XAI2CSI: Interpreting CSI with eXplainable AI for Human Activity Recognition

Idio Guarino*, Alfredo Nascita†, Domenico Ciuonzo†, Damiano Carra‡, Antonio Pescapé†
*University of Bologna, Italy, †University of Napoli Federico II, Italy, ‡University of Verona, Italy,
idio.guarino@unibo.it, damiano.carra@univr.it, {alfredo.nascita, domenico.ciuonzo, pescape}@unina.it,

***Abstract*—Wi-Fi Channel State Information (CSI) has emerged as a key enabler for device-free Human Activity Recognition (HAR), enabling low-cost, unobtrusive sensing using existing communication infrastructure. However, Deep Learning (DL) models trained on CSI data often struggle to generalize across users, environments, and device setups due to the context sensitivity of wireless propagation. Despite this challenge, limited attention has been devoted to understanding model decisions and generalization failures. This paper introduces XAI2CSI, a framework that leverages eXplainable Artificial Intelligence (XAI) to analyze DL-based CSI sensing systems. XAI2CSI employs SAGE, a model-agnostic explainability method, to quantify temporal, spectral, and spatial CSI contributions to HAR decisions under nominal and cross-context evaluations on IEEE 802.11ax data. Our analysis reveals that the considered DL model exhibits limited robustness to unseen conditions due to an over-reliance on context-specific CSI patterns, causing models to misinterpret the underlying signal dynamics when deployment conditions change. The proposed methodology and findings provide a reference framework to explore alternative solutions and guide the development of robust, transparent Wi-Fi sensing systems.**

I. Introduction

In recent years, Deep Learning (DL) has been increasingly adopted to leverage wireless signals for environmental sensing [1]. A rich source of information is the Channel State Information (CSI) from Wi-Fi signals, which has become a key enabler for device-free Human Activity Recognition (HAR) [2]. By exploiting subtle disturbances induced by human movement on radio-frequency propagation, CSI enables low-cost and non-intrusive sensing using existing commercial-off-the-shelf communication hardware, thus avoiding wearable-device limitations such as limited battery life, user discomfort, and complex deployment. Despite its promise, the *main obstacle* to widespread adoption is *generalizability*: DL models trained on the CSI data may *struggle* when deployed across different users, environments, or device setups due to wireless propagation's high sensitivity to context [3]. Multipath effects, where signals reflect off walls, furniture, and people, create a unique CSI "fingerprint" for each specific setting. Consequently, even minor changes to the physical environment alter this fingerprint, degrading model performance and preventing the development of a universal, "one-size-fits-all" model that can be trained once and deployed anywhere without extensive retraining. Although numerous studies have documented these performance degradations, they largely treat generalization as a performance problem, leaving open a key question: *why do CSI-based HAR models fail under unseen conditions*?

Answering this question is challenging because DL models operate as *black boxes*, making it *unclear* which CSI components drive successful predictions and how their contribution changes across deployment contexts. eXplainable Artificial Intelligence (XAI) provides tools to interpret complex models by identifying the features that most influence their decisions. However, despite the acknowledged challenges in CSI-based sensing, the systematic application of XAI to understand model behavior and generalization remains *largely unexplored*.

This paper bridges this gap through **XAI2CSI**, a framework that applies XAI to analyze CSI-based HAR under nominal and cross-context conditions. Specifically, XAI2CSI leverages SAGE [4], a novel model-agnostic explainability method, to quantify temporal, spectral, and spatial CSI contributions to HAR decisions. Through rigorous evaluations on IEEE 802.11ax data [5], we investigate the models' failures in cross-user scenarios by analyzing feature importance across the spatial, temporal, and spectral dimensions of the CSI input. Our analysis reveals limited robustness to unseen conditions due to an over-reliance on context-specific CSI patterns, causing models to misinterpret the underlying signal dynamics when deployment conditions change. These findings provide actionable insights toward robust, transparent, and interpretable Wi-Fi sensing systems. Our **contributions** are: (*i*) we evaluate DL-based HAR models under standalone and combined CSI representations (amplitude, phase difference, and their combination) using 802.11ax hardware; (*ii*) we perform systematic cross-domain evaluations across users, environments, and receiver node positions to assess model generalization; (*iii*) we apply SAGE to provide a comprehensive XAI analysis, characterizing temporal, spectral, and spatial CSI contributions under distribution shifts.

The paper is organized as follows: Sec. II reviews related works. Sec. III details our XAI methodology. Secs. IV and V present the setup and results, while Sec. VI concludes it.

II. Related Works

Research on leveraging CSI for human sensing has grown rapidly, evolving from coarse metrics to fine-grained subcarrier-level data. By capturing multipath propagation, CSI has enabled tasks ranging from presence detection to HAR and vital signs monitoring [1]. DL has further advanced

Table I: Related CSI-based HAR works and positioning of this work.

