

Zero-Fi: Zero-Shot Wi-Fi-Based Human Activity Recognition via Contrastive Signal-Language Alignment

Yitong Shen^{1*}, Cheng Guo^{2*}, Peiliang Wang¹, Jingzhe Zhang¹, Yi Sheng¹, Haopeng Zhang²,
Hongfei Xue^{2†}, Yili Ren^{1†}

¹University of South Florida

²University of North Carolina at Charlotte

shen202@usf.edu, cguo3@charlotte.edu, peiliangwang@usf.edu, jingzhe@usf.edu, sheng1@usf.edu,
haopeng.zhang@charlotte.edu, hongfei.xue@charlotte.edu, yiliren@usf.edu

Abstract

Wi-Fi-based human activity recognition has advanced substantially, but most existing methods assume a closed set of activities and require labeled Wi-Fi samples for every target class, limiting their ability to recognize unseen activities. We present Zero-Fi, a contrastive signal-language alignment framework for zero-shot Wi-Fi-based human activity recognition. Zero-Fi learns unified representations from complementary Wi-Fi signal features and aligns them with the semantic representations of natural-language activity descriptions in a shared embedding space. This cross-modal alignment enables Zero-Fi to recognize new activity classes without requiring labeled Wi-Fi samples or model adaptation for those classes. Experiments on large-scale public benchmark datasets demonstrate effective zero-shot recognition of held-out activity classes, highlighting the potential of signal-language alignment to extend Wi-Fi sensing beyond predefined activity classes.

Introduction

Human activity recognition (HAR) is an important technology that automatically detects and identifies human physical movements (Kaur, Rani, and Kumar 2024). Over the decades, HAR has been extensively studied (Zhang et al. 2020) and has demonstrated broad applicability in various areas such as healthcare (Ge et al. 2022), security (Jiang et al. 2013), human-computer interaction (Kellogg, Talla, and Gollakota 2014), and smart homes (Du, Lim, and Tan 2019).

Owing to its substantial practical value, HAR has been realized through various conventional sensing modalities, including cameras (Tran et al. 2018), inertial measurement units (IMUs) (Zhang et al. 2022), wearable sensors (Ordóñez and Roggen 2016), millimeter-wave (mmWave) radar (Singh et al. 2019), and LiDAR (Luo, Poslad, and Bodanese 2020). However, each modality has inherent limitations. Specifically, IMU- and wearable-based systems require users to continuously carry or wear dedicated devices/sensors, which can be intrusive and inconvenient. Radar- and LiDAR-based systems (Cui and Dahnoun 2021; Bouazizi, Lorite Mora, and Ohtsuki 2023) often incur non-negligible hardware costs and

deployment complexity. The performance of camera-based HAR systems (Zhang et al. 2019; Liu et al. 2019) is highly sensitive to illumination conditions and may deteriorate substantially in low-light or dark environments.

In recent years, Wi-Fi has emerged as a promising sensing modality for HAR. Human movements alter the propagation paths of Wi-Fi signals through reflection, diffraction, and scattering, producing measurable variations in signals (Tan et al. 2022). These variations encode motion-related information that can be analyzed to distinguish different human activities. Compared with conventional sensing modalities, Wi-Fi enables contactless sensing and remains robust under varying ambient illumination conditions. Moreover, existing Wi-Fi infrastructure and widely available Wi-Fi devices can be repurposed for HAR, thereby reducing deployment costs. The pervasive coverage of Wi-Fi signals in indoor environments further provides a practical foundation for scalable and widely deployable HAR systems.

Owing to these advantages, Wi-Fi-based HAR has attracted growing research interest, leading to the development of numerous systems (Adib and Katabi 2013; Wang et al. 2015; Zheng et al. 2019; Xiao et al. 2021; Li et al. 2021; Zhang et al. 2026). However, most existing systems have limited ability to generalize to previously unseen activity categories. They typically rely on annotated training data and strong supervision under closed-set assumptions, and thus cannot recognize activities that are absent from the training set. This limitation reduces their practical applicability because deployed HAR systems may encounter activity categories that are absent from the training data. Given the breadth and diversity of human behavior, collecting and annotating representative Wi-Fi samples for every possible activity category is prohibitively expensive and ultimately infeasible. The ability to recognize unseen activities is therefore essential for Wi-Fi-based HAR systems.

In this paper, we present Zero-Fi, a novel framework for zero-shot Wi-Fi-based human activity recognition that can effectively recognize unseen activities. Realizing this capability, however, requires addressing three key challenges.

First, zero-shot Wi-Fi-based HAR requires learning representations from Wi-Fi signals that generalize beyond the activity categories observed during training. The key challenge

*These authors contributed equally.

†Corresponding author.

is to learn Wi-Fi representations that capture shared, transferable motion characteristics, rather than features specific to a fixed set of predefined activities. To address this challenge, we align Wi-Fi representations with the semantic space of natural language using contrastive learning, which offers a scalable interface for describing human activities (Guadarrama et al. 2013), including categories absent from the Wi-Fi training data. Unlike discrete labels that treat each activity as an independent category, language descriptions explicitly encode motion attributes, such as body parts, movement directions, and speeds, that are shared across activities. Although Wi-Fi signals and language differ substantially in form, both characterize the same underlying human motion: Wi-Fi signals capture the physical effects of body movements on wireless signal propagation, while language expresses those movements at a semantic level. Aligning the two modalities therefore enables the model to associate Wi-Fi signals with these shared motion attributes. Since unseen activities can often be characterized as new combinations of attributes already observed in seen activities, the resulting signal-language correspondence naturally extends beyond the training classes, enabling zero-shot recognition.

