UNIVERSITY OF WEST LONDON
The Career University

UWL REPOSITORY

repository.uwl.ac.uk

CIG-MAE: Cross-modal information-guided masked autoencoder for self-supervised Wi-Fi sensing

Liu, Gang, Hao, Yanling and Zou, Yixuan (2026) CIG-MAE: Cross-modal information-guided masked autoencoder for self-supervised Wi-Fi sensing. IEEE Internet of Things Journal, 13 (16). 35614 - 35624.

https://doi.org/10.1109/JIOT.2026.3696504

This is the Published Version of the final output.

UWL repository link: https://repository.uwl.ac.uk/id/eprint/15773/

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Rights Retention Statement:

arXiv:2512.04723v2 [eess.SP] 19 Jul 2026

CIG-MAE: Cross-Modal Information-Guided Masked Autoencoder for Self-Supervised WiFi Sensing

Gang Liu, Yanling Hao, and Yixuan Zou

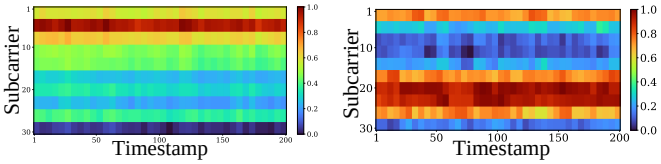
***Abstract*—Human Action Recognition using WiFi Channel State Information (CSI) has emerged as an attractive alternative to vision-based methods due to its ubiquity, device-agnostic nature, and inherent privacy-preserving capabilities. However, the high cost of manual annotation and the limited scale of publicly available CSI datasets restrict the performance of supervised approaches. Self-supervised learning (SSL) offers a promising avenue, but existing contrastive paradigms rely on data augmentations that conflict with the physical semantics of radio signals and require large-batch training, making them poorly suited for CSI. To overcome these challenges, we introduce CIG-MAE—a Cross-Modal Information-Guided Masked Autoencoder—that reconstructs both the amplitude and phase of CSI using a symmetric dual-stream architecture with a high masking ratio. Specifically, we propose an Adaptive Information-Guided Masking strategy that dynamically allocates attention to time–frequency regions with high information density to improve learning efficiency, and we incorporate a Barlow Twins regularizer to align cross-modal representations without negative samples. Experiments on three public datasets show that CIG-MAE consistently outperforms SOTA SSL methods and even surpasses a fully supervised baseline, demonstrating superior data efficiency, robustness, and representation generalization.**

***Index Terms*—Human activity recognition, WiFi sensing, self-supervised learning, masked autoencoder, cross-modal learning, Adaptive Information-Guided Masking.**

I. Introduction

Human Action Recognition (HAR) has garnered significant attention for its broad applications in health monitoring, elderly care, and smart homes [1]–[5]. Among various sensing modalities, HAR based on WiFi Channel State Information (CSI-HAR) has emerged as a particularly promising approach due to its ubiquity, device-agnostic nature, privacy-preserving characteristics, and capability to operate in non-line-of-sight (NLOS) environments [6], [7].

Despite its potential, the practical deployment of CSI-HAR is significantly hampered by the high cost and difficulty of acquiring large-scale, accurately labeled datasets. The performance of conventional supervised learning models is fundamentally constrained by data scarcity and the challenge of covering diverse subjects and environments, which can lead to poor generalization and domain shift issues [8], [9]. To mitigate this dependency on labeled data, Self-Supervised Learning (SSL) has been introduced as a powerful paradigm capable of leveraging large volumes of unlabeled CSI to learn transferable representations and enhance robustness in unseen scenarios [10], [11].



(a) Amp Reconstruction Heatmap (b) Phase Reconstruction Heatmap

Fig. 1. Reconstruction error heatmaps for amplitude and phase on the subcarrier-time plane. A larger reconstruction error (warmer color) is interpreted as higher information density. Both modalities show high difficulty on specific subcarrier indices, but their high-difficulty regions do not fully overlap, reflecting modal complementarity and non-uniform information distribution.

However, mainstream SSL paradigms, predominantly developed for computer vision and natural language processing, often prove suboptimal when directly applied to WiFi CSI data. In particular, contrastive learning depends heavily on data augmentations that lack consistency with radio propagation physics and typically requires large-batch training, which is incompatible with the small-dataset reality of WiFi sensing. These mismatches can corrupt semantic structure and undermine the ability to learn invariant features.

A key factor that has been largely overlooked is the intrinsic complementary nature of the amplitude and phase components of CSI. While much existing SSL work relies solely on amplitude [12], [13], phase has been shown to be highly sensitive to micro-motion and propagation geometry. Yet, even when phase is used, it is often treated as a unidirectional prediction target rather than a jointly learned modality, failing to capture their inherent coupling [14]. Moreover, as illustrated in Fig. 1, the information density—interpreted as reconstruction difficulty—exhibits strong non-uniformity across the time–frequency axis for both modalities, and

(Corresponding author: Yanling Hao.)

Gang Liu and Yixuan Zou are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: gangliu2004@outlook.com; yixuan.zou@qmul.ac.uk).

Yanling Hao is with the School of Computing and Engineering, University of West London, London W5 5RF, U.K. (e-mail: yanling.hao@uwl.ac.uk).

Copyright (c) 2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

critically, their high-density regions only partially overlap. This indicates that amplitude and phase provide complementary physical information and that a well-designed SSL framework must explicitly model and leverage this complementarity to obtain discriminative and physically meaningful representations.

To address these challenges, we propose the Cross-Modal Information-Guided Masked Autoencoder (CIG-MAE), a self-supervised framework tailored to the physical structure and sensing properties of CSI. CIG-MAE adopts a symmetric dual-stream architecture that jointly reconstructs amplitude and phase under a high masking ratio, eliminating the need for problematic augmentations and large batch sizes. Furthermore, we introduce an Adaptive Information-Guided Masking (AIM) mechanism that allocates visibility to time–frequency regions with higher information density, thereby enhancing representation learning efficiency and boosting model performance. The main contributions of this work are summarized as follows:

- We propose a symmetric dual-stream self-supervised reconstruction framework, CIG-MAE, tailored to the physical properties of CSI. This framework reconstructs the amplitude and phase components of CSI with a high masking ratio and introduces a non-contrastive Barlow Twins (BT) regularizer to align the dual-stream representations and eliminate feature redundancy. This design explicitly models the complementarity between amplitude and phase, effectively avoiding the reliance of existing methods on data augmentations that are inconsistent with physical semantics and on large-batch configurations.
- We design an AIM strategy to enhance data efficiency and representation quality. This strategy introduces a learnable policy network that can evaluate the information value of each time-frequency region, thereby dynamically focusing the model's "visibility budget" on regions with higher information density that are more critical for learning discriminative features, addressing the inefficiency of random masking from a mechanistic perspective.
- We conduct systematic evaluations on three public datasets. The experimental results show that under the standard linear probing protocol, the performance of the proposed CIG-MAE not only significantly surpasses various advanced self-supervised baselines but even exceeds a fully supervised model in certain scenarios. Furthermore, exhaustive ablation studies and parameter analyses have verified the effectiveness of our model's components and its overall robustness.

The remainder of this paper is organized as follows: Section II reviews related work. Section III elaborates on our proposed CIG-MAE framework. Section IV reports detailed experimental settings and results. Section V discusses broader implications and potential extensions of CIG-MAE. Finally, Section VI concludes the paper and discusses future research directions.

II. Related Work

SSL has achieved remarkable success across various domains [15]–[18] and is gradually being adopted for signal processing tasks [19]–[21]. Early research on SSL for HAR explored various pretext tasks. Saeed et al. [20] systematically designed eight different pretext tasks such as feature prediction and transformation recognition for diverse signals. Subsequently, they extended SSL to federated settings [22], proposing scalogram-signal correspondence learning via wavelet transforms. As research deepened, methods specifically tailored to the physical structure of WiFi CSI have emerged, broadly categorized into discriminative (contrastive/predictive) and generative paradigms.

A. Contrastive and Predictive Paradigms

The fundamental idea of contrastive learning is to maximize representation consistency between different views of the same sample while minimizing consistency between different samples [23]–[26]. In WiFi sensing, current works optimize this paradigm primarily through view construction and objective function design.

Regarding *view construction*, Lau et al. [27] treated CSI from spatially separated receivers as positive pairs. Liu et al. [28] introduced STFNets to mine time-frequency features, demonstrating the importance of frequency domain augmentations. Song et al. [21] focused on multi-representation consistency, utilizing a translator-predictor structure to map distinct radio frequency (RF) features like Angle of Arrival (AoA) and Doppler Frequency Shift (DFS) into a unified space. Xu et al. [29] employed a dual-stream architecture to collaboratively model spatial and channel features. Lyons et al. [30] proposed WiFiAct, which combines an environment-invariant preprocessing pipeline with a Bayesian CNN for generalized learning, though its partial reliance on labels distinguishes it from pure SSL.