Paper	802.11	CSI data	Cross Eval.			XAI
			🏠	👤	🚶	
Yousefi et al. [6]	n	AMP	○	○	○	○
Zou et al. [7]	n	AMP	○	○	○	○
Wang et al. [8]	n	STFT	●	●	●	○
Yang et al. [9]	n	PHD	●	●	○	○
Chen et al. [10]	n	AMP	●	○	○	○
Yang et al. [11]	n	AMP	○	○	○	○
Zhang et al. [12]	n	AMP	●	●	●	○
Yang et al. [13]	n	AMP	○	○	○	○
Yang et al. [14]	n	BV	●	●	○	○
Zhang and Jiao [15]	n	AMP	○	○	○	○
Meneghello et al. [16]	ac	DS	●	●	●	○
Cominelli et al. [5]	ax	DS	●	●	○	○
Chen et al. [17]	n	PHD	●	●	○	○
Hussain et al. [18]	n	AMP	○	○	○	○
This work	ax	AMP, PHD AMP&PHD	●	●	●	●

CSI Data: Amplitude (**AMP**), Phase Difference (**PHD**), Doppler Shift (**DS**), Body Velocity (**BV**), Short-Time Fourier Transform (**STFT**). *Cross-Evaluation* scenarios: Environments (🏠), Users (👤), Receiver Node Location (🚶), present (●), absent (○). *XAI* Analysis: present (●), absent (○).

these capabilities, yet several key challenges remain. Table I summarizes representative HAR works by Wi-Fi version, signal representation, cross-evaluation scenarios, and whether explainability was considered.

A key limitation is the lack of native CSI extraction on commodity devices [19]. Early works relied on specialized tools such as the *Linux 802.11n CSI Tool*, thus constraining research to outdated 802.11n hardware. Although newer frameworks—e.g., *Nexmon* for 802.11ac and *AX-CSI* for 802.11ax—enable richer data collection, many recent works still rely on legacy tools, thereby limiting their relevance to modern Wi-Fi deployment [19]. Only a few studies leverage the wider bandwidths supported by modern Wi-Fi standards [5, 16]. Most HAR approaches rely on CSI amplitude for its simplicity and real-time suitability [1]. Alternatives such as phase differences, Doppler spectrum, short-time Fourier transform, or body velocity estimates capture richer motion information but increase computational cost, require calibration, and may be affected by orientation and multipath distortions [1]. A major challenge across these works is generalization: CSI strongly depends on users, environments, and hardware, often causing severe performance drops when models are tested across different settings [2]. Although this problem is acknowledged, several studies still evaluate their models in the same context in which they were trained [6, 7, 11, 13, 15, 18]. In contrast, other works explicitly assess generalization by testing models on users, environments, or device placements that are not seen during training [5, 8, 9, 10, 12, 14, 16, 17]. To address these challenges, prior works have explored two main strategies. The first involves signal preprocessing to isolate motion-related components and filter out environment-dependent multipath effects [14, 16]. While improving stability, these methods may also discard valuable cues together with noise. The second uses learning-based adaptation strategies, e.g., transfer-learning and few-shot learning [9, 17], which improve cross-domain performance but require additional data collection in each new context, limiting real-world deployability. Recent evaluations further show that, despite these efforts, generalization across unseen domains remains limited [5]. Finally, while DL-based approaches become the dominant paradigm, they typically behave as black boxes, offering little insight into their decision-making process. To the best of our knowledge, no prior works have explored explainability beyond performance metrics. Our work fills this gap by leveraging SAGE to systematically analyze model behavior, thus uncovering the intrinsic mechanisms that govern decision-making and paving the way for more transparent Wi-Fi sensing systems.

Positioning. Our work advances prior research in *three directions*: (*i*) evaluating HAR models under standalone and combined CSI representations (amplitude, phase difference, and their integration via "early" fusion) using modern Wi-Fi 802.11ax hardware; (*ii*) performing systematic cross-domain evaluations across users, environments, and receiver node locations to assess model generalization; and (*iii*) applying SAGE (the core of XAI2CSI) for a comprehensive XAI analysis, providing a global view of model decision mechanisms.

III. Methodology

CSI background. CSI characterizes the signal propagation between transmitter and receiver, encapsulating reflection, scattering, diffraction, and multipath effects. It supports communication optimization by enabling tasks such as beamforming and rate adaptation [1]. In an Orthogonal Frequency Division Multiplexing (OFDM) system such as Wi-Fi, CSI provides a detailed estimate of the wireless channel in the frequency domain. Thanks to the wide bandwidth of Wi-Fi signals and the fine granularity of OFDM subcarriers, this information is valuable for both communication optimization and device-free sensing applications, e.g., indoor localization, and human activity monitoring. Formally, in an OFDM system with $\mathtt{F}$ subcarriers and a single spatial stream, the channel can be modeled in the frequency domain as $\mathbf{r} = \mathbf{h} \otimes \mathbf{s} + \mathbf{n}$, where $\mathbf{r} \in \mathbb{C}^{\mathtt{F}}$ is the received OFDM symbol, $\mathbf{h} \in \mathbb{C}^{\mathtt{F}}$ is the channel frequency response (i.e., the CSI), $\mathbf{s} \in \mathbb{C}^{\mathtt{F}}$ is the transmitted OFDM symbol, $\mathbf{n} \in \mathbb{C}^{\mathtt{F}}$ is the noise vector, and $\otimes$ denotes the *Hadamard* product. The channel response $\mathbf{h}$ captures the amplitude attenuation and phase shift of each subcarrier. The receiver can estimate the channel ($\mathbf{c} \simeq \mathbf{h}$) by computing the ratio between the received $\mathbf{r}$ and transmitted $\mathbf{s}$ signals.