Second, Wi-Fi signals are inherently sensitive to environmental variations. Since the signal component reflected off the human body is often superimposed with signals reflected from surrounding objects, even minor environmental changes (e.g., furniture relocation) can cause the received signals to differ substantially for the same activity. In addition, Wi-Fi devices introduce hardware-induced phase noise, such as random phase offset and carrier frequency offset (Ma, Zhou, and Wang 2019), which further distorts the signal. The challenge is therefore to reliably extract activity-relevant features from Wi-Fi signals that remain robust to such environmental and hardware noise. To address these, we introduce a domain discriminator that encourages the feature encoder to suppress environment-specific information while preserving motion patterns. To suppress hardware-induced noise, we leverage the fact that distortions are shared across antennas on the same Wi-Fi device, and cancel them out by computing the conjugate multiplication of signals received across antennas, thereby removing the common noise component while preserving activity-relevant signal variations.

Third, existing Wi-Fi HAR datasets typically provide only simple activity labels. Directly aligning Wi-Fi representations with such brief text labels is unlikely to yield satisfactory generalization, as simple labels fail to capture detailed motion attributes and the subtle relationships among activities. For example, “answering the phone” and “combing hair” are semantically distant as class names, yet both activities involve moving the arm toward the head. The challenge is therefore to obtain semantically rich language descriptions that explicitly characterize the underlying human motion. To address this, we leverage the broad semantic knowledge encoded in large language models (LLMs) to decompose each activity label into motion attributes associated with the torso, head, arms, hands, legs, and overall trajectory. This allows us to explicitly model the semantic relationships between activities that share common motion components: although “answering the phone” and “combing hair” are distinct as

class names, their attribute-level descriptions both involve moving the arm toward the head, resulting in high similarity between their attribute embeddings.

We evaluate Zero-Fi on public datasets under a strict zero-shot setting, in which the Wi-Fi signals and language descriptions of all test activity categories are unseen during training. Results demonstrate that Zero-Fi effectively recognizes unseen activities, achieving an average accuracy of 69.58%, and consistently outperforms existing baselines. Our main contributions are summarized as follows:

- We present Zero-Fi, a contrastive framework for zero-shot Wi-Fi-based human activity recognition that aligns Wi-Fi signals with semantically rich language descriptions.
- We design mechanisms to extract robust Wi-Fi features and to construct rich, attribute-level language descriptions for each activity, enabling cross-modal alignment.
- We conduct extensive experiments on multiple public datasets, demonstrating the effectiveness and generalization of our work under strict zero-shot settings.

Related Work

Human Activity Recognition

HAR has been implemented using diverse sensing modalities (Kaur, Rani, and Kumar 2024). For example, IF-ConvTransformer (Zhang et al. 2022) fuses measurements from multiple IMU sensors and employs a convolutional Transformer for activity recognition. TSAM (Li et al. 2025a) augments a pretrained visual backbone with a sequential perceiver adapter to capture spatial and temporal features. mm-CLIP (Cao et al. 2024) aligns mmWave radar heatmaps with textual embeddings to recognize unseen activities, while Luo et al. (Luo, Poslad, and Bodanese 2020) combine LSTM and temporal convolutional networks to classify activities from LiDAR point clusters. However, these modalities face inherent limitations in deployment cost, hardware requirements, and robustness to illumination conditions. In contrast, Wi-Fi-based HAR offers a practical and scalable alternative by leveraging the ubiquitous Wi-Fi devices and signals.

Wi-Fi Sensing

Wi-Fi sensing has been widely studied for object sensing (Ren et al. 2020; Wang et al. 2024c), localization (Kotaru et al. 2015; Fan et al. 2024), pose estimation (Ren et al. 2022; Yan et al. 2024), security (Jiang et al. 2013; Lin et al. 2023), healthcare (Ge et al. 2022; Wang et al. 2024b), smart homes (Du, Lim, and Tan 2019; Li et al. 2025c), and HAR (Li et al. 2021; Xiao et al. 2021; Li et al. 2025b; Zhang et al. 2026). For HAR, THAT (Li et al. 2021) models time-channel dependencies using a two-stream convolution-augmented Transformer, while OneFi (Xiao et al. 2021) applies meta-learning (Hospedales et al. 2021) for one-shot gesture adaptation. However, they cannot support strict zero-shot recognition. Wi-Chat (Ren et al. 2025) uses an LLM to infer activities without task-specific training, but its reliance on manually designed signal-description prompts limits its demonstrated zero-shot capability to four activity categories.

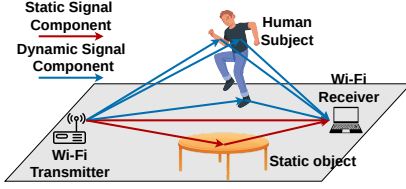


Figure 1: Wi-Fi sensing illustration.

Preliminary

Wi-Fi Sensing Basics

Wi-Fi has evolved beyond its traditional role as a communication technology to become a promising sensing modality (Tan et al. 2022). As illustrated in Figure 1, a Wi-Fi transmitter emits signals that propagate through the environment, interact with the human body through reflection, diffraction, and scattering, and are subsequently captured by a Wi-Fi receiver. Human movements alter these propagation paths, producing measurable variations in the received signals. Because different activities generate distinct signal patterns, these variations can be analyzed to infer human motion and recognize activities.

Wi-Fi Channel State Information

To capture Wi-Fi signal changes caused by a target, we utilize the Channel State Information (CSI). It describes how a transmitted signal propagates through the wireless medium and captures the effects of human movements and the environment. Wi-Fi CSI could be expressed as $H(f, t) = \sum_i^N A_i e^{-j2\pi \frac{d_i(t)}{\lambda}}$. Here, A_i is the complex attenuation, $d_i(t)$ is the length of the i^{th} path, N is the total number of paths, and λ is the wavelength. As shown in Figure 1, the received Wi-Fi signal can be divided into static and dynamic components. Specifically, static components are the signal reflections from static objects (e.g., furniture) and dynamic components are the signals reflected off the dynamic human body. Thus, CSI can be further expressed as $H(f, t) = H_s(f, t) + H_{dyn}(f, t) = H_s(f, t) + a(f, t)e^{-j2\pi \frac{d(t)}{\lambda}}$, where $H_s(f, t)$ and $H_{dyn}(f, t)$ denote the static and dynamic components, respectively. $a(f, t)$ represents the amplitude attenuation of the dynamic component, $e^{-j2\pi \frac{d(t)}{\lambda}}$ denotes the corresponding phase, $d(t)$ is the propagation path length of the dynamic component, and λ is the signal wavelength.

