Breaking Coordinate Overfitting: Geometry-Aware WiFi Sensing for Cross-Layout 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 cross-layout 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 subjects, 5 scenes, 18 actions, and 7 device layouts. Experiments show that PerceptAlign reduces in-domain error by 12.3% 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

1 Introduction

In recent years, 3D human pose estimation has served as a core technology for numerous applications [3, 28, 57], including health monitoring and natural human–computer interaction. While vision-based methods have long dominated this field [30, 33, 34, 42], 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 [13, 51, 65]. 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. [21, 43, 44, 53, 54, 66]. 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 [21] incorporates prior knowledge of human skeleton structure, HPE-Li [6] leverages attention mechanisms, and GoPose [39] 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 [48] or detailed hand skeletons [20].

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 [21, 53, 54, 66]. 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 [59] for richer feature extraction or alignment losses [67] for improved supervision, they remain unable to handle the domain shifts frequently encountered in practice, such as changes in transceiver layouts.
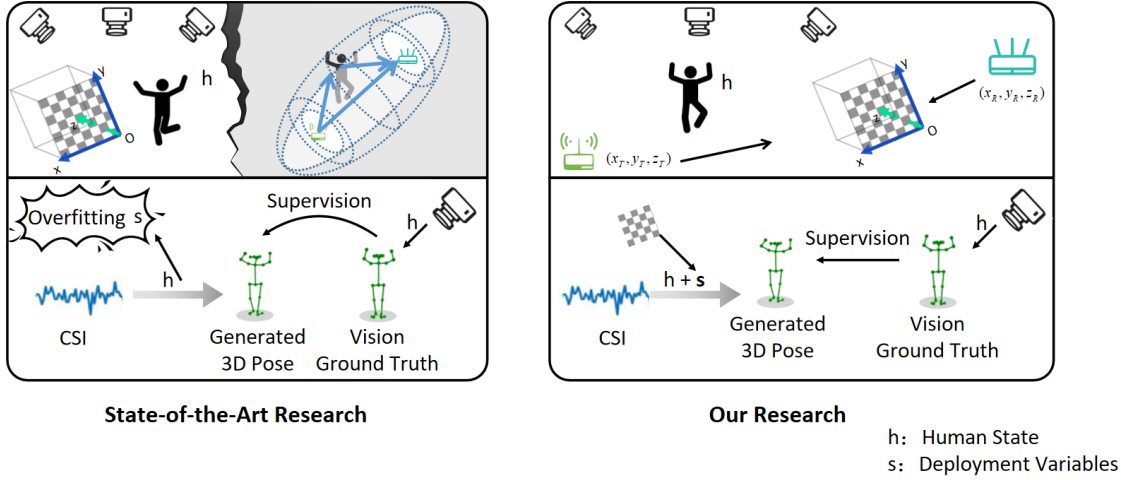
Figure 1: Left: Conventional pipelines implicitly memorize the geometric layout of WiFi devices, conflating it with target knowledge and leading to coordinate overfitting. Right: PerceptAlign explicitly makes the model aware that WiFi transceiver geometry is a conditional factor rather than knowledge to be memorized.

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 [50, 60, 63] 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 compelled to entangle deployment specific layouts with human pose representations, which prevents effective generalization to unseen layouts and thereby limits its practical applicability [5, 44, 55].

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 idea is straightforward: explicitly exposing device geometry as a varying condition rather than allowing the model to memorize it as hidden background. 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:

Figure 2: Sensing in vision and WiFi. The Vision system establishes an absolute world coordinate system rooted in the checkerboard B. The WiFi system establishes relative sensing coordinate system based on the transceiver locations. The colored ellipsoids represent the Fresnel Zones corresponding to each TX-RX pairs.

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 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. And 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-based pose estimation dataset to date, covering diverse participants, scenes, actions, and device setups with detailed geometric calibration. Extensive experiments demonstrate that PerceptAlign reduces in-domain error by 12.3% 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 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:

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

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:

$$\mathbf{X}_c = R\mathbf{X}_{\mathbb{B}} + \mathbf{t},$$

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

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:

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

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 [8, 26, 49]. We model the various factors in WiFi-based human pose estimation that influence WiFi signal propagation as follows:

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

where h denotes the dynamic human state and s denotes the static deployment variables [14] (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:

$$H_t \in \mathbb{C}^{A \times M \times T}, \quad (6)$$

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:

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

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 [15, 17, 18, 29, 46]:

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

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 [41, 64], 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 [50, 61] 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:

$$\mathbf{y} \overset{\mathbb{B}}{\longleftarrow} \mathcal{G} \overset{\mathcal{F}}{\longrightarrow} H, \quad (9)$$

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.