In Multiple-Input Multiple-Output (MIMO) systems with $\mathtt{N_{tx}}$ transmit antennas and $\mathtt{N_{rx}}$ receive antennas, the CSI at time t can be represented as a tensor $\mathbf{H}(t) \in \mathbb{C}^{\mathtt{N_{tx}} \times \mathtt{N_{rx}} \times \mathtt{F}}$, which captures the channel response for every transmit–receive antenna pair across all subcarriers. Specifically, its amplitude and phase components are defined as: $\mathbf{A}(t) = ||\mathbf{H}(t)||$ and $\boldsymbol{\phi}(t) = \angle\mathbf{H}(t)$, respectively, where $\mathbf{A}(t), \boldsymbol{\phi}(t) \in \mathbb{R}^{\mathtt{N_{tx}} \times \mathtt{N_{rx}} \times \mathtt{F}}$.

CSI is collected by capturing Wi-Fi packets, reflecting both static factors (e.g., room layout) and dynamic ones (e.g., human presence and movement). While primarily used in

transceivers for tasks like equalization and beamforming, CSI can also be extracted via dedicated tools. These tools—for each Tx-Rx antenna pair—record the channel estimate across all available subcarriers at a given sampling rate $\mathtt{S_x}$, producing a CSI time series that reflects wireless channel evolution.

DL-based HAR. We use CSI measurements across a $1 \times \mathtt{N_{rx}}$ MIMO setup to detect human activities, since different actions uniquely perturb wireless propagation [7]. We compare *three* input configurations: (*i*) amplitude ($\mathbf{A}$) only, (*ii*) phase difference ($\mathbf{\Delta}\boldsymbol{\phi}$) only, and (*iii*) their combination ($\mathbf{A}\&\mathbf{\Delta}\boldsymbol{\phi}$) via *early-fusion*, where both are concatenated as input features. To mitigate hardware impairments (e.g., *CFO* or *SFO*) [6], phase differences $\Delta\phi_{j-r} = \phi_j - \phi_r$ are computed across receiving antenna pairs relative to a reference ($R_{x,r} : j \neq r$) and over time via complex conjugate products of raw CSI samples, preserving relative phase shifts and signal integrity. The resulting time series is segmented into overlapping *frames* $X = \{\bar{\boldsymbol{X}}_i\}_{i=1}^{N}$, where $\bar{\boldsymbol{X}}_i \in \mathbb{R}^{\mathtt{D}\times\mathtt{F}\times\mathtt{W}}$ captures dynamics over a window $\mathtt{W}$.[1] The spatial dimension $\mathtt{D}$ is $\mathtt{N_{rx}}$ for $\mathbf{A}$, $\mathtt{N_{rx}}$-1 for $\mathbf{\Delta}\boldsymbol{\phi}$, and $2\mathtt{N_{rx}}$-1 for $\mathbf{A}\&\mathbf{\Delta}\boldsymbol{\phi}$. Each frame $\bar{\boldsymbol{X}}_i$ is fed into a DL model $\mathcal{M}_{\boldsymbol{\theta}} : \bar{\boldsymbol{X}}_i \rightarrow [0,1]^{|\mathcal{C}|}$ that outputs activity class probabilities over $\mathcal{C}$. The predicted label is obtained as $\hat{y} = \arg\max_{c\in\mathcal{C}} \mathcal{M}_{\boldsymbol{\theta},c}(\bar{\boldsymbol{X}}_i)$, where $\mathcal{M}_{\boldsymbol{\theta},c}(\bar{\boldsymbol{X}}_i)$ is the probability assigned to class c. Model training optimizes $\boldsymbol{\theta}$ using the *Categorical Cross-Entropy* loss $\mathcal{L}(y,\hat{y}) = -\sum_{j=1}^{|\mathcal{C}|} \mathbf{1}_{[y=j]} \ln\left(\mathcal{M}_{\boldsymbol{\theta},j}(\bar{\boldsymbol{X}}_i)\right)$, where $y \in \mathcal{C}$ is the true label.

Interpreting CSI-based Human Sensing via SAGE. We evaluate the importance of CSI components using *SAGE* [4], a model-agnostic approach that, unlike local methods (e.g., LIME, SHAP), quantifies global loss degradation across entire distributions. SAGE quantifies the (global) contribution of each feature by comparing the model's loss $\mathcal{L}(\cdot,\cdot)$ with and without the feature, thereby measuring its impact on predictive performance. Formally, for a model $\mathcal{M}(\cdot)$ predicting label y from features $\boldsymbol{x}$ (the vectorized frame $\bar{\boldsymbol{X}}_i$ in our case), SAGE evaluates performance on different feature subsets $\boldsymbol{x}_S$ for $S \subseteq \mathcal{D}$, where $\mathcal{D}$ is the set of all feature indices. The predictive power $\mathcal{M}(\cdot)$ derives from a subset $\boldsymbol{x}_S$ is defined by the function $v_{\mathcal{M}} : \mathcal{P}(\mathcal{D}) \rightarrow \mathbb{R}$:

$$v_{\mathcal{M}}(S) = \mathbb{E}[\mathcal{L}(\mathcal{M}_{\varnothing}(\boldsymbol{x}_{\varnothing}), Y)] - \mathbb{E}[\mathcal{L}(\mathcal{M}_S(\boldsymbol{x}_S), Y)] \quad (1)$$