Methodology

As shown in Figure 2, Zero-Fi comprises signal, text, and signal-text contrastive learning modules.

Signal Module

Activity Feature Extractor. Commodity Wi-Fi devices introduce hardware noise including random phase offset (RPO) and carrier frequency offset (CFO) (Kotaru et al. 2015; Li et al. 2016). In particular, RPO and CFO induce time-varying phase offsets that distort the temporal phase variations of CSI, thereby degrading the fidelity of the extracted signal

representations. Accordingly, Wi-Fi CSI can be modeled as $H(f, t) = e^{-j\theta}(H_s(f, t) + a(f, t)e^{-j2\pi \frac{d(t)}{\lambda}})$, where $e^{-j\theta}$ represents the aggregation of these phase offsets. Because these phase offsets are the same across the antennas of the same Wi-Fi device, conjugate multiplication of CSIs between antennas can be applied to suppress them:

$$\begin{aligned} H_{cm}(f, t) &= H_1(f, t)\bar{H}_2(f, t) \\ &= (e^{-j\theta}(H_{1,s}(f, t) + a_1(f, t)e^{-j2\pi \frac{d_1(t)}{\lambda}})) \\ &\quad (e^{j\theta}(\bar{H}_{2,s}(f, t) + a_2(f, t)e^{j2\pi \frac{d_2(t)}{\lambda}})) \\ &= H_{1,s}(f, t)\bar{H}_{2,s}(f, t) \quad \dots \quad \textcircled{1} \\ &\quad + a_1(f, t)a_2(f, t)e^{-j2\pi \frac{d_1(t)-d_2(t)}{\lambda}} \quad \dots \quad \textcircled{2} \\ &\quad + \bar{H}_{2,s}(f, t)a_1(f, t)e^{-j2\pi \frac{d_1(t)}{\lambda}} \quad \dots \quad \textcircled{3} \\ &\quad + H_{1,s}(f, t)a_2(f, t)e^{j2\pi \frac{d_2(t)}{\lambda}} \quad \dots \quad \textcircled{4} \end{aligned} \quad (1)$$

Here, $H_{cm}(f, t)$ denotes the result of conjugate multiplication, while $H_1(f, t)$ and $\bar{H}_2(f, t)$ denote the CSI obtained from one antenna and the complex-conjugated CSI obtained from another antenna, respectively. Term $\textcircled{1}$ is motion-invariant and removed by high-pass filtering. Term $\textcircled{2}$ has a small magnitude and is therefore negligible, and terms $\textcircled{3}$ and $\textcircled{4}$ retain the dominant motion-induced variations. The denoised CSIs are subsequently used for feature extraction.

We then extract three complementary representations from Wi-Fi signals: Doppler frequency shift (DFS), phase, and amplitude. DFS captures the time-frequency dynamics of human motion, phase preserves fine-grained activity variations, and amplitude reflects coarse-grained motion patterns (Qian et al. 2017; Ren et al. 2020). For DFS extraction, we first compute the maximum mean-to-variance ratio of the amplitude for each transmit-receive antenna pair’s CSI. The CSIs with higher amplitude values and lower amplitude variance tend to contain less dynamic information (Qian et al. 2017). Therefore, removing these CSIs helps retain signals with richer motion-related dynamics. The Short-Time Fourier Transform (STFT) is subsequently applied to obtain the time-frequency representation, from which the frequency bins within the target DFS range are retained (Zheng et al. 2019). An additional ℓ_1 normalization step is applied to the extracted DFS representations to reduce scale differences across samples. The signal phase can be extracted using the $\text{angle}(\cdot)$ function as follows $H_P = \text{unwrap}(\text{angle}(H_{cm}(f, t)))$. Because phase measurements are periodic with a period of 2π , the $\text{unwrap}(\cdot)$ function is applied to eliminate discontinuities caused by phase wrapping. Then, a sliding-window smoothing filter is applied to H_P to reduce residual fluctuations and improve the stability of the extracted phase. The signal amplitude H_A can be calculated as $H_A = \text{abs}(H(f, t))$. We also apply a sliding-window smoothing filter to H_A to improve the stability of the amplitude. The extracted Wi-Fi representations for human activities are shown in Figure 3.

Domain Discriminator. CSIs are affected by both dataset-specific acquisition conditions and environment-specific characteristics. Because these factors are often coupled in public Wi-Fi sensing datasets, we treat each unique dataset-environment configuration as an independent domain. To

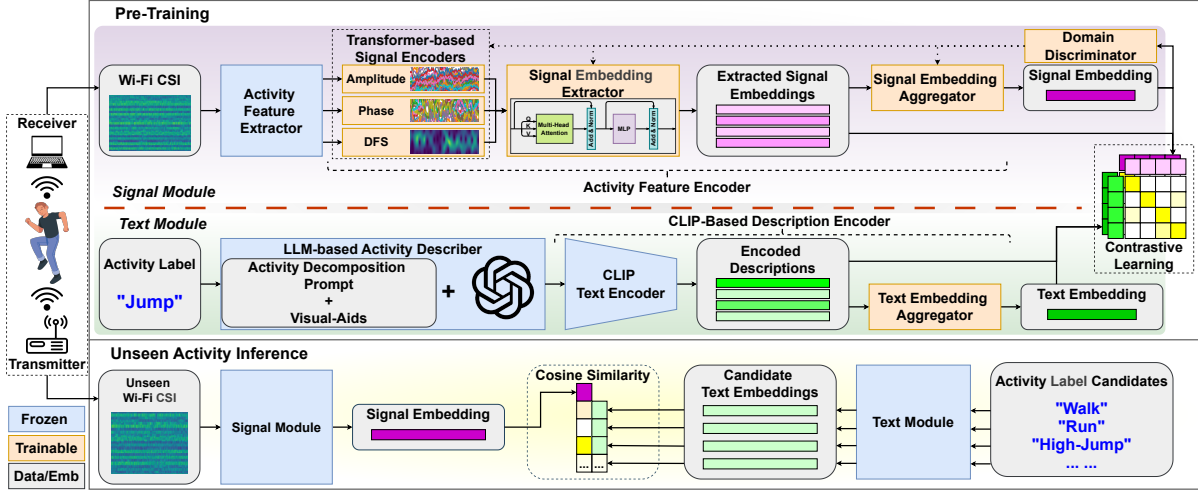


Figure 2: Overview of Zero-Fi.

