Wi-CBR: Salient-aware Adaptive WiFi Sensing for Cross-domain Behavior Recognition

Ruobei Zhang[1], **Shengeng Tang**[1*], **Huan Yan**[2], **Xiang Zhang**[3], **Jiabao Guo**[1*]
[1] Hefei University of Technology
[2] Guizhou Normal University
[3] Tianjin University
zrb@mail.hfut.edu.cn, {tangsg, garbo_guo}@hfut.edu.cn, yh1995.cs@gmail.com, zhangxiang@ieee.org

Abstract

The challenge in WiFi-based cross-domain Behavior Recognition lies in the significant interference of domain-specific signals on gesture variation. However, previous methods alleviate this interference by mapping the phase from multiple domains into a common feature space. If the Doppler Frequency Shift (DFS) signal is used to dynamically supplement the phase features to achieve better generalization, it enables the model to not only explore a wider feature space but also to avoid potential degradation of gesture semantic information. Specifically, we propose a novel *Salient-aware Adaptive **Wi**Fi Sensing for **C**ross-domain **B**ehavior **R**ecognition* (**Wi-CBR**), which constructs a dual-branch self-attention module that captures temporal features from phase information reflecting dynamic path length variations while extracting kinematic features from DFS correlated with motion velocity. Moreover, we design a Saliency Guidance Module that employs group attention mechanisms to mine critical activity features and utilizes gating mechanisms to optimize information entropy, facilitating feature fusion and enabling effective interaction between salient and non-salient behavioral characteristics. Extensive experiments on two large-scale public datasets (Widar3.0 and XRF55) demonstrate the superior performance of our method in both in-domain and cross-domain scenarios.

Code — https://github.com/zrbwsw/Wi-CBR

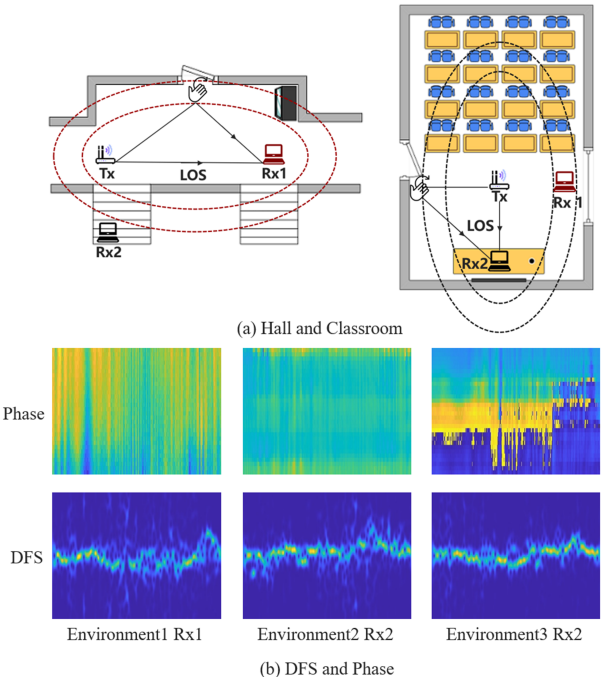


Figure 1: When the user and behavior remain unchanged, relative to the location of transmitter–receiver (Tx-Rx) pairs or the overall environment changes, the phase changes significantly, and DFS has a certain degree of domain independence.

Introduction

Human Behavior Recognition (HBR) enables systems to intelligently interpret human actions and is widely applied in areas like human-computer interaction (Tang et al. 2025a,b), Interactive integration (Song et al. 2024, 2023), action understanding (Xu et al. 2025; Zhang et al. 2025b), and accessible communication (Tang et al. 2022a,b). WiFi-based HBR, leveraging Channel State Information (CSI) and Received Signal Strength Indicator (RSSI), has gained attention for its ability to model motion-induced signal changes (Wang, Cang, and Yu 2019; Yadav et al. 2021). While coarse-grained activities are easily detectable, fine-grained gestures remain challenging due to subtle movements (Corballis 2010). By mapping CSI variations to behaviors, WiFi sensing supports applications in homes, workplaces, and other indoor environments (Abdelnasser, Youssef, and Harras 2015; Tan and Yang 2016).

An important challenge in WiFi-based behavior recognition is environmental dependence, as wireless signals often carry environment-specific information that is unrelated to human behavior(Zhang et al. 2025a). Cross-domain recognition accuracy significantly decreases due to differences in data distribution, noise, and changes in the propagation path. Traditional methods are divided into Modeling-Based and Learning-Based approaches (Pu et al. 2013; Wang et al. 2014). Modeling-Based methods use handcrafted features that are robust to environmental variation but lose raw signal richness, limiting performance in complex tasks. Learning-Based

*Corresponding authors.

methods leverage deep neural networks to learn directly from raw CSI but are hindered by environmental sensitivity and a lack of domain-invariant guidance. Therefore, how to integrate original features with domain-independent features is key to achieving effective cross-domain generalization (Guo et al. 2024; Zhang, Wu, and Han 2025; Wu and Deng 2025; Yin et al. 2022; Guo et al. 2025), bridging the gap between controlled training environments and real-world applications.

Our observations find that, in wireless communication systems, the characteristics of the received signal are influenced by various factors, including the positions of the transmitter (Tx) and receiver (Rx), the transmission path, and the dynamics of the overall environment. As shown in Fig. 1, significant variations in phase values occur when the receiver's position changes, even if the user and behavior remain constant. Figure (b) further demonstrates that phase values are highly sensitive to environmental changes, exhibiting unstable patterns across different environments and receiver positions. In contrast, the DFS shows greater domain independence and stability, with its features generally remaining consistent across varying environments and receiver locations. Building on these observations, the proposed method leverages the stability of DFS features while addressing the sensitivity of phase values. By extracting and aligning domain-invariant DFS features, the approach enhances the adaptability to dynamic and multi-domain scenarios. Furthermore, we aim to preserve the rich temporal-spatial patterns of raw signals while embedding domain-independent priors that facilitate robust cross-domain generalization. This hybrid motivation directly informs the network design, enabling complementary feature learning and improving cross-domain performance.