Here, $\mathcal{M}_S(\boldsymbol{x}_S) = \mathbb{E}[\mathcal{M}(\boldsymbol{X}) \mid \boldsymbol{X}_S = \boldsymbol{x}_S]$ is the conditional expectation denoting the model's prediction when only features in S are known, marginalizing over the rest. Conversely, $\mathcal{M}_{\varnothing}(\boldsymbol{X}_{\varnothing}) = \mathbb{E}[\mathcal{M}(X)]$ is the mean prediction across all features. The *SAGE value* $\phi_i(v_{\mathcal{M}})$ is the Shapley value of the predictive power function $v_{\mathcal{M}}$, attributing a contribution to each feature x_i to the overall model predictive performance. SAGE values represent the expected per-instance SHAP values computed w.r.t. the loss $\mathcal{L}$. For a specific sample $(\boldsymbol{x}, y)$, this is expressed as $v_{\mathcal{M}}^{(\boldsymbol{x},y)}(S) = \mathcal{L}(\mathcal{M}_{\varnothing}(\boldsymbol{x}_{\varnothing}), y) - \mathcal{L}(\mathcal{M}_S(\boldsymbol{x}_S), y)$ which quantifies how the prediction loss changes when conditioning on the subset of features S compared to using no features at all. Finally, the *global importance* of each feature is then obtained by averaging these per-instance contributions across the entire dataset $\phi_i(v_{\mathcal{M}}) = \mathbb{E}_{\boldsymbol{X}Y}[\phi_i(v_{\mathcal{M}}^{(\boldsymbol{X},Y)})]$.

[1] With $\mathbf{A}\&\mathbf{\Delta}\boldsymbol{\phi}$, the first amplitude sample is discarded to ensure temporal alignment with the $\mathbf{\Delta}\boldsymbol{\phi}$ series.

IV. Experimental Setup

This section details the experimental setup: the dataset, preprocessing, the DL model, and the evaluation procedure.

Dataset. We use the dataset from [5], collected with the `AX-CSI` tool supporting the IEEE 802.11ax standard. Each trace records a specific user performing a given activity for 80s in a particular environment on a given day. We restrict our analysis to 5 activities (Empty Room, Jumping, Running, Sitting, Walking) covering static, repetitive, and dynamic Doppler profiles. The collection was performed over two consecutive days in three furniture-rich environments (*Lab*, *Office*, and *Hall*) involving three users $U \equiv \{U_k\}_{k=1}^{3}$, enabling evaluation across different settings. CSI was captured in the 5GHz band at $\mathtt{S_x} = 150$Hz via a 1×4 Single-Input Multiple-Output (SIMO) setup ($\{R_{x,j}\}_{j=1}^{4}$) across three receiver nodes $\mathcal{N} \equiv \{N_i\}_{i=1}^{3}$ (≤ 160MHz bandwidth, ≤ 2048 subcarriers/antenna).

Pre-processing. Following [5], we apply a series of trace-level processing steps, before feeding the CSI data into the DL model: (*i*) CSI data is downsampled to 80MHz channels with $\mathtt{F} = 1024$ subcarriers by selecting central frequency bins, exploiting the redundancy in high-resolution CSI and reducing computational load, and non-informative subcarriers are discarded using a predefined filter ($\mathtt{F}$: 1024→994). (*ii*) Values are normalized per receiver antenna by their mean amplitude. (*iii*) For amplitude features, a *Hampel filter* [19] is applied along the time dimension to mitigate outliers. (*iv*) The time series is segmented into overlapping windows of $\mathtt{W} = 50$ packets ($\Delta \approx 333$ms) with a stride of 25 (i.e., 50% overlap). Finally, Min–Max scaling is applied to the receiver antenna, with parameters computed solely from the training set.

Model Considered. We employ a `2D-CNN` consisting of three 2D convolutional layers with kernel sizes 7, $(5,3)$, and $(3,3)$ and strides $(3,1)$, $(2,2)$, and 1, respectively. Each layer has ReLU activation and is followed by a 2×2 max-pooling layer. The output of the convolutional block is flattened and passed through a dense layer with 128 units and ReLU activation, followed by a softmax-activated dense layer for classification with $\mathcal{C}$ units. To regularize training and reduce overfitting, a dropout (rate of 0.3) is applied after both the last pooling and the first dense layer. The frame $\bar{\boldsymbol{X}}_i$ is structured as a $(\mathtt{D}, \mathtt{F}, \mathtt{W})$ tensor, treating $\mathtt{D}$ as the spatial-feature dimension. This arrangement enables the `2D-CNN` to capture joint correlations across antennas, subcarriers, and time, while facilitating the interpretability of individual components. Following [13], we adopt this architecture for its optimal accuracy-complexity trade-off compared to more complex models (e.g., ResNets, ViT), ensuring real-time execution on edge devices.

Evaluation Procedure. To ensure reproducible evaluation while preserving causal relationships, each trace is split along the time axis into training and test with an $80/20\%$ ratio. We

further reversed 20% of the training set for validation. Model training was performed for up to 200 epochs using the SGD optimizer. To compute SAGE values, each frame is flattened ($\mathcal{D} = \{1, \ldots, \mathtt{D} \times \mathtt{F} \times \mathtt{W}\}$) and conditional expectations are approximated using a background of 40 training samples/class with a convergence threshold of 0.05. This preserves the original frame structure, enabling the grouping of features by subcarrier, antenna, or time index for interpretable analysis. We release our code at https://github.com/IdioGuarino/XAI2CSI.