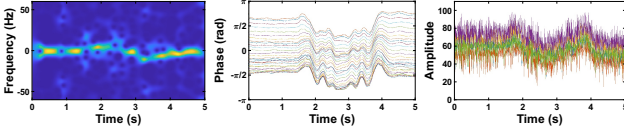


Figure 3: DFS, phase, and amplitude features.

mitigate the influence of configuration-specific characteristics, we introduce a domain discriminator that encourages the activity feature encoder in Figure 2 to learn domain-invariant activity representations. Each training sample x_i is therefore associated with a domain label $d_i \in \{1, \dots, N_d\}$, where N_d denotes the number of domains represented in the training data. Let $\mathbf{z}_i = F_{AFE}(x_i)$ denote the signal embedding produced by the activity feature encoder $F_{AFE}(\cdot)$. We first calculate its cosine similarity $\text{sim}(\cdot)$ to the text embedding $\mathbf{t}_c$ of each seen activity class: $\mathbf{q}_i = \text{softmax}(\exp(\tau) \text{sim}(\bar{\mathbf{z}}_i, \bar{\mathbf{t}}_c))$, where $\bar{\mathbf{z}}_i$ and $\bar{\mathbf{t}}_c$ denote ℓ_2 -normalized signal and text embeddings, respectively, τ is the learnable log-temperature parameter, and $\mathbf{q}_i$ is the predicted activity distribution over the seen classes. Only descriptions of the seen activities are used to calculate $\mathbf{q}_i$ during training.

The domain discriminator is conditioned on both the signal representation and its predicted activity distribution (Zhao et al. 2017): $\hat{\mathbf{p}}_i^d = D_\phi(\text{GRL}_{\gamma_d}(\mathbf{z}_i) \oplus \text{sg}(\mathbf{q}_i))$, where $\text{GRL}_{\gamma_d}(\cdot)$ denotes a gradient-reversal layer with coefficient γ_d , $\text{sg}(\cdot)$ denotes the stop-gradient operation, $\oplus$ denotes concatenation, and D_ϕ denotes the domain discriminator. Conditioning the discriminator on the activity prediction helps account for activity-dependent differences between domains, while stopping the gradient through $\mathbf{q}_i$ prevents the domain objective from directly altering the activity predictions. The discriminator is trained using the domain-classification loss

$$\mathcal{L}_{\text{domain}} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^{N_d} \mathcal{I}[d_i = j] \log \hat{p}_{i,j}^d, \quad (2)$$

where B denotes the batch size, $\mathcal{I}[\cdot]$ is the indicator function,

and $\hat{p}_{i,j}^d$ is the predicted probability that sample i belongs to dataset-environment configuration j . We use domain-balanced mini-batches to reduce the influence of domain imbalance. The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{domain}}. \quad (3)$$

During backpropagation, the gradient-reversal layer multiplies the gradient from $\mathcal{L}_{\text{domain}}$ to the activity feature encoder by $-\gamma_d$. Thus, the domain discriminator is optimized to correctly identify the dataset-environment configuration, whereas the activity feature encoder is optimized to make this prediction difficult while preserving signal-text alignment. This adversarial objective encourages the learned signal representations to become less dependent on configuration-specific acquisition and propagation characteristics.

Activity Feature Encoder. We design three distinct, representation-specific Transformer-based signal encoders, one for each extracted CSI representation: amplitude, phase, and DFS. For the amplitude and phase encoders, the two-dimensional convolutional layer commonly adopted in standard Vision Transformer (ViT) architectures (Dosovitskiy et al. 2020) is replaced with a one-dimensional convolutional layer. This design captures the temporal structure of the amplitude and phase representations while preserving dependencies among their constituent measurements, which contain discriminative activity-related information. Specifically, given the amplitude and phase inputs $H_A, H_P \in \mathcal{R}^{t \times N_{sc}}$, where t denotes the temporal length and N_{sc} denotes the number of subcarriers, the corresponding encoders apply one-dimensional convolutional patch embedding, temporal positional encoding, and self-attention to each representation. The resulting embeddings for the encoded amplitude and phase representations are denoted by $\text{Emb}_A, \text{Emb}_P \in \mathcal{R}^{[t/s] \times d_{SE}}$, respectively, where s denotes the temporal downsampling factor and d_{SE} denotes the output feature dimension of the signal encoders. For the DFS encoder, we retain the standard ViT architecture because the STFT-derived DFS jointly captures both temporal and frequency-domain features. For the DFS input

$H_D \in \mathbb{R}^{t \times N_f \times 1}$, where N_f denotes the number of frequency bins, a two-dimensional convolutional patch embedding with a window size of $w \times h$ is applied. The result is projected to an embedding $Emb_D \in \mathbb{R}^{\lfloor t/w \rfloor \times d_{SE}}$.