Regarding *objective functions*, Yang et al. [31] proposed AutoFi, which enhances consistency by minimizing the Kullback-Leibler (KL) divergence of probability distributions alongside mutual information and geometric structure constraints. Chen et al. [13] addressed the instability of KL divergence by maximizing Jensen-Shannon (JS) divergence and introducing Gaussian regularization to improve generalization. Recently, Xiao et al. [12] incorporated diffusion models, specifically Denoising Diffusion Probabilistic Models (DDPM) to generate high-quality, temporally specific augmentations and designed adaptive weights based on activity information.

In predictive learning, the model learns representations that capture the dynamics and latent structures of the data by predicting future segments, missing context, or certain transformation properties [32], [33]. Haresamudram et al. [34]

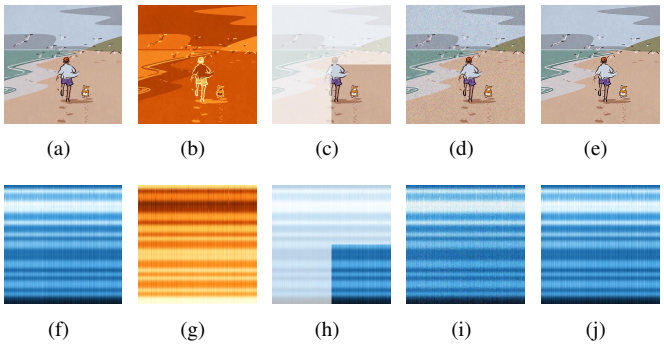


Fig. 2. Illustration of five visualization/augmentation operators applied to a natural image (top row) and to CSI data (bottom row). From left to right: Original, Color Mapping, Crop, Gaussian Noise, and Unsharp Mask. Directly transferring vision augmentations to CSI can misalign with RF physical semantics.

applied Contrastive Predictive Coding (CPC) to wearable sensors. In CSI-HAR, Barahimi et al. [35] proposed CAPC, combining CPC with BT to leverage uplink/downlink views for capturing temporal context and view consistency.

However, these discriminative paradigms face two inherent limitations in the CSI context. First, they rely heavily on view construction via augmentations (e.g., cropping, noise injection) that are often semantically inconsistent with the physical properties of radio signals [36], [37], as illustrated in Fig. 2. While solutions like using multiple receivers or specialized augmentations exist [21], [27], [35], [38], they often introduce extra hardware costs or complexity. Second, contrastive learning typically requires large batches and numerous negative samples to perform well [23], [26], a condition that large-scale image datasets like ImageNet can satisfy [39]. This is fundamentally at odds with the CSI domain, where public datasets are typically small, containing only hundreds to thousands of samples [40]. Predictive methods, which learn by predicting future or masked segments [35], often still rely on contrastive objectives and are sensitive to hyperparameters, magnifying training instability under small-sample, high-noise conditions.

B. Generative and Reconstructive Paradigms

Generative learning, which forces models to learn intrinsic structures by reconstructing masked inputs, offers a solution more suitable for small-sample scenarios with specific physical structures [41]. In general sensor domains, Cheng et al. proposed MaskCAE [19], utilizing sparse convolutions to process masked sensor data without explicit augmentation. Miao et al. [42] proposed STMAE for wearable devices, employing spatial-temporal masking and an asymmetric autoencoder to model device correlations. In WiFi-specific contexts, Yang et al. [43] proposed MaskFi, which transforms multi-modal (WiFi-Vision) data into discrete tokens via a VQ-VAE and reconstructs them using a shared Transformer. Ji et al. [44] proposed SiFall, framing fall detection as anomaly detection by using a VAE to model non-fall activities and utilizing reconstruction error as the anomaly signal. Focusing on modal correlation, Gao et al. [14] proposed AutoSen, a cross-modal autoencoder that reconstructs phase from amplitude to capture intrinsic CSI semantics.

Despite avoiding complex augmentations, existing generative methods exhibit structural shortcomings. Approaches like MaskFi rely on heavy backbones (e.g., Transformers) that are computationally prohibitive and unstable on small, noisy CSI datasets [45]. Asymmetric designs, such as AutoSen's unidirectional prediction, create an information bottleneck by assuming amplitude fully specifies phase. Moreover, standard uniform random masking ignores the non-uniform information density of CSI, potentially wasting learning budgets on redundant regions [46].

III. Methodology

To address the challenges of data scarcity and non-uniform information density in WiFi sensing, we propose CIG-MAE, a self-supervised framework tailored to the physical representation of CSI. As illustrated in Fig. 3, the proposed framework processes synchronized dual-stream CSI through a *Filter-then-Reconstruct* paradigm. The pipeline is composed of three integral modules: an *AIM* mechanism that actively identifies information-dense signal patches; a *Dual-Stream Convolutional Backbone* that reconstructs the signal from these focused regions; and a *Cross-Modal Representation Alignment* module that enforces semantic consistency between modalities. The specific design motivations and details are described below.

A. Preliminaries: Signal Formulation and Properties

WiFi CSI characterizes the wireless channel at a fine-grained level using Orthogonal Frequency Division Multiplexing (OFDM). Mathematically, the CSI for subcarrier k at time t is represented as a complex value:

$$H_k(t) = A_k(t)e^{j\phi_k(t)} \tag{1}$$

where $A_k(t)$ is the amplitude and $\phi_k(t)$ is the phase. Consequently, the input to our model is defined as a synchronized dual-stream tensor $(\mathbf{X}^A, \mathbf{X}^P)$, where each $\mathbf{X}^m \in \mathbb{R}^{N\times S\times T}$ for modality $m \in \{A, P\}$, with N, S, and T denoting antennas, subcarriers, and time steps, respectively.

This formulation underscores two critical physical properties that inform our architectural design. First, the coupling of $A_k(t)$ and $\phi_k(t)$ implies that single-stream models inherently discard complementary modal information, necessitating a **symmetric dual-stream architecture**. Second, human activities typically manifest as locally structured patterns on the time-frequency plane ($S \times T$). Unlike vision tasks, CSI is heavily contaminated by background. This renders standard random masking inefficient, motivating the introduction of an adaptive masking strategy.

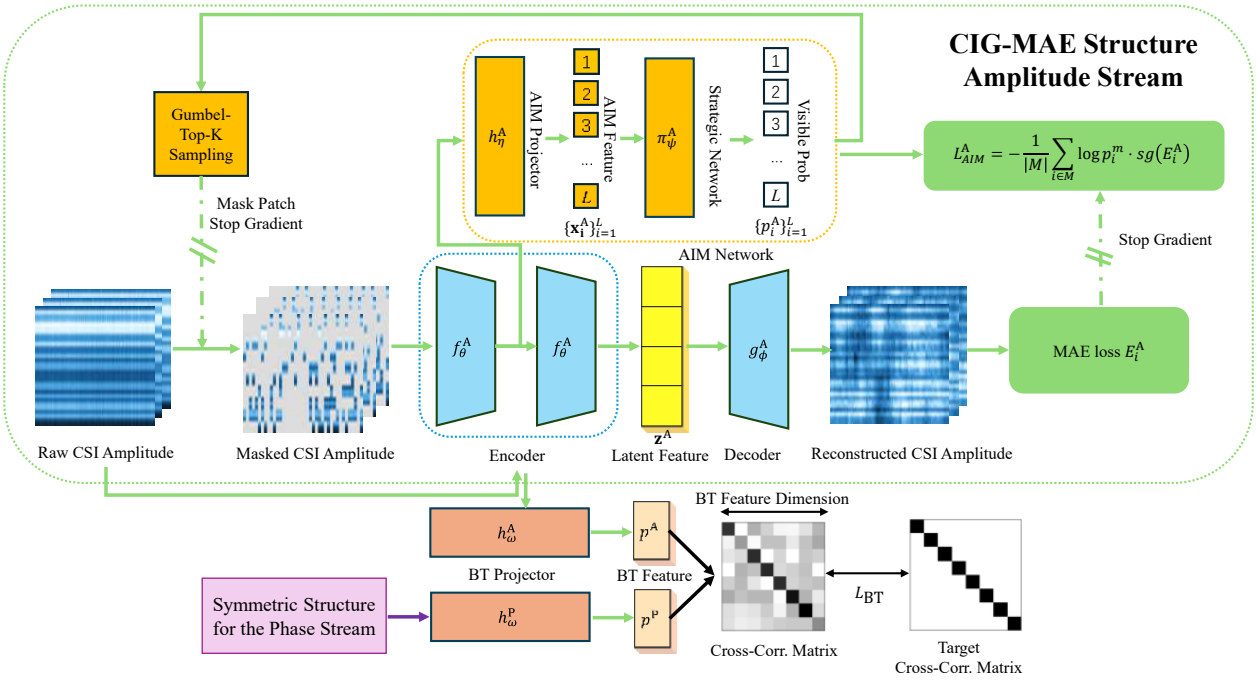


Fig. 3. Schematic overview of the CIG-MAE framework. The architecture integrates three core modules corresponding to Section III: (1) **Adaptive Information-Guided Masking (Sec. III-B)**: The policy network evaluates information density to generate non-uniform masks on the $S \times T$ plane; (2) **Dual-Stream Convolutional Backbone (Sec. III-C)**: Symmetric CNN encoder-decoder networks reconstruct the signal from masked amplitude and phase streams; (3) **Cross-Modal Representation Alignment (Sec. III-D)**: A Barlow Twins regularizer aligns unmasked representations to ensure semantic consistency.