V. Experimental Evaluation

In this section, we first evaluate model performance and generalization under contextual changes. Then, we employ SAGE to investigate how input features influence decisions.

Assessing Model Performance under Context-Shifts. We evaluate our CSI-based sensing under *three* context shifts: (*i*) unseen users (*cross-user*), (*ii*) unseen environments (*cross-env*), and (*iii*) changes in node position (*cross-node*). In each case, we adopt a leave-one-out cross-validation scheme, where a model trained on all but one setting (viz. $\mathcal{M}_X^*$) is tested on the unseen one and compared to a model trained on all settings (viz. $\mathcal{M}_{all}$), serving as an upper bound.

Fig. 1 compares the accuracy of $\mathcal{M}_X^*$ and $\mathcal{M}_{all}$ across the above conditions based on input type: *amplitude* (A), *phase difference* ($\Delta\phi$), and *their* early-fusion *combination* ($A\&\Delta\phi$). *Cross-user* (resp. *cross-env*) results are obtained by focusing on the *Lab* environment (resp. user U_1) while including all nodes. In *cross-node*, two nodes are used for training and the remaining one for testing, using data from user U_1 in the *Lab*, thus emulating an unseen node location.

As expected, $\mathcal{M}_{all}$ models consistently achieve high accuracy (88-98%), with amplitude-based inputs (A and $A\&\Delta\phi$) outperforming the phase-only counterpart ($\Delta\phi$). Conversely, for $\mathcal{M}_X^*$ models, performance degradation heavily depends on both the context shift and the input type. Under *cross-user* conditions, despite the same environment, the accuracy of $\mathcal{M}_X^*$ drops to at most 65% (U_2), with amplitude-based models (A and $A\&\Delta\phi$) yielding the highest robustness. However, this trend inverts under *cross-env* and *cross-node* conditions, where phase-based models ($\Delta\phi$ and $A\&\Delta\phi$) achieve slightly better performance, despite not exceeding 28% in *cross-env* (*Office*). This underscores that amplitude is more sensitive to environmental changes and transmitter-receiver geometry, whereas phase difference offers greater cross-domain stability.

Focusing on *cross-user* settings, Fig. 2 shows the confusion matrices for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_2\}}$ models on the unseen user U_2, highlighting misclassification patterns across the different input types. For $\mathcal{M}_{all}$, *Empty* and *Sitting* activities are consistently recognized, while more dynamic activities such as *Run* and *Walk* are often misclassified as each other, especially when leveraging only phase difference (Fig. 2b). Conversely, $\mathcal{M}_X^*$ performance highly depends on the input type. When considering A or $\Delta\phi$ standalone, *Run* and *Sitting* prove to be the most challenging classes (confused with *Walk* and *Jump*, respectively), with the latter never correctly classified. Notably, with the same inputs, the model frequently misclassifies *Empty*, which is particularly critical as it implies mistaking an absent user for one performing an activity. Interestingly, this misclassification pattern is resolved for *Empty*, which the model with $A\&\Delta\phi$ recognizes with near-perfect accuracy. Conversely, for *Run* and *Sitting*, early fusion only partially mitigates the issue, as their classification remains challenging.

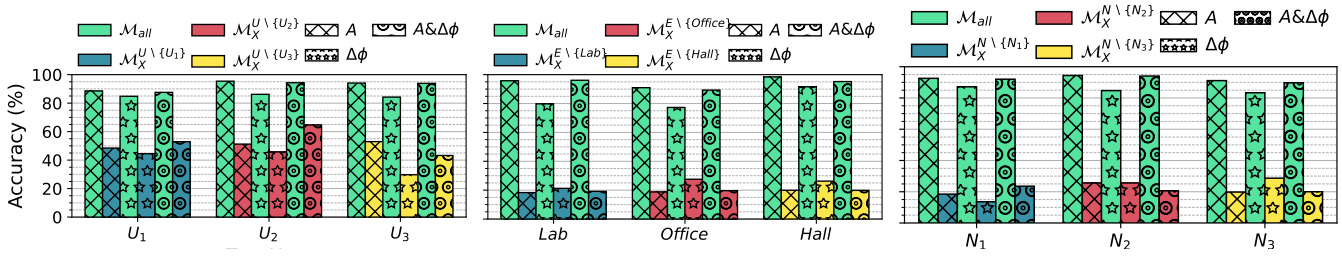
***Take-Away**: Cross-domain generalizability heavily depends on the input representation. While standalone features suffer severe degradation under context shifts, their early fusion only partially mitigates the issue, resolving macroscopic errors (e.g., empty environments) while leaving dynamic and complex static activities an open challenge.*

Analyzing feature effects on models' behavior with SAGE. To gain insight into the behavior of the models and understand the reasons behind their failures in cross-user scenarios, we analyze the feature importance for each target user U_k, considering two subsets: $\mathcal{T}_s$ containing samples correctly classified by both $\mathcal{M}_{all}$ and $\mathcal{M}_X^*$, and $\mathcal{T}_u$ containing samples correctly classified by $\mathcal{M}_{all}$ but misclassified by $\mathcal{M}_X^*$. This separation allows us to distinguish features that consistently support correct predictions from those associated with failures of $\mathcal{M}_X^*$. Then, we analyze how the different dimensions of the input frame, i.e., receiving antennas, packet arrival order, and subcarriers, affect the models' decisions.

