Breaking Coordinate Overfitting: Geometry-Aware WiFi Sensing for Cross-Domain 3D Pose Estimation

Abstract

WiFi-based 3D human pose estimation offers a low-cost and privacy-preserving alternative to vision-based systems for smart interaction. However, existing approaches rely on visual 3D poses as supervision and directly regress CSI to a camera-based coordinate system. We find that this practice leads to coordinate overfitting: models memorize deployment-specific WiFi transceiver layouts rather than only learning activity-relevant representations, resulting in severe generalization failures. To address this challenge, we present PerceptAlign, the first geometry-conditioned framework for WiFi-based pose estimation. PerceptAlign introduces a lightweight coordinate unification procedure that aligns WiFi and vision measurements in a shared 3D space using only two checkerboards and a few photos. Within this unified space, it encodes calibrated transceiver positions into high-dimensional embeddings and fuses them with CSI features, making the model explicitly aware of device geometry as a conditional variable. This design forces the network to disentangle human motion from deployment layouts, enabling robust and, for the first time, layout-invariant WiFi pose estimation. To support systematic evaluation, we construct the largest cross-domain 3D WiFi pose estimation dataset to date, comprising 21 participants, 3 environments, 18 actions, and multiple device configurations. Experiments show that PerceptAlign reduces in-domain error by 38% and cross-domain error by more than 60% compared to state-of-the-art baselines. These results establish geometry-conditioned learning as a viable path toward scalable and practical WiFi sensing.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing**; **Interaction techniques**; • **Networks →** *Wireless access points, base stations and infrastructure.*

Keywords

Human pose estimation, WiFi Sensing, Cross-Domain

ACM Reference Format:

. 2018. Breaking Coordinate Overfitting: Geometry-Aware WiFi Sensing for Cross-Domain 3D Pose Estimation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, 3D human pose estimation has served as a core technology for numerous applications [3, 21, 49], including health monitoring and natural human–computer interaction. While vision-based methods have long dominated this field [23, 25, 26, 34], WiFi sensing has recently emerged as a compelling alternative. Unlike cameras, WiFi sensing operates without line-of-sight, preserves user privacy, and leverages existing infrastructure, making it highly attractive for real-world deployment [12, 43, 57]. The prevailing WiFi-based pose estimation pipeline is conceptually straightforward. They typically collect synchronized WiFi Channel State Information (CSI) streams together with human poses obtained from calibrated camera systems, and then train a deep network under visual supervision to predict human pose directly from the CSI inputs. [17, 35, 36, 45, 46, 58]. Most existing work following this paradigm focuses on improving accuracy and producing smoother pose estimates through tailored loss functions or specialized signal acquisition setups. For example, WiPose [17] incorporates prior knowledge of human skeleton structure, HPE-Li [6] leverages attention mechanisms, and GoPose [31] employs a customized antenna array to extract 2D angles of arrival. Other studies push beyond skeleton-level estimation and aim for finer-grained representations, such as reconstructing full 3D human meshes [40] or detailed hand skeletons [16]. More related works please refer to Appendix A

These solutions have demonstrated the feasibility of WiFi-based human pose estimation and achieved competitive accuracy when training and testing are performed under the same deployment conditions [17, 45, 46, 58]. However, current methods exhibit striking brittleness when applied to new settings. Although some studies attempt to address environmental (e.g., room) variation through techniques such as multiresolution convolution [51] for richer feature extraction or alignment losses [59] for improved supervision, they

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

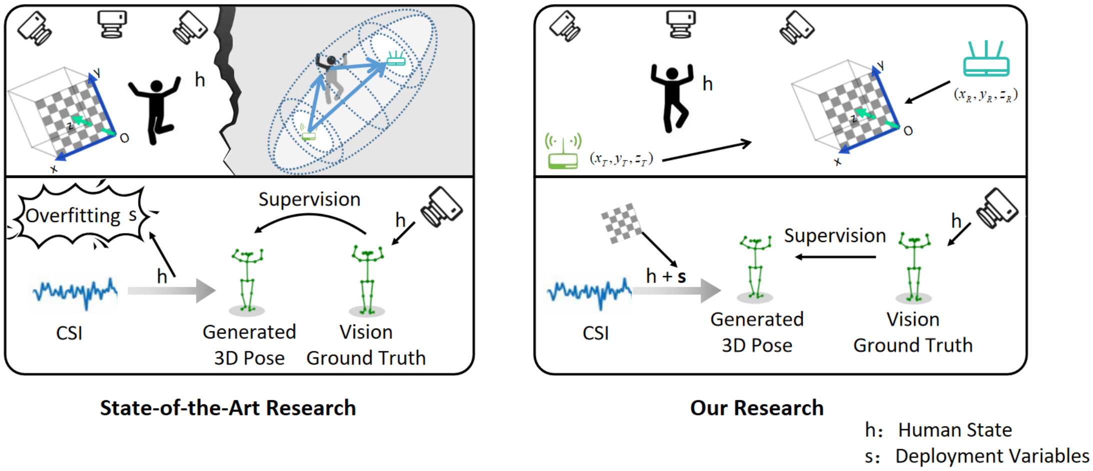


Figure 1: Left: Conventional pipelines implicitly memorize the geometric layout of WiFi devices, conflating it with target knowledge and leading to coordinate overfitting. Right: PerceptAlign explicitly makes the model aware that WiFi transceiver geometry is a conditional factor rather than knowledge to be memorized.

remain unable to handle the domain shifts frequently encountered in practice, such as changes in transceiver layouts. We argue that the root cause lies in what we term coordinate overfitting. By enforcing a direct regression from CSI to camera generated skeletons, existing pipelines implicitly entangle CSI measurements with deployment-specific WiFi transceiver layouts. As illustrated in Fig.1, visual annotations and WiFi measurements used in conventional methods originate from different coordinate systems. State-of-the-art multi-camera systems generate 3D pose labels in a Cartesian space anchored to a calibration checkerboard, whereas WiFi sensing is inherently tied to the Fresnel zones [42, 52, 55] defined by transmitter and receiver positions, capturing variations in propagation paths. Training models on CSI using only visual labels as supervision, without accounting for the relationship between WiFi transceiver geometry and the coordinate system of visual ground truth, prevents the model from recognizing that device layout is a conditional variable rather than knowledge to be learned. As a result, the network is forced to memorize deployment-specific layouts together with human poses. Consequently, the trained system is often effective only for CSI collected under previously memorized WiFi device layouts and suffers severe generalization failures when device placements or user orientations change, thereby limiting its practical applicability [5, 36, 47].

To overcome this limitation, we propose PerceptAlign, a geometry conditioned framework that disentangles human motion from deployment-specific artifacts in WiFi-based 3D pose estimation. Our key insight is that visual labels inherently filter out environmental and layout factors, preserving only human motion, whereas WiFi signals encode a composite of human dynamics, transceiver geometry, and multipath reflections. When trained solely with visual supervision, models cannot distinguish which components of CSI correspond to human motion and which stem from deployment conditions. As a result, they often treat transceiver layouts as fixed priors, leading to coordinate overfitting. PerceptAlign tackles this issue with a straightforward yet effective idea: explicitly exposing device geometry as a varying condition rather than allowing the model to memorize it as hidden background knowledge. Specifically, we first unify heterogeneous WiFi and camera coordinate systems into a shared 3D space through a lightweight coordinate unification procedure. The registered transceiver positions are then encoded as high-dimensional spatial priors and integrated into the learning process, enabling the model to separate human-dependent signals from deployment-dependent factors. This design compels the network to learn features that remain stable under layout changes, ultimately yielding more robust and generalizable WiFi-based 3D pose estimation.

The workflow of PerceptAlign consists of two main components:

Lightweight Coordinate Unification Procedure. The goal of this step is to align the coordinate system of WiFi transceivers with that of the multi-camera system, thereby establishing a unified coordinate system. In current multi-camera pose estimation systems, cameras must first capture images of a checkerboard $\mathbb{B}$ placed in the scene and then perform parameter calibration using predefined algorithms in order to obtain accurate 3D human poses. After calibration, all cameras share a world coordinate system defined

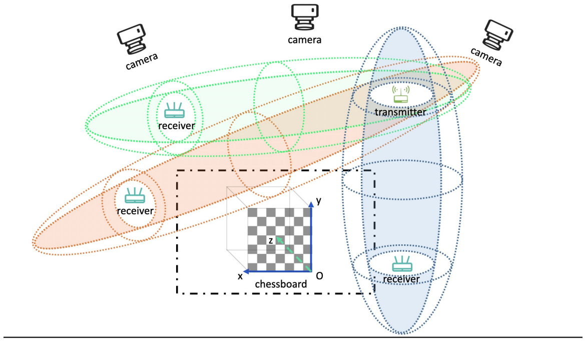


Figure 2: Sensing in vision and WiFi.

by the $\mathbb{B}$’s origin and axes, and a transformation matrix is obtained that map the world coordinates into each camera’s imaging coordinate system. After camera calibration, our coordinate unification procedure uses an additional calibration checkerboard $\mathbb{B}_1$ placed between the middle of the WiFi transceiver, with its x-axis aligned along the transceiver line-of-sight (LOS). This checkerboard coordinate system then represents the transceiver coordinate system. Then, we use a camera $\mathbb{C}$ simultaneously observes $\mathbb{B}$ and $\mathbb{B}_1$, allowing us to compute transformation matrices from each checkerboard to the camera. Using the camera as an intermediary, the WiFi transceiver coordinates can thus be mapped into the world coordinate system quickly and conveniently. ($\mathbb{B}_1$->$\mathbb{C}$->$\mathbb{B}$).