In this work, we propose Wi-CBR to leverage the full spectrum of raw data while ensuring the integration of domain-independent features. Specifically, our method introduces a novel multimodal collaborative awareness framework that efficiently combines phase data, which captures dynamic path length changes, and DFS data, which reflects frequency shifts tied to gesture speed. To capture spatial-temporal patterns within each modality, we employ a Two-branch self-attention module, enabling the system to focus on important temporal and spatial features within each signal type. A group attention mechanism is then applied to the concatenated phase and DFS features, allowing the model to identify key group features that are essential for behavior recognition. Finally, a gating mechanism is used to divide the fused features into enhancement and suppression branches, optimizing information entropy and facilitating collaborative complementarity. This fusion of multiple data sources, along with the innovative use of attention and gating mechanisms, enables more accurate and robust behavior recognition, particularly in cross-domain scenarios. Our main contributions are summarized as follows:

- We propose Wi-CBR, a novel Salient-aware Adaptive WiFi Sensing framework for Cross-domain Behavior Recognition. Wi-CBR proposes a two-branch self-attention module that captures temporal features from phase information, reflecting dynamic path length variations, while extracting kinematic features from DFS, which are correlated with motion velocity.
- We design a Saliency Guidance Module that leverages group attention mechanisms to identify critical activity features. This module incorporates gating mechanisms to optimize information entropy, facilitating feature fusion and enabling effective interaction between salient and non-salient behavioral characteristics.
- Extensive experiments on two large-scale public datasets demonstrate the superior performance of our method in both in-domain and cross-domain scenarios.

Related Work

Modeling-Based HBR

Modeling-based approaches preprocess raw CSI data, extract manual features (e.g., velocity statistics), and use machine learning for gesture recognition (Tan and Yang 2016; Meng et al. 2021; Regani, Wang, and Liu 2021; Gao et al. 2022; Gu et al. 2023; Zhang et al. 2023). WiGest (Abdelnasser, Youssef, and Harras 2015) relies on coarse-grained RSS, limiting accuracy. WiMU (Venkatnarayan, Page, and Shahzad 2018) struggles with scalability, and WiDraw (Sun et al. 2015) requires over 25 transceivers, making it impractical. QGesture (Yu et al. 2018) uses two antennas but depends on prior hand position knowledge. (Gao et al. 2022) introduces dynamic phase exponential error for gesture quality, while Wi-NN(Zhang et al. 2023) applies time-domain feature selection with KNN classification. These methods link WiFi signals to gestures but neglect environmental impact, as gesture performance in varying environments alters WiFi waveforms (Chen, Zhou, and Lin 2023; Kang, Zhang, and Huang 2021). Widar 3.0 (Zheng et al. 2019) introduces domain-agnostic BVP features, and WiHF (Li, Liu, and Cao 2020) derives domain-independent motion patterns; however, handcrafted features limit the capture of spatiotemporal cues. WiGesture (Gao et al. 2021) focuses on position-independent Motion Navigation Primitives (MNP). WiGNN(Chen and Huang 2024) focuses on graph modeling for multi-receiver topologies through GNN-based temporal-frequency aggregation and data augmentation.

Learning-Based HBR

Learning-based methods directly process raw CSI data, such as amplitude and phase, for automatic pattern recognition. Wikey (Ali et al. 2015) enables keystroke recognition but is highly sensitive to environmental changes. WiSign (Zhang, Zhang, and Zheng 2020) extracts spatio-temporal features for sign language recognition but requires extensive domain-specific training. Tong et al.(Tong et al. 2023) introduced a CNN-GRU-Attention (CGA) model with phase correction and gesture truncation for improved data validity, while Yang et al.(Yang et al. 2019) proposed a CNN-RNN architecture for enhanced spatiotemporal pattern learning. WiHGR (Meng et al. 2021) uses a phase difference matrix and an improved ABGRU for feature extraction. To address cross-domain challenges, CROSSGR (Li et al. 2021b) extracts gesture-related features independent of users. WiGr (Zhang et al. 2021) uses query-class prototype similarity to mitigate cross-domain variations. WIGRUNT (Gu et al. 2022) applies a spatio-temporal dual-attention network with ResNet for feature

extraction, while Wi-SFDAGR (Yan et al. 2025) addresses cross-domain issues using Unsupervised Domain Adaptation (UDA) for unlabeled test data scenarios

Method

As shown in Fig. 2, Wi-CBR consists of four components: signal preprocessing, two-branch self-attention learning, saliency guidance, and classification prediction. In signal preprocessing, the CSI-ratio model handles denoising, while DFS is extracted using STFT. The phase and DFS matrices are visualized as images for deep learning. The network employs two-branch self-attention and pre-trained ResNet18 for initial feature extraction, followed by feature fusion via a cross-model interactive module. Finally, behavior prediction is achieved using a classifier with dual-loss constraints.

Task Definition

WiFi CSI describes the signal's attenuation on its propagation paths, such as scattering, multipath fading or shadowing, and power decay over distance. It can be characterized as:

$$\mathbf{Y} = \mathbf{H} \cdot \mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are the received and transmitted signal vectors, respectively. $\mathbf{N}$ is the additive white Gaussian noise, and $\mathbf{H}$ is the channel matrix representing CSI. CSI is a superposition of signals of all the propagation paths, and its channel frequency response (CFR) can be represented as:

$$H(f,t) = \sum_{m \in \Phi} a_m(f,t) e^{-j2\pi \frac{d_m(t)}{\lambda}}, \tag{2}$$

where f and t represent center frequency and time stamp, m is the multipath component. $a_m(f,t)$ and $d_m(t)$ denote the complex attenuation and propagation length of the mth multipath component, respectively. Φ denotes the set of multipath components, and λ is the signal wavelength. In the case of CSI-based gesture recognition, the multipath component m consists of dynamic and static paths:

$$\begin{aligned} H(f,t) = &\sum_{m_s \in \Phi_s} a_{m_s}(f,t) e^{-j2\pi \frac{d_{m_s}(t)}{\lambda}} \\ &+ \sum_{m_d \in \Phi_d} a_{m_d}(f,t) e^{-j2\pi \frac{d_{m_d}(t)}{\lambda}} \end{aligned} \tag{3}$$

CSI Denoising Preprocessing

As demonstrated in the previous section, the gesture can be portrayed by the change of phase shift in CSI. Unfortunately, for commodity WiFi devices, as the transmitter and receiver are not synchronized, there exists a time-varying random phase offset $e^{-j\theta_{\text{offset}}}$:

$$\begin{aligned} H(f,t) = e^{-j\theta_{\text{offset}}} &(a_{m_s}(f,t) e^{-j2\pi \frac{d_{m_s}(t)}{\lambda}} \\ &+ a_{m_d}(f,t) e^{-j2\pi \frac{d_{m_d}(t)}{\lambda}}), \end{aligned} \tag{4}$$

where $e^{-j2\pi \frac{d(t)}{\lambda}}$, and $d(t)$ denote phase-shift, and path length of dynamic components, respectively. This random offset thus prevents us from directly using the CSI phase information.

