	98.80 ±0.11	96.94 ±0.94	96.94 ±0.95
MLP [34]	98.21 ±0.28	98.29 ±0.18	99.13 ±0.11	99.13 ±0.11	98.19 ±0.36	98.19 ±0.36
PatchTST [30]	98.01 ±0.69	98.28 ±0.54	98.59 ±0.36	98.59 ±0.36	97.49 ±0.71	97.49 ±0.72
ResNet18 [16]	99.86 ±0.11	99.86 ±0.11	99.73 ±0.05	99.73 ±0.05	99.48 ±0.32	99.48 ±0.32
TimeSformer-1D [8]	96.56 ±0.64	96.92 ±0.56	98.68 ±0.18	98.69 ±0.18	97.32 ±0.31	97.31 ±0.32
Transformer [36]	98.73 ±0.62	98.80 ±0.55	98.63 ±0.17	98.63 ±0.17	98.08 ±0.55	98.08 ±0.55
ViT [11]	98.38 ±0.87	98.41 ±0.81	99.27 ±0.32	99.27 ±0.32	98.10 ±0.45	98.10 ±0.45

Table 10: **Breathing Detection** performance comparison of supervised models. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Easy		Medium		Hard	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	99.11 ±0.17	99.11 ±0.17	98.61 ±0.13	98.61 ±0.13	98.08 ±0.28	98.08 ±0.28
MLP [34]	98.54 ±0.14	98.54 ±0.14	97.67 ±0.15	97.67 ±0.15	96.46 ±0.13	96.46 ±0.13
PatchTST [30]	99.20 ±0.06	99.20 ±0.06	98.77 ±0.19	98.77 ±0.19	98.49 ±0.22	98.49 ±0.22
ResNet18 [16]	98.94 ±0.17	98.94 ±0.17	98.42 ±0.16	98.42 ±0.16	98.32 ±0.25	98.32 ±0.25
TimeSformer-1D [8]	99.05 ±0.22	99.05 ±0.22	98.29 ±0.31	98.29 ±0.31	98.60 ±0.23	98.60 ±0.23
Transformer [36]	98.23 ±0.24	98.23 ±0.24	97.31 ±0.47	97.31 ±0.47	97.54 ±0.31	97.54 ±0.31
ViT [11]	99.56 ±0.08	99.56 ±0.08	99.41 ±0.08	99.41 ±0.08	99.17 ±0.11	99.17 ±0.11

Table 10 presents breathing detection results, where all models maintain high accuracy and F1-scores (>96%) across tiers. ViT performs best, achieving over 99% F1 consistently. LSTM and PatchTST follow closely, especially in the Easy setting. Even in the Hard tier, model performance drops only slightly. ResNet18 and TimeSformer-1D also generalize well, with minimal performance variance. The results suggest that breathing patterns are relatively easier to model and robust to environmental changes.

Table 11 demonstrates that localization is a highly separable task. Most models—including PatchTST, ResNet18, and TimeSformer-1D—achieve perfect scores in the Easy and Medium tiers and retain

Table 13: **Human Activity Recognition** cross-domain performance. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Cross-Device		Cross-Env		Cross-User	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	60.57 ±2.12	57.04 ±2.32	53.65 ±0.89	46.22 ±0.72	53.33 ±2.11	45.70 ±2.01
MLP [34]	56.33 ±1.23	50.79 ±1.11	52.15 ±0.85	43.45 ±1.40	52.06 ±0.54	42.05 ±0.97
PatchTST [30]	61.61 ±1.81	58.05 ±1.54	56.85 ±0.63	49.55 ±0.47	56.44 ±1.47	49.25 ±1.33
ResNet18 [16]	66.21 ±1.96	63.57 ±1.90	57.98 ±0.87	50.90 ±0.96	59.24 ±1.47	52.07 ±1.53
TimeSformer-1D [8]	60.24 ±1.00	55.70 ±1.20	54.65 ±0.93	46.63 ±0.79	54.95 ±0.84	45.74 ±0.79
Transformer [36]	61.82 ±0.95	57.80 ±0.78	54.92 ±0.98	47.17 ±1.12	54.72 ±0.84	46.67 ±1.00
ViT [11]	66.33 ±1.73	63.65 ±1.69	58.87 ±1.12	51.86 ±1.31	59.00 ±1.36	51.48 ±1.26