Geometry-Conditioned Learning. Once physical-space coordinate unification is established, the WiFi transceiver coordinates are encoded as high-dimensional spatial embeddings. These embeddings are fused with CSI features extracted by a CNN encoder, and the combined representations are processed by an attention-based fusion backbone that jointly reasons about spatial and temporal evidence to predict 3D human poses. By explicitly incorporating transceiver layout as conditional knowledge, the model avoids implicitly memorizing it as background information, thereby mitigating overfitting and substantially improving generalization.

Our contributions are as follows:

- We reveal coordinate overfitting as the fundamental bottleneck in WiFi-based 3D pose estimation. Existing pipelines directly regress from CSI to camera-generated skeletons, implicitly memorizing deployment-specific layouts, and thus failing to generalize across domains.
- We propose PerceptAlign, the first geometry conditioned framework for 3D WiFi pose estimation. It introduces a lightweight coordinate unification procedure that aligns WiFi and vision into a shared space, and a geometry conditioned learning strategy that explicitly encodes transceiver layouts as conditional priors. This design disentangles human motion from device layouts, achieving the first robust WiFi-based pose estimation across transceiver layouts.
- We construct the largest cross-domain WiFi–vision dataset to date, covering diverse participants, environments, actions, and device setups with detailed geometric calibration. Extensive experiments demonstrate that PerceptAlign reduces in-domain error by 38% and cross-domain error by more than 60% over state-of-the-art baselines.

2 Preliminary

In this section, we outline how the vision system produces human pose estimates and how the corresponding WiFi sensing system captures motion-related information. We then analyze the inherent overfitting problem in existing approaches and conclude with the motivation for our proposed framework.

2.1 Vision-based 3D pose estimation

Owing to their unparalleled accuracy and robustness, current SOTA vision-based approaches for 3D human pose estimation typically rely on multi-camera systems. We also use poses generated by such systems serve as high-quality ground truth for our WiFi-based method. As shown in Figure 2, this requires the multi-camera system to first calibrate both intrinsic and extrinsic parameters using a checkerboard $\mathbb{B}$. This calibration process involves capturing an image of current scene containing $\mathbb{B}$ and then computing the intrinsic and extrinsic parameters through predefined algorithms.

A calibrated camera is characterized by an intrinsic parameter matrix

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

together with lens-distortion parameters κ when applicable. These parameters specify the focal lengths, principal point coordinates, and distortion coefficients of the camera. The calibration also provides extrinsic parameters:

$$(R, \mathbf{t}) \in \mathrm{SO}(3) \times \mathbb{R}^3, \tag{2}$$

which define the rotation R and translation $\mathbf{t}$ of each camera relative to a world coordinate system, and SO(3) means special orthogonal group. The world coordinate system is typically established by the checkerboard $\mathbb{B}$: one chosen corner of the board serves as the origin, the grid directions define the $x_{\mathbb{B}}$ and $y_{\mathbb{B}}$ axes, and $z_{\mathbb{B}}$ is set orthogonal to the board plane. With the extrinsic parameters, a camera can project any 3D point $\mathbf{X}_{\mathbb{B}} = [X, Y, Z]^\top$ defined in the world coordinate system to an ideal image point $\tilde{\mathbf{x}} = [\tilde{u}, \tilde{v}]^\top$ according to:

$$\begin{aligned} \mathbf{X}_c &= R\,\mathbf{X}_{\mathcal{B}} + \mathbf{t}, \\ \tilde{\mathbf{x}} = \Pi(\mathbf{X}_c) &= \begin{bmatrix} f_x \frac{X_c}{Z_c} + c_x \\ f_y \frac{Y_c}{Z_c} + c_y \end{bmatrix}, \end{aligned} \tag{3}$$

and the observed pixel coordinates $\mathbf{x}$ are related to $\tilde{\mathbf{x}}$ via distortion model $\mathbf{x} = \mathcal{D}(\tilde{\mathbf{x}}; \kappa)$. Camera calibration [1] estimates $K, \kappa, R, \mathbf{t}$ from images of the known checkerboard pattern.

After completing multi-camera calibration, we employ the state-of-the-art EasyMocap framework [1] to obtain 3D visual human poses. EasyMocap[1] applies multi-view 2D keypoint detection followed by geometric triangulation to obtain 3D skeleton coordinates. Let $\{\mathbf{x}_i^{(v)}\}_{v=1}^{V}$ denote the detected 2D location of joint i in the v-th calibrated views. EasyMocap recovers the corresponding 3D keypoint $\mathbf{y}_i \in \mathbb{R}^3$ through triangulation by solving a small linear (or nonlinear reprojection-error) problem:

$$\hat{\mathbf{y}}_i \;=\; \arg\min_{\mathrm{Y}} \sum_{v=1}^{V} \left\| \mathbf{x}_i^{(v)} - \Pi_v\big(R_v \mathrm{Y} + \mathbf{t}_v; K_v, \kappa_v\big) \right\|^2, \tag{4}$$

where $(K_v, R_v, \mathbf{t}_v, \kappa_v)$ are the intrinsics/extrinsics/distortion for view v. The resulting 3D skeleton $\mathbf{y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_J^\top]^\top$ is expressed in the world coordinate system.

Clearly, with the aid of the checkerboard, the 3D human poses produced by a multi-camera system are disentangled from environmental and camera layout factors. The system outputs the absolute coordinates of all human skeleton joints in a predefined world coordinate system, thereby providing high-quality ground truth for training WiFi-based pose estimation models.

2.2 WiFi-based Motion Sensing

WiFi-based sensing perceives both the environment and human activities by capturing variations in the propagation paths of electromagnetic signals between transmitters and receivers [20, 41]. We model the various factors in WiFi-based human pose estimation that influence WiFi signal propagation as follows:

$$\mathcal{G} \;=\; (h, s), \tag{5}$$

where h denotes the dynamic human state and s denotes the static deployment variables [13] (TX/RX 3D poses, static scene scatterers, and hardware-specific parameters). Within a synchronized vision–WiFi multimodal sensing setup, the WiFi measurements corresponding to each camera frame are represented as:

$$H_t \in \mathbb{C}^{A \times M \times T}, \tag{6}$$

where A denotes the number of antenna pairs, M the number of subcarriers, and T the number of sampled time points. For the WiFi sensing system, we define a sensing operator $\mathcal{F}$ that maps the $\mathcal{G}$ to the measured CSI H_t:

$$H_t \;=\; \mathcal{F}(\mathcal{G}; t) + \varepsilon, \;=\; \mathcal{F}(h, s; t) + \varepsilon. \tag{7}$$

where ε models measurement noise and other unmodeled effects. As a result, the measured CSI H_t is a composite signal that inherently couples human motion with deployment-specific device layout geometry and environmental factors. Under a finite-path approximation the H_t is presented as [22, 38]:

$$H(f_m, t) \;=\; \sum_{k=1}^{K} \alpha_k(h, s, t)\, e^{-j2\pi f_m \tau_k(h,s,t)}, \tag{8}$$

where τ_k and α_k are path delays and complex amplitudes that depend jointly on h and s, f_m is the frequency of the m-th subcarrier.

According to Fresnel-zone sensing theory [33, 56], the impact of object motion on WiFi CSI can be characterized by a Fresnel model in which the transmitter and receiver serve as the focal points of an ellipsoid. As shown in Figure 2, the influence of motion on the sensed signal is modeled as variations in the path and distance relative to these two foci. Thus, the coordinate system of WiFi sensing can be geometrically defined by the positions of the transceivers. In WiFi sensing, the absence of intrinsic and extrinsic calibration methods, as found in vision systems, means that there is no unified world coordinate system. As a result, the sensing operator $\mathcal{F}$ is influenced not only by human activity but also by device layout, environmental structures, and measurement noise. Among these factors, changes in transceiver layout significantly alter the propagation paths affected by human motion, creating substantial barriers to generalization. Prior studies [42, 53] have also noted that variations in the geometric relationship between device placement and human orientation or position can induce large shifts in the WiFi sensing signal sequences.

2.3 Coordinate Overfitting in WiFi-based 3D Pose Estimation

Current SOTA WiFi-based 3D pose estimation systems all rely on visual labels for training. The relationship among the actual human pose, the visual annotations, and the observed WiFi CSI can be expressed as follows:

$$\mathbf{y} \xleftarrow{\mathbb{B}} \mathcal{G} \xrightarrow{\mathcal{F}} H, \tag{9}$$

where $\mathbf{y}$ denotes visual 3D annotations (Sec. 2.1), $\mathcal{G} = (h, s)$ is the complete geometric state with dynamic human component h and static/deployment component s (Sec. 2.2), and $\mathcal{F}$ is the RF forward operator producing CSI H. From this formulation, we can draw the following conclusion:

(1) **Asymmetric dependencies.** Thanks to intrinsic and extrinsic calibration, the visual labels $\mathbf{y} = \mathcal{T}(h)$ depend (after calibration) only on the dynamic state h, whereas CSI satisfies $H = \mathcal{F}(h, s)$ and depends jointly on h and the deployment variables s.

(2) **Implicit entanglement during learning.** During training, a WiFi-based 3D pose estimation system employs a blind regressor $f_\theta : H \mapsto \mathbf{y}$ that attempts to approximate $p(\mathbf{y} \mid H)$. This process can be expressed as:

$$p(\mathbf{y} \mid H) = \iint p(\mathbf{y} \mid h)\, p(h \mid H, s)\, p(s \mid H)\, \mathrm{d}h\, \mathrm{d}s, \quad (10)$$

the learner must marginalize over the unobserved and varying s. With finite data, the model commonly exploits spurious correlations between s and $\mathbf{y}$ in the training set, effectively learning a mapping $f_\theta(H) \approx G_\theta(h, s)$ that depends on s.