Therefore, we need to eliminate $e^{-j\theta_{\text{offset}}}$. Fortunately, for commodity WiFi cards, this random offset remains the same across different antennas on the same WiFi network interface card (NIC) as they share the same RF oscillator. Thus, it can be eliminated by the CSI-ratio model:

$$H_r(f,t) = \frac{H_1(f,t)}{H_2(f,t)} = \frac{e^{-j\theta_{\text{offset}}}(H_{s,1} + H_{d,1})}{e^{-j\theta_{\text{offset}}}(H_{s,2} + H_{d,2})} \tag{5}$$

where $H_1(f,t)$ and $H_2(f,t)$ are the CSI of two receiving antennas. When two antennas are close to each other, Δd can be regarded as a constant. According to Mobius transformation (Zeng et al. 2019), (5) represents transformations such as scaling and rotation of the phase-shift $e^{-j2\pi \frac{d_1(t)}{\lambda}}$ of antenna 1 in the complex plane, and these transformations will not affect the changing trend of the phase-shift(Kotaru et al. 2015; Li et al. 2016; Zeng et al. 2018).

CSI to Phase. Phase ratio P extracted from H_r can be used to describe gestures:

$$\mathrm{P} = angle(H_r), \tag{6}$$

where $angle(\cdot)$ denotes the phase extraction function. For a complex $z = \text{abs}(z) \cdot e^{j\theta}$, we can use $angle(\cdot)$ to obtain the phase of z, $\theta = angle(z)$.

CSI to DFS. To extract Doppler Frequency Shift (DFS) features, we take the raw CSI signal received by a specific Tx–Rx antenna pair, denoted as $H_q(f,t)$, as an example. This stream contains temporal variations caused by human motion. Unlike $H_r(f,t)$, which is a ratio of two antenna CSI streams designed to remove phase offset for phase analysis, $H_q(f,t)$ retains the original frequency content, making it more suitable for Doppler-based gesture analysis. The DFS reflects frequency changes due to gesture motion, such as hand speed or direction.

$$S_q(\tau,\omega) = \int_{-\infty}^{\infty} H_q(f,t) w(t-\tau) e^{-j\omega t} dt\,, \tag{7}$$

where $w(t-\tau)$ is a window function (e.g., Hanning window) to segment the signal, τ is the time localization, and ω is the frequency in radians, tied to the Doppler shift. The power spectrogram is then computed:

$$\mathrm{D} = |S_q(\tau,\omega)|^2 = \left| \int_{-\infty}^{\infty} H_q(f,t) w(t-\tau) e^{-j\omega t} dt \right|^2 \tag{8}$$

In this spectrogram, the frequency ω at each time τ corresponds to the Doppler shift $f_d = \omega/(2\pi)$. f_d is proportional to the gesture speed $v(t)$. Positive frequencies indicate motion toward the receiver, while negative frequencies indicate motion away. This time-frequency representation enables detailed analysis of gesture dynamics.

For the Widar3.0 dataset, each DFS file is a $6 \times 121 \times \mathrm{T}$ matrix, where the first dimension represents the 6 receivers, the second dimension represents the 121 frequency segments ranging from $[-60, 60]$ Hz, and the third dimension represents the timestamps with a sampling rate of 1000 Hz.

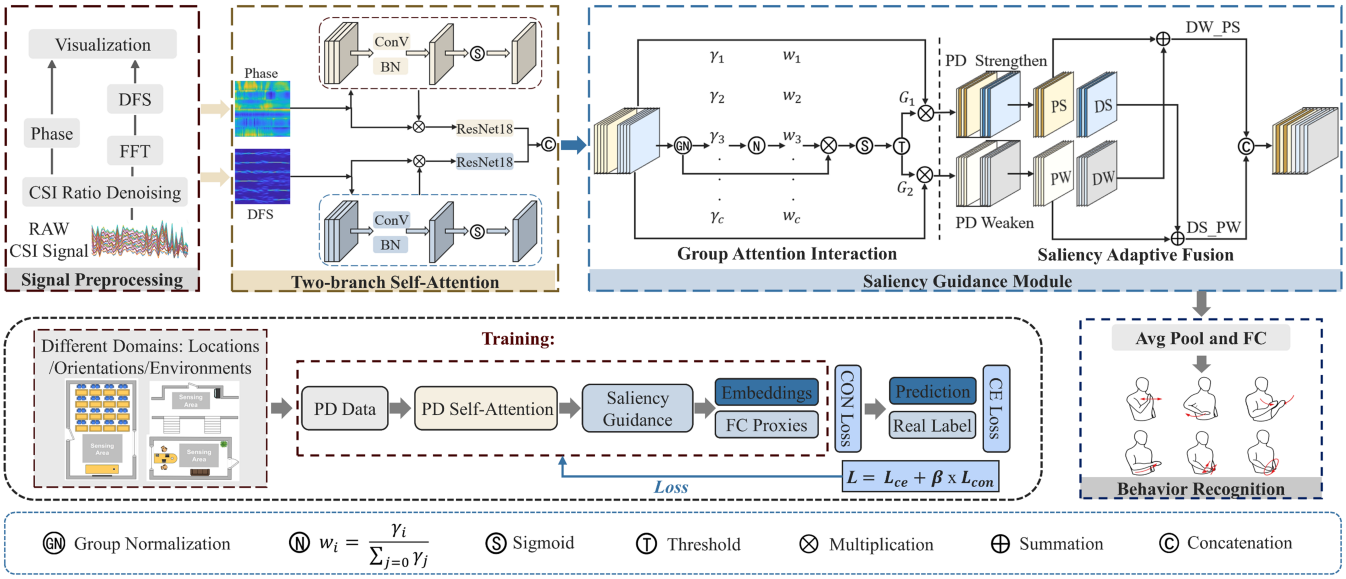


Figure 2: Framework Overview of Wi-CBR. The proposed Wi-CBR framework integrates Signal Preprocessing, a Two-branch Self-attention Module, and a Saliency Guidance Module to achieve robust behavior recognition across domains. It leverages phase information for temporal dynamics and Doppler Frequency Shift (DFS) for spatial motion characteristics. A Group Attention Interaction mechanism identifies salient features, while Saliency Adaptive Fusion effectively combines critical and non-critical behavior features. The system is designed to generalize across different environments, orientations, and locations, ensuring superior performance in both in-domain and cross-domain scenarios.