Fig. 3 compares the relative importance of the receiving antennas $\{R_{x,j}\}_{j=1}^4$ (normalized by the sum of their absolute values) for the combined-input ($A\&\Delta\phi$) variants of $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_k\}}$. Results are broken down by the target user U_k, isolating the individual contributions of amplitude $\{R_{x,j}^A\}_{j=1}^4$ and phase difference $\{R_{x,j-1}^{\Delta\phi}\}_{j=2}^4$.

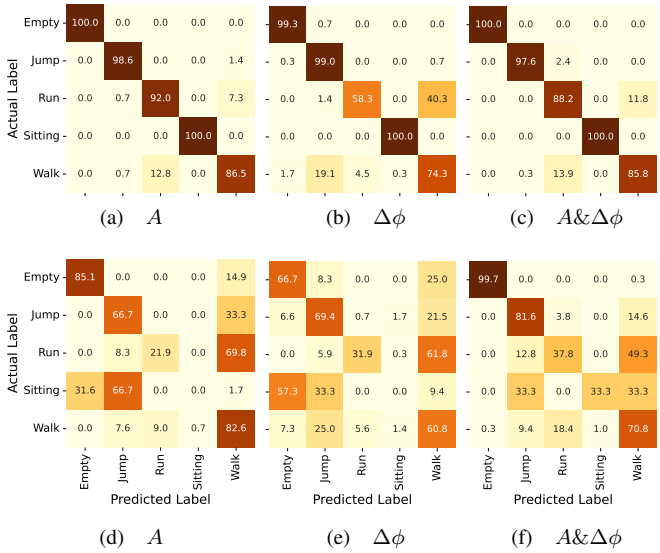
Specifically, $\mathcal{M}_{all}$ (viz., hatched bars) exhibits a consistent behavior across users and sample sets. Specifically, the model maintains a stable feature hierarchy in which amplitude components are most important, while phase differences are negligible, with values remaining close to zero. Among the amplitudes, $R_{x,1}^A$ and $R_{x,2}^A$ stand out as the most important, while $R_{x,4}^A$ consistently emerges as the least influential. This overall trend reveals that the model exploits amplitudes as a leaky shortcut, leading to cross-domain overfitting. In fact, while amplitude provides more stable patterns than volatile phase difference, it remains strictly dependent on the specific user and environment. Consequently, for $\mathcal{M}_X$, the antennas show a highly unstable behavior depending on the unseen user. For instance, for U_1, $\mathcal{M}_X$ over-relies on $R_{x,1}^A$, driving the misclassifications on $\mathcal{T}_u$, while underweighting $R_{x,2}^A$, i.e., the most relevant feature for $\mathcal{M}_{all}$. Similarly, for U_3, while $R_{x,1}^A$ and $R_{x,4}^A$ support correct predictions in $\mathcal{M}_X$, $R_{x,4}^A$ mainly drives its misclassifications, despite consistently representing the least influential structural component for $\mathcal{M}_{all}$.

Focusing on U_1 for brevity, as other users yield similar results, Fig. 4 shows feature importance along the time-axis for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_1\}}$ on $\mathcal{T}_s$ and $\mathcal{T}_u$, computed over consecutive packet pairs within a frame duration $\mathtt{W}$. In $\mathcal{T}_s$, both models mainly rely on earlier (i.e., left-handed) or central packets,



(a) *Cross-user* (b) *Cross-env* (c) *Cross-node*

Fig. 1. $\mathcal{M}^*_X$ vs. $\mathcal{M}_{all}$ under (a) *cross-user*, (b) *cross-env* and, (c) *cross-node* shifts based on input type: *amplitude* (**A**), *phase difference* ($\mathbf{\Delta}\boldsymbol{\phi}$), and *their combination* ($\mathbf{A}\&\mathbf{\Delta}\boldsymbol{\phi}$). Superscript in $\mathcal{M}^*_X$ denotes training set composition. x**-axis**: tested user, environment, or node.



(a) A (b) $\Delta\phi$ (c) $A\&\Delta\phi$

(d) A (e) $\Delta\phi$ (f) $A\&\Delta\phi$

Fig. 2. Misclassification errors of $\mathcal{M}_{all}$ (*top*) and $\mathcal{M}_X^{U\setminus\{U_2\}} \triangleright U_2$ (*bottom*) when testing on U_2 based on input type.