2.4 Our Motivation

From the above analysis, the only theoretically robust link between the visual and WiFi modalities in WiFi based pose estimation is h, while the factors in s act as confounders that are not observable in the visual domain. By explicitly treating the dominant component of s, namely the transceiver geometry, as conditional information analogous to intrinsic or extrinsic parameters, the model can avoid memorizing layout specific patterns that are not transferable. To operationalize this idea, two key steps are required: **coordinate unification**, which embeds transceiver coordinates into a unified world coordinate system, and **geometry conditioned learning**, which enforces the use of transceiver geometry as conditional knowledge rather than a learnable knowledge.

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, $T_{C \to B}$. Using the camera as an intermediary, the mapping from the WiFi coordinate system to the world coordinate system can be efficiently established. Consequently, the WiFi transceiver coordinates can be expressed in the world coordinate system.

3.2 Geometry-Conditioned Learning.

For geometry-conditioned learning, we replace the blind regressor by a conditioned estimator:

$$\hat{\mathbf{y}} = g_\theta(H, P), \tag{11}$$

where P reduces the 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 conditions. 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.

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
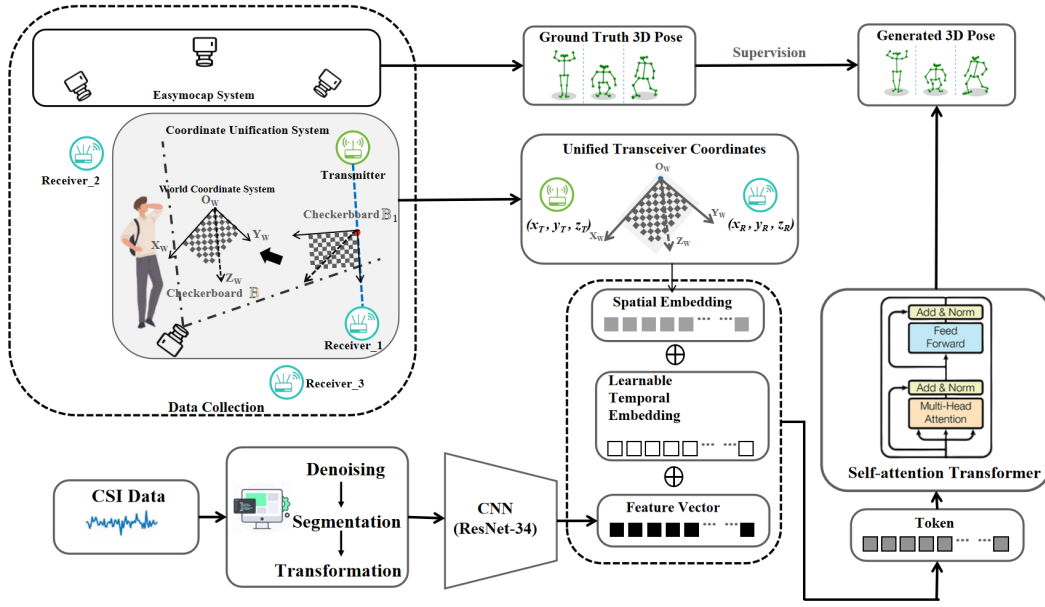
Figure 3: System overview.

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. An example of coordinate unification for the TX–RX1 pair is illustrated in the upper left of Figure 3.

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:

$$\mathbf{T}_{X \to Y} = \begin{bmatrix} \mathbf{R}_{X \to Y} & \mathbf{t}_{X \to Y} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3), \quad (12)$$

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:

$$\mathbf{p}_{B_1,T} = \begin{bmatrix} -L \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{p}_{B_1,R} = \begin{bmatrix} L \\ 0 \\ 0 \end{bmatrix}. \quad (13)$$

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:

$$L = \frac{1}{2} \cdot \left(\text{measured distance} \right) = \frac{1}{2} \cdot \begin{cases} S & \text{(direct)} \\ g\,d & \text{(grid units)} \\ p\,\rho & \text{(pixels)} \end{cases}. \quad (14)$$