B. Adaptive Information-Guided Masking (AIM)

Standard masked autoencoders typically employ uniform random masking. However, given the high noise floor in CSI data, such a strategy usually fails to distinguish between information-dense and redundant regions, leading to a significant waste of computational resources on meaningless reconstruction tasks [47], [48]. To mitigate this, we incorporate the AIM strategy [49] to dynamically allocate visibility to regions with high information density.

Implicit Patch Embedding. We first discretize the input $\mathbf{X}^m$ into a grid of patches. We define the patch indices as $\mathcal{L} = \{1, \ldots, L\}$, where each patch has dimensions (s_p, t_p). To avoid the computational overhead of a separate tokenizer, we strictly align the kernel size and stride of the encoder's first convolutional layer with (s_p, t_p). This yields native patch embeddings, which are projected to a policy feature space $\{\mathbf{x}_i^m\}_{i=1}^L$ via a linear layer h_η^m.

Policy Network and Probabilistic Sampling. A lightweight Policy Network π_ψ^m, implemented as a single-layer Transformer Encoder, evaluates the contextual importance of each patch. It aggregates global information via multi-head self-attention (MHA) and produces hidden states $\mathbf{U}$ via an MLP:

$$\mathbf{Z} = \{\mathbf{x}_i^m\} + \mathrm{MHA}(\mathrm{LN}(\{\mathbf{x}_i^m\})), \quad \mathbf{U} = \mathbf{Z} + \mathrm{MLP}(\mathrm{LN}(\mathbf{Z})). \tag{2}$$

The importance probability p_i^m (i.e., the probability of being **visible**) for the i-th patch is then derived via a softmax projection:

$$p_i^m = \frac{\exp(\mathbf{w}^\top \mathbf{U}_i)}{\sum_{j=1}^L \exp(\mathbf{w}^\top \mathbf{U}_j)}. \tag{3}$$

Based on $\{p_i^m\}$, we employ Gumbel-Top-K sampling to partition the patch set $\mathcal{L}$ into a masked set $\mathcal{M}$ (where $|\mathcal{M}| = \rho L$, mask ratio $\rho \in (0, 1)$) and a visible set $\mathcal{V}$. This generates the binary mask Π applied to the input: $\tilde{\mathbf{X}}^m = \mathbf{X}^m \odot \Pi$.

Policy Optimization via Reconstruction Reward. The core innovation of AIM lies in how we define *valuable* patches without ground-truth labels. We propose a hypothesis rooted in information theory: *patches that are difficult to reconstruct from their context contain higher information density (e.g., human activity), whereas patches that are easily inferable correspond to redundant background.*

Guided by this intuition, we model the masking process as learning a mask distribution induced by the policy network. Let $q_\psi(\mathcal{M} \mid \mathbf{X}^m)$ denote the distribution over the masked

set $\mathcal{M}$ generated by the Gumbel-Top-K sampler from the policy scores. For a sampled mask, we define the patch-wise reconstruction error E_i^m as the learning signal. Unlike adversarial approaches that simply mask salient regions, our goal is to identify high-error regions and increase their likelihood of being preserved in future iterations. Given the high masking ratio (e.g., 95%) in WiFi sensing, random masking risks removing the sparse activity signal entirely. By explicitly encouraging the network to select difficult-to-reconstruct regions into the visible set $\mathcal{V}$, we ensure the model focuses its limited visibility budget on capturing the discriminative "skeleton" of the signal.

Accordingly, we optimize the policy using a REINFORCE-style objective [50], [51]. Under a fixed masking budget, Eq. (4) acts as a tractable single-sample surrogate that assigns larger gradients to high-error patches:

$$\mathcal{L}_{\mathrm{AIM}}^m = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_i^m \cdot \mathrm{stopgrad}(E_i^m), \tag{4}$$

where p_i^m is the policy score used by the sampler. The $\mathrm{stopgrad}(\cdot)$ operator prevents the reconstruction backbone from artificially inflating the reward signal, so that the encoder-decoder is trained only to minimize reconstruction loss, while the policy network independently learns which regions should remain visible.

Clarification. AIM is *RL-inspired* rather than a full sequential reinforcement learning formulation. We use a REINFORCE-style score-function estimator only to optimize the discrete mask selection induced by Gumbel-Top-K sampling. AIM does not involve multi-step trajectories, delayed rewards, or an independently evolving environment. Therefore, Eq. (4) is used directly as a synchronous reconstruction-derived training signal within the same self-supervised optimization loop, and we do not employ additional RL-specific stabilization techniques such as experience replay, baseline subtraction, or reward normalization.

Theoretical Justification. We further clarify the connection between Eq. (4) and mutual information maximization. Let X denote the full input, and let X_V and X_M denote the visible and masked subsets under a sampled mask. Since $H(X)$ is constant for a given sample, maximizing the mutual information between the visible subset and the full signal is equivalent to minimizing the conditional uncertainty of the full signal given the visible subset:

$$\max I(X_V; X) = H(X) - H(X \mid X_V) \propto \min H(X \mid X_V). \tag{5}$$

Under the common patch-wise approximation, we write$H(X \mid X_V) \approx \sum_{i \in \mathcal{M}} H(X_i \mid X_V)$, where X_i denotes the i-th masked patch. Moreover, minimizing the L_1 reconstruction loss is equivalent to maximum likelihood estimation under a Laplace conditional model, $X_i \mid X_V \sim \mathrm{Laplace}(\mu_i, b_i)$, for which the conditional entropy satisfies$H(X_i \mid X_V) = 1 + \ln(2b_i)$. Since the optimal expected L_1 error is proportional to b_i, a larger reconstruction error implies a larger conditional entropy. Therefore, maximizing $I(X_V; X)$ can be relaxed into minimizing the *expected* entropy (or equivalently, reconstruction error) of the masked subset:

$$\max I(X_V; X) \Longrightarrow \min_{\psi} \mathbb{E}_{\mathcal{M} \sim q_\psi(\mathcal{M}|X)} \left[\sum_{i \in \mathcal{M}} E_i \right]. \tag{6}$$

Applying the score-function estimator to Eq. (6) yields

$$\nabla_\psi \mathcal{J}(\psi) \propto \mathbb{E}_{\mathcal{M} \sim q_\psi} \left[\left(\sum_{i \in \mathcal{M}} \mathrm{stopgrad}(E_i) \right) \nabla_\psi \log q_\psi(\mathcal{M} \mid X) \right]. \tag{7}$$

Eq. (4) is the practical single-sample surrogate of Eq. (7) under our fixed-budget Gumbel-Top-K sampling scheme. Hence, the policy is encouraged to reduce the likelihood that high-error patches are selected into the masked set, thereby increasing their tendency to be preserved in the visible set under the fixed-budget sampling scheme. In this sense, AIM does not optimize mutual information in closed form; rather, it optimizes a reconstruction-driven surrogate whose effect is to reduce $H(X \mid X_V)$ and thus increase $I(X_V; X)$.

C. Dual-Stream Convolutional Reconstruction Backbone

The core of CIG-MAE is the reconstruction of the original signal from the partial input filtered by AIM. We deliberately employ a pure Convolutional Neural Network (CNN) encoder-decoder architecture. This choice is substantiated by both physical characteristics and practical constraints:

1) **Inductive Bias Match:** CNNs capture local structural and translation-equivariant features, which align with CSI patterns [52]. In contrast, Transformer-based backbones (ViTs) model global dependencies and require massive datasets to learn such local priors [53]. However, existing CSI datasets are limited in scale, often ranging from hundreds to thousands of samples [54]–[57], with only a few reaching tens of thousands [58]. This scarcity is insufficient to train data-hungry Transformers effectively.
2) **Deployment Efficiency:** WiFi sensing applications are typically deployed on resource-constrained edge devices [59]. As empirically validated in Section IV-G, our CNN backbone requires significantly fewer FLOPs and parameters compared to ViTs, making real-time inference feasible without sacrificing accuracy.