Two-branch Self-attention Learning

Two-branch Self-attention. The proposed dual-path spatial attention mechanism processes phase ($\mathbf{P}_x$) and Doppler ($\mathbf{D}_x$) features through independent attention branches. Let $f_{\text{conv}}^{(k)}$ denote a convolution operation with kernel size $k \times k$, and BN represent batch normalization. The attention weights are computed as:

$$\mathbf{A}_p = \sigma\left(\text{BN}\left(f_{\text{conv}}^{(7)}\left(f_{\text{conv}}^{(7)}(\mathbf{P}_x)\right)\right)\right). \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. $\mathbf{A}_d$ is computed in the same way using $\mathbf{D}_x$ as input. The refined features are obtained through:

$$\mathbf{P}_{\text{out}} = \mathbf{P}_x \otimes \mathbf{A}_p \oplus \mathbf{P}_x; \quad \mathbf{D}_{\text{out}} = \mathbf{D}_x \otimes \mathbf{A}_d \oplus \mathbf{D}_x \tag{10}$$

where $\otimes$ is element-wise multiplication, $\oplus$ is element-wise summation. Each branch maintains independent convolution parameters, with channel dimensions preserved in the first convolution ($3 \to 3$) and reduced to a single channel ($3 \to 1$) in the second convolution.

Two-branch Feature Extraction. Independent spatial attention parameters for P/D branches; Separate batch normalization statistics for each modality; Symmetric padding (3 pixels) maintained in all convolutions; No parameter sharing between ϕ_{ResNet}^{P} and ϕ_{ResNet}^{D}.

$$\mathbf{X}_{\text{PD}} = \text{Concat}(\phi_{\text{ResNet}}^{P}(\mathbf{P}_{\text{out}}), \phi_{\text{ResNet}}^{D}(\mathbf{D}_{\text{out}})) \in \mathbb{R}^{1024\times7\times7} \tag{11}$$

Saliency Guidance Module

Group Attention Interaction. To make the extracted Phase and DFS features interact and merge, we use group normalization to achieve the attention of different channels. The generated attention maps are thresholded to get the strengthened and weakened maps, and then the strengthened and weakened Phase and DFS features are obtained. Based on retaining important features and attenuating minor features (Li, Wen, and He 2023), the feature space redundancy is reduced while utilizing both features.

Separate operation aims to separate those informative feature maps from less informative ones corresponding to the spatial content. We leverage the scaling factors in Group Normalization (GN) (Wu and He 2018) layers to assess the informative content of different feature maps. To be concrete, given an intermediate feature map $X_{\text{PD}} \in \mathbb{R}^{N\times C\times H\times W}$, where N is the batch axis, C is the channel axis, H and W are the spatial height and width axes. We first standardize the input feature X by subtracting mean μ and dividing by standard deviation σ as follows:

$$X_{\text{gn}} = \text{GN}(\mathbf{X}_{\text{PD}}) = \gamma\frac{X_{\text{PD}} - \mu}{\sqrt{\sigma^2 + \epsilon}} + z \tag{12}$$

where μ and σ are the mean and standard deviation of X_{PD}, ϵ is a small positive constant added for numerical stability, γ and z are trainable affine transformation parameters.

Noted that we leverage the trainable parameters $\gamma \in \mathbb{R}^C$ in GN layers as a way to measure the variance of spatial pixels for phase and DFS. The richer spatial information reflects more variation in spatial pixels contributing to a larger γ. The

normalized correlation weights $W_\gamma \in \mathbb{R}^C$ are obtained by equation (14), which indicates the importance of different phase and DFS feature maps.

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}, \quad i, j = 1, 2, \ldots, C \tag{13}$$

Then the weight values of feature maps reweighted by W_γ are mapped to the range (0, 1) by the sigmoid function and gated by a threshold. We set those weights above the threshold to 1 to obtain the informative weights G_1 while setting them to 0 to gain the non-informative weights G_2 (the threshold is set to 0.5 in the experiments):

$$G = \text{Gate}(\text{Sigmoid}(W_\gamma(GN(X_{\text{PD}})))) \tag{14}$$

We multiply input features X_{PD} by G_1 and G_2 respectively, yielding two weighted features: the strengthen ones X_{PD}^S and less informative ones X_{PD}^W.Thus, we completed the interaction between the input phase and DFS features. The attention weights are obtained by learning the variance and bias through group normalization, which is gated to obtain the strengthened and weakened attention maps. Two components obtained are as follows: X_{PD}^S has informative and expressive spatial content and is strengthened, while X_{PD}^W has little or no information, which is considered redundant and weakened.

Saliency Adaptive Fusion. We propose a fusion operation to achieve synergistic utilization. An information-rich feature after strengthening is added to a feature with less information after weakening. New features with richer information are generated, i.e., one feature dominates while the other aids. Instead of adding the two components directly, we use a cross-fertilization operation to thoroughly combine the two weighted different information features to enhance the information flow between them. The cross-rendered features $Y_{\text{PS\_DW}}$ and $Y_{\text{PW\_DS}}$ are then stitched together to obtain a spatially fine feature mapping of X_{out}. The whole process of Fusion operation can be expressed as :

$$\begin{cases} X_{\text{PD}}^S = G_1 \otimes X_{\text{PD}}, X_{\text{PD}}^W = G_2 \otimes X_{\text{PD}}; \\ Y_{\text{DS\_PW}} = X_{\text{D}}^S \oplus X_{\text{P}}^W, Y_{\text{DW\_PS}} = X_{\text{D}}^W \oplus X_{\text{P}}^S; \\ X_{\text{out}} = \text{Concat}(Y_{\text{DS\_PW}}, Y_{\text{DW\_PS}}) \end{cases} \tag{15}$$

where $\otimes$ is element-wise multiplication, $\oplus$ is element-wise summation, Concat is concatenation. When DFS is remarkable as a domain-independent feature, we use the enhanced DFS feature, assisted by the weakened detailed phase feature, namely $Y_{\text{DS\_PW}}$. When DFS is weak as a domain-independent feature, we use the weakened DFS feature and use the enhanced detailed phase feature as a supplement, that is, $Y_{\text{DW\_PS}}$. After the Saliency Guidance Module is applied to the input features X_{PD}, not only do we separate the informative features from less informative ones, but also we reconstruct them to enhance the representative features and suppress the redundant features in spatial dimension.