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

$$\mathbf{T}_{B_1 \to B} = \mathbf{T}_{C \to B}^{-1}\,\mathbf{T}_{C \to B_1}. \quad (15)$$
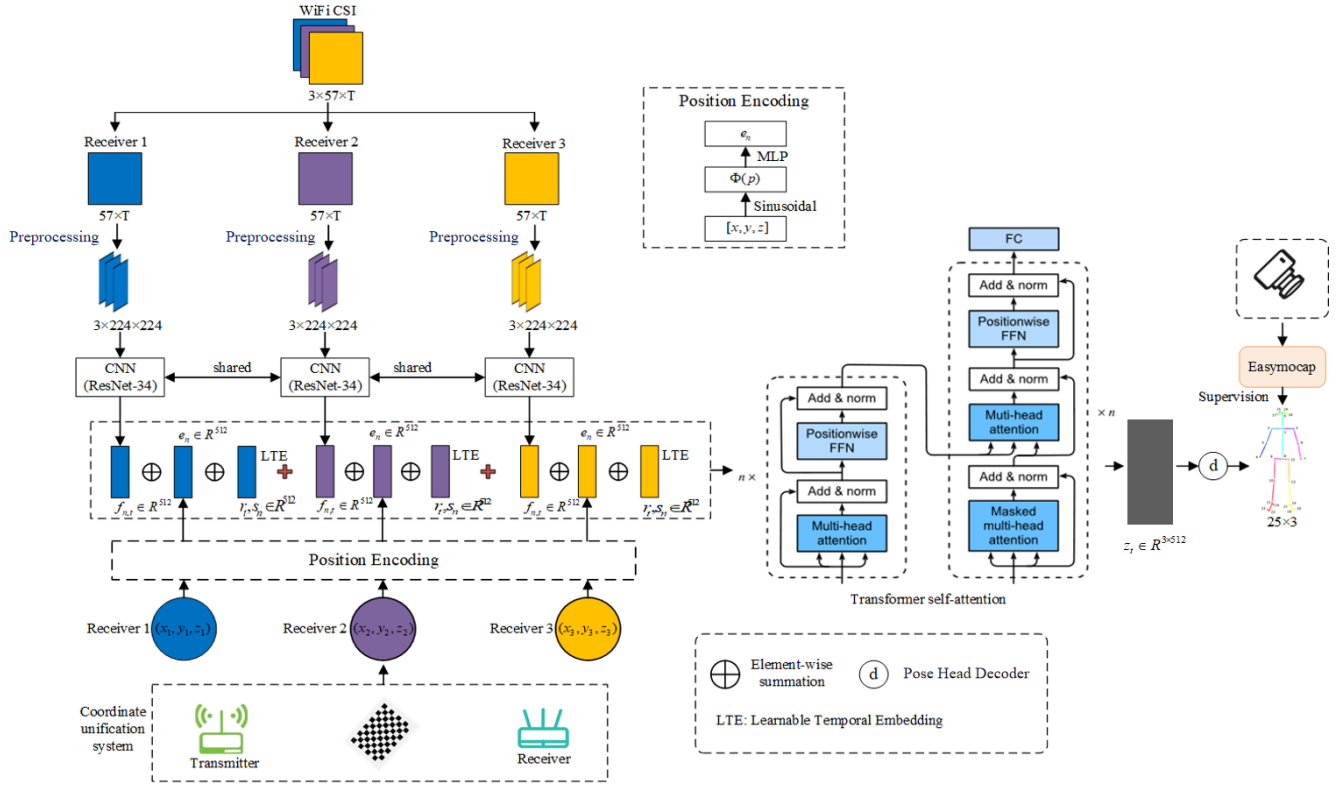
Figure 4: Geometry-conditioned learning network architecture.

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

$$\tilde{\mathbf{p}}_{B,\alpha} = \mathbf{T}_{B_1 \to B} \, \tilde{\mathbf{p}}_{B_1,\alpha}, \qquad \alpha \in \{T, R\}, \qquad (16)$$

or in non-homogeneous form

$$\mathbf{p}_{B,\alpha} = \mathbf{R}_{B_1 \to B} \, \mathbf{p}_{B_1,\alpha} + \mathbf{t}_{B_1 \to B}. \qquad (17)$$

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 (15). The result is a set of calibrated device coordinates expressed in the unified world coordinate system:

$$P = \{\mathbf{p}_{B,\text{tx}}, \mathbf{p}_{B,\text{rx}_1}, \mathbf{p}_{B,\text{rx}_2}, \dots \}. \qquad (18)$$

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 (13) and transform to the world coordinate system using (15).

(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 [58]:

$$\tilde{c}_n(t, f) = \frac{c_{n,1}(t, f)}{c_{n,2}(t, f)}, \tag{19}$$

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) [36] along the temporal axis. These features are concatenated following the procedure in WiGRUNT [12, 61], 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 receivers is processed through a shared CNN encoder E_θ to extract features. Specifically, the CSI streams collected at the three receivers are fed into the same encoder E_θ, which is fine-tuned across all inputs. In this work, E_θ is implemented as a pretrained ResNet-34 truncated before the final classification layer. It can be expressed as:

$$\mathbf{f}_{n,t} = \text{Pool}\big(E_\theta(\mathbf{X}_{n,t})\big) \in \mathbb{R}^D, \tag{20}$$

where $\mathbf{f}_{n,t}$ is the feature, which is taken from the output of the final convolutional layer of the CNN encoder. $\mathbf{X}_{n,t}$ denotes the input tensor for receiver n at frame t, $n \in \{1, \ldots, N_r\}$, and $t \in \{1, \ldots, T\}$.