Reconstruction Process. The masked input $\tilde{\mathbf{X}}^m$ is mapped by the encoder f_θ^m to a latent representation $\mathbf{z}^m$, and subsequently reconstructed by the decoder g_ϕ^m to yield $\hat{\mathbf{X}}^m$. To quantify quality, we compute the Mean Absolute Error (MAE) exclusively on the masked pixels $\Omega(\Pi) = \{(n, s, t) \mid \Pi(s, t) = 0\}$:

$$\mathcal{L}_{\mathrm{rec}}^m = \frac{1}{|\Omega(\Pi)|} \sum_{(n,s,t) \in \Omega(\Pi)} \left| \hat{\mathbf{X}}_{n,s,t}^m - \mathbf{X}_{n,s,t}^m \right|. \tag{8}$$

The choice of an unnormalized reconstruction target combined with the MAE function is deliberately tailored to the physical properties of CSI. First, regarding the loss metric, we employ MAE over the standard Mean Squared Error (MSE). CSI signals are prone to sporadic, high-amplitude noise spikes due to hardware imperfections and multipath interference. MSE imposes a quadratic penalty on such outliers, which can dominate the gradient and destabilize training. In contrast, MAE's linear penalty provides robustness, ensuring the model focuses on learning the underlying activity structure rather than fitting these transient artifacts. Second, regarding the target, we avoid per-sample normalization during loss computation. Normalization tends to compress the signal's dynamic range, potentially obscuring the semantic distinction between high-magnitude activity bursts and low-magnitude background fluctuations. By preserving the full dynamic range, we force the model to prioritize the reconstruction of significant, activity-induced signal variations. The empirical superiority of this design is validated in Section IV-E.

D. Cross-Modal Representation Alignment

While the dual streams independently reconstruct their respective modalities, it is crucial that their learned latent representations are semantically consistent. To achieve this without relying on negative samples, we introduce a BT regularizer [24].

To ensure the statistical stability of the correlation estimates, this alignment process is performed on the original, *unmasked* input $\mathbf{X}^m$, paralleling the reconstruction path. For each modality $m \in \{A, P\}$, the projection vector $\mathbf{p}^m$ is obtained through the cascaded mapping of the encoder f_θ^m and the BT projection head h_ω^m:

$$\mathbf{p}^m = h_\omega^m(f_\theta^m(\mathbf{X}^m)). \tag{9}$$

Within a data batch of size B, we assess the alignment between the two streams by computing the cross-correlation matrix $\mathbf{C}$:

$$\mathbf{C}_{ij} = \frac{1}{B}\sum_{b=1}^{B} \tilde{\mathbf{p}}_{b,i}^A \, \tilde{\mathbf{p}}_{b,j}^P, \tag{10}$$

where $\tilde{\mathbf{p}}$ denotes the representations after batch normalization. The BT loss is then defined to strictly penalize deviations from the identity matrix:

$$\mathcal{L}_{\text{BT}} = \sum_i (1 - \mathbf{C}_{ii})^2 + \lambda \sum_{i \neq j} \mathbf{C}_{ij}^2. \tag{11}$$

The first term (invariance) forces the diagonal elements to 1, aligning the representations of amplitude and phase. The second term (redundancy reduction) forces off-diagonal elements to 0, decorrelating the feature dimensions to maximize information content.

Algorithm 1 CIG-MAE Pre-training Procedure

Require: Data $(\mathbf{X}^A, \mathbf{X}^P)$, Backbones $\Theta = \{f_\theta, g_\phi, h_\omega\}$, Policy Nets $\Psi = \{\pi_\psi, h_\eta\}$
Ensure: Optimized Encoders f_θ^A, f_θ^P

while not converged **do**
  Sample batch $(\mathbf{X}^A, \mathbf{X}^P)$
  // Phase 1: Adaptive Masking (Sec. III-B)
  // Policy network π_ψ estimates importance; $\mathcal{M}$ is sampled via Gumbel-Top-K
  $\mathcal{M}^A \leftarrow \text{AIM}(\mathbf{X}^A, \Psi^A)$; $\mathcal{M}^P \leftarrow \text{AIM}(\mathbf{X}^P, \Psi^P)$
  // Phase 2: Dual-Stream Reconstruction (Sec. III-C)
  // Apply mask and reconstruct using Encoder f_θ and Decoder g_ϕ
  $\tilde{\mathbf{X}}^m \leftarrow \mathbf{X}^m \odot (\mathbf{1} - \mathcal{M}^m)$ for $m \in \{A, P\}$
  $\hat{\mathbf{X}}^m \leftarrow g_\phi^m(f_\theta^m(\tilde{\mathbf{X}}^m))$ for $m \in \{A, P\}$
  $\mathcal{L}_{\text{rec}} \leftarrow \text{MaskedMAE}(\hat{\mathbf{X}}^{A,P}, \mathbf{X}^{A,P}, \mathcal{M}^{A,P})$
  // Phase 3: Cross-Modal Alignment (Sec. III-D)
  // Project unmasked features via h_ω for Barlow Twins
  $\mathbf{p}^A \leftarrow h_\omega^A(f_\theta^A(\mathbf{X}^A))$; $\mathbf{p}^P \leftarrow h_\omega^P(f_\theta^P(\mathbf{X}^P))$
  $\mathcal{L}_{\text{BT}} \leftarrow \text{BTLoss}(\mathbf{p}^A, \mathbf{p}^P)$
  // Phase 4: Decoupled Parameter Optimization
  // 1. Update Policy Ψ: maximize reconstruction error of masked patches
  $E^{A,P} \leftarrow \text{PixelWiseError}(\hat{\mathbf{X}}^{A,P}, \mathbf{X}^{A,P})$
  $\mathcal{L}_{\text{AIM}} \leftarrow \text{PolicyGradient}(E^{A,P}, \mathcal{M}^{A,P}, \Psi)$
  $\Psi \leftarrow \text{Optimizer}(\nabla_\Psi(w_{\text{aim}}\mathcal{L}_{\text{AIM}}))$
  // 2. Update Backbone Θ: minimize reconstruction and redundancy
  $\Theta \leftarrow \text{Optimizer}(\nabla_\Theta(\mathcal{L}_{\text{rec}} + w_{\text{bt}}\mathcal{L}_{\text{BT}}))$
end while
return Optimized encoders f_θ^A, f_θ^P

E. Overall Training Objective

The total loss is a weighted sum of the three objectives:

$$\mathcal{L}_{\text{CIG-MAE}} = \mathcal{L}_{\text{rec}} + w_{\text{aim}}\mathcal{L}_{\text{AIM}} + w_{\text{bt}}\mathcal{L}_{\text{BT}}. \tag{12}$$

We employ a **decoupled update strategy** to distinguish the optimization targets of the reconstruction and masking-policy components. Specifically, the backbone parameters $\Theta = \{f_\theta, g_\phi, h_\omega\}$ are updated by minimizing the reconstruction and alignment losses ($\mathcal{L}_{\text{rec}} + w_{\text{bt}}\mathcal{L}_{\text{BT}}$), while the policy parameters $\Psi = \{\pi_\psi, h_\eta\}$ are updated solely by the policy gradient loss $\mathcal{L}_{\text{AIM}}$. This separation ensures that the policy network independently learns to identify information-dense regions without interference from the reconstruction objective. The complete pre-training procedure is summarized in Algorithm 1.

IV. Experiments

This section systematically evaluates the proposed CIG-MAE framework. First, we validate its effectiveness by comparing it with several SOTA self-supervised learning

TABLE I
NETWORK ARCHITECTURE USED IN CIG-MAE. E, D, AND F REPRESENT ENCODER, DECODER, AND CLASSIFIER, RESPECTIVELY.

Layer	E **(Encoder)**	D **(Decoder)**	F **(Classifier)**
Input CSI: $3(6) \times 30 \times 200$ (antenna $\times$ subcarrier $\times$ timestamp)			
1	Conv, $C = 128$, $K = (3, 5)$, $S = (3, 5)$	MLP: $256 \to 512 \times 5 \times 10$	FC: $512 \to C$, Softmax
2	Conv, $C = 256$, $K = (2, 2)$, $S = (2, 2)$	DeConv, $C = 512$, $K = (1, 2)$, $S = (1, 2)$	—
3	Conv, $C = 512$, $K = (1, 2)$, $S = (1, 2)$	DeConv, $C = 256$, $K = (2, 2)$, $S = (2, 2)$	—
4	MLP: $512 \times 5 \times 10 \to 256$	DeConv, $C = 128$, $K = (3, 5)$, $S = (3, 5)$	—

methods on three public CSI datasets. Subsequently, the model's behavior and performance are analyzed through ablation studies, hyperparameter sensitivity analysis, and settings with varying amounts of labeled data.