Contrastive Loss Optimization

Building upon the aforementioned modules, we obtain the Cross-Model Fusion feature representation X_{out}. Most existing works directly feed the global representation X_{out} into the classifier (i.e., a fully-connected layer with a Softmax) to predict the probability of gestures $\hat{y}$. The model is then trained by minimizing the corresponding loss L_{ce} between the prediction values $\hat{y}$ and their ground truths y.

$$\mathcal{L}_{ce} = \mathcal{L}_{ce}(y, \hat{y}) = -\frac{1}{M} \sum_{m=1}^{M} \sum_{s=1}^{S} y_{m,s} \log(\hat{y}_{m,s}) \tag{16}$$

where L_{ce} is a classification cross-entropy loss, and M is the number of data samples, S is the number of gestures.

Contrastive Loss. optimizes the objective by learning a distance measure based on multiple positive and negative sample-to-sample pairs. The key idea behind this is to learn an embedding space where similar pairs of samples are close to each other and dissimilar pairs are far apart. Thus, we can obtain an invariant representation across different environments for the same instance, i.e., a domain-independent characterization. Several elements can influence the variation patterns of the CSI signal, with varying degrees of effect. For example, a user's influence on signal patterns is less significant compared to changes in location and orientation. Importantly, positive sample pairs across diverse environments differ, and some may be challenging to match due to substantial data discrepancies. Perfectly aligning all samples might limit the model's ability to generalize. To mitigate this, we utilize class proxies to symbolize each gesture, ideally enhancing resilience across samples from varied settings. Formally, these proxy vectors are defined as the weights of the final fully connected layer in the classifier. To further improve semantic consistency, we implement a **proxy-based** contrastive loss that utilizes the connections between class proxies and samples to foster robust representations (Yao et al. 2022). Given the representation x_i of the ith sample, we select its class proxy w_c in place of positive samples x_+ to form proxy-to-sample positive pairs. The contrast loss is incorporated into the overall loss function $\mathcal{L}$:

$$\mathcal{L}_{con} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{(w_c^T xi)/\tau}}{e^{(w_c^T x_i)/\tau} + \sum_{k=1, k \neq c}^{R} e^{(w_k^T x_i)/\tau}};$$
$$\mathcal{L} = \mathcal{L}_{ce} + \beta \times \mathcal{L}_{con} \tag{17}$$

where w_c represents the class proxy corresponding to class c, R is the total number of classes, and τ serves as the temperature parameter. We aim to minimize the following final loss function, where $\mathcal{L}_{ce}$ is the cross-entropy loss and $\mathcal{L}_{con}$ is the proxy-based contrastive loss. β is the trade-off parameter.

Experiments

Experimental Setup

We train and evaluate our proposed Wi-CBR on the two largest human sensing multimodal public datasets.

Gesture Recognition. To evaluate the effectiveness of our model for cross-domain gesture recognition, we conducted extensive experiments on the Widar 3.0 and XRF55 datasets. For in-domain, cross-location, and orientation evaluations on Widar 3.0, we used 80% of the data for training and 20% for testing with five-fold cross-validation. For cross-location evaluation, one location was used for testing and the remaining

Method	Processing Flow	Widar3.0			Mean	XRF55
		CL	CO	CE		
EI (Jiang et al. 2018)	CSI→Amplitude	73.33	79.70	–	–	–
Widar3.0 (Zheng et al. 2019)	CSI→DFS→BVP	90.48	81.58	83.30	85.12	–
WiHF (Li, Liu, and Cao 2020)	CSI→DFS→MCP	91.22	80.64	–	–	–
THAT (Li et al. 2021a)	Raw CSI	71.56	81.76	49.71	67.68	23.23
WIGRUNT (Gu et al. 2022)	CSI→Phase	97.08	93.39	95.36	95.28	55.92
AaD (Yang, Jui, and Van De Weijer 2022)	CSI→Phase	95.90	95.38	93.00	94.83	55.64
ImgFi (Zhang and Jiao 2023)	CSI→STFT, RT image	39.58	38.12	40.37	39.36	31.90
WiSR (Liu et al. 2023)	CSI image	67.73	69.74	52.77	63.41	26.66
Recurrent ConFormer (Shang and Hong 2023)	Raw CSI	73.84	85.88	50.38	70.03	16.54
WiDual (Dai et al. 2023)	CSI→Phase	97.39	94.87	–	–	–
WIGNN (Chen and Huang 2024)	CSI→DFS	95.20	93.30	–	–	–
Wi-SFDAGR (Yan et al. 2025)	CSI→Phase	97.30	**97.17**	95.52	96.66	57.99
Wi-CBR (Ours)	**CSI→Phase**, **DFS**	**98.34**	96.30	**96.87**	**97.17**	**66.05**

Table 1: THE accuracy of Wi-CBR under CL(Cross-Location), CO(Cross-Orientation), and CE(Cross-Environment) settings in the WIDAR3.0 dataset and the XRF55 dataset.

four for training. The in-domain and cross-direction evaluations followed a similar approach. For cross-environment evaluation, we used data from three environments, totaling 12,750 samples, with training on two environments and testing on the third using triple cross-validation. For XRF55, the cross-environment evaluation involved four scenarios, with quad-fold cross-validation on 6,240 samples.

Activity Recognition. For the fairness of the experimental evaluation, the implementation details are exactly the same as those of XRF55. Note that the quad-fold cross-validation is not used here. At this time, due to the lack of multi-environment data support, the cross-domain migration capability of the model is highly required. The training set of scene 1 is used for training, that is, the first 14 times of each behaviour of each user. The test set is tested with the samples of scenes 2, 3, and 4, respectively. There are 33000 samples in the training set and 3300 samples in each scene in the test set, including 55 daily human behaviours.

Implementation Details. We used MATLAB to preprocess CSI data and generate 224×224 RGB images. After obtaining phase and DFS images, all models were implemented in PyTorch 1.13.1. The network architecture and data dimensions are shown in Fig. 2. Wi-CBR employs ResNet-18 (He et al. 2016) with pre-trained ImageNet weights as the feature extractor. The Cross-Model Interaction module set the threshold to 0.5. The contrast loss weight β and temperature are both set to 0.1. During training, the model is optimized using Adam with a learning rate of 0.0001, a batch size of 10, and 30 epochs. We used the same network structure for both the Widar 3.0 and XRF55 datasets, ensuring robustness across datasets. Full connection layer classification header: in gesture recognition, it is 6 to 9 in widar3.0 dataset, 8 in XRF55 dataset, and 55 in activity recognition. A random seed of 42 was set for reproducibility.