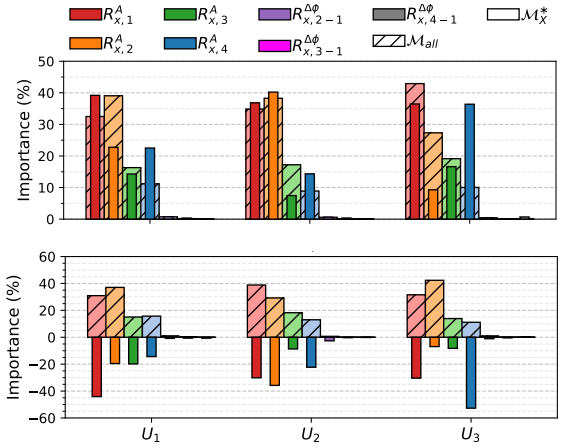


Fig. 3. Importance of receiving antenna $R_{x,j}$ ($A\&\Delta\phi$ input) for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_k\}}$ on $\mathcal{T}_s$ (**top**) and $\mathcal{T}_u$ (**bottom**) samples across target users U_k. $\mathbf{R}^A_{x,j}$: *amplitude*. $\mathbf{R}^{\Delta\phi}_{x,j-1}$: *phase difference* w.r.t $R_{x,1}$.

indicating that HAR is driven by stable motion patterns rather than recent observations, which do not aid accurate predictions. Conversely, in $\mathcal{T}_u$, packets that previously contributed positively now negatively impact $\mathcal{M}_X^{U\setminus\{U_1\}}$. This indicates that $\mathcal{M}_X^{U\setminus\{U_1\}}$ misinterprets temporal patterns that $\mathcal{M}_{all}$ correctly utilizes, showing its difficulty in generalizing to unseen users.

Lastly, Fig. 5 shows subcarrier importance for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_k\}}$ on $\mathcal{T}_s$ and $\mathcal{T}_u$, computed by grouping left and right subcarriers of the central DC carrier into 15 bins each. Specifically, on $\mathcal{T}_s$, importance values are broadly distributed across bins with most subcarriers contributing positively in both models and ranging in 0-15%. This indicates that successful classifications rely on information extracted from a wide spectral range. Moreover, for U_3, importance is more dispersed across subcarriers, revealing greater disagreement between models even on correctly classified samples. In contrast, the analysis on $\mathcal{T}_u$ shows that, while $\mathcal{M}_{all}$ maintains positive, distributed importance, $\mathcal{M}^*_X$ assigns predominantly negative importance to nearly all subcarriers. This indicates that $\mathcal{M}^*_X$ misinterprets frequency-domain information, perceiving spectral features as misleading and thus driving misclassifications. Across all users and sample sets, the DC subcarrier has no effect on the decisions of both models.

***Take-Away**: The analysis indicates that $\mathcal{M}^*_X$ struggles to generalize to unseen users by over-relying on the most domain-dependent amplitudes while ignoring the potentially more generalizable, yet volatile, phase differences. Conversely, $\mathcal{M}_{all}$ effectively utilizes information across antennas, packets, and subcarriers, whereas $\mathcal{M}^*_X$ misinterprets these structural signals, leading to severe misclassifications.*

VI. Discussion and Conclusions

This paper presented XAI2CSI, a framework embedding SAGE as its core to systematically apply XAI to CSI-based HAR and investigate the behavior of DL models under realistic conditions. Using different CSI input representations (amplitude, phase difference, and their combination) collected from modern IEEE 802.11ax hardware, we first evaluated model generalization to unseen users, environments, and receiver positions, showing that performance drops significantly when the deployment context changes. This highlights that cross-domain robustness in Wi-Fi sensing remains an open

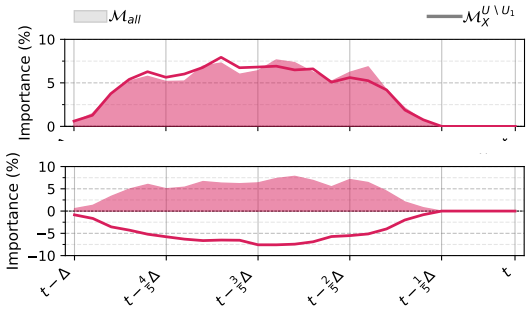


Fig. 4. Packet arrival order importance ($A\&\Delta\phi$ input) for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_1\}}$ on $\mathcal{T}_s$ (**top**) and $\mathcal{T}_u$ (**bottom**) samples of user U_1. x-axis: chronologically sorted packet indices (2 packets/bin; t is most recent). $\mathcal{M}_{all}$ (areas) vs. $\mathcal{M}_X^{U\setminus\{U_k\}}$ (lines).

challenge. Focusing on the *cross-user* scenario, we applied SAGE to analyze the contribution of temporal, spectral, and spatial CSI components to model decisions. Our analysis revealed that generalization failures stem from over-reliance on user-specific patterns, which are misinterpreted when the model encounters a new user. Overall, this study showed how explainability can provide a principled understanding of DL behavior in Wi-Fi sensing, paving the way toward transparent, interpretable, and robust CSI-based systems. *Future works* will include diverse XAI methods and DL models, mitigating CSI artifacts via domain adaptation, and enhancing multimodal fusion for cross-domain robustness.

REFERENCES

[1] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity WiFi Sensing in Ten Years: Status, Challenges, and Opportunities," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17 832–17 843, 2022.

[2] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Zhu, "Deep Learning for Radio-Based Human Sensing: Recent Advances and Future Directions," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 995–1019, 2021.

[3] I. Guarino, D. Carra, M. Cominelli, F. Gringoli, and R. Lo Cigno, "A survey on CSI-based Wi-Fi sensing datasets and models with a focus on reproducibility," *Elsevier Comput. Commun.*, vol. 249, p. 108431, 2026.

[4] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *NeurIPS'20*.

[5] M. Cominelli, F. Gringoli, and F. Restuccia, "Exposing the CSI: A Systematic Investigation of CSI-based Wi-Fi Sensing Capabilities and Limitations," in *IEEE PerCom'23*, pp. 81–90.