A. Experimental Setup

Datasets. We evaluate CIG-MAE on three datasets representing diverse sensing scenarios:

- **SignFi** [55] (Sign Language): A subset from User 5 (home environment), comprising 2,760 samples across 276 classes, collected using an 802.11n CSI tool [60]. The input tensor shape is $3 \times 30 \times 200$.
- **NLOS** [54] (Daily Activities): Data from Environment 1, consolidated into 6 classes (3,000 samples) following [61], [62]. The subjects in Environment 1 consist of 10 males with distinct ages, weights, and heights. Variable durations $[754, 1601]$ were downsampled to 200.
- **HTHI** [63] (Human-Interaction): Contains 4,800 trials of 12 interaction types from 40 pairs. The 40 pairs were randomly formed from a pool of 66 subjects with varying ages, weights, and heights. Durations $[1040, 2249]$ were downsampled to 200.

Protocol & Implementation. We implement the model in PyTorch on an NVIDIA RTX 5090. Datasets are stratified (8:2 train/test split), so the training splits of SignFi, NLOS and HTHI contain 2,208, 2,400 and 3,840 unlabeled sequences, respectively. Raw CSI phase is linearly calibrated [64], and both modalities are z-score normalized. We adhere to the standard k-shot linear probing protocol: pre-trained encoders are frozen, and concatenated representations are fed to a linear classifier. **Pre-training** runs for 300 epochs (batch size 256) using AdamW ($lr = 1 \times 10^{-4}$, weight decay 0.01, $\beta = (0.9, 0.95)$). We set the mask ratio $\rho = 0.95$, patch size $(3, 5)$, and loss weights $w_{\text{bt}} = 0.2, w_{\text{aim}} = 1 \times 10^{-4}$. The AIM policy features $\mathbf{x}^m$ and encoder latent features $\mathbf{z}^m$ are both 256. Critically, the BT projection head is a 3-layer MLP with width 1024 (SignFi/HTHI) or 5096 (NLOS). **Downstream** evaluation uses 1-shot (SignFi) or 10-shot (NLOS/HTHI) learning for 100 epochs (batch size 32, $lr = 1 \times 10^{-3}$). Architecture details are provided in Table I. We adopt this mid-sized dual-stream CNN backbone to practically balance representation expressiveness and deployment efficiency. Smaller variants lack sufficient receptive field and channel capacity to capture the spatio-temporal structure of $3 \times 30 \times 200$ CSI maps, whereas larger models incur prohibitive latency and memory overheads on edge devices with diminishing empirical gains.

B. Baselines

We compare CIG-MAE against a fully supervised upper bound and three SOTA self-supervised methods:

- **Supervised Model**: Trained end-to-end on the full labeled dataset using the exact same backbone as CIG-MAE, serving as the performance upper bound.
- **AutoFi** [31]: A contrastive method that jointly optimizes probabilistic consistency, mutual information maximization, and geometric structure consistency.
- **AutoSen** [14]: A cross-modal autoencoder that learns joint representations through the asymmetric proxy task of reconstructing phase from amplitude.
- **CAPC** [35]: Combines CPC with BT. It predicts future latent representations to capture temporal context while using BT to reduce feature redundancy.

C. Performance Comparison

Table II compares CIG-MAE against SOTA SSL methods and a supervised baseline. CIG-MAE consistently outperforms all SSL counterparts. Notably, on SignFi, it achieves 98.49% accuracy, surpassing even the supervised baseline (98.19%), indicating superior generalization. This advantage widens in complex scenarios (NLOS and HTHI), where CIG-MAE maintains robustness with accuracies of 59.80% and 38.90% respectively, establishing a clear margin over all SSL baselines.

TABLE II
PERFORMANCE (MEAN $\pm$ STD) COMPARISON OF DIFFERENT METHODS ON SIGNFI, NLOS, AND HTHI DATASETS WITH A LINEAR CLASSIFIER.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Supervised	98.19 ± 2.56	97.88 ± 3.01	99.66 ± 0.23	99.67 ± 0.27	97.76 ± 0.37	97.76 ± 0.37
AutoFi	94.60 ± 0.21	94.41 ± 0.06	48.59 ± 0.83	47.73 ± 1.49	30.99 ± 1.54	29.88 ± 2.20
AutoSen	88.59 ± 1.28	87.85 ± 1.27	48.08 ± 5.30	48.19 ± 3.56	30.95 ± 2.33	30.10 ± 1.63
CAPC	95.56 ± 1.92	95.63 ± 1.46	46.91 ± 0.82	46.02 ± 2.21	34.06 ± 2.35	35.05 ± 1.81
CIG-MAE	98.49 ± 0.27	98.42 ± 0.20	59.80 ± 3.48	57.93 ± 3.30	38.90 ± 1.93	39.43 ± 1.57

The performance gap stems from the misalignment of baselines with CSI characteristics:

- **AutoFi** falters in multi-user settings (NLOS, HTHI) as its contrastive paradigm relies on simplistic augmentations insufficient to handle high intra-class variance.
- **CAPC** imposes a restrictive temporal continuity bias, causing it to neglect self-contained spatio-temporal patterns critical for fine-grained actions.
- **AutoSen**'s asymmetric design creates an information bottleneck by assuming amplitude fully specifies phase, thereby losing unique phase-dependent information.

In contrast, CIG-MAE overcomes these limitations through a synergistic design. Its masked autoencoding paradigm captures holistic intrinsic structures rather than brittle inter-sample relations. The symmetric, BT-regularized architecture ensures complete, decorrelated fusion of amplitude and phase. Crucially, AIM dynamically focuses the learning budget on discriminative regions, ensuring robustness against noise where other models fail.

D. Ablation Study

We validate the contribution of each component by systematically dismantling the full model (Table III).

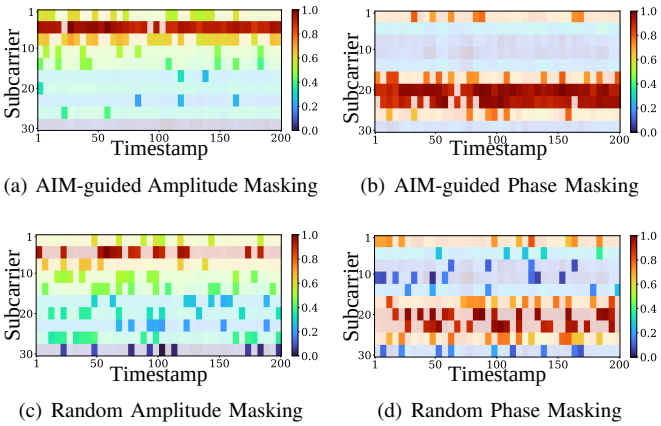
TABLE III
THE PERFORMANCE (MEAN ± STD) WITH DIFFERENT DESIGN CHOICES FOR SIGNFI, NLOS, AND HTHI DATA.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Single-Stream (Amp.)	96.43±0.73	96.20±0.67	56.53±3.40	54.35±3.24	37.03±0.73	37.90±0.83
Dual-Stream MAE	97.48±0.97	97.35±1.03	52.99±4.07	53.78±3.04	37.03±0.52	37.74±0.47
w/o BT	97.88±0.73	97.56±0.82	52.33±3.58	54.05±3.16	36.82±1.02	37.30±0.92
w/o AIM	97.70±0.69	97.45±0.66	57.50±3.64	57.33±3.85	37.29±2.10	38.18±2.30
CIG-MAE	**98.49±0.27**	**98.42±0.20**	**59.80±3.48**	**57.93±3.30**	**38.90±1.93**	**39.43±1.57**

Impact of AIM. Replacing AIM with random masking (*w/o AIM*) leads to a consistent performance degradation across all datasets. Specifically, removing the AIM strategy results in an accuracy drop of 2.30% on NLOS (from 59.80% to 57.50%) and 1.61% on HTHI (from 38.90% to 37.29%). This confirms that in low-SNR and complex interaction scenarios, strictly random masking is insufficient. By focusing on information-dense regions, AIM enables the model to learn more discriminative features from limited visibility, as visualized in Fig. 4.

Impact of BT Regularizer. BT proves essential for coordinating the dual streams, especially in noisy environments. Removing it (*w/o BT*) causes a significant 7.47% accuracy drop on NLOS. Notably, on the challenging NLOS dataset, *w/o BT* performs even worse than the baseline *Dual-Stream MAE*. This suggests that without the alignment constraint provided by BT, the AIM-guided dual streams may diverge towards modality-specific artifacts or noise, failing to learn a coherent representation. Fig. 5 demonstrates how BT effectively aligns representations and reduces redundancy.