The encoded amplitude, phase, and DFS embeddings have identical dimensions, $\mathbb{R}^{T \times d_{SE}}$, where $T = \lfloor t/w \rfloor$ denotes the temporal length. All three representations preserve temporal information, and the amplitude and phase are inherently sequential; their encoded embeddings are concatenated along the feature dimension while maintaining temporal correspondence across representations. A learnable positional encoding $PE(\cdot)$ is subsequently applied along the temporal dimension to capture temporal dependencies: $Emb = PE(Emb_D \oplus Emb_P \oplus Emb_A)$. Here, Emb denotes the concatenated embedding, and $\oplus$ denotes concatenation along the feature dimension. Consequently, the resulting embedding Emb has dimensions $(T, 3d_{SE})$. Next, a signal embedding extractor (SEE) is used to learn shared representations from the concatenated modality embeddings. The SEE employs multi-head self-attention to model cross-modal interactions before forwarding the resulting features to the subsequent aggregation stage. The attention output is further processed using layer normalization $LN(\cdot)$ and a multilayer perceptron $MLP(\cdot)$, together with a residual connection: $z = Attention(Q, K, V)$, $Emb_{SEE} = MLP(LN(z)) + z$. Here, the query Q , key K , and value V are obtained through separate linear projections of the concatenated embedding Emb . Finally, we employ a single transformer layer, termed the signal embedding aggregator (SEA), to further aggregate and refine the representations produced by the SEE: $Emb_{signal} = Linear(SEA(Emb_{SEE}))$. Here, $Linear(\cdot)$ denotes a linear projection layer. The resulting representation Emb_{signal} serves as the final CSI-derived embedding.

Text Module

LLM-Based Activity Descriptor. Effective signal-language alignment requires richer semantic supervision than that provided by a single activity label. We therefore employ an LLM, pretrained on large-scale text corpora, to construct the activity descriptor. The architecture of this component is illustrated in Figure 4. To constrain and standardize its output, we design a structured prompt that instructs the LLM to expand each activity label into a detailed description of body-part movements, including those of the head, arms, hands, torso, and legs, as well as the overall movement direction and trajectory. Specifically, the activity label and the structured prompt are jointly provided to the LLM to generate the corresponding activity description. To further improve descriptive accuracy, we exploit the multimodal reasoning capabilities of contemporary LLMs by supplying relevant visual references, such as images and videos, when available. These descriptions provide semantically informative representations of both seen and unseen daily activities, thereby facilitating subsequent alignment with the corresponding signal features.

CLIP-Based Description Encoder. Because language descriptions and Wi-Fi CSI measurements are not inherently aligned, their representations must be projected into a shared latent space. CLIP (Radford et al. 2021) was originally trained to align textual and visual representations through

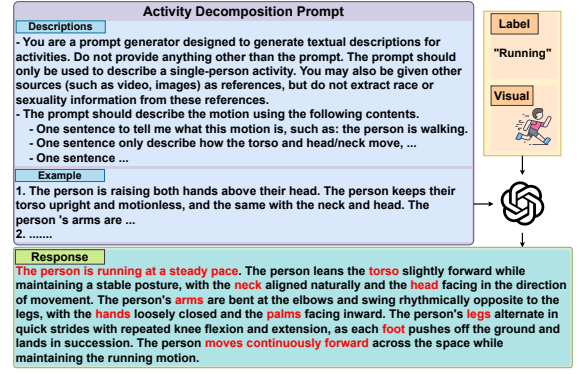


Figure 4: LLM-based activity descriptor.

contrastive learning. Its text encoder therefore captures visually grounded semantic information related to objects, actions, and human motion, which can provide a useful supervisory representation for activity-related signal features. Accordingly, we employ the frozen CLIP text encoder to encode the generated activity descriptions for subsequent contrastive alignment with the CSI representations. The resulting text embeddings are denoted by Emb_{t-enc} . In addition, we introduce a single transformer layer, termed the Text Embedding Aggregator (TEA), to aggregate and refine the embeddings of multiple descriptions into a unified representation, denoted by Emb_{t-text} , for downstream tasks.

Signal-Text Alignment

The objective of this stage is to train the overall framework to produce semantically meaningful signal embeddings and align them with the corresponding textual representations. This alignment is achieved through contrastive learning, which increases the similarity between positive signal-text pairs associated with the same activity while decreasing the similarity between negative pairs associated with different activities. Specifically, we employ the InfoNCE objective (Radford et al. 2021) to encourage bidirectional alignment between the signal and textual embeddings. The signal embedding Emb_{signal} and text embedding Emb_{text} are first ℓ_2 -normalized. Their pairwise similarities are computed as $\ell_{i,j} = \exp(\tau) \text{sim} \left(\frac{Emb_{signal}^i}{\|Emb_{signal}^i\|_2}, \frac{Emb_{text}^j}{\|Emb_{text}^j\|_2} \right)$, where τ is a learnable log-temperature parameter, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and $\ell_{i,j}$ denotes the similarity between the i th signal sample and the j th text sample. The resulting similarity matrix has dimensions $B \times B$. To reduce the occurrence of false-negative pairs caused by multiple samples from the same activity class, we employ class-aware sampling to maximize activity-class diversity within each batch. Consequently, most batches contain at most one sample from each activity class, although duplicate classes may occasionally occur because of sampling constraints.

Cross-entropy losses are then computed in both alignment directions: text-to-signal and signal-to-text. The overall contrastive loss is defined as

$$L_{con} = \frac{1}{2}(aL_{CE,t2s} + bL_{CE,s2t}), \quad (4)$$

where $L_{CE,t2s}$ and $L_{CE,s2t}$ denote the text-to-signal and signal-to-text contrastive losses, respectively. The weighting coefficients a and b are set to 1 by default. Each directional loss is implemented using cross-entropy: $L_{CE} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B y_{i,j} \log \hat{y}_{i,j}$, where B denotes the batch size, $y_{i,j} = 1$ when the i th signal and j th text form the paired sample, i.e., $i = j$, and $y_{i,j} = 0$ otherwise. The predicted probability $\hat{y}_{i,j}$ is obtained by applying softmax along the corresponding signal-to-text or text-to-signal direction of the similarity matrix. The contrastive loss is applied not only to the final aggregated embeddings but also to the intermediate signal and text embeddings, Emb_{SEE} and Emb_{t-enc} , respectively. This intermediate-level supervision enables the signal encoders to learn more directly from the textual embeddings. Through bidirectional contrastive supervision at both the intermediate and final embeddings, the model receives semantic guidance from both the less-processed features and the refined embeddings.