(3) **Device-geometry sensitivity.** Although the learning process of a WiFi-based pose estimation model is influenced by all factors in s, device geometry is the most sensitive component. According to Fresnel-zone theory, variations in transceiver distance and orientation may cause substantial shifts in the signal propagation patterns induced by the same human motion. Even small changes in transmitter–receiver placements can lead to large, non-local shifts in the distribution of H. Since $\mathbf{y}$ remains fixed in $\mathbb{B}$, a regressor that has internalized a deployment-specific alignment will be systematically biased when the transceiver layout changes at test time, resulting in the large in-domain versus cross-domain performance gap observed empirically.

In summary, while current approaches have achieved promising progress in refining pose estimation through advanced neural architectures and sensing system designs, they remain fundamentally limited by a severe generalization challenge caused by memorizing device geometry. We refer to this phenomenon as coordinate overfitting, where the model encodes deployment-specific alignments rather than learning deployment-invariant representations of human motion.

2.4 Our Motivation

From the above analysis, it is clear that in WiFi-based pose estimation the only theoretically robust link between the visual and WiFi modalities is h. The factors in s act as confounding variables that are imperceptible to the vision modality. However, if we explicitly inform the model that the dominant component of s (the transceiver geometry in this paper) should be treated as conditional information, akin to additional intrinsic or extrinsic parameters. This reduces the marginalization burden on the learner and prevents deployment-specific memorization. To realize this idea, two essential procedures are required:

(1) **Coordinate Unification.** To enable the WiFi 3D pose estimation model to recognize transceiver layout relative to visual annotations, the coordinates of WiFi transceivers must first be incorporated into the unified world coordinate system used by the multi-camera setup.

(2) **Geometry-conditioned learning.** After coordinate unification, the model must be guided to recognize transceiver geometry as conditional knowledge rather than a system requirement, thereby encouraging it to factor out this information during learning and acquire more robust features.

3 System Overview

In this section, we present an overview of PerceptAlign, our cross-domain robust WiFi-based 3D pose estimation system. The framework consists of two key components: **Lightweight Coordinate Unification** and **Geometry Conditioned Learning**. Specifically, PerceptAlign first performs lightweight coordinate unification in the physical world. This procedure requires only two calibration boards and a few photos, providing a simple and efficient means of aligning coordinate systems. The WiFi sensing system then collects CSI data, which undergoes preprocessing steps including denoising, segmentation, and temporal alignment. A CNN encoder is applied to extract features from the preprocessed data, which are subsequently integrated with high-dimensional embeddings of the WiFi transceiver geometry. The fused representation is finally used to infer the 3D human pose.

3.1 Lightweight Coordinate Unification.

The most straightforward approach for coordinate unification would be to manually measure transceiver positions with a ruler and then design a transformation matrix, but this is impractical in real applications due to time and labor requirements. In our method, we introduce an additional checkerboard $\mathbb{B}_1$ to represent the WiFi transceiver coordinate system in the physical world. Using standard camera calibration, we simultaneously establish the transformation from the camera coordinate system to the WiFi checkerboard coordinate system, $T_{C\to B_1}$, as well as the transformation from the camera to the world coordinate system, $T_{C\to B}$. With the camera as an intermediary, the mapping from the WiFi coordinate system to the world coordinate system can then be efficiently derived as

$$T_{B_1\to B} = T_{C\to B}^{-1}\, T_{C\to B_1}, \quad (11)$$

and the WiFi transceiver coordinates $P = \{\mathbf{p}_{\mathrm{tx}}, \mathbf{p}_{\mathrm{rx}_n}\}$ can be expressed in the unified world coordinate frame $\mathbb{B}$.

3.2 Geometry-Conditioned Learning.

For geometry-conditioned learning, we replace the blind regressor $f_\theta(H)$ by a conditioned estimator:

$$\hat{\mathbf{y}} = g_\theta(H, P), \quad (12)$$

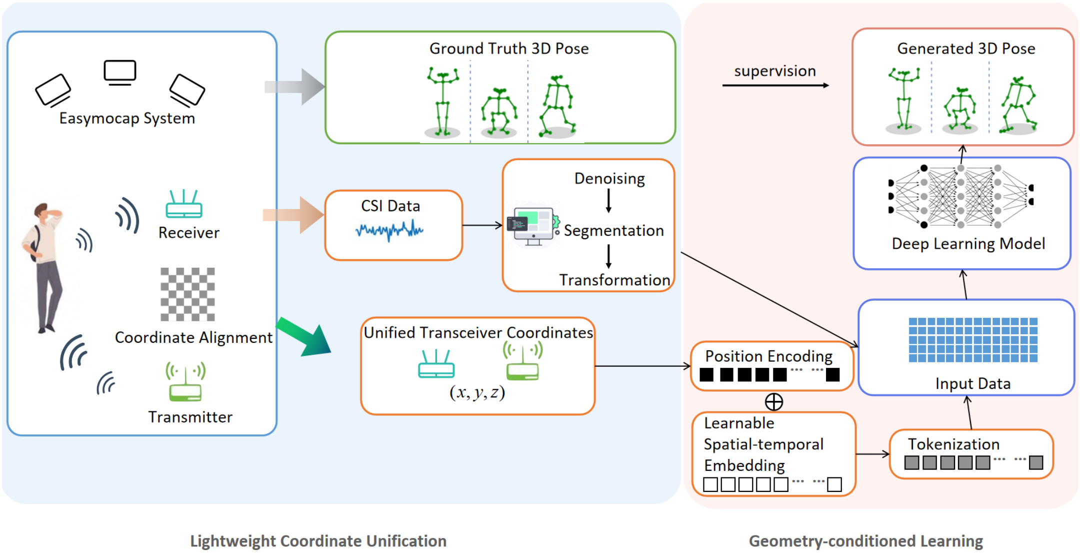


Figure 3: System overview.

where P reduces the effective uncertainty about s. Conditioning on P changes the posterior from $p(\mathbf{y} \mid H)$ to $p(\mathbf{y} \mid H, P)$, which is typically much sharper and more stable when P captures the dominant geometric degrees of freedom. Operationally, this reduces the need for the network to memorize deployment-specific transforms and yields markedly improved cross-deployment generalization. Specifically, we augments CSI features with compact spatial embeddings of receiver/transmitter locations and applies a lightweight fusion backbone to jointly reason about spatial and temporal evidence.

4 Method

In this section, we provide a detailed description of the implementation of the PerceptAlign framework.

4.1 Lightweight Coordinate Unification

The goal of this component is to conveniently transform the 3D spatial coordinates of WiFi transceivers into the world coordinate system B defined by the checkerboard $\mathbb{B}$. We achieve this by:(i) introducing an additional checkerboard $\mathbb{B}_1$ to align the sensing coordinate system B_1 of the WiFi transceiver pair, (ii) expressing transceiver device offsets in B_1 using measured inter-device distance, and (iii) capturing both $\mathbb{B}$ and $\mathbb{B}_1$ with a camera to obtain the corresponding transformation matrices, then using the camera as an intermediary to map WiFi transceiver coordinates from B_1 to B.

We begin with the following notation: for a point expressed in coordinate system X, its Euclidean coordinate is denoted as $\mathbf{p}_X \in \mathbb{R}^3$, and its homogeneous form is $\tilde{\mathbf{p}}_X = [\mathbf{p}_X^\top \ 1]^\top$. The homogeneous form vectorizes the 3D coordinate, allowing it to be manipulated in matrix form. The transformation of coordinates between reference frames using extrinsic parameters, i.e., a rigid transform, can be expressed as:

$$\mathbf{T}_{X\to Y} = \begin{bmatrix} \mathbf{R}_{X\to Y} & \mathbf{t}_{X\to Y} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathrm{SE}(3), \tag{13}$$

with $\tilde{\mathbf{p}}_Y = \mathbf{T}_{X\to Y}\,\tilde{\mathbf{p}}_X$.