Comparisons to Prior SOTA Results

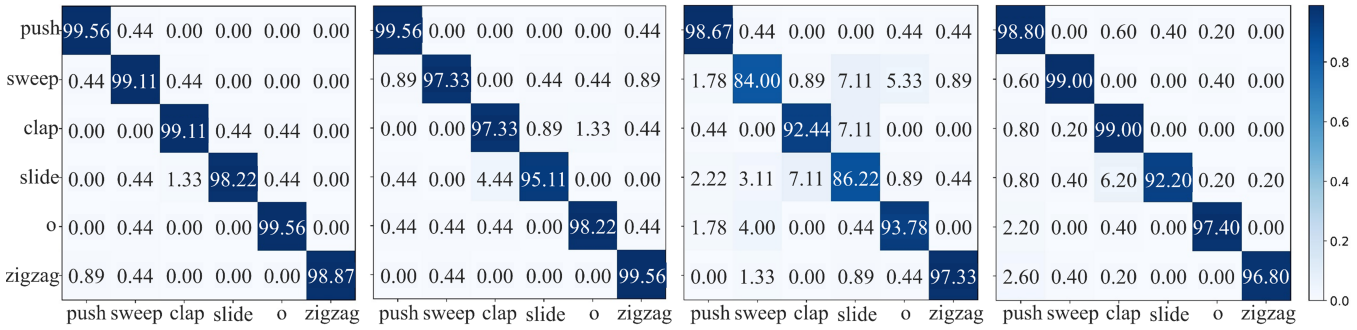
We compare our method with the baseline method in downstream human perception tasks, including fine-grained HGR and human HAR. We mainly study single-factor cross-domain, that is, only one domain factor we have never seen in location, orientation, and environment. We trained in some specific domains and test in an unknown domain.

CE	Method	few-shot		
		zero	one	two
2	WIGRUNT(Gu et al. 2022)	21.82	41.34	50.17
	XRF55(Wang et al. 2024)	2.52	49.83	57.51
	Wi-CBR (Ours)	**31.33**	**50.92**	**58.96**
3	WIGRUNT(Gu et al. 2022)	13.12	41.34	56.53
	XRF55(Wang et al. 2024)	2.14	50.85	**63.30**
	Wi-CBR (Ours)	**28.60**	**51.74**	61.01
4	WIGRUNT(Gu et al. 2022)	14.55	42.60	53.81
	XRF55(Wang et al. 2024)	2.03	50.94	61.41
	Wi-CBR (Ours)	**31.30**	**52.70**	**62.66**

Table 2: 55 kinds of human daily behaviors cross-domain recognition on XRF55 under few-shot.

HGR Comparison with SOTA Methods. We compare the cross-domain performance with SOTA methods across modeling-based, low-level semantic, and high-level semantic approaches (Tab. 1). Modeling-based methods (e.g., Widar3.0 (Zheng et al. 2019), WiHF (Li, Liu, and Cao 2020)) extract DFS and domain-independent features, achieving 85.12% on Widar3.0. Low-level semantic methods (e.g., ImgFi (Zhang and Jiao 2023), WiSR (Liu et al. 2023)) rely on raw CSI or visualizations, with performance ranging from 40% to 70% on Widar3.0 and up to 31.90% on XRF55. High-level semantic methods (e.g., Wi-SFDAGR (Yan et al. 2025)) utilize CSI-ratio denoising and advanced networks like ResNet18, reaching 96.66% on Widar3.0 through optimized feature aggregation.

Wi-CBR outperforms state-of-the-art on both datasets. In addition, we observe an interesting phenomenon: The perfor-



(a)In-Domain (99.04%) (b) CL (97.85%) (c) CO (92.07%) (d) CE (97.47%)

Figure 3: The confusion matrices in the ID (repetition 1), CL (location 1), CO (orientation 1), and CE (environment 3).

Methods	ID	CL	CO	CE
Without DFS	99.42	96.07	92.76	93.76
Without Phase	97.13	95.04	87.09	94.93
Without CL	99.49	97.26	95.56	96.08

Table 3: Effectiveness of each component.

Methods	ID	CL	CO	CE
Channel Attention	99.11	97.07	95.07	95.93
PS_DS and PW_DW	99.54	97.14	95.13	96.28
Wi-CBR	**99.54**	**98.34**	**96.30**	**96.57**

Table 4: Effectiveness of Fusion.

mance across environments on the XRF55 dataset is lower than that of Widar3.0, regardless of which method is used. This is because the XRF55 dataset has a lower sampling rate than Widar3.0, and the number of receiver arrangements in the environment is half that of Widar3.0. Wi-CBR, on the other hand, demonstrates robustness across datasets, and our work still achieves 66.05% accuracy despite changes in conditions such as receiver and sampling rate.

HAR included Comparison on XRF55. We further evaluate the recognition of human daily activities. We initially train the model using training samples from environment 1, and subsequently finetune it using 0/1/2 samples (per subject per action) from environment 2. We then test the finetuned model with the remaining 20/19/18 samples (per subject per action) in environment 2. Tab. 2 indicates the model's best transferability to unseen environment. In the case of zero-shot, most models can not effectively learn the domain invariance feature, and the model is close to a random guess. The advantage of Wi-CBR is that it can recognize the good features learned by multimodal collaborative sensing even if it is only learned in one environment, without using any samples of an unknown environment.

Ablation Study

We evaluated the impact of different components and fusion methods on the experiment. The results in Tab. 3 show that DFS greatly improves cross-environment performance, while phase mainly affects in-domain and cross-domain performance within the environment. Additionally, omitting contrast loss guidance led to a noticeable drop in cross-domain performance, while in-domain performance remained stable, indicating that contrast loss helps the model focus on gesture-related features rather than environmental noise. The results in Tab. 4 show that easy channel attention using WiGRUNT (Gu et al. 2022) channel-gate module will continue to reduce the accuracy of the model, highlighting the importance of interaction for balancing phase and DFS data. We also tried different fusion methods. In fact, when there is only one kind of data as input, it is equivalent to self-interaction, such as using **PS_PW**. When using the two kinds of data as input, the respective enhanced addition **PS_DS** exists feature redundancy and cannot play a complementary effect. Overall, Wi-CBR effectively leverages full CSI raw data and DFS for improved cross-domain performance.