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

$$\Phi(\mathbf{p}) = \Big[\sin(2^k \pi \mathbf{p}), \cos(2^k \pi \mathbf{p}) \Big]_{k=0}^{K-1}, \tag{21}$$

Here, K denotes the number of frequency bands utilized for spectral expansion, and D_p represents the dimensionality of the resulting coordinate embedding vector ($K = 10$ in this paper, and $D_p = 6K$ for 3D coordinates). This vector is subsequently projected by a small MLP g_ψ to obtain the final spatial embedding

$$\mathbf{e}_n = g_\psi\big(\Phi(\mathbf{p}_n)\big) \in \mathbb{R}^D. \tag{22}$$

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 (LTE) $\mathbf{r}_t \in \mathbb{R}^D$ (shared across receivers). Specifically, we initialize a parameter matrix $\mathbf{R} \in \mathbb{R}^{T_{seq} \times D}$, where T_{seq} denotes the temporal window size and D is the embedding dimension. For each time step t, the corresponding vector $\mathbf{r}_t$ provides a unique, dense positional signature. During training, these vectors are optimized via backpropagation to encode relative temporal dependencies, thereby enabling the self-attention mechanism to distinguish and order sequential motion cues. A small, learnable receiver-specific bias $\mathbf{s}_n \in \mathbb{R}^D$ is also added to capture static hardware idiosyncrasies.

Token construction. To integrate the learned features and the spatial-temporal embedding, we employ a projection-and-summation fusion strategy. Specifically, $\mathbf{f}_{n,t}$ and the spatial embedding $\mathbf{e}_n$ are first projected into the same dimension. They are then combined with $\mathbf{r}_t$ and $\mathbf{s}_n$ via element-wise summation to form the tokens, as formulated below:

$$\mathbf{u}_{n,t}^{(0)} = \text{LayerNorm}\big(\mathbf{W}_f \mathbf{f}_{n,t} + \mathbf{W}_e \mathbf{e}_n + \mathbf{r}_t + \mathbf{s}_n\big) \in \mathbb{R}^D, \tag{23}$$

where $\mathbf{W}_f, \mathbf{W}_e$ are projection functions. The final input to the Transformer encoder is formed by concatenating the tokens extracted from each receiver, ordered according to the receiver index and the temporal sequence:

$$\mathcal{U}^{(0)} = \big\{\mathbf{u}_{n,t}^{(0)}\big\}_{n=1, t=1}^{N_r, T}, \tag{24}$$

of length $N_r * T$. In the final token, the spatial embedding provides 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.

Spatio-temporal Transformer. The Transformer encoder we used consists of $L = 6$ stacked layers. Each layer comprises two sub-layers: a Multi-Head Self-Attention (MSA) layer and a Feed-Forward Network (FFN). Layer Normalization (LN) is applied before the input of each sub-layer, followed by a residual connection. The MSA utilizes 8 heads with a feature dimension of 512. The FFN consists of two linear transformations with GeLU activation and a dropout rate of 0.1, expanding the intermediate dimension to 2048.
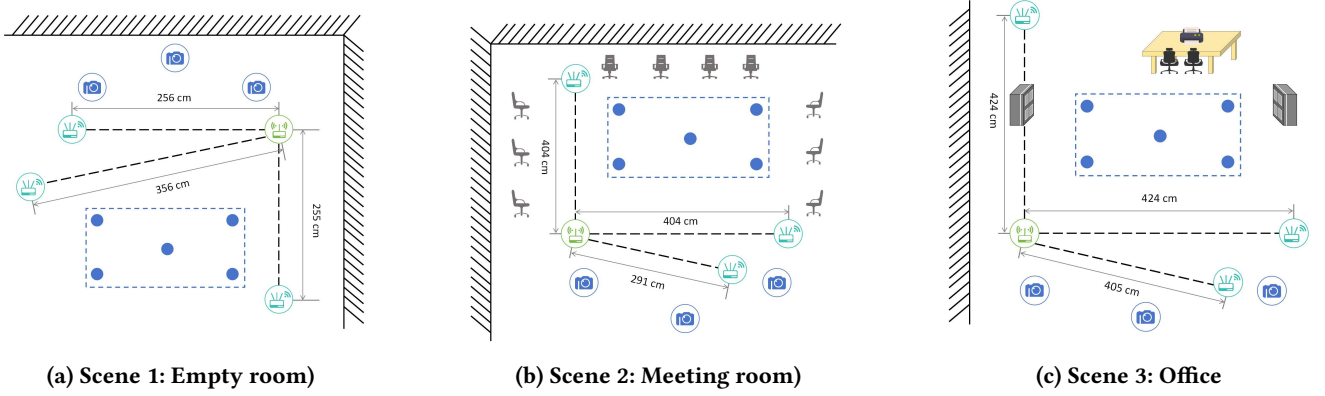
(a) Scene 1: Empty room)   (b) Scene 2: Meeting room)   (c) Scene 3: Office

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

Decoding and pose estimation. The input $\mathcal{U}^{(0)}$ is first processed by the spatiotemporal Transformer encoder, which produces contextualized tokens denoted as $\mathbf{u}_{n,t}^{(L)}$. Then, to synthesize multi-view information, we reshape this output. Specifically, for each frame t, the N_r tokens $\{\mathbf{u}_{n,t}^{(L)}\}_{n=1}^{N_r}$ are combined to form a frame-level representation $\mathbf{z}_t \in \mathbb{R}^{N_r \times D}$, where D denotes the token dimension. A pose head decoder h_ϕ then regresses the predicted joints:

$$\hat{\mathbf{y}}_t = h_\phi(\mathbf{z}_t) \in \mathbb{R}^{J \times 3}. \tag{25}$$

The pose head decoder h_ϕ is a lightweight MLP designed to map the final feature $\mathbf{z}_t$ to skeletal keypoint coordinates. Concretely, it progressively reduces the feature dimensionality through a sequence of fully connected layers, following the architecture $N_r \times D \to 1024 \to 512 \to J \times 3$, where J denotes the number of skeletal joints.