Experiment

Datasets

We evaluate our method on two of the largest Wi-Fi datasets, WiDAR 3.0 (Zheng et al. 2019) and XRF55 (Wang et al. 2024a), which are combined for training and zero-shot testing. **WiDAR 3.0** is one of the largest publicly available datasets for human activity recognition. It contains approximately 270,000 samples covering 22 activities, 17 participants, and 3 environments. **XRF55** is a large-scale and complex dataset for Wi-Fi-based human activity recognition. It contains 128,700 samples covering 55 human daily activities, 31 participants, and 4 environments. In our experiments, we retain all XRF55 activities and remove two overlapping activity classes from WiDAR 3.0, resulting in a total of 75 activity classes.

Data Processing

We follow the procedure described in the activity feature extractor to preprocess the CSI measurements and extract the corresponding signal representations using MATLAB. To enable batch processing and ensure consistent model inputs, all samples are resampled to a fixed temporal length of 1,000 Wi-Fi packets before being fed into the activity feature encoder. After preprocessing and flattening, the amplitude and phase representations have dimensions of 1000×90 , whereas the DFS representation has dimensions of $1000 \times 121 \times 1$.

Model and Environment Settings

As described in Section , each signal encoder comprises two ViT-style Transformer layers, whereas the signal embedding extractor contains four self-attention blocks. The signal embedding aggregator and text embedding aggregator each consist of a single Transformer layer. The signal embedding aggregator includes an additional linear projection layer. Both the signal and text modules produce 512-dimensional output embeddings. For the LLM-based activity describer, we employ ChatGPT 5.5 as the backbone model. The visual aids are provided with the original datasets and are used only as class-level references for generating activity descriptions.

During pretraining, the model is optimized for 60,000 iterations using Adam (Kingma and Ba 2014) with a learning rate of 10^{-4} and a batch size of 32. Training requires approximately 7 hours. The framework is implemented in PyTorch 2.0.2 with Python 3.9.18 and CUDA 11.8. All experiments are conducted on a Linux 24.04 platform equipped with an Intel(R) Core(TM) i9-13900 CPU, an NVIDIA GeForce RTX 4090 GPU with 24 GB of video memory.

Zero-Shot Evaluation Protocol

We evaluate all zero-shot-compatible methods under a class-disjoint protocol. Specifically, in each trial, the activity categories are partitioned into a seen-class set and an unseen-class set, with no overlap between them. All model parameters are trained exclusively using Wi-Fi samples from the seen classes. Wi-Fi samples from the unseen classes are used only during final evaluation and are not involved in model training, hyperparameter selection, receiver selection, or checkpoint selection. Additionally, no classifier is trained or adapted using unseen-class Wi-Fi samples. During inference, the descriptions of the unseen activities are introduced as fixed candidate semantic representations, and only to classify the held-out Wi-Fi samples. Participants and environments are not treated as additional holdout factors in the current evaluation; therefore, the reported results specifically measure generalization across activity categories rather than cross-subject or cross-environment generalization. We designate seven activity categories out of all 75 classes as unseen classes. The unseen categories are randomly selected in six independent trials.

We use classification accuracy as the evaluation metric for the unseen-activity classification task, measuring the proportion of unseen activity samples that are correctly classified. To reduce the influence of an exceptionally easy or difficult class partition, we report a one-sample-per-tail trimmed mean: for each method, the highest and lowest split accuracies are removed, and the remaining accuracies are averaged.

Baselines

We compare Zero-Fi with five representative baselines on the Wi-Fi-based HAR task. The baselines include **THAT** (Li et al. 2021), a Transformer-based HAR system under supervised learning that achieves strong recognition performance. **OneFi** (Xiao et al. 2021), a Transformer-based meta-learning framework designed to recognize previously unseen activities in a one-shot setting. **CLAR** (Xiao et al. 2024), a diffusion-based contrastive learning system for HAR. **FM-ZSL-IoT** (Xue et al. 2024) and **Wi-CLIP** (Zhang et al. 2025), both employ Transformer-based contrastive learning for zero-shot Wi-Fi-based HAR.

For all baselines, we implement the architectures and training procedures based on their original papers and publicly released code, when available. Necessary modifications are made to accommodate our datasets and zero-shot evaluation protocol. For baselines that natively support signal-language zero-shot inference, i.e., FM-ZSL-IoT and Wi-CLIP, we retain their original semantic-inference procedures while adopting our activity splits and input data settings.

THAT, OneFi, and CLAR do not natively provide a mechanism for predicting unseen activity labels. They demonstrate that potentially transferable signal representations are insufficient for semantic zero-shot recognition when the system lacks a mechanism for associating unseen signal patterns with unseen activity labels. Nevertheless, we pretrain their original architectures on the seen activities, use the resulting models as frozen signal encoders, and attach untrained, randomly initialized classifiers whose output dimensions correspond to the held-out activity classes. No Wi-Fi sample from unseen activities is used to optimize these classifiers.

Results

Method	Accuracy (%)
THAT (Li et al. 2021)	16.60
OneFi (Xiao et al. 2021)	11.33
CLAR (Xiao et al. 2024)	15.79
FM-ZSL-IoT (Xue et al. 2024)	23.86
Wi-CLIP (Zhang et al. 2025)	26.79
Zero-Fi	69.58

Table 1: Performance comparisons on unseen activities.