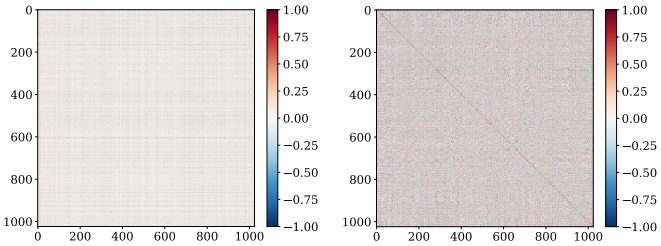
Dual-Stream Architecture. Comparison between *Single-Stream* and *Dual-Stream MAE* reveals that naively combining amplitude and phase improves performance on SignFi but degrades it on NLOS. This conflict underscores the necessity of the BT regularizer to resolve feature redundancy, validating the synergistic design of the full CIG-MAE model.



(a) AIM-guided Amplitude Masking (b) AIM-guided Phase Masking

(c) Random Amplitude Masking (d) Random Phase Masking

Fig. 4. Demonstration of AIM's visibility allocation. Compared to random masking, AIM suppresses low-information regions and increases the visibility of high-information regions, concentrating the learning budget on critical time-frequency blocks related to the activity, thereby obtaining more discriminative representations.



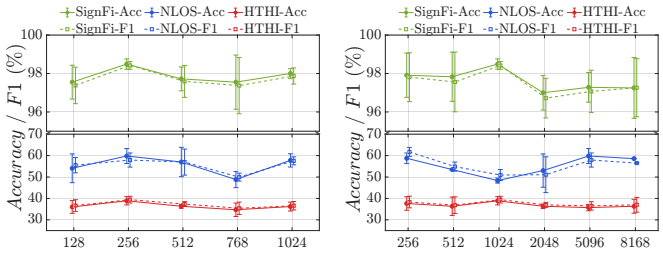
(a) Amplitude-Phase Correlation Matrix without BT Constraint (b) Amplitude-Phase Correlation Matrix with BT Constraint

Fig. 5. Demonstration of the alignment and decorrelation effect of BT. After introducing BT, the diagonal elements of the correlation matrix approach 1, and the off-diagonal elements approach 0, reflecting enhanced consistency and reduced redundancy in the corresponding dimensions of the two streams.

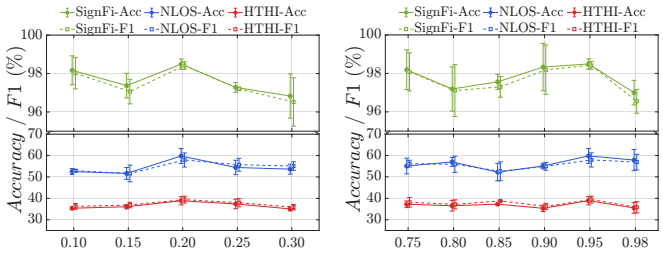
E. Sensitivity Analysis

We analyze key hyperparameters in Fig. 6 and reconstruction objectives in Table IV.

Hyperparameters. The AIM feature dimension ($\mathbf{x}^m$) consistently peaks at 256; lower dimensions lack expressive capacity, while higher ones risk overfitting to spurious correlations. The BT loss weight (w_{bt}) stabilizes at 0.2, striking a critical balance: lower weights provide insufficient regularization, while higher weights over-optimize decorrelation at the expense of preserving fine-grained signal details. In contrast, the projection head dimension (ω^m) adapts to dataset complexity. While 1024 suffices for SignFi/HTHI, the noisier NLOS dataset requires a larger capacity (5096) to effectively disentangle features. Finally, a high mask ratio ($\rho = 0.95$) is consistently optimal, confirming that forcing reconstruction from sparse inputs (5%) compels the model to learn deep intrinsic structures rather than memorizing superficial patterns.



(a) AIM Feature Dimension ($\mathbf{x}^m$) (b) Projection Dimension (ω^m)



(c) BT Loss Weight (w_{bt}) (d) Mask Ratio (ρ)

Fig. 6. Sensitivity analysis of CIG-MAE's key hyperparameters across the SignFi, NLOS, and HTHI datasets. The results show that while some parameters are stable, others adapt to data characteristics, demonstrating the model's robustness and flexibility.

TABLE IV
IMPACT OF DIFFERENT RECONSTRUCTION LOSS FUNCTIONS ON CIG-MAE PERFORMANCE (MEAN ± STD) FOR SIGNFI, NLOS, AND HTHI DATA.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
MAE (w/o Norm.)	**98.49±0.27**	**98.42±0.20**	**59.80±3.48**	**57.93±3.30**	**38.90±1.93**	**39.43±1.57**
MAE (w/ Norm.)	97.09±1.03	96.98±0.97	54.39±1.44	55.43±1.38	35.97±0.77	36.32±1.11
MSE (w/o Norm.)	95.82±1.03	95.63±1.25	51.83±5.57	52.13±3.51	37.08±1.29	37.92±1.53
MSE (w/ Norm.)	95.38±0.38	95.15±0.56	55.33±2.08	54.46±1.96	36.15±1.59	36.97±1.47

Loss Function. Table IV shows that *MAE (w/o Norm.)* consistently yields the best performance. MAE is preferred over MSE because its linear penalty is robust to sporadic high-amplitude noise spikes (outliers), avoiding the gradient dominance caused by quadratic penalties. Furthermore, omitting normalization preserves the original dynamic range, which is crucial for distinguishing high-magnitude activity bursts from low-magnitude background fluctuations.

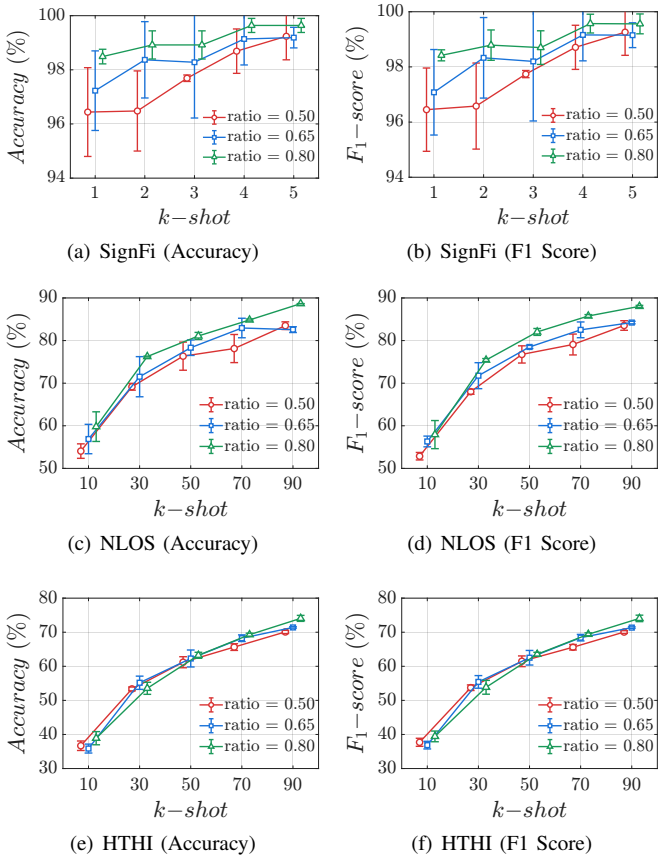
F. Data Efficiency Analysis

We analyze the impact of labeled fine-tuning data size (k) and unlabeled pre-training data scale (50%, 65%, 80%) in Fig. 7.

First, performance monotonically improves with labeled samples (k) across all settings. This confirms that CIG-MAE learns a meaningful, linearly separable feature space where the classifier can effectively leverage additional supervision to refine decision boundaries.

Second, the impact of pre-training data volume depends heavily on dataset complexity. On the relatively clean SignFi dataset, increasing the unlabeled pre-training data from 50% to 80% yields only marginal gains, suggesting early saturation. This is likely because the SignFi subset used in this work (User 5 in a home environment) was collected in a comparatively controlled setting, so even 50% of the unlabeled data already covers most of the dominant signal manifold.

By contrast, the benefit is much more pronounced on NLOS. In the NLOS 10-shot setting, for example, increasing the pre-training ratio from 50% to 80% improves the accuracy from 54.06% to 59.80%. This is consistent with the sensitivity analysis in Fig. 6, where NLOS requires a substantially larger BT projection dimension (5096 versus 1024 for SignFi/HTHI), suggesting a more complex representation structure. In addition, NLOS contains 10 subjects with distinct ages, weights, and heights, introducing greater subject-dependent variability. Together with stronger background interference and propagation uncertainty, this makes additional unlabeled data particularly valuable for learning invariant and noise-robust features.