In the physical setup, after the multi-camera system has been calibrated using $\mathbb{B}$, we deploy an additional checkerboard $\mathbb{B}_1$ between the WiFi transceivers. We place the origin of $\mathbb{B}_1$ at the midpoint of the line joining a transmitter T and a receiver R (or at a convenient marker near that midpoint), and align the x-axis of $\mathbb{B}_1$ with the T–R direction. In this way, the checkerboard coordinate system coincides with that of the WiFi sensing setup. Let the physical Euclidean distance between T and R be denoted as S (in meters). Writing $L = S/2$ for the half-distance, the device coordinates (offsets) in B_1 can then be expressed as:

$$\mathbf{p}_{B_1,T} = \begin{bmatrix} -L \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{p}_{B_1,R} = \begin{bmatrix} L \\ 0 \\ 0 \end{bmatrix}. \tag{14}$$

In practice S may be obtained either by direct physical measurement (tape measure) or by image-based measurement using the auxiliary camera that images the checkerboard. If the distance is measured in board grid units (for example g checkerboard squares) and each square has side length d (in metres), then $S = g \cdot d$ and the half-distance is $L = \frac{gd}{2}$. Equivalently, if the pixel distance between device markers is p and the pixel-to-metre scale on the checkerboard is ρ

(meters per pixel, obtained from board calibration), then $S = p \cdot \rho$ and $L = \frac{p\rho}{2}$. These conversion formulae give a single, implementation-ready rule:

$$L \;=\; \frac{1}{2}\cdot\Big(\text{measured distance}\Big) \;=\; \frac{1}{2}\cdot\begin{cases} S & \text{(direct)} \\ g\,d & \text{(grid units)} \\ p\,\rho & \text{(pixels)} \end{cases}. \tag{15}$$

Let $\mathbb{C}$ denote the auxiliary camera coordinate system that observes both $\mathbb{B}$ and $\mathbb{B}_1$. From checkerboard calibration we obtain camera-to-checkerboard transforms $\mathbf{T}_{C\to B}$ and $\mathbf{T}_{C\to B_1}$. The rigid transform that maps coordinates expressed in B_1 into the world coordinate system B is

$$\mathbf{T}_{B_1\to B} \;=\; \mathbf{T}_{C\to B}^{-1}\,\mathbf{T}_{C\to B_1}. \tag{16}$$

Therefore the device homogeneous coordinates in the world coordinate system B are obtained by:

$$\tilde{\mathbf{p}}_{B,\alpha} \;=\; \mathbf{T}_{B_1\to B}\,\tilde{\mathbf{p}}_{B_1,\alpha}, \qquad \alpha\in\{T,R\}, \tag{17}$$

or in non-homogeneous form

$$\mathbf{p}_{B,\alpha} \;=\; \mathbf{R}_{B_1\to B}\,\mathbf{p}_{B_1,\alpha} + \mathbf{t}_{B_1\to B}. \tag{18}$$

The procedure above is applied for every transceiver pair. If multiple receivers are observed simultaneously in a single auxiliary view, their positions can be read off in the same B_1 coordinate system and converted en masse via (16). The result is a set of calibrated device coordinates expressed in the unified world coordinate system:

$$P \;=\; \{\mathbf{p}_{B,\text{tx}},\ \mathbf{p}_{B,\text{rx}_1},\ \mathbf{p}_{B,\text{rx}_2},\dots\}. \tag{19}$$

We summarize the implementation summary as follows.

(1) Place the auxiliary board $\mathbb{B}_1$ so its origin is at the midpoint of the device pair and align its x-axis with the device line.
(2) Acquire images of $\mathbb{B}_1$ with the auxiliary camera and detect board corners to obtain $\mathbf{T}_{C\to B_1}$; likewise obtain $\mathbf{T}_{C\to B}$.
(3) Measure the inter-device distance in board-grid units g (so $S = gd$), or in pixels p (so $S = p\rho$); compute $L = S/2$.
(4) Form $\mathbf{p}_{B_1,\{T,R\}}$ via (14) and transform to the world coordinate system using (16).
(5) Repeat for all devices to obtain the full set P expressed in B.

By explicitly measuring P in the world coordinate system B, we make the previously invisible device geometry perceptible to the vision-based annotation system. This enables transceiver coordinates to be introduced as conditional knowledge within a unified reference frame, thereby guiding the model to learn deployment-invariant features. Our proposed lightweight unification strategy requires only the placement of two checkerboards and the capture of a few photos, allowing cross-modal coordinate alignment to be completed efficiently within a short time.

4.2 Geometry-conditioned learning

We implement the geometry-conditioned learning illustrated in Fig. 4. Each receiver's CSI tensor is encoded by a shared CNN; calibrated coordinates P are encoded via multi-band sinusoidal features and an MLP to form spatial tokens that are concatenated with CNN features. A Transformer-based spatio-temporal encoder fuses these tokens across receivers and time, and a lightweight decoder outputs camera-frame 3D skeleton joints.

Preprocessing. For each receiver R_n we compute the complex *CSI ratio* between antenna 1 and antenna 2 for denoising [50]:

$$\tilde{c}_n(t,f) \;=\; \frac{c_{n,1}(t,f)}{c_{n,2}(t,f)}, \tag{20}$$

where $c_{n,a}(t,f)$ denotes the CSI at receiver n, antenna a, time index t and subcarrier f. The ratio eliminates amplitude and phase noise introduced by hardware. After denoising, the CSI stream is segmented and temporally synchronized with each camera frame to serve as the network input. Specifically, we first divide the CSI sequence into several groups, where each group sequentially contains $G = \lfloor N_c/N_f \rfloor$ CSI samples, with N_c denoting the total number of CSI samples and N_f the number of video frames. Thus, each frame is paired with a fixed-length CSI group. From each group, we extract magnitude, phase, and Doppler Frequency Shift (DFS) [28] along the temporal axis. These features are concatenated following the procedure in WiGRUNT [11, 53], resized to $1 \times 224 \times 224$, and then replicated across three channels to form a $3 \times 224 \times 224$ tensor. This process is repeated for all WiFi transceivers to construct the CSI inputs corresponding to the visual annotations.

Shared CNN encoder. CSI input from all transceiver pairs is processed through a shared CNN encoder E_θ to extract features. In this work, E_θ is implemented as a pretrained ResNet-34 truncated before the final classification layer. It can be expressed as:

$$\mathbf{f}_{n,t} \;=\; \mathrm{Pool}\big(E_\theta(\mathbf{X}_{n,t})\big) \in \mathbb{R}^D, \tag{21}$$

where $\mathbf{f}_{n,t}$ is the feature, and $\mathbf{X}_{n,t}$ denotes the input tensor for receiver n at frame t, $n \in \{1,\dots,N_r\}$, and $t \in \{1,\dots,T\}$.

Position encoding. We represent the geometry of each antenna pair using the coordinate offset of the receiver relative to the transmitter in the world coordinate system. Let $\mathbf{p}_n = (x_n, y_n, z_n) \in \mathbb{R}^3$ denote the 3D geometric relation of receiver R_n. Each receiver coordinate $\mathbf{p}_n$ is then lifted

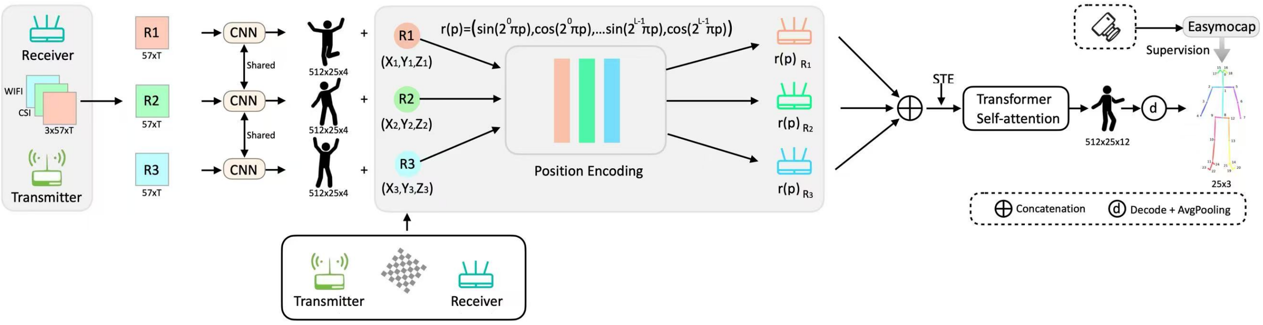


Figure 4: Geometry-conditioned learning network architecture.

into a high-dimensional spatial embedding through a multi-frequency mapping $\Phi : \mathbb{R}^3 \to \mathbb{R}^{D_p}$:

$$\Phi(\mathbf{p}) \;=\; \Big[\sin(2^k\pi\mathbf{p}),\; \cos(2^k\pi\mathbf{p})\Big]_{k=0}^{K-1}, \tag{22}$$

and projected by a small MLP g_ψ to obtain the spatial embedding

$$\mathbf{e}_n \;=\; g_\psi\big(\Phi(\mathbf{p}_n)\big) \in \mathbb{R}^D. \tag{23}$$

This step is designed to enrich the geometric representation by expanding the $\mathbf{p}_n$ across multiple frequencies, enabling the model to capture more complex patterns. It also prevents the model from relying solely on shallow features such as simple relative displacements, and further encodes the geometry into an embedding space with a dimensionality comparable to CSI features, which facilitates effective feature fusion.

Temporal encoding. To represent temporal order we associate each frame t with a learnable temporal embedding (STE) $\mathbf{r}_t \in \mathbb{R}^D$ (shared across receivers). A small, receiver-specific bias $\mathbf{s}_n \in \mathbb{R}^D$ is also maintained to capture per-receiver idiosyncrasies.

Token construction. For every receiver n and frame t we form a token that fuses CNN features with spatial and temporal encodings:

$$\mathbf{u}_{n,t}^{(0)} \;=\; \mathrm{LayerNorm}\big(\mathbf{W}_f\mathbf{f}_{n,t} + \mathbf{W}_e\mathbf{e}_n + \mathbf{r}_t + \mathbf{s}_n\big) \in \mathbb{R}^D, \tag{24}$$

where $\mathbf{W}_f, \mathbf{W}_e$ are learned linear projections. The full input sequence to the decoder is the concatenation of these tokens over receivers and time:

$$\mathcal{U}^{(0)} \;=\; \{\mathbf{u}_{n,t}^{(0)}\}_{n=1,t=1}^{N_r,T}, \tag{25}$$

of length $N_r * T$. In the final token, the spatial embedding provides the model with geometric priors of the transceivers, while the temporal embedding enables the attention layers to condition temporal evidence on explicit spatial priors at each timestep. This design encourages the model to learn time-varying observation-to-geometry mappings, leading to improved temporal consistency and higher pose estimation accuracy.