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

$$\mathcal{L}_{\text{MSE}} = \frac{1}{LJ} \sum_{l=1}^{L} \sum_{j=1}^{J} \left\| \hat{\mathbf{y}}_{l,j} - \mathbf{y}_{l,j} \right\|_2^2. \tag{26}$$

Once the system is trained, deploying it in a new scene 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, and it is available at https://github.com/Trymore-lab/PerceptAlign.

5.1 Dataset overview

The dataset contains recordings from 21 participants across three indoor scenes. As shown in Fig.5, these revscenes include an empty room, a meeting room, and an office. Data collection spanned six months, resulting in 483.22 GB of recordings and over 98 hours of video. The dataset comprises 7,243,590 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. More details please refer to Appendix A.

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	7243.5k	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 [53], 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-subject generalization. Although the number of participants is smaller than in MM-Fi [54], our dataset places greater emphasis on the cross-domain challenges most relevant to real-world applications. Unlike prior datasets, which include cross-scene data collection but use the same WiFi device layout within each scene, our dataset assigns a different device layout to every scenario. This design enables rigorous evaluation under simultaneous scene and layout shifts, a condition that is far more common in real-world applications. We also include multiple layouts within the same scene to analyze layout-induced variations, and capture extensive data across diverse subject 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

Hardware configuration. Our hardware system consists of a WiFi and a visual sensing platform, with detailed specifications summarized in Table 2. The WiFi sensing platform includes one transmitter and three receivers, all equipped with Intel 5300 Network Interface Cards and configured to collect CSI using the picoscenes tool. The transmitter operates with a single antenna, while each receiver is equipped with three antennas. For visual data acquisition, we employ Intel RealSense D435 depth cameras to capture RGB video at a resolution of 1920 × 1080 and a frame rate of 30 fps.

Table 2: Acquisition hardware, software and nominal parameters.

Modality	Hardware	Key Parameters
RGB Video	Intel RealSense D435	1920 × 1080, 30 fps, RGB-D aligned
WiFi CSI	Intel 5300 + CSI Tool	1 TX, 3 RX (3 antennas/RX); 57 subcarriers
		Bandwidth: 20 MHz; Freq: 5.2 GHz; Rate: 810 Hz
		Antenna: Omnidirectional
Calibration	Checkerboard (11 × 8)	Square size: 30 mm; EasyMocap-compatible
Ground-truth	EasyMocap	Camera-centric 3D skeleton reconstruction

Model settings. Our model is implemented using the PyTorch framework. All training and inference processes are conducted on one NVIDIA GeForce RTX 4090 GPU. All experiments follow a fixed training recipe and model configuration. 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) and PCK@20/50(%).

To comprehensively evaluate PerceptAlign, we compare it with two state-of-the-art methods:

Table 3: Comprehensive quantitative comparison across different domain shifts. We report MPJPE (mm, ↓), PCK@20 (%, ↑), and PCK@50 (%, ↑).

Evaluation Setting	Person-in-WiFi 3D			DT-Pose			PerceptAlign (Ours)		
	MPJPE ↓	PCK@20 ↑	PCK@50 ↑	MPJPE ↓	PCK@20 ↑	PCK@50 ↑	MPJPE ↓	PCK@20 ↑	PCK@50 ↑
In-Domain	221.0	18.5	48.2	156.5	32.4	68.5	**137.2**	**55.2**	**88.7**
Cross-Location	253.1	14.2	41.5	220.0	21.8	55.4	**144.6**	**50.1**	**86.3**
Cross-Orientation	254.0	13.8	40.8	255.7	17.5	48.9	**147.7**	**49.5**	**85.1**
Cross-Subject	266.7	12.1	39.5	260.5	16.2	46.5	**145.3**	**51.3**	**84.8**
Cross-Layout	649.3	2.5	10.4	583.2	5.8	18.2	**170.2**	**46.8**	**80.1**
Cross-Scene	717.2	1.8	8.5	571.1	6.5	20.1	**181.5**	**44.2**	**79.5**

- **Person-in-WiFi 3D (CVPR 2024) [53]**: A supervised framework that employs Transformers for global correlation modeling. We adopt it as a strong baseline due to its demonstrated superiority over prior approaches and the availability of its open-source implementation.
- **DT-Pose (CVPR 2025) [4]**: A self-supervised approach that alleviates cross-domain discrepancies through temporally consistent contrastive masking. We include this method to highlight the benefits of our explicit geometry conditioning compared to implicit representation learning in the presence of domain shifts.