(a) SignFi (Accuracy) (b) SignFi (F1 Score)

(c) NLOS (Accuracy) (d) NLOS (F1 Score)

(e) HTHI (Accuracy) (f) HTHI (F1 Score)

Fig. 7. Impact of unlabeled pre-training data size and labeled fine-tuning data size on model performance. The x-axis represents the number of labeled samples per class used in k-shot learning, and different curves correspond to different pre-training data ratios (50%, 65%, 80%).

HTHI shows an intermediate trend: the gain from more pre-training data is limited in the 10- and 30-shot settings, but

becomes clearer from 50-shot onward. Since HTHI contains 40 interaction pairs randomly formed from 66 subjects with varying ages, weights, and heights, its representation space is also highly diverse. We conjecture that under extremely scarce supervision, the linear classifier is still the bottleneck, whereas with more labeled samples it can better exploit the improved pre-trained representation.

G. Backbone Comparison and Analysis

To validate the suitability of our design for practical IoT deployment, we benchmark the proposed CNN backbone against a ViT-based counterpart. Specifically, we replaced the CNN encoder-decoder with a ViT equivalent following the standard Masked Autoencoder design [16], [53], while keeping other components unchanged. The ViT baseline consists of an encoder with 6 blocks (1024 embedding dim, 16 heads) and a decoder with 5 blocks (512 dim, 8 heads), determined as the best-performing configuration by sequentially sweeping decoder depth over $\{1\text{–}6\}$, encoder depth over $\{1\text{–}7\}$, and embedding dimension over $\{256, 512, 1024\}$ on the NLOS dataset. The comprehensive comparison is detailed in Table V.

TABLE V
PERFORMANCE AND EFFICIENCY COMPARISON BETWEEN THE PROPOSED CNN BACKBONE AND THE VIT BASELINE.

Method	SignFi		NLOS		HTHI	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
CNN (Ours)	**98.49 ± 0.27**	**98.42 ± 0.20**	**59.80 ± 3.48**	**57.93 ± 3.30**	**38.90 ± 1.93**	**39.43 ± 1.57**
ViT	91.76 ± 0.13	91.09 ± 0.01	51.67 ± 4.71	43.00 ± 2.26	37.86 ± 0.96	36.14 ± 0.01
Params	**CNN: 14.03M**			ViT: 152.08M		
FLOPs	**CNN: 0.14G**			ViT: 129.98G		
Memory	**CNN: 131.97MB**			ViT: 580.13MB		

Performance Superiority. The CNN-based model substantially outperforms the ViT counterpart across all datasets. This gap stems from the inherent inductive biases of CNNs, such as locality and translation equivariance, which are highly compatible with the structured features of CSI time-frequency maps. In contrast, ViT lacks these priors and struggles to learn effectively from the limited scale of CSI datasets.

Edge Deployment Feasibility. Critically, the CNN architecture demonstrates superior efficiency, utilizing approximately $10\times$ fewer parameters and $1000\times$ fewer FLOPs than the ViT (Table V). For practical context, commercial WiFi access points (e.g., Qualcomm IPQ4019 used by Nexmon [65]–[67]) typically offer a general-purpose compute budget of only a few GFLOPs/s and 1 GB of DRAM. Requiring about 0.14 GFLOPs per inference, our CNN can theoretically sustain real-time sensing at roughly 20 FPS while comfortably fitting within the memory budget. In contrast, the ViT baseline's massive computational and memory demands render it entirely infeasible on such low-power SoCs without dedicated accelerators.

Overfitting in Self-Supervised Pre-training. Although the training splits contain only a few thousand unlabeled sequences (e.g., roughly 2.5k for SignFi), the masked autoencoding objective provides significantly more effective constraints than the raw sequence count suggests. Specifically, each sequence is partitioned into 400 patches. With a masking ratio of $\rho = 0.95$, the model must reconstruct 380 patches per sequence. Over 300 pre-training epochs, this combinatorial effect yields approximately $2.5\text{k} \times 380 \times 300 \approx 2.8 \times 10^8$ patch-level predictions per stream. For our 14M-parameter CNN backbone, this equates to roughly ten supervisory constraints per parameter. Combined with convolutional weight sharing and the redundancy-reduction loss, this massive volume of fine-grained supervision acts as a strong regularizer. Consequently, increasing the pre-training data volume consistently improves test accuracy, confirming that overfitting is not dominant in the self-supervised stage.

V. DISCUSSION

Although this work is evaluated on HAR, CIG-MAE is not inherently tied to activity semantics. Since CSI amplitude and phase jointly encode wireless propagation characteristics, the dual-stream masked reconstruction objective can also be viewed as learning propagation-aware priors from unlabeled channel observations. This creates a natural connection to broader wireless channel modeling tasks, such as radio map construction and channel twinning [68]–[70]. In particular, AIM may help prioritize informative time-frequency regions for propagation modeling, while cross-modal alignment may support consistency across heterogeneous channel views [71]. We emphasize that such extensions would require additional task-specific designs, such as spatial conditioning or map-level supervision, and are therefore beyond the scope of the present study [72]. From a practical perspective, resource-constrained deployment is also an important consideration. In this setting, the CNN backbone adopted in CIG-MAE provides a favorable basis for pruning, quantization, and distillation, while the pretrained encoder supports lightweight downstream adaptation [73], [74]. More broadly, a lightweight CSI-native pretraining framework for small-scale, noisy, and deployment-sensitive scenarios may also be relevant to future wireless foundation-model pipelines [75].

VI. CONCLUSION

This paper proposes CIG-MAE, a generative self-supervised framework tailored for CSI-HAR. By integrating dual-stream masked reconstruction, AIM, and BT regularization, CIG-MAE effectively leverages the complementarity of amplitude and phase. This design eliminates the reliance on problematic data augmentations and negative samples while capturing intrinsic signal structures. Experiments on three public datasets demonstrate that CIG-MAE consistently achieves SOTA performance. Critically, our comparison with ViT backbones confirms that the proposed CNN-based architecture offers superior accuracy with orders of magnitude lower computational cost ($1000\times$ fewer FLOPs), establishing

a feasible path for deploying advanced sensing algorithms on resource-constrained edge devices. Future work will focus on several directions: evaluating scalability on larger, diverse datasets to further test generalization, and extending the framework to complex multi-person scenarios. Additionally, we plan to explore model compression techniques to further optimize real-time inference on commercial WiFi hardware and investigate optimization refinements for the discrete AIM policy, such as variance-reduction techniques, to further improve training efficiency.

REFERENCES

[1] Z. Wang, J. A. Zhang, M. Xu, and Y. J. Guo, "Single-Target Real-Time Passive WiFi Tracking," *IEEE Trans. Mob. Comput.*, vol. 22, no. 6, pp. 3724–3742, Jun. 2023.

[2] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "WiFi-Based Real-Time Breathing and Heart Rate Monitoring During Sleep," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[3] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "BodyCompass: Monitoring sleep posture with wireless signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, p. 66, Jun. 2020.

[4] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices," *IEEE Trans. Mob. Comput.*, vol. 16, no. 2, pp. 511–526, Feb. 2017.

[5] L. Guo, Z. Lu, S. Zhou, X. Wen, and Z. He, "Emergency Semantic Feature Vector Extraction From WiFi Signals for In-Home Monitoring of Elderly," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1423–1438, Nov. 2021.

[6] C. Wu, B. Wang, O. C. Au, and K. J. R. Liu, "Wi-fi can do more: Toward ubiquitous wireless sensing," *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 42–49, Jun. 2022.

[7] Y. Ma, G. Zhou, and S. Wang, "WiFi Sensing with Channel State Information: A Survey," *ACM Comput. Surv.*, vol. 52, no. 3, p. 46, Jun. 2019.

[8] S. G. Dhekane and T. Ploetz, "Transfer Learning in Human Activity Recognition: A Survey," 2024, arXiv preprint arXiv:2401.10185.

[9] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards Environment Independent Device Free Human Activity Recognition," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 289–304.

[10] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024.

[11] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Evaluating self-supervised learning for wifi csi-based human activity recognition," *ACM Trans. Sen. Netw.*, vol. 21, no. 2, pp. 21:1–21:38, Mar. 2025.

[12] C. Xiao, Y. Han, W. Yang, Y. Hou, F. Shi, and K. Chetty, "Diffusion-Model-Based Contrastive Learning for Human Activity Recognition," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 33 525–33 536, Oct. 2024.

[13] B. Chen, J. Wang, Y. Lv, Q. Gao, M. Pan, and Y. Fang, "Device-Free Wireless Sensing With Few Labels Through Mutual Information Maximization," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 10 513–10 524, Mar. 2024.

[14] Q. Gao, Y. Hao, and Y. Liu, "Autosen: Improving automatic wifi human sensing through cross-modal autoencoder," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 8235–8239.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988.