Decoding and pose estimation. The token sequence $\mathcal{U}^{(0)}$ is processed by a Transformer-based encoder with L layers of multi-head self-attention:

$$\mathcal{U}^{(\ell)} \;=\; \mathcal{TF}^{(\ell)}\big(\mathcal{U}^{(\ell-1)}\big), \qquad \ell = 1, \ldots, L, \tag{26}$$

producing contextualized tokens $\mathbf{u}_{n,t}^{(L)}$. For each frame t the N_r tokens $\{\mathbf{u}_{n,t}^{(L)}\}_{n=1}^{N_r}$ are aggregated to produce a frame-level feature $\mathbf{z}_t \in \mathbb{R}^D$. A pose head h_ϕ then regresses the predicted joints:

$$\hat{\mathbf{y}}_t \;=\; h_\phi(\mathbf{z}_t) \in \mathbb{R}^{J\times 3}. \tag{27}$$

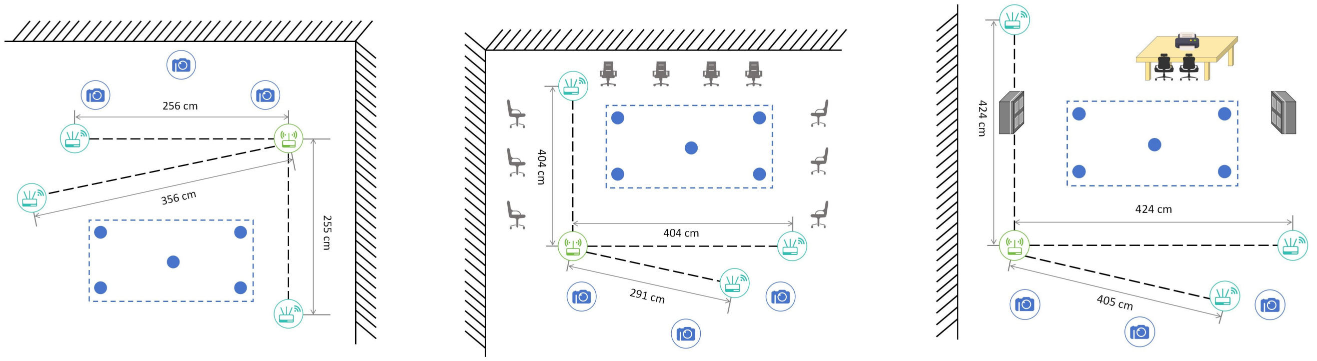
Loss function. We train the network using a simple mean-squared-error (MSE) loss between the predicted and ground-truth skeleton joint coordinates. Let $\hat{\mathbf{y}}_{l,j} \in \mathbb{R}^3$ and $\mathbf{y}_{l,j} \in \mathbb{R}^3$ denote the predicted and ground-truth positions of joint j at video frame l, respectively, with $l = 1, \ldots, L$ and $j = 1, \ldots, J$. The training objective is

$$\mathcal{L}_{\mathrm{MSE}} \;=\; \frac{1}{LJ}\sum_{l=1}^{L}\sum_{j=1}^{J}\big\|\hat{\mathbf{y}}_{l,j} - \mathbf{y}_{l,j}\big\|_2^2. \tag{28}$$

Once the system is trained, deploying it in a new environment requires only minimal setup. The user places $\mathbb{B}$ at any convenient location and $\mathbb{B}_1$ at the new transceiver pair to complete the position encoding process. By providing these encodings as conditional parameters to the model, the pre-trained system can be directly applied in the new settings.

5 Datasets

To validate our approach, we design and collect the largest cross-domain WiFi-based 3D pose estimation dataset. The dataset will be released publicly upon acceptance of this paper.



(a) scene_1 (empty room) **(b) scene_2 (meeting room)** **(c) scene_3 (office)**

Figure 5: Environments and device placements used for data collection. Scene_3 contains multiple layouts (A/B/C) as documented in the dataset README file.

5.1 Dataset overview

The dataset contains recordings from 21 participants across three indoor environments. As shown in Fig.5, these environments include an empty room, a meeting room, and an office. Data collection spanned six months, resulting in 370.66 GB of recordings and over 82 hours of video. The dataset comprises 6,369,186 synchronized frames. Each capture includes synchronized RGB video and raw WiFi CSI from three receivers, calibration metadata that registers WiFi devices into the world coordinate system, and multi-view 3D joint annotations reconstructed with EasyMocap. Three WiFi transceivers were used in this dataset. More details please refer to Appendix A.1

5.2 Action taxonomy

We record 18 daily-action categories designed to cover a wide range of limb motions and dynamics. The actions are: (1) left arm stretch, (2) right arm stretch, (3) both-arms stretch, (4) left lateral raise, (5) right lateral raise, (6) left forward lunge, (7) right forward lunge, (8) left side lunge, (9) right side lunge, (10) jump, (11) pick-up, (12) clockwise spin, (13) counterclockwise spin, (14) jumping jack, (15) squat, (16) left rotation, (17) right rotation, and (18) directional hops (forward/back/left/right).

5.3 Calibration and ground-truth

We register WiFi devices into the world coordinate system using a lightweight coordinate unification method described in Section 4.1. We recorded the unification result for each placement as a calibrated set of transmitter/receiver coordinates stored in the dataset metadata. Visual ground-truth are estimated using EasyMocap; frames with EasyMocap/OpenPose confidence < 0.8 are filtered and the remaining annotations are spot-checked manually.

5.4 Domain splits

To facilitate reproducible evaluation, we provide standardized splits under different testing scenarios. **Per-scene:** 80/20 train/test split by action instances for each scene. **Cross-user:** leave one subject out across 21 participants. **Cross-scene:** leave one scene out (train on two scenes, test on the third). **Cross-setup (scene_3):** leave one setup-out among Setup A/B/C. **Cross-orientation :** leave one orientation-out among the three orientations. **Cross-Location:** leave one location out among the five capture points in each scene. A detailed description of the partitioning scheme also included in the dataset's README file.

Table 1: Dataset comparison,and PiW means Person-in-WiFi.

Dataset	#Subjects	#Actions	#Frames	Multi-layout
MM-Fi	40	16	320k	No
PiW-3D	7	8	97k	No
Ours	21	18	637k	Yes

5.5 Dataset comparison

Table 1 compares our dataset with existing large-scale WiFi-based human pose estimation datasets. Compared to Person-in-WiFi-3D [45], our dataset significantly expands both subject diversity and overall scale. This broader coverage mitigates overfitting to individual motion styles and provides a stronger basis for evaluating cross-user generalization. Although the number of participants is smaller than in MM-Fi [46], our dataset places greater emphasis on the cross-domain challenges most relevant to real-world applications. Unlike prior datasets, which include cross-environment data collection but use the same WiFi device layout within each environment, our dataset assigns a different device layout to every scenario. This design enables rigorous evaluation under simultaneous environment and layout shifts, a condition that is far more common in real-world applications. We also include multiple layouts within the same environment to analyze layout-induced variations, and capture extensive data across diverse user orientations and positions to support more comprehensive generalization studies.

Beyond pose estimation, our dataset can also support cross-domain activity recognition research, as it encompasses a wide variety of actions. Beyond scale and diversity, a key advantage of our release is the inclusion of detailed geometric information of the sensing system, along with scene photographs capturing the calibration checkerboards. Our work demonstrates that such geometric information plays a crucial role in enabling the development of more generalizable WiFi sensing systems. We therefore believe that this new dataset, with its comprehensive geometric annotations, will greatly benefit the research community.

5.6 Limitations and usage notes

The dataset targets single-person indoor pose estimation and cross-domain evaluation; it does not include multi-person or outdoor scenes. The effective temporal resolution for pose recovery is bounded by commodity CSI packet rates and by the grouping strategy. Users should consult the provided metadata to select samples with desired temporal granularity. Consent and ethics: participants signed informed-consent forms; visual data are provided under a controlled-access policy (details in the release).

6 Evaluations

Model settings. All experiments follow a fixed training recipe and model configuration. The CNN encoder is a pre-trained ResNet-34, with its pooled output projected to a 512-dimensional feature space. WiFi transceiver coordinates are encoded using a multi-frequency mapping with $K = 8$ frequency bands and similarly projected to dimension 512. Decoding is performed by a 6-layer Transformer encoder with 8 attention heads and a hidden size of 512. Models are trained end-to-end using Adam with an initial learning rate of 1×10^{-4}, decayed to 1×10^{-6} via a cosine-annealing schedule over 200 epochs. The batch size is 64, with weight decay set to 1×10^{-5}. Training minimizes MSE loss on 3D skeleton coordinates, and evaluation is reported using MPJPE (mm). For comparison, we evaluate the state-of-the-art method Person-in-WiFi-3D [45] (using the author-released weights). We choose this baseline because many recent studies have not released their code, whereas Person-in-WiFi-3D provides open-source implementations and reports performance surpassing prior work.