Comparisons with Baselines. Following the zero-shot evaluation protocol described above, all methods are evaluated on identical seen-unseen activity-class splits. The trimmed mean accuracy across these splits is reported in Table 1. Among the methods that directly support zero-shot inference, Zero-Fi achieves the best performance, with an accuracy of 69.58%, substantially outperforming Wi-CLIP and FM-ZSL-IoT. The performance of THAT, OneFi, and CLAR under this setting reflects the absence of a learned correspondence between their extracted signal representations and the held-out activity labels. Zero-Fi achieves this performance by explicitly aligning Wi-Fi signal representations with natural-language activity descriptions. Although FM-ZSL-IoT achieves an accuracy of 23.86%, its fine-tuning stage relies on GAN-generated synthetic data, which may introduce distributional discrepancies and artifacts that limit further performance improvements (Gong et al. 2025). Wi-CLIP (Zhang et al. 2025) achieves an accuracy of 26.79%. However, its reliance on relatively uninformative single-sentence descriptions and a BERT-based text encoder that is not explicitly optimized for signal-language alignment (Koroteev 2021) limits its zero-shot inference capability.

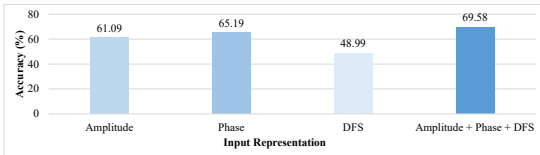


Figure 5: Ablation study: impact of signal representations.

Ablation Studies. We first evaluate the contribution of individual CSI representations and their combinations through an ablation study. As shown in Figure 5, amplitude achieves

an accuracy of 61.09%, phase achieves the highest accuracy among the individual representations at 65.19%, and DFS achieves an accuracy of 48.99%. Combining all three representations yields the best accuracy of 69.58%, improving upon the strongest individual representation by 4.39%. Amplitude and phase jointly capture the overall characteristics of human motion, while DFS contributes complementary frequency-domain information that further improves zero-shot recognition accuracy for unseen activities. Together, amplitude, phase, and DFS provide complementary motion information, yielding a more comprehensive representation that improves zero-shot recognition of unseen activities.

Text format	Accuracy (%)
Simple Activity Label	48.49
Language Description	69.58

Table 2: Ablation study: LLM-based activity describer.

We further evaluate the effectiveness of the LLM-based activity describer. Specifically, we directly feed simple activity labels, such as “walking” and “running,” into the CLIP-based text encoder while keeping all other model components unchanged. As shown in Table 2, replacing the generated activity descriptions with simple labels leads to a substantial decrease in unseen-activity classification accuracy. This result indicates that the enriched descriptions provide more informative semantic supervision by explicitly characterizing activity-related motion attributes, thereby improving the model’s ability to transfer to held-out activity classes.

Domain Discriminator	Accuracy (%)
Off	63.68
On	69.58

Table 3: Ablation study: domain discriminator.

Moreover, we evaluate the contribution of the domain discriminator by removing it while keeping all other experimental settings and hyperparameters unchanged. As shown in Table 3, disabling the domain discriminator reduces the recognition accuracy from 69.58% to 63.68%. This result demonstrates that the domain discriminator improves unseen-activity recognition, suggesting that reducing dataset and environment-specific variations benefits the learned activity embeddings.

Conclusion

In this work, we propose Zero-Fi, a contrastive signal-language alignment framework for zero-shot Wi-Fi-based human activity recognition. The framework leverages semantically rich textual representations and aligns them with CSI-derived signal embeddings in a shared latent space, thereby enabling the recognition of previously unseen activity categories without requiring labeled samples from those classes. Extensive experiments demonstrate that Zero-Fi achieves strong zero-shot recognition performance and exhibits robust generalization to unseen activities.