To analyze recognition accuracy and misclassification rates, we present confusion matrices for test location 1, orientation 1, and environment 3 on Widar 3.0, as shown in Fig. 3. In cross-location and cross-orientation tests, "push" and "zigzag" achieve the best accuracy, while "slide" and "sweep" show the lowest, respectively. In the cross-environment test, "sweep" and "clap" perform best. Interestingly, "slide" is frequently misclassified as "clap." This likely stems from similar motion trajectories, as both gestures involve inclined or sliding movements, making them harder to distinguish.

Conclusion

In this work, we propose Wi-CBR, a WiFi-based cross-domain human behavior recognition framework with strong cross-domain performance. It integrates model-based and learning-based methods, using phase data (path length changes) and DFS data (frequency shifts) for robust cross-domain performance. It employs a two-branch self-attention module to extract spatio-temporal features and a saliency-aware adaptive mechanism to enhance phase features based on DFS significance. This time-frequency fusion focuses on behavior while reducing environmental interference, improving recognition accuracy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 62502144, 61932009, 62462015), the Anhui Provincial Natural Science Foundation, China (Grant No. 2408085QF191), the Young Elite Scientist Sponsorship Program by Gast (Grant No. GASTYESS202429), and the Fundamental Research Funds for the Central Universities (Grants No. JZ2024HGTA0178, PA2025IISL0112). The computation is completed on the HPC Platform of Hefei University of Technology.

References

Abdelnasser, H.; Youssef, M.; and Harras, K. A. 2015. Wigest: A ubiquitous WiFi-based gesture recognition system. In *INFOCOM*, 1472–1480. IEEE.

Ali, K.; Liu, A. X.; Wang, W.; and Shahzad, M. 2015. Keystroke recognition using WiFi signals. In *MobiCom*, 90–102.

Chen, C.; Zhou, G.; and Lin, Y. 2023. Cross-domain WiFi sensing with channel state information: A survey. *ACM Computing Surveys*, 55(11): 1–37.

Chen, Y.; and Huang, X. 2024. WiGNN: WiFi-based cross-domain gesture recognition inspired by dynamic topology structure. *IEEE Wireless Communications*, 31(3): 249–256.

Corballis, M. C. 2010. The gestural origins of language. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1): 2–7.

Dai, M.; Cao, C.; Liu, T.; Su, M.; Li, Y.; and Li, J. 2023. WiDual: User Identified Gesture Recognition Using Commercial WiFi. In *CCGrid*, 673–683.

Gao, R.; Li, W.; Xie, Y.; Yi, E.; Wang, L.; Wu, D.; and Zhang, D. 2022. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. *IMWUT*, 6(1): 1–26.

Gao, R.; Zhang, M.; Zhang, J.; Li, Y.; Yi, E.; Wu, D.; Wang, L.; and Zhang, D. 2021. Towards position-independent sensing for gesture recognition with WiFi. *IMWUT*, 5(2): 1–28.

Gu, Y.; Yan, H.; Zhang, X.; Wang, Y.; Huang, J.; Ji, Y.; and Ren, F. 2023. Attention-based gesture recognition using commodity WiFi devices. *IEEE Sensors Journal*, 23(9): 9685–9696.

Gu, Y.; Zhang, X.; Wang, Y.; Wang, M.; Yan, H.; Ji, Y.; and Dong, M. 2022. WiGRUNT: WiFi-enabled gesture recognition using dual-attention network. *IEEE Transactions on Human-Machine Systems*, 52(4): 736–746.

Guo, J.; Liu, A.; Diao, Y.; Zhang, J.; Ma, H.; Zhao, B.; Hong, R.; and Wang, M. 2025. Domain Generalization for Face Anti-spoofing via Content-aware Composite Prompt Engineering. *IEEE Transactions on Multimedia*.

Guo, J.; Liu, H.; Luo, Y.; Hu, X.; Zou, H.; Zhang, Y.; Liu, H.; and Zhao, B. 2024. Style-conditional prompt token learning for generalizable face anti-spoofing. In *ACM International Conference on Multimedia*, 994–1003.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE.

Jiang, W.; Miao, C.; Ma, F.; Yao, S.; Wang, Y.; Yuan, Y.; and Su, L. 2018. Towards environment independent device free human activity recognition. In *MobiCom*, 289–304.

Kang, H.; Zhang, Q.; and Huang, Q. 2021. Context-aware wireless-based cross-domain gesture recognition. *IEEE Internet of Things Journal*, 8(17): 13503–13515.

Kotaru, M.; Joshi, K.; Bharadia, D.; and Katti, S. 2015. SpotFi: Decimeter level localization using wifi. In *ACM Conference on Special Interest Group on Data Communication*, 269–282. ACM.

Li, B.; Cui, W.; Wang, W.; Zhang, L.; Chen, Z.; and Wu, M. 2021a. Two-Stream Convolution Augmented Transformer for Human Activity Recognition. *AAAI Conference on Artificial Intelligence*, 35: 286–293.

Li, C.; Liu, M.; and Cao, Z. 2020. Wihf: Enable user identified gesture recognition with wifi. In *INFOCOM*, 586–595.

Li, J.; Wen, Y.; and He, L. 2023. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *CVPR*, 6153–6162. IEEE/CVF.

Li, X.; Chang, L.; Song, F.; Wang, J.; Chen, X.; Tang, Z.; and Wang, Z. 2021b. CrossGR: Accurate and low-cost cross-target gesture recognition using Wi-Fi. *IMWUT*, 5(1): 1–23.

Li, X.; Li, S.; Zhang, D.; Xiong, J.; Wang, Y.; and Mei, H. 2016. Dynamic-MUSIC: Accurate device-free indoor localization. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 196–207. ACM.

Liu, S.; Chen, Z.; Wu, M.; Liu, C.; and Chen, L. 2023. WiSR: Wireless domain generalization based on style randomization. *IEEE Transactions on Mobile Computing*, 23(5): 4520–4532.

Meng, W.; Chen, X.; Cui, W.; and Guo, J. 2021. WIHGR: A robust WiFi-based human gesture recognition system via sparse recovery and modified attention-based BGRU. *IEEE Internet of Things Journal*, 9(12): 10272–10282.

Pu, Q.; Gupta, S.; Gollakota, S.; and Patel, S. 2013. Whole-home gesture recognition using wireless signals. In *MobiCom*, 27–38.

Regani, S. D.; Wang, B.; and Liu, K. R. 2021. WiFi-based device-free gesture recognition through-the-wall. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 8017–8021. IEEE.

Shang, M.; and Hong, X. 2023. Recurrent conformer for WiFi activity recognition. *IEEE/CAA Journal of Automatica Sinica*, 10(6): 1491–1493.