Evaluation setup. We evaluate pose estimation performance using the standard Mean Per-Joint Position Error (MPJPE). Let $\hat{\mathrm{y}}_{l,j} \in \mathbb{R}^3$ and $\mathrm{y}_{l,j} \in \mathbb{R}^3$ denote the predicted and ground-truth 3D positions of joint j at frame l, respectively, with $l = 1, \ldots, L$ and $j = 1, \ldots, J$. The MPJPE (reported in millimeters) is computed as the average Euclidean distance between predictions and ground truth over all joints and frames:

$$\text{MPJPE} \;=\; \frac{1}{LJ} \sum_{l=1}^{L} \sum_{j=1}^{J} \left\| \hat{\mathrm{y}}_{l,j} - \mathrm{y}_{l,j} \right\|_2. \tag{29}$$

6.1 In-domain Evaluation

We report in-domain MPJPE for models trained and tested within the same scene in Table 2 and Figure 6. For brevity, we present the overall mean MPJPE across the three scenes. All reported models share the same training and evaluation protocol as well as the same backbone encoder architecture. Each scene is split 80/20 by action instances, and errors are averaged across all test frames. Compared to Person-in-WiFi-3D, PerceptAlign reduces the overall in-domain MPJPE from 221.0mm to 137.2mm, corresponding to a relative improvement of approximately 38%. This sizable gain demonstrates that incorporating WiFi device geometry as prior conditional knowledge can significantly improve 3D pose estimation accuracy. We attribute this to the fact that, even though device layouts remain unchanged in in-domain settings, our approach provides the model with explicit awareness of the transceiver configuration. This allows the model to interpret CSI perturbations in a well-defined spatial context, thereby enhancing its ability to map CSI variations caused by human activity to accurate 3D poses.

6.2 Cross-Domain Evaluation

To evaluate system performance under realistic deployment variations, we conduct cross-domain experiments targeting different sources of distribution shift: spatial location, subject orientation, environment, subject identity, and device layout. Each protocol follows a leave-one-out design, and results are summarized in Table 2. Representative samples are shown in Fig. 6.

Cross-Location. This section evaluates how model performance changes when subject positions vary within the same room and device layout. In each round, one position is held out for testing while the remaining positions are used for training. Compared to in-domain settings, Person-in-WiFi-3D shows significant degradation (253.1 mm), whereas PerceptAlign also experiences some decline but reduces the error

Table 2: Evaluation Results (MPJPE (mm)).

Method	In-doamin	Location	Orientation	Environment	Layout	User
Person-in-WiFi 3D	221.0	253.1	254	717.2	649.3	266.7
PerceptAlign	**137.2**	**144.6**	**147.7**	**181.5**	**170.2**	**145.3**

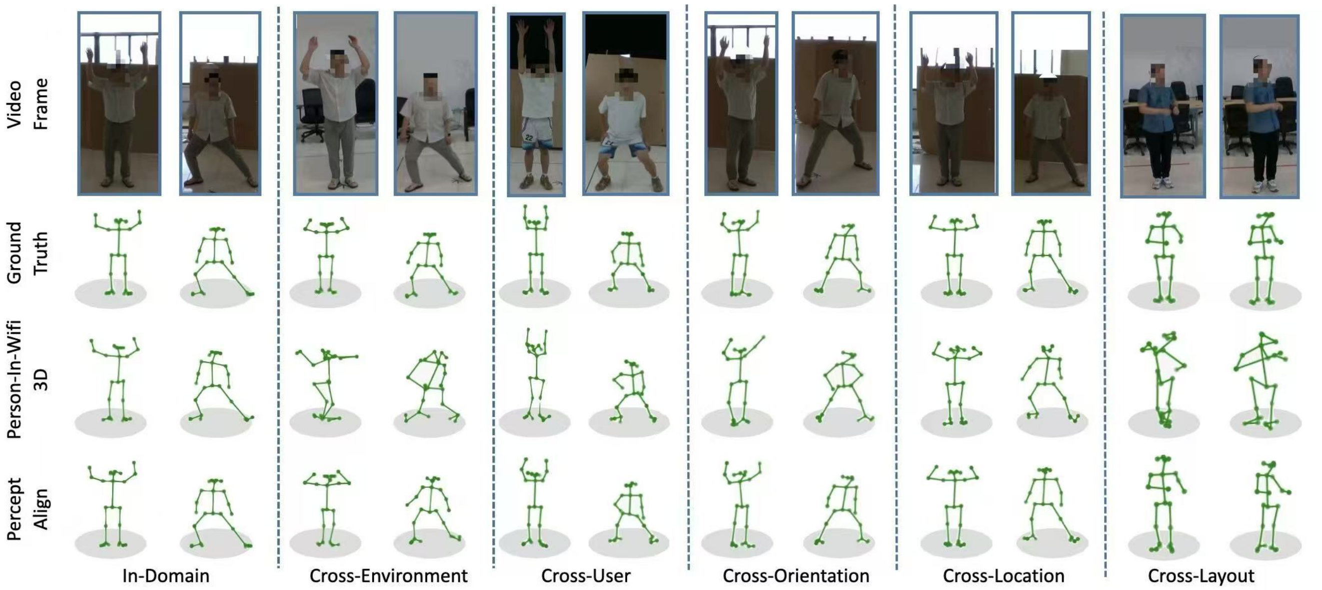


Figure 6: Illustrative examples of constructed skeletons across different cross-domain settings.

to 144.6 mm (43% improvement) by incorporating geometric conditioning.

Cross-Orientation. For variations in subject orientation, we evaluate performance using a leave-one-orientation-out scheme while keeping other conditions fixed. Person-in-WiFi-3D reaches 254 mm, indicating that orientation changes hinder generalization. In contrast, PerceptAlign achieves 147.7 mm, a 42% reduction in error, demonstrating the effectiveness of our strategy under this setting.

Cross-User. In cross-user evaluation, Person-in-WiFi-3D and our method achieve MPJPEs of 266.7 mm and 145.3 mm, respectively. These results highlight the non-negligible impact of user variation on pose estimation performance, while PerceptAlign maintains robustness through its geometry-conditioned learning strategy.

Cross-Layout. The cross-layout setting is a central focus of our work and a key motivation, since layout variations frequently occur in real-world deployments. Results confirm that changes in WiFi device layout severely degrade the performance of existing pipelines: even the state-of-the-art Person-in-WiFi-3D becomes nearly unusable under this setting, with an average error of 649.3 mm. In contrast, our method achieves an MPJPE of 170.2 mm, demonstrating strong robustness to layout changes and validating the effectiveness of the proposed geometry-aware approach.

Cross-Environment. Our cross-environment evaluation is more challenging, as it jointly considers changes in both indoor scenes and device layouts—variations that are common in practical deployments. Results show that Person-in-WiFi-3D suffers severe degradation, with an average error of 717.2 mm. Although PerceptAlign also experiences some decline, it maintains an MPJPE of only 181.5 mm, demonstrating substantially improved robustness in pose estimation.

Overall, these results validate our earlier claim that existing methods overfit by implicitly memorizing WiFi device layouts, conflating them with useful motion cues. The more than 60% improvement over the SOTA method across diverse cross-domain settings further demonstrates the effectiveness of our approach: by making device geometry an explicit condition, the model is able to disentangle deployment factors from human motion, leading to significantly stronger generalization.

Table 3: Ablation results (MPJPE (mm)). "No Align" omits geometry-conditioned spatial position embedding; "No Spatial PE" injects raw 3D device vectors instead of high-dimensional spatial embeddings.

Variant	In-domain	Cross-envir	Cross-layout
PerceptAlign	**137.2**	**181.5**	**170.2**
No Align	279.0	729.5	687.0
No Spatial PE	297.2	744.0	692.3

6.3 Ablation study

Table 3 evaluates the contribution of the two core components of PerceptAlign: the geometry-conditioned spatial embedding and the spatial positional encoding (PE) method, which lifts 3D device coordinates into a higher-dimensional representation before fusion.

1) Coordinate unification. Removing the geometry-conditioned spatial embedding, i.e., training without calibrated device layouts, increases in-domain MPJPE from 137.2 mm to 279.0 mm and cross-domain errors by over 300% (e.g., cross-environment 181.5 → 729.5 mm). This shows that omitting device geometry forces the model to overfit to layout-specific cues: it can fit a single deployment but fails catastrophically under deployment shifts. These results confirm the necessity of incorporating device geometry as explicit conditional knowledge to achieve robust generalization.

2) High-dimensional spatial encoding. Replacing the high-dimensional spatial positional encoding with raw 3D coordinate concatenation leads to severe degradation, with in-domain MPJPE rising from 137.2 mm to 297.2 mm and cross-scene error from 181.5 mm to 744.0 mm. This shows that simply appending coordinates as low-dimensional side information is ineffective: raw vectors lack the nonlinear basis to capture spatial patterns and disrupt feature extraction due to scale mismatch with CSI features. In contrast, lifting coordinates into high-dimensional embeddings provides richer representations that fuse smoothly with CSI, enabling the model to learn geometry-conditioned patterns and maintain robustness under deployment shifts.

7 Limitations and Future Work

Limitations. Despite these strengths, several limitations remain. First, CSI remains noisy and temporally aliased at commodity packet rates; the method's temporal resolution is therefore bounded by capture hardware and the grouping strategy used to align CSI with video frames. Second, while positional encoding and Transformer fusion improve generalization, deployment changes that radically alter the set of effective links (for example, extreme TX–RX reconfiguration or highly cluttered scenes) still cause noticeable performance degradation; targeted domain-adaptation or lightweight calibration transfer techniques may be required to fully close this gap (similar challenges have been observed for other WiFi-based systems). Finally, the present evaluation focuses on single-person scenarios; extending the approach to robust multi-person 3D human pose estimation raises additional challenges (data association, overlapping multipath), as identified in recent multi-person WiFi work. In practice, these limitations point to several concrete research avenues: (i) increasing CSI temporal resolution or incorporating complementary RF modalities (e.g., ultra-wide-band or mmWave) to better capture fast motions; (ii) developing calibration-light domain adaptation or self-supervised fine-tuning routines so that minimal labeled data in a new deployment suffice to restore performance; and (iii) scaling the dataset and model design to support multi-person scenarios while preserving computational efficiency.