[6] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A Survey on Behavior Recognition Using WiFi Channel State Information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, 2017.

[7] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "DeepSense: Device-Free Human Activity Recognition via Autoencoder Long-Term Recurrent Convolutional Network," in *IEEE ICC'18*, 2018.

[8] F. Wang, W. Gong, and J. Liu, "On Spatial Diversity in WiFi-Based Human Activity Recognition: A Deep Learning-Based Approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2035–2047, 2019.

[9] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning Gestures From WiFi: A Siamese Recurrent Convolutional Architecture," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 763–10 772, 2019.

[10] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM," *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2714–2724, 2019.

[11] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "EfficientFi: Toward Large-Scale Lightweight WiFi Sensing via CSI Compression," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13 086–13 095, 2022.

[12] Y. Zhang, Y. Chen, Y. Wang, Q. Liu, and A. Cheng, "CSI-Based Human Activity Recognition With Graph Few-Shot Learning," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4139–4151, 2022.

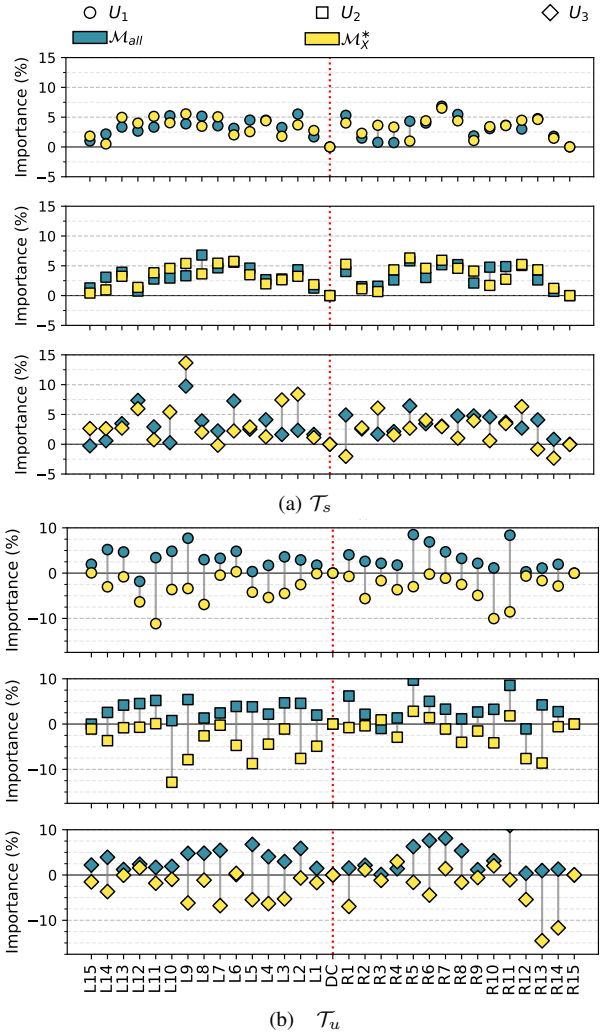


Fig. 5. Subcarrier importance ($A\&\Delta\phi$ input) for $\mathcal{M}_{all}$ and $\mathcal{M}_X^{U\setminus\{U_k\}}$ on (a) $\mathcal{T}_s$ and (b) $\mathcal{T}_u$ samples across target users U_k (marked differently). Left/right subcarriers are grouped into 15 bins ($\{\mathtt{L}_s\}_{s=1}^{15}$ and $\{\mathtt{R}_s\}_{s=1}^{15}$) around the DC carrier (*dotted red line*).

[13] J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Cell Press Patterns*, vol. 4, no. 3, p. 100703, 2023.

[14] Z. Yang, Y. Zhang, and Q. Zhang, "Rethinking Fall Detection With Wi-Fi," *IEEE Trans. Mob. Comput.*, vol. 22, no. 10, pp. 6126–6143, 2023.

[15] C. Zhang and W. Jiao, "ImgFi: A High Accuracy and Lightweight Human Activity Recognition Framework Using CSI Image," *IEEE Sensors J.*, vol. 23, no. 18, pp. 21 966–21 977, 2023.

[16] F. Meneghello, D. Garlisi, N. D. Fabbro, I. Tinnirello, and M. Rossi, "SHARP: Environment and Person Independent Activity Recognition With Commodity IEEE 802.11 Access Points," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 6160–6175, 2023.

[17] X. Chen, Y. Zou, C. Li, and W. Xiao, "A Deep Learning Based Lightweight HAR System Using Reconstructed WiFi CSI," *IEEE Trans. Human-Mach. Syst.*, vol. 54, no. 1, pp. 68–78, 2024.

[18] A. Hussain, Y. Chen, A. Ullah, and S. Zhang, "WiSigPro: Transformer for elevating CSI-based human activity recognition through attention mechanisms," *Expert Systems with Apps.*, vol. 258, p. 124976, 2024.

[19] J. A. Armenta-Garcia, F. F. Gonzalez-Navarro, and J. Caro-Gutierrez, "Wireless sensing applications with Wi-Fi Channel State Information, preprocessing techniques, and detection algorithms: A survey," *Elsevier Comput. Commun.*, 2024.