Song, P.; Guo, D.; Yang, X.; Tang, S.; and Wang, M. 2024. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 33: 1122–1135.

Song, P.; Guo, D.; Yang, X.; Tang, S.; Yang, E.; and Wang, M. 2023. Emotion-prior awareness network for emotional video captioning. In *ACM International Conference on Multimedia*, 589–600.

Sun, L.; Sen, S.; Koutsonikolas, D.; and Kim, K.-H. 2015. Widraw: Enabling hands-free drawing in the air on commodity WiFi devices. In *MobiCom*, 77–89.

Tan, S.; and Yang, J. 2016. WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition. In *MobiHoc*, 201–210.

Tang, S.; Guo, D.; Hong, R.; and Wang, M. 2022a. Graph-Based Multimodal Sequential Embedding for Sign Language Translation. *IEEE Transactions on Multimedia*, 4433–4445.

Tang, S.; He, J.; Cheng, L.; Wu, J.; Guo, D.; and Hong, R. 2025a. Discrete to Continuous: Generating Smooth Transition Poses from Sign Language Observations. In *CVPR*, 3481–3491.

Tang, S.; He, J.; Guo, D.; Wei, Y.; Li, F.; and Hong, R. 2025b. Sign-IDD: Iconicity Disentangled Diffusion for Sign Language Production. In *AAAI Conference on Artificial Intelligence*, 7266–7274.

Tang, S.; Hong, R.; Guo, D.; and Wang, M. 2022b. Gloss Semantic-Enhanced Network with Online Back-Translation for Sign Language Production. In *ACM International Conference on Multimedia*, 5630–5638.

Tong, G.; Li, Y.; Zhang, H.; and Xiong, N. 2023. A fine-grained channel state information-based deep learning system for dynamic gesture recognition. *Information Sciences*, 636: 118912.

Venkatnarayan, R. H.; Page, G.; and Shahzad, M. 2018. Multi-user gesture recognition using WiFi. In *MobiSys*, 401–413.

Wang, F.; Lv, Y.; Zhu, M.; Ding, H.; and Han, J. 2024. Xrf55: A radio frequency dataset for human indoor action analysis. *IMWUT*, 8(1): 1–34.

Wang, Y.; Cang, S.; and Yu, H. 2019. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, 137: 167–190.

Wang, Y.; Liu, J.; Chen, Y.; Gruteser, M.; Yang, J.; and Liu, H. 2014. E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures. In *MobiCom*, 617–628.

Wu, A.; and Deng, C. 2025. Percept, Memory, and Imagine: World Feature Simulating for Open-Domain Unknown Object Detection. In *CVPR*, 4682–4691.

Wu, Y.; and He, K. 2018. Group normalization. In *European Conference on Computer Vision*, 3–19. Springer.

Xu, H.; Cheng, L.; Wang, Y.; Tang, S.; and Zhong, Z. 2025. Towards Fine-Grained Emotion Understanding via Skeleton-Based Micro-Gesture Recognition. *arXiv preprint arXiv:2506.12848*.

Yadav, S. K.; Tiwari, K.; Pandey, H. M.; and Akbar, S. A. 2021. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223: 106970.

Yan, H.; Zhang, X.; Huang, J.; Feng, Y.; Li, M.; Wang, A.; and Liu, Z. 2025. Wi-SFDAGR: WiFi-Based Cross-Domain Gesture Recognition via Source-Free Domain Adaptation. *IEEE Internet of Things Journal*.

Yang, J.; Zou, H.; Zhou, Y.; and Xie, L. 2019. Learning gestures from WiFi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, 6(6): 10763–10772.

Yang, S.; Jui, S.; and Van De Weijer, J. 2022. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35: 5802–5815.

Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; and Yu, B. 2022. PCL: Proxy-based Contrastive Learning for Domain Generalization. In *CVPR*, 7097–7107. IEEE/CVF.

Yin, Y.; Zhu, B.; Chen, J.; Cheng, L.; and Jiang, Y.-G. 2022. Mix-dann and dynamic-modal-distillation for video domain adaptation. In *ACM International Conference on Multimedia*, 3224–3233.

Yu, N.; Wang, W.; Liu, A. X.; and Kong, L. 2018. Qgesture: Quantifying gesture distance and direction with WiFi signals. *IMWUT*, 2(1): 1–23.

Zeng, Y.; Wu, D.; Gao, R.; Gu, T.; and Zhang, D. 2018. FullBreathe: Full Human Respiration Detection Exploiting Complementarity of CSI Phase and Amplitude of WiFi Signals. *IMWUT*, 2(3): 1–19.

Zeng, Y.; Wu, D.; Xiong, J.; Yi, E.; Gao, R.; and Zhang, D. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *IMWUT*, 3(3): 1–26.

Zhang, C.; and Jiao, W. 2023. Imgfi: A high accuracy and lightweight human activity recognition framework using csi image. *IEEE Sensors Journal*, 23(18): 21966–21977.

Zhang, L.; Zhang, Y.; and Zheng, X. 2020. Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices. *ACM Transactions on Intelligent Systems and Technology*, 11(3): 1–24.

Zhang, X.; Huang, J.; Yan, H.; Feng, Y.; Zhao, P.; Zhuang, G.; Liu, Z.; and Liu, B. 2025a. Wiopen: A robust wi-fi-based open-set gesture recognition framework. *IEEE Transactions on Human-Machine Systems*.

Zhang, X.; Tang, C.; Yin, K.; and Ni, Q. 2021. WiFi-based cross-domain gesture recognition via modified prototypical networks. *IEEE Internet of Things Journal*, 9(11): 8584–8596.

Zhang, Y.; Yuan, B.; Yang, Z.; Li, Z.; and Liu, X. 2023. Winn: Human gesture recognition system based on weighted KNN. *Applied Sciences*, 13(6): 3743.

Zhang, Z.; Li, K.; Tang, S.; Wei, Y.; Wang, F.; Zhou, J.; and Guo, D. 2025b. Temporal Boundary Awareness Network for Repetitive Action Counting. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4): 1–22.

Zhang, Z.; Wu, A.; and Han, Y. 2025. Style Evolving along Chain-of-Thought for Unknown-Domain Object Detection. In *CVPR*, 14225–14234.

Zheng, Y.; Zhang, Y.; Qian, K.; Zhang, G.; Liu, Y.; Wu, C.; and Yang, Z. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *MobiSys*, 313–325.