Future Work. In future work, we aim to develop a calibration framework for WiFi sensing systems analogous to the intrinsic and extrinsic calibration of multi-camera systems using checkerboards. For intrinsic calibration, our goal is to eliminate deployment-independent hardware biases so that CSI measurements more closely approximate "ideal physical propagation plus minor noise." To this end, we plan to experiment with direct transceiver connections via coaxial cables as well as antenna array parameter estimation techniques. For extrinsic calibration, our objective is to rapidly unify different WiFi coordinate systems, similar to how cameras establish a consistent mapping from the world coordinate system to individual camera frames. Possible directions include extending our current checkerboard-based vision-assisted calibration method or leveraging the correlation between CSI and controlled motion trajectories. We believe this line of research is critical for bringing WiFi sensing systems closer to practical deployment.

8 Conclusion

This work identified coordinate overfitting as the main bottleneck in WiFi-based 3D pose estimation, where models memorize deployment-specific layouts. We introduced PerceptAlign, a geometry-conditioned framework that unifies WiFi and vision into a shared 3D space and encodes transceiver positions as priors, disentangling motion from deployment artifacts and enabling layout-invariant features. We also built the largest ross-domain 3D WiFi-based human pose estimation dataset and showed that PerceptAlign reduces in-domain error by 38% and cross-domain error by over 60% compared to state-of-the-art baselines. Future work will explore standardized intrinsic and extrinsic calibration protocols to support scalable deployment.

Acknowledgments

Thanks...

References

[1] 2021. EasyMoCap - Make human motion capture easier. Github. https://github.com/zju3dv/EasyMocap

[2] Xiaoqi An, Lin Zhao, Chen Gong, Jun Li, and Jian Yang. 2025. Pre-training a Density-Aware Pose Transformer for Robust LiDAR-based 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1755–1763.

[3] Andrea Avogaro, Federico Cunico, Bodo Rosenhahn, and Francesco Setti. 2023. Markerless human pose estimation for biomedical applications: a survey. *Frontiers in Computer Science* 5 (2023), 1153160.

[4] Yang Chen, Jingcai Guo, Song Guo, Jingren Zhou, and Dacheng Tao. 2025. Towards Robust and Realistic Human Pose Estimation via WiFi Signals. *arXiv preprint arXiv:2501.09411* (2025).

[5] Yi-Chung Chen, Zhi-Kai Huang, Lu Pang, Jian-Yu Jiang-Lin, Chia-Han Kuo, Hong-Han Shuai, and Wen-Huang Cheng. 2023. Seeing the unseen: Wifi-based 2D human pose estimation via an evolving attentive spatial-Frequency network. *Pattern Recognition Letters* 171 (2023), 21–27.

[6] Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen. 2024. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In *European Conference on Computer Vision*. Springer, 93–111.

[7] Jie Deng, Kaiqi Chen, Pengsen Jing, Guannan Dong, Min Yang, Aichun Zhu, and Yifeng Li. 2025. CSI-Channel Spatial Decomposition for WiFi-Based Human Pose Estimation. *Electronics* 14, 4 (2025), 756.

[8] Jiaqi Geng, Dong Huang, and Fernando De la Torre. 2022. Densepose from wifi. *arXiv preprint arXiv:2301.00250* (2022).

[9] Toan D Gian, Tien-Hoa Nguyen, Nhan Thanh Nguyen, and Van-Dinh Nguyen. 2024. WiLHPE: WiFi-enabled Lightweight Channel Frequency Dynamic Convolution for HPE Tasks. In *2024 Tenth International Conference on Communications and Electronics (ICCE)*. IEEE, 516–521.

[10] Yangyang Gu, Jing Chen, Congrui Chen, Kun He, Jia Ju, Yebo Feng, Ruiying Du, and Cong Wu. 2025. CSIPose: Unveiling Human Poses Using Commodity WiFi Devices Through the Wall. *IEEE Transactions on Mobile Computing* (2025).

[11] Yu Gu, Xiang Zhang, Yantong Wang, Meng Wang, Huan Yan, Yusheng Ji, Zhi Liu, Jianhua Li, and Mianxiong Dong. 2022. WiGRUNT: WiFi-enabled gesture recognition using dual-attention network. *IEEE transactions on human-machine systems* 52, 4 (2022), 736–746.

[12] Yu Gu, Xiang Zhang, Huan Yan, Jingyang Huang, Zhi Liu, Mianxiong Dong, and Fuji Ren. 2023. WiFE: WiFi and vision based unobtrusive emotion recognition via gesture and facial expression. *IEEE Transactions on Affective Computing* 14, 4 (2023), 2567–2581.

[13] Xueqiang Han, Tianyue Zheng, Tony Xiao Han, and Jun Luo. 2025. RayLoc: Wireless Indoor Localization via Fully Differentiable Ray-tracing. *arXiv preprint arXiv:2501.17881* (2025).

[14] Ting-Wei Hsu and Hung-Yun Hsieh. 2024. Robust Multi-User Pose Estimation Based on Spatial and Temporal Features from WiFi CSI. In *ICC 2024-IEEE International Conference on Communications*. IEEE, 1600–1605.

[15] Huakun Huang, Peiliang Wang, Lingjun Zhao, Zeyang Dai, Guowei Liu, and Honghao Gao. 2025. WiPE: Privacy-friendly WiFi-based Human Pose Estimation on Consumer Platform. *IEEE Transactions on Consumer Electronics* (2025).

[16] Sijie Ji, Xuanye Zhang, Yuanqing Zheng, and Mo Li. 2023. Construct 3d hand skeleton with commercial wifi. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 322–334.

[17] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[18] Hyeon-Ju Lee and Seok-Jun Buu. 2025. Wi-Fi-enabled Vision via Spatially-variant Pose Estimation based on Convolutional Transformer Network. *IEEE Access* (2025).

[19] Wouter Lemoine, Nabeel Nisar Bhat, Jakob Struye, Andrey Belogaev, Jesus Omar Lacruz, Joerg Widmer, and Jeroen Famaey. 2025. WIP: Distributed inference for human pose estimation using mmWave Wi-Fi. In *2025 IEEE 26th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 141–144.

[20] Xin Li, Jingzhi Hu, Hongbo Wang, Zhe Chen, and Jun Luo. 2025. Enabling Ultra-Wideband Wi-Fi Sensing via Sparse Channel Sampling. *IEEE Journal on Selected Areas in Communications* (2025).

[21] Jianchu Lin, Shuang Li, Hong Qin, Hongchang Wang, Ning Cui, Qian Jiang, Haifang Jian, and Gongming Wang. 2023. Overview of 3d human pose estimation. *CMES-Computer Modeling in Engineering & Sciences* 134, 3 (2023).

[22] Jinyi Liu, Wenwei Li, Tao Gu, Ruiyang Gao, Bin Chen, Fusang Zhang, Dan Wu, and Daqing Zhang. 2023. Towards a dynamic fresnel zone model to wifi-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–24.

[23] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. 2025. Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5478–5486.

[24] Yuquan Luo, Yuqiang He, Yaxin Li, Huaiqiang Liu, Jun Wang, and Fei Gao. 2025. A Sliding Window-Based CNN-BiGRU Approach for Human Skeletal Pose Estimation Using mmWave Radar. *Sensors* 25, 4 (2025), 1070.

[25] Soroush Mehraban, Vida Adeli, and Babak Taati. 2024. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 6920–6930.

[26] Rama Bastola Neupane, Kan Li, and Tesfaye Fenta Boka. 2024. A survey on deep 3D human pose estimation. *Artificial Intelligence Review* 58, 1 (2024), 24.

[27] Xuan Hoang Nguyen, Van-Dinh Nguyen, Quang-Trung Luu, Toan Dinh Gian, and Oh-Soon Shin. 2025. Robust wifi sensing-based human pose estimation using denoising autoencoder and cnn with dynamic subcarrier attention. *IEEE Internet of Things Journal* (2025).

[28] Kai Niu, Xuanzhi Wang, Fusang Zhang, Rong Zheng, Zhiyun Yao, and Daqing Zhang. 2022. Rethinking Doppler effect for accurate velocity estimation with commodity WiFi devices. *IEEE Journal on Selected Areas in Communications* 40, 7 (2022), 2164–2178.

[29] Yanyi Qu, Haoyang Ma, and Wenhui Xiong. 2025. MultiFormer: A Multi-Person Pose Estimation System Based on CSI and Attention Mechanism. *arXiv preprint arXiv:2505.22555* (2025).

[30] Yiming Ren, Xiao Han, Chengfeng Zhao, Jingya Wang, Lan Xu, Jingyi Yu, and Yuexin Ma. 2024. Livehps: Lidar-based scene-level human pose and shape estimation in free environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1281–1291.

[31] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. GoPose: 3D human pose estimation using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.

[32] Souvik Sen, Jeongkeun Lee, Kyu-Han Kim, and Paul Congdon. 2013. Avoiding multipath to revive inbuilding WiFi localization. In *Proceeding*

of the 11th annual international conference on Mobile systems, applications, and services. 249–262.

[33] Wenchao Song, Zhu Wang, Yifan Guo, Zhuo Sun, Zhihui Ren, Chao Chen, Bin Guo, Zhiwen Yu, Xingshe Zhou, and Daqing Zhang. 2024. FinerSense: A Fine-Grained Respiration Sensing System Based on Precise Separation of Wi-Fi Signals. *IEEE Transactions on Mobile Computing* (2024).

[34] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 2023. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4713–4725.

[35] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–34.