References

- Adib, F.; and Katabi, D. 2013. See through walls with WiFi! In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, 75–86.
- Bouazizi, M.; Lorite Mora, A.; and Ohtsuki, T. 2023. A 2D-Lidar-equipped unmanned robot-based approach for indoor human activity detection. *Sensors*, 23(5): 2534.
- Cao, Q.; Xue, H.; Liu, T.; Wang, X.; Wang, H.; Zhang, X.; and Su, L. 2024. mmclip: Boosting mmwave-based zero-shot har via signal-text alignment. In *Proceedings of the 22nd ACM conference on embedded networked sensor systems*, 184–197.
- Cui, H.; and Dahnoun, N. 2021. High precision human detection and tracking using millimeter-wave radars. *IEEE Aerospace and Electronic Systems Magazine*, 36(1): 22–32.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Du, Y.; Lim, Y.; and Tan, Y. 2019. A novel human activity recognition and prediction in smart home based on interaction. *Sensors*, 19(20): 4474.
- Fan, N.; Tian, Z.; Dubey, A.; Deshmukh, S.; Murch, R.; and Chen, Q. 2024. Multitarget device-free localization via cross-domain Wi-Fi RSS training data and attentional prior fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 91–99.
- Ge, Y.; Taha, A.; Shah, S. A.; Dashtipour, K.; Zhu, S.; Cooper, J.; Abbasi, Q. H.; and Imran, M. A. 2022. Contactless WiFi sensing and monitoring for future healthcare-emerging trends, challenges, and opportunities. *IEEE Reviews in Biomedical Engineering*, 16: 171–191.
- Gong, C.; Liang, B.; Gao, W.; and Xu, C. 2025. Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 209–222.
- Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, 2712–2719.
- Hospedales, T.; Antoniou, A.; Micaelli, P.; and Storkey, A. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169.
- Jiang, Z.; Zhao, J.; Li, X.-Y.; Han, J.; and Xi, W. 2013. Rejecting the attack: Source authentication for wi-fi management frames using csi information. In *2013 Proceedings IEEE INFOCOM*, 2544–2552. IEEE.
- Kaur, H.; Rani, V.; and Kumar, M. 2024. Human activity recognition: A comprehensive review. *Expert Systems*, 41(11): e13680.
- Kellogg, B.; Talla, V.; and Gollakota, S. 2014. Bringing gesture recognition to all devices. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, 303–316.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Koroteev, M. V. 2021. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.
- Kotaru, M.; Joshi, K.; Bharadia, D.; and Katti, S. 2015. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM conference on special interest group on data communication*, 269–282.
- Li, B.; Cui, W.; Wang, W.; Zhang, L.; Chen, Z.; and Wu, M. 2021. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 286–293.
- Li, B.; Liu, M.; Wang, G.; and Yu, Y. 2025a. Frame order matters: A temporal sequence-aware model for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18218–18226.
- Li, R.; Deng, T.; Feng, S.; Sun, M.; and Jia, J. 2025b. Consense: Continually sensing human activity with wifi via growing and picking. In *Proceedings of the AAAI Conference on Artificial intelligence*, volume 39, 14292–14300.
- Li, S.; Liu, Z.; Lv, Q.; Zou, Y.; Zhang, Y.; and Zhang, D. 2025c. WiLife: Long-term daily status monitoring and habit mining of the elderly leveraging ubiquitous Wi-fi signals. *ACM Transactions on Computing for Healthcare*, 6(1): 1–29.
- Li, X.; Li, S.; Zhang, D.; Xiong, J.; Wang, Y.; and Mei, H. 2016. Dynamic-MUSIC: Accurate device-free indoor localization. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 196–207.
- Lin, C.; Wang, P.; Ji, C.; Obaidat, M. S.; Wang, L.; Wu, G.; and Zhang, Q. 2023. A contactless authentication system based on WiFi CSI. *ACM Transactions on Sensor Networks*, 19(2): 1–20.
- Liu, Y.; Lu, Z.; Li, J.; Yang, T.; and Yao, C. 2019. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29: 3168–3182.
- Luo, F.; Poslad, S.; and Bodanese, E. 2020. Temporal convolutional networks for multiperson activity recognition using a 2-d lidar. *IEEE Internet of Things Journal*, 7(8): 7432–7442.
- Ma, Y.; Zhou, G.; and Wang, S. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)*, 52(3): 1–36.
- Ordóñez, F. J.; and Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1): 115.
- Qian, K.; Wu, C.; Zhou, Z.; Zheng, Y.; Yang, Z.; and Liu, Y. 2017. Inferring motion direction using commodity Wi-Fi for interactive exergames. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 1961–1972.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Ren, Y.; Tan, S.; Zhang, L.; Wang, Z.; Wang, Z.; and Yang, J. 2020. Liquid level sensing using commodity wifi in a smart home environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1): 1–30.
- Ren, Y.; Wang, Z.; Wang, Y.; Tan, S.; Chen, Y.; and Yang, J. 2022. GoPose: 3D human pose estimation using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2): 1–25.
- Ren, Y.; Zhang, H.; Yuan, H.; Zhang, J.; and Shen, Y. 2025. Wi-Chat: Large Language Model-powered Wi-Fi-based Human Activity Recognition. In *Proceedings of the International Workshop on Environmental Sensing Systems for Smart Cities*, 8–13.
- Singh, A. D.; Sandha, S. S.; Garcia, L.; and Srivastava, M. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 51–56.
- Tan, S.; Ren, Y.; Yang, J.; and Chen, Y. 2022. Commodity WiFi sensing in ten years: Status, challenges, and opportunities. *IEEE Internet of Things Journal*, 9(18): 17832–17843.
- Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6450–6459.
- Wang, F.; Lv, Y.; Zhu, M.; Ding, H.; and Han, J. 2024a. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1): 1–34.
- Wang, M.; Huang, J.; Zhang, X.; Liu, Z.; Li, M.; Zhao, P.; Yan, H.; Sun, X.; and Dong, M. 2024b. Target-oriented wifi sensing for respiratory healthcare: from indiscriminate perception to in-area sensing. *IEEE Network*, 39(5): 201–208.
- Wang, W.; Liu, A. X.; Shahzad, M.; Ling, K.; and Lu, S. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*, 65–76.
- Wang, X.; Wang, J.; Niu, K.; Xiong, J.; Zhang, F.; Yi, E.; Yu, A.; Yao, Z.; and Zhang, D. 2024c. Wi2DMeasure: WiFi-based 2D object size measurement. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 253–266.
- Xiao, C.; Han, Y.; Yang, W.; Hou, Y.; Shi, F.; and Chetty, K. 2024. Diffusion-model-based contrastive learning for human activity recognition. *IEEE Internet of Things Journal*, 11(20): 33525–33536.
- Xiao, R.; Liu, J.; Han, J.; and Ren, K. 2021. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 206–219.
- Xue, D.; Fan, X.; Chen, T.; Lan, G.; and Song, Q. 2024. Leveraging foundation models for zero-shot IoT sensing. *arXiv preprint arXiv:2407.19893*.
- Yan, K.; Wang, F.; Qian, B.; Ding, H.; Han, J.; and Wei, X. 2024. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 969–978.
- Zhang, H.; Guo, Y.; Wang, Z.; Sun, Z.; Guo, B.; and Yu, Z. 2025. Wi-CLIP: Toward Zero-Shot Air Gesture Recognition Based on RF-Text Foundation Model. In *International Conference on Artificial Intelligence of Things and Systems*, 174–189. Springer.
- Zhang, H.-B.; Zhang, Y.-X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.-X.; and Chen, D.-S. 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5): 1005.
- Zhang, R.; Tang, S.; Yan, H.; Zhang, X.; and Guo, J. 2026. Wi-CBR: Salient-aware adaptive WiFi sensing for cross-domain behavior recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 1552–1560.
- Zhang, Y.; Wang, L.; Chen, H.; Tian, A.; Zhou, S.; and Guo, Y. 2022. IF-ConvTransformer: A framework for human activity recognition using IMU fusion and ConvTransformer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2): 1–26.
- Zhang, Y.; Wang, X.; Wang, Y.; and Chen, H. 2020. Human activity recognition across scenes and categories based on CSI. *IEEE Transactions on Mobile Computing*, 21(7): 2411–2420.
- Zhao, M.; Yue, S.; Katabi, D.; Jaakkola, T. S.; and Bianchi, M. T. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International conference on machine learning*, 4100–4109. PMLR.
- 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 *Proceedings of the 17th annual international conference on mobile systems, applications, and services*, 313–325.