All baseline methods are retrained from scratch on our dataset and fully converged, while strictly adhering to the same training and testing splits as PerceptAlign.

Evaluation setup. We employ two standard metrics to evaluate performance:

- **Mean Per Joint Position Error(MPJPE(mm)):** The average Euclidean distance between the predicted and ground-truth joint coordinates.
- **Percentage of Correct Keypoints(PCK@σ(%)):** The percentage of predicted joints whose Euclidean error is below a threshold σ. We report PCK@20 and PCK@50, representing the accuracy with strict thresholds of 20mm and 50mm, respectively.

6.1 In-domain Evaluation

We report in-domain MPJPE and PCK@20/50 for models trained and tested within the same scene in Table 3. For brevity, we present the overall mean MPJPE and PCK@20/50 across the three scenes. All reported models share the same training and evaluation protocol. Each scene is split 80/20 by action instances, and errors are averaged across all test frames. Compared with Person-in-WiFi-3D, PerceptAlign markedly improves in-domain performance by reducing MPJPE from 221.0 mm to 137.2 mm and increasing PCK@20/50 from 18.5/48.2 to 55.2/88.7. Relative to DT-Pose, it further lowers MPJPE from 156.5 mm to 137.2 mm, a 12.3% improvement,

while achieving the same substantial gains in PCK@20/50. This 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, scene, subject identity, and device layout. Each protocol follows a leave-one-out design, and results are summarized in Table 3. And 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 with in-domain settings, Person-in-WiFi-3D exhibits substantial performance degradation, with MPJPE increasing to 253.1 mm and PCK@20/50 dropping to 14.2/41.5, respectively. DT-Pose also suffers notable degradation, reaching an MPJPE of 220.0 mm with PCK@20/50 of 21.8/55.4. In contrast, although PerceptAlign experiences some performance decline under cross-domain conditions, it significantly mitigates the degradation by incorporating geometric conditioning, reducing the MPJPE to 144.6 mm, which corresponds to a 34.4% improvement, while raising PCK@20/50 to 50.1/86.3, respectively.

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 attains an MPJPE of 254 mm with PCK@20/50 of
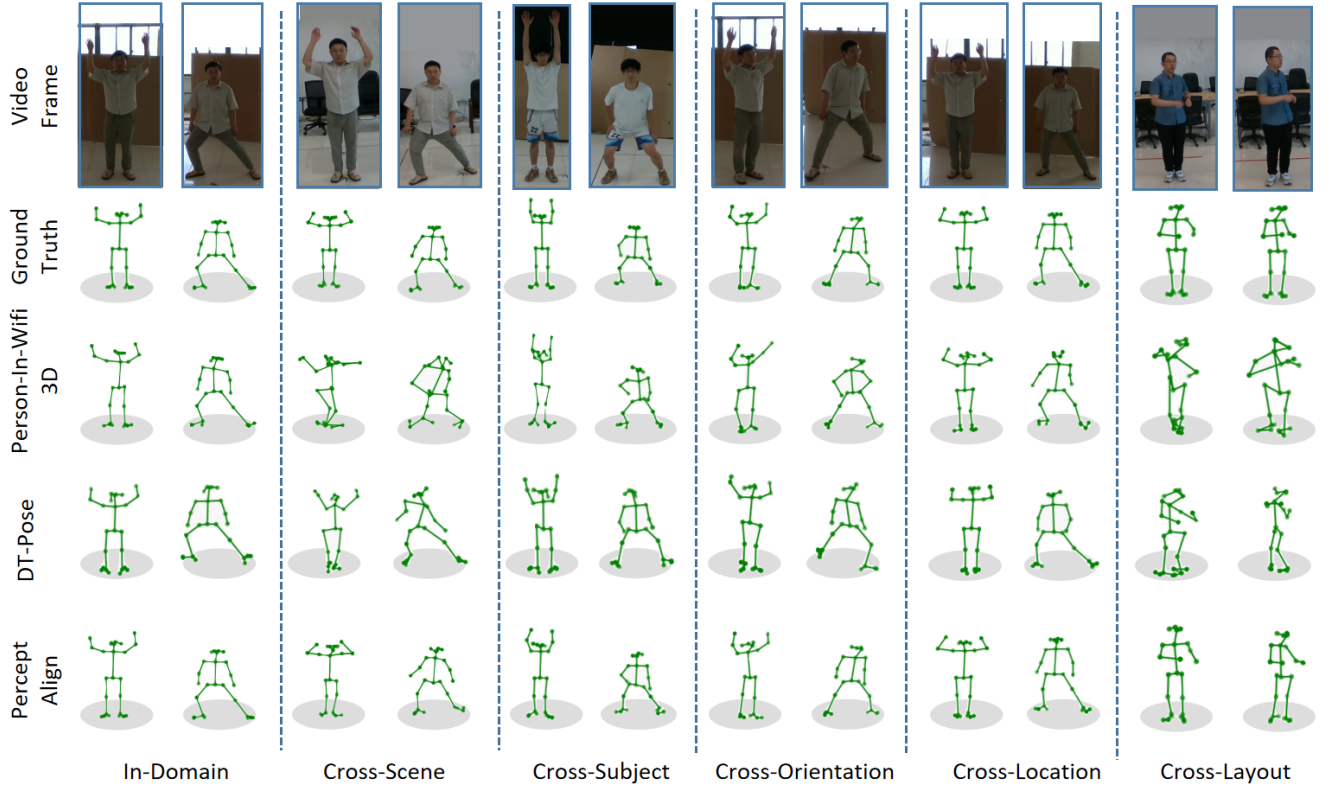
Figure 6: Illustrative examples of constructed skeletons across different cross-domain settings.