[36] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-Grained Person Perception Using WiFi. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2019).

[37] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5452–5461.

[38] Hongbo Wang, Xin Li, Jiachun Li, Haojin Zhu, and Jun Luo. 2025. VR-Fi: Positioning and Recognizing Hand Gestures via VR-embedded Wi-Fi Sensing. *IEEE Transactions on Mobile Computing* (2025).

[39] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 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.

[40] Yichao Wang, Yili Ren, and Jie Yang. 2024. Multi-Subject 3D Human Mesh Construction Using Commodity WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–25.

[41] Yuxuan Weng, Tianyue Zheng, Yanbing Yang, and Jun Luo. 2025. FM-Fi 2.0: Foundation Model for Cross-Modal Multi-Person Human Activity Recognition. *IEEE Transactions on Mobile Computing* (2025).

[42] Dan Wu, Youwei Zeng, Fusang Zhang, and Daqing Zhang. 2022. WiFi CSI-based device-free sensing: from Fresnel zone model to CSI-ratio model. *CCF Transactions on Pervasive Computing and Interaction* 4, 1 (2022), 88–102.

[43] Huan Yan, Xiang Zhang, Jinyang Huang, Yuanhao Feng, Meng Li, Anzhi Wang, Weihua Ou, Hongbing Wang, and Zhi Liu. 2025. Wi-sfdagr: Wifi-based cross-domain gesture recognition via source-free domain adaptation. *IEEE Internet of Things Journal* (2025).

[44] Huan Yan, Yong Zhang, Yujie Wang, and Kangle Xu. 2019. WiAct: A passive WiFi-based human activity recognition system. *IEEE Sensors Journal* 20, 1 (2019), 296–305.

[45] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. 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.

[46] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2023. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems* 36 (2023), 18756–18768.

[47] Jianfei Yang, Yunjiao Zhou, He Huang, Han Zou, and Lihua Xie. 2022. MetaFi: Device-free pose estimation via commodity WiFi for metaverse avatar simulation. In *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. IEEE, 1–6.

[48] Dongqiangzi Ye, Yufei Xie, Weijia Chen, Zixiang Zhou, Lingting Ge, and Hassan Foroosh. 2024. LPFormer: LiDAR pose estimation transformer with multi-task network. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 16432–16438.

[49] Calvin Yeung, Tomohiro Suzuki, Ryota Tanaka, Zhuoer Yin, and Keisuke Fujii. 2025. AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5945–5956.

[50] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[51] Lei Zhang, Haoran Ning, Jiaxin Tang, Zhenxiang Chen, Yaping Zhong, and Yahong Han. 2025. WiViPose: A Video-aided Wi-Fi Framework for Environment-Independent 3D Human Pose Estimation. *IEEE Transactions on Multimedia* (2025).

[52] Xiang Zhang, Yu Gu, Huan Yan, Yantong Wang, Mianxiong Dong, Kaoru Ota, Fuji Ren, and Yusheng Ji. 2023. Wital: A COTS WiFi devices based vital signs monitoring system using NLOS sensing model. *IEEE Transactions on Human-Machine Systems* 53, 3 (2023), 629–641.

[53] Xiang Zhang, Jinyang Huang, Huan Yan, Yuanhao Feng, Peng Zhao, Guohang Zhuang, Zhi Liu, and Bin Liu. 2025. Wiopen: A robust wi-fi-based open-set gesture recognition framework. *IEEE Transactions on Human-Machine Systems* (2025).

[54] Xinyu Zhang, Zhonghao Ye, Jingwei Zhang, Xiang Tian, Zhisheng Liang, and Shipeng Yu. 2025. VST-Pose: A Velocity-Integrated Spatiotem-poral Attention Network for Human WiFi Pose Estimation. *arXiv preprint arXiv:2507.09672* (2025).

[55] Xiang Zhang, Jie Zhang, Zehua Ma, Jinyang Huang, Meng Li, Huan Yan, Peng Zhao, Zijian Zhang, Bin Liu, Qing Guo, et al. 2025. Camlopa: A hidden wireless camera localization framework via signal propagation path analysis. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3653–3671.

[56] Xiang Zhang, Jie Zhang, Huan Yan, Jinyang Huang, Zehua Ma, Bin Liu, Meng Li, Kejiang Chen, Qing Guo, Tianwei Zhang, and Zhi Liu. 2025. DiffLoc: WiFi Hidden Camera Localization Based on Electromagnetic Diffraction. In *The 34th USENIX Security Symposium*. Seattle, WA, USA.

[57] Peng Zhao, Jinyang Huang, Xiang Zhang, Zhi Liu, Huan Yan, Meng Wang, Guohang Zhuang, Yutong Guo, Xiao Sun, and Meng Li. 2025. Wi-Pulmo: Commodity WiFi Can Capture Your Pulmonary Function Without Mouth Clinging. *IEEE Internet of Things Journal* 12, 1 (2025), 854–868.

[58] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. 2023. MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal* 10, 16 (2023), 14128–14136.

[59] Yunjiao Zhou, Jianfei Yang, He Huang, and Lihua Xie. 2024. AdaPose: Toward Cross-Site Device-Free Human Pose Estimation With Commodity WiFi. *IEEE Internet of Things Journal* 11, 24 (2024), 40255–40267.

A Related Work

RF-based human pose estimation fall into (1) specialized radar or mmWave imaging, which yields high spatial resolution but relies on dedicated sensors [2, 19, 24, 30, 48]; and (2) commodity WiFi CSI-based methods, which trade some physical resolution for ubiquity and low cost.

Table 4: Acquisition hardware and nominal parameters.

Modality	Hardware	Key parameters
RGB video	Intel RealSense 435D	1980×1080, 30 fps
WiFi CSI	Intel 5300 + CSI tool	1 TX, 3 RX; 3 antennas / RX; 57 subcarriers
Calibration board	Checkerboard (12×9)	square size 30 mm; EasyMocap-compatible
Ground-truth	EasyMocap	camera-centric 3D skeleton reconstruction

Early RF studies demonstrated robust activity recognition and coarse localization using RSSI/CSI features and classical machine learning[32, 39, 44]. As research advanced, deep learning models began to extract richer spatio-temporal patterns from CSI, enabling finer-grained tasks such as pose estimation [8, 10, 17].

(1) Extending from 2D to 3D: Wang et al. [37] first used Wi-Fi devices to achieve human pose estimation by combining joint heat maps (JHMs) and body part affinity fields (PAFs) from OpenPose to supervise deep learning models. Later, Yan et al. [45] expanded to 3D pose estimation via a architecture comprising a Wi-Fi encoder, pose decoder, and fine decoder. (2) Multi-person estimation: Qu et al. [29], Hsu et al. [14], and Yan et al. [45] explored multi-user pose estimation via improved resolution, loss functions, and user separation. (3) Improving accuracy: Huang et al. [15] leveraged the sparsity of joint heat maps and introduced a 3D streaming signal fusion module. Nguyen et al. [27] proposed an autoencoder denoiser and an estimator focusing on informative OFDM subcarriers. Deng et al. [7] applied CSI spatial decomposition to observe spatial and channel-sensitive views. Zhang et al. [54] introduced Vista-Former with dual-stream spatiotemporal attention. Lee et al. [18] combined CNNs and transformers for spatiotemporal feature extraction. Gian et al.[6] proposed a multi-branch CNN with selective kernel attention, and extended it in [9] with a teacher-student framework to enhance resolution efficiently. (4) Cross-domain generalization:Several recent papers propose domain-invariant representations, topology- or physics-informed regularizers, or time–frequency fusion strategies to improve robustness. Chen et al.[4] learned domain-invariant features with topology constraints. Zhang et al. [51] employed cross-layer optimization and bilinear time-spectral fusion.

However, prior work relies on vision for cross-modal supervision, which causes overfitting between visual and Wi-Fi perspectives and lacks systematic cross-domain evaluation with rich device metadata. Existing datasets such as MM-Fi [46] and Person-in-WiFi-3D [45] are limited in scale, subject diversity, scene complexity, and within-scene position/orientation variations, and they omit unified Wi-Fi–camera coordinate calibration, leading to poor generalization. Our dataset addresses these gaps by including more subjects, larger frame counts, richer position and orientation sampling, multiple device layouts, and explicit calibration metadata, providing a stronger benchmark for cross-domain evaluation.

A.1 More details of our dataset

The hardware specifications are listed in Table4, and the recording protocol are as follows:

- Each non-rotation action is performed at five predefined spatial points within the activity area. At each point the subject faces one of three orientations (frontal, +45°, −45°). For each (point, orientation) the subject repeats the action 3 times; each repetition lasts 5s.
- Rotation actions (clockwise / counterclockwise) and a few high-variation hops may vary slightly in duration depending on subject habit; these rotation actions are recorded without fixed point/orientation constraints.
- Scene_3 contains three device-placement configurations (Setups A/B/C). Each setup differs in TX–RX relative distances (the exact layouts are illustrated in the dataset release figures); angles (facing orientations) are kept unchanged across setups. Two subjects were recorded per setup.

The public release will include:

- Raw CSI traces (per receiver, per session) and per-session metadata (timestamps, packet indices).
- Synchronized RGB video clips and timestamps.
- Calibration metadata: recorded checkerboard poses, computed transform matrix, and unified TX/RX coordinates .
- Ground-truth 3D human pose annotations from EasyMocap.
- Standardized domain split definitions and data-loading / synchronization scripts.
- Illustrative figures showing device-placement schematics for Setups A/B/C in scene_3.
- Code of PerceptAlign and a list of filtered frames.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009