[17] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *International Conference on Learning Representations (ICLR)*, 2022.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[19] D. Cheng, L. Zhang, L. Qin, S. Wang, H. Wu, and A. Song, "MaskCAE: Masked convolutional autoencoder via sensor data reconstruction for self-supervised human activity recognition," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 5, pp. 2687–2698, May 2024.

[20] A. Saeed, V. Ungureanu, and B. Gfeller, "Sense and Learn: Self-supervision for omnipresent sensors," *Mach. Learn. Appl.*, vol. 6, p. 100152, 2021.

[21] R. Song, D. Zhang, Z. Wu, C. Yu, C. Xie, S. Yang, Y. Hu, and Y. Chen, "RF-URL: unsupervised representation learning for RF sensing," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 282–295.

[22] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien, "Federated Self-Supervised Learning of Multisensor Representations for Embedded Intelligence," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1030–1040, Jan. 2021.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 9726–9735.

[24] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 12 310–12 320.

[25] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2022.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[27] M. J. Bocus, H.-S. Lau, R. McConville, R. J. Piechocki, and R. Santos-Rodriguez, "Self-Supervised WiFi-Based Activity Recognition," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 552–557.

[28] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher, "Contrastive Self-Supervised Representation Learning for Sensing Signals from the Time-Frequency Perspective," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–10.

[29] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-Stream Contrastive Learning for Channel State Information Based Human Activity Recognition," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 329–338, Jan. 2023.

[30] N. Lyons, A. Santra, V. K. Ramanna, K. Uln, R. Taori, and A. Pandey, "WiFiAct: Enhancing Human Sensing Through Environment Robust Preprocessing and Bayesian Self-Supervised Learning," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2024, pp. 13 391–13 395.

[31] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "AutoFi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7416–7425, Apr. 2023.

[32] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[33] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6775–6794, 2024.

[34] H. Haresamudram, I. Essa, and T. Ploetz, "Contrastive predictive coding for human activity recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–25, 2021.

[35] B. Barahimi, H. Tabassum, M. Omer, and O. Waqar, "Context-aware predictive coding: A representation learning framework for wifi

sensing," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 6119–6134, 2024.

[36] J. Strohmayer and M. Kampel, *Data Augmentation Techniques for Cross-Domain WiFi CSI-Based Human Activity Recognition*, 2024, pp. 42–56.

[37] O. G. Serbetci, J.-H. Lee, D. Burghal, and A. F. Molisch, "Simple and effective augmentation methods for csi based indoor localization," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Dec. 2023, pp. 3947–3952.

[38] W. Hou and C. Wu, "RFBoost: Understanding and boosting deep WiFi sensing via physical data augmentation," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–26, May 2024.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] A. Y. Radwan, M. Yildirim, N. Hasanzadeh, H. Tabassum, and S. Valaee, "A Tutorial-cum-Survey on Self-Supervised Learning for Wi-Fi Sensing: Trends, Challenges, and Outlook," *IEEE Commun. Surv. Tuts.*, pp. 1–1, 2025.

[41] X. Kong and X. Zhang, "Understanding masked image modeling via learning occlusion invariant feature," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 6241–6251.

[42] S. Miao, L. Chen, and R. Hu, "Spatial-Temporal Masked Autoencoder for Multi-Device Wearable Human Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, p. 172, Jan. 2024.

[43] J. Yang, S. Tang, Y. Xu, Y. Zhou, and L. Xie, "MaskFi: Unsupervised learning of WiFi and vision representations for multimodal human activity recognition," 2024, arXiv preprint arXiv:2402.19258.

[44] S. Ji, Y. Xie, and M. Li, "SiFall: Practical online fall detection with RF sensing," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2022, pp. 563–577.

[45] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. De Nadai, "Efficient training of visual transformers with small datasets," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[46] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, *What to Hide from Your Students: Attention-Guided Masked Image Modeling*, 2022, pp. 300–318.

[47] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "SLNet: A spectrogram learning neural network for deep wireless sensing," in *Proc. 20th USENIX Symp. Netw. Syst. Des. Implementation (NSDI 23)*, Apr. 2023, pp. 1221–1236.

[48] Y. Taso, S.-C. Yeh, Y.-Y. Liang, C.-H. Wang, and S.-H. Fang, "Subcarrier selection for efficient CSI-based indoor localization," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 383, no. 1, p. 012017, Jul. 2018.

[49] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel, "Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 507–14 517.

[50] Z. Zheng, J. Oh, and S. Singh, "On learning intrinsic rewards for policy gradient methods," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 4649–4659.

[51] M. J. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 18 097–18 107.

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

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[54] B. A. Alsaify, M. M. Almazari, R. Alazrai, and M. I. Daoud, "A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments," *Data in Brief*, vol. 33, p. 106534, 2020.

[55] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 23, Mar. 2018.

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

[57] S. Huang, K. Li, D. You, Y. Chen, A. Lin, S. Liu, X. Li, and J. A. McCann, "Wimans: A benchmark dataset for wifi-based multi-user activity sensing," in *Computer Vision – ECCV 2024: 18th European Conference on Computer Vision*, 2024, pp. 72–91.

[58] Z. Yang, Y. Zhang, G. Zhang, Y. Zheng, and G. Chi, "Widar 3.0: WiFi-based activity recognition dataset," IEEE Dataport, 2020.

[59] O. Jouini, K. Sethom, A. Namoun, N. Aljohani, M. H. Alanazi, and M. N. Alanazi, "A Survey of Machine Learning in Edge Computing: Techniques, Frameworks, Applications, Issues, and Research Directions," *Technologies*, vol. 12, no. 6, p. 81, 2024.

[60] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, Jan. 2011.

[61] B. A. Alsaify, M. M. Almazari, R. Alazrai, S. Alouneh, and M. I. Daoud, "A CSI-Based Multi-Environment Human Activity Recognition Framework," *Appl. Sci.*, vol. 12, no. 2, p. 930, 2022.

[62] B. A. Alsaify, M. M. Almazari, R. Alazrai, and M. I. Daoud, "Exploiting Wi-Fi Signals for Human Activity Recognition," in *Proc. 12th Int. Conf. Inf. Commun. Syst. (ICICS)*, May 2021, pp. 245–250.

[63] R. Alazrai, A. Awad, B. Alsaify, M. Hababeh, and M. I. Daoud, "A dataset for wi-fi-based human-to-human interaction recognition," *Data in Brief*, vol. 31, p. 105668, 2020.

[64] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are facing the Mona Lisa: Spot localization using PHY layer information," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, 2012, pp. 183–196.

[65] M. Schulz, D. Wegemer, and M. Hollick. (2017) Nexmon: The c-based firmware patching framework. [Online]. Available: https://nexmon.org

[66] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your csi: A channel state information extraction platform for modern wi-fi chipsets," ser. WiNTECH '19, 2019, p. 21–28.

[67] E. T. Ltd., "Qualcomm wireless module ipq4019 wi-fi soc," https://www.excelpoint.com/product/qualcomm-wireless-module-ipq-4019-wifi-soc/, 2026, accessed: 2026-02-14.

[68] X. Wang, Y. Pan, and N. Cheng, "A tutorial on learning-based radio map construction: Data, paradigms, and physics-awarenes," 2026. [Online]. Available: https://arxiv.org/abs/2603.17499

[69] Y. Wang, Z. Su, S. Guo, M. Dai, T. H. Luan, and Y. Liu, "A survey on digital twins: Architecture, enabling technologies, security and privacy, and future prospects," *IEEE Internet of Things Journal*, vol. 10, no. 17, pp. 14 965–14 987, Sep. 2023.

[70] L. U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, and C. S. Hong, "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2230–2254, Fourthquarter 2022.

[71] R. J. Williams, M. S. Abouamer, and P. Popovski, "Digital twin-assisted measurement design and channel statistics prediction," 2026. [Online]. Available: https://arxiv.org/abs/2603.23787

[72] H. Luo, S. R. Khosravirad, and A. Alkhateeb, "Wireless digital twin calibration: Refining dft-domain channel information," 2026. [Online]. Available: https://arxiv.org/abs/2603.16126

[73] S. M. Hernandez and E. Bulut, "Wifi sensing on the edge: Signal processing techniques and challenges for real-world systems," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 46–76, Firstquarter 2023.

[74] X. Wang, Q. Li, and W. Jia, "Cognitive edge computing: A comprehensive survey on optimizing large models and ai agents for pervasive deployment," 2025. [Online]. Available: https://arxiv.org/abs/2501.03265

[75] B. Guler, G. Geraci, and H. Jafarkhani, "A multi-task foundation model for wireless channel representation using contrastive and masked autoencoder learning," 2025. [Online]. Available: https://arxiv.org/abs/2505.09160