Table 14: **Human Identification** cross-domain performance. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Cross-Device	
	Acc	F1
LSTM [17]	59.25 ±1.69	59.32 ±1.72
MLP [34]	57.31 ±1.61	57.15 ±1.45
PatchTST [30]	60.45 ±1.07	60.56 ±1.17
ResNet18 [16]	68.07 ±1.93	68.21 ±1.97
TimeSformer-1D [8]	60.84 ±0.81	61.00 ±0.79
Transformer [36]	59.94 ±0.77	59.81 ±0.96
ViT [11]	69.37 ±1.53	69.55 ±1.61

near-perfect performance in the Hard tier. ViT, Transformer, and LSTM also show strong results (F1 $> 98\%$). MLP consistently underperforms, particularly in the Hard tier (F1: 80.03%), likely due to limited spatial modeling. Overall, most models handle localization with high reliability. These results indicate that CSI-based localization is a highly separable task, and that most temporal or spatially-aware models can solve it with high reliability.

Table 12 shows consistently high motion source recognition performance across all difficulty levels. Most models achieve F1-scores above 96%, with ViT and ResNet18 exceeding 99% even in the Hard setting. MLP, PatchTST, and Transformer also perform well, indicating the task is relatively easy to separate. Performance variance remains low, suggesting stable generalization.

C.2 Evaluation on OOD Splits

Tables 13-15 present the performance of supervised models under three cross-domain generalization settings—Cross-Device, Cross-Environment, and Cross-User—for Human Activity Recognition, Human Identification, and Proximity Recognition, respectively. Across all tasks, ViT consistently achieves the highest performance, with the best F1-scores in most OOD settings.

For **Human Activity Recognition** (Table 13), performance drops significantly under all OOD axes, particularly in the Cross-Environment and Cross-User settings, where even the top-performing models (ViT and ResNet18) show F1-scores below 53%. This highlights the challenge of domain shifts in activity classification.

In **Human Identification** results (Table 14), ViT again leads with a 69.55% F1 under Cross-Device, followed closely by ResNet18, suggesting strong person-specific feature learning.

Table 15: **Proximity Recognition** cross-domain performance. Accuracy (Acc) and F1-score are reported as mean ± std (%) over three runs.

Model	Cross-Device		Cross-Env		Cross-User	
	Acc	F1	Acc	F1	Acc	F1
LSTM [17]	24.89 ±2.97	24.29 ±3.02	28.64 ±1.08	26.76 ±1.02	29.20 ±0.55	23.83 ±1.34
MLP [34]	28.73 ±0.89	27.31 ±0.84	25.76 ±0.77	20.86 ±1.41	26.19 ±0.57	17.32 ±0.74
PatchTST [30]	28.13 ±2.17	26.60 ±1.38	26.42 ±1.88	25.35 ±1.63	28.86 ±0.67	23.15 ±0.76
ResNet18 [16]	31.19 ±5.81	27.93 ±4.95	30.67 ±3.06	28.01 ±3.78	32.67 ±1.50	27.64 ±2.53
TimeSformer-1D [8]	27.95 ±2.04	25.85 ±2.67	29.73 ±2.99	27.93 ±3.97	31.19 ±0.87	26.98 ±1.48
Transformer [36]	30.68 ±3.11	28.76 ±3.51	29.67 ±1.63	27.12 ±1.79	30.26 ±1.94	25.97 ±2.26
ViT [11]	32.04 ±1.95	30.11 ±2.12	30.83 ±2.36	28.62 ±2.51	31.54 ±1.66	26.94 ±1.77

Lastly, **Proximity Recognition** (Table 15) is the most challenging task, with all models performing poorly across OOD conditions. Even the best-performing ViT model achieves only around 30% F1, and large variances are observed, indicating poor robustness and generalization.

Overall, these results reveal that while certain models like ViT and ResNet18 show relative resilience, significant performance degradation remains under distribution shifts, underscoring the need for more robust domain generalization strategies in CSI-based sensing tasks.