13.8/40.8, while DT-Pose reports an MPJPE of 255.7 mm and PCK@20/50 of 17.5/48.9, indicating that orientation changes substantially hinder generalization. In contrast, PerceptAlign reduces MPJPE to 147.7 mm, corresponding to a 42% error reduction, demonstrating its effectiveness under this setting.

Cross-Subject. In the cross-subject evaluation, Person-in-WiFi-3D, DT-Pose, and our method achieve MPJPEs of 266.7 mm, 260.5 mm, and 145.3 mm, respectively. Correspondingly, their PCK@20/50 scores are 12.1/39.5, 16.2/46.5, and 51.3/84.8. These results highlight the non-negligible impact of subject variation on pose estimation performance, while PerceptAlign maintains robustness through its geometry-conditioned learning strategy. More details please refer to Appendix B.

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 and DT-Pose becomes nearly unusable under this setting, with average MPJPE values of 649.3 mm and 583.2, and PCK@20/50 scores of 2.5/10.4 and 5.8/18.2, respectively. In contrast, our method achieves an MPJPE of

170.2 mm and PCK@20/50 scores of 46.8/80.1, demonstrating strong robustness to layout changes and validating the effectiveness of the proposed geometry-aware approach.

Cross-Scene. Our cross-scene 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 and DT-Pose suffers severe degradation,with average MPJPE values of 717.2 mm and 571.1 mm, and PCK@20/50 scores of 1.8/8.5 and 6.5/20.1, respectively.. Although PerceptAlign also experiences some decline, it maintains an MPJPE of 181.5 mm and PCK@20/50 scores of 44.2/79.5, demonstrating substantially improved robustness in pose estimation.

Overall, these results validate 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. Moreover, we compare different methods in terms of

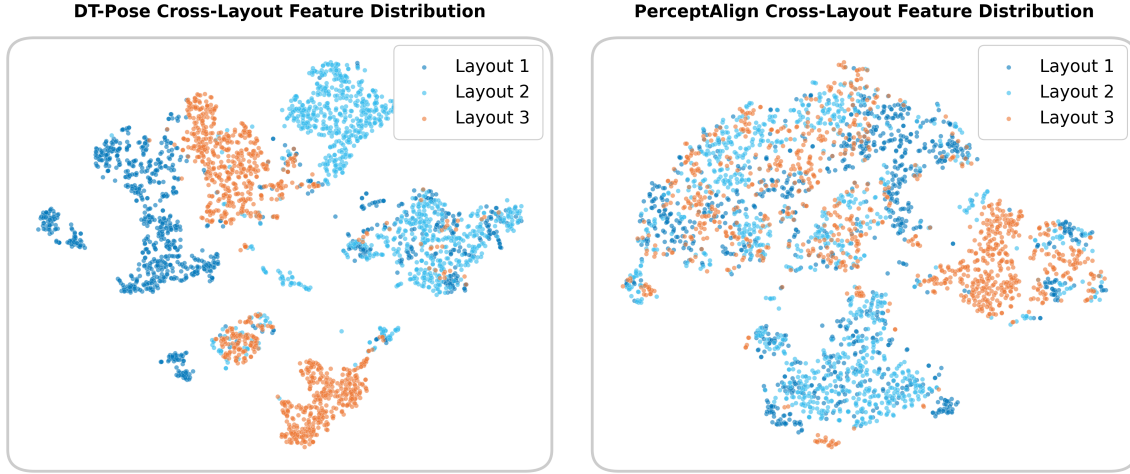**DT-Pose Cross-Layout Feature Distribution**   **PerceptAlign Cross-Layout Feature Distribution**

Figure 7: t-SNE visualization of feature distributions under the Cross-Layout setting. Blue shades denote training layouts (L1/L2), and orange denotes the unseen test layout (L3).

model size and inference speed to evaluate their deployment feasibility. Person in WiFi 3D [53] contains 20.4 million parameters and achieves an inference speed of 54 FPS. The recent DT Pose [4] employs 34.5 million parameters, with an estimated inference speed of 41 FPS. In comparison, our proposed PerceptAlign consists of 29.71 million parameters. Owing to the incorporation of position and temporal encoding, PerceptAlign introduces a moderate computational overhead and attains an inference speed of 37 FPS. This modest reduction in inference speed enables substantially improved generalization performance, particularly under cross-layout and cross-scene settings.

6.3 Evidence of Coordinate Overfitting

To validate our coordinate overfitting hypothesis and demonstrate the effectiveness of PerceptAlign, we visualize the output features of DT-Pose and PerceptAlign using t-SNE [32] under the cross layout setting of Scene 3. All models are trained on Layouts 1 and 2 and evaluated on the unseen Layout 3. The visualized features are taken from the input to the decoder pose head. The results are shown in Figure 7. As observed, the features produced by DT-Pose (left) form layout specific clusters, indicating that the model overfits to deployment geometry. As a result, samples from the unseen Layout 3, shown in orange, occupy a disjoint region in the feature space. In contrast, the features learned by PerceptAlign (right) exhibit minimal dependence on the layout, substantially alleviating coordinate overfitting. These results suggest that existing state-of-the-art methods implicitly memorize transceiver geometry, which leads to poor generalization, whereas the geometry conditioned learning strategy adopted by PerceptAlign effectively mitigates this issue.

Table 4: 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

6.4 Ablation study

Table 4 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-scene $181.5 \rightarrow 729.5$ 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
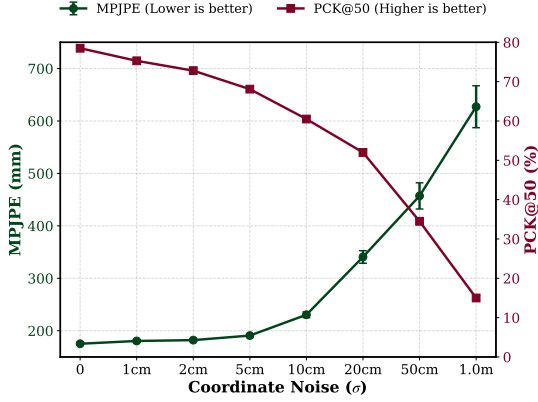
Figure 8: **Influence of coordinate estimation errors.**

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.

6.5 Influence of Coordinate Errors

Our lightweight coordinate unification procedure may introduce inaccuracies when estimating the positions of WiFi transceivers. To quantify the impact of such errors on system performance, we conduct a sensitivity analysis under the cross layout setting. Specifically, we inject random perturbations with magnitude σ into the three dimensional coordinates of the WiFi transceivers, simulating coordinate estimation errors ranging from the centimeter scale ($\sigma = 1$ cm) to the meter scale ($\sigma = 1.0$ m). The results are shown in Figure 8. The results indicate that the proposed model is highly tolerant to centimeter level coordinate errors, while performance degradation becomes pronounced once the error exceeds the decimeter scale. Notably, when $\sigma = 1.0$ m, the MPJPE approaches that of the baselines, suggesting that when geometric priors become unreliable, the model effectively degenerates into a geometry agnostic configuration. In practical deployments, coordinate unification is typically performed using checkerboards that are widely adopted for precise camera calibration. As a result, the induced coordinate estimation errors are usually small and have a negligible impact on overall system performance.

6.6 More Generalization Evaluation

To further evaluate the generalization capability of PerceptAlign under previously unseen spatial configurations, we collected an additional test set in two distinct environments,

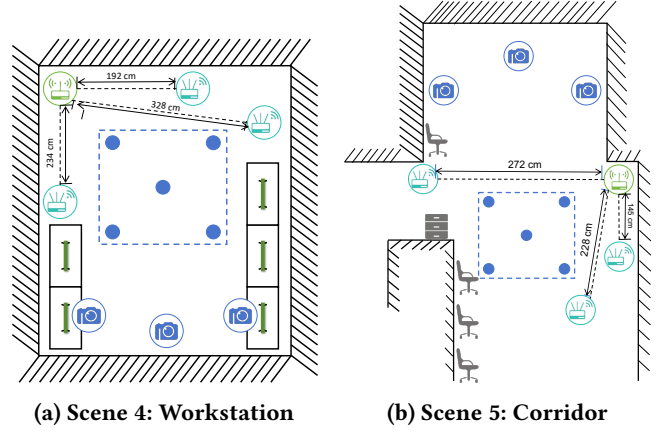**(a) Scene 4: Workstation**   **(b) Scene 5: Corridor**

Figure 9: **Visual illustration of testing scenes.**

namely **Scene 4 (Workstation)** and **Scene 5 (Corridor)**, as shown in Figure 9. Compared with Scenes 1-3, these environments introduce substantially larger structural variations. Moreover, to impose a more stringent generalization setting, this test set was captured at a frame rate of 22 fps rather than the 30 fps used in Scenes 1-3. The motion categories in this new evaluation set are consistent with those in the earlier data collection. For each scene, the dataset includes recordings from 3 subjects, 5 locations, and 3orientations.

Results and Analysis. All models were trained using data from the first three scenes and evaluated on the newly collected test set. Table 5 summarizes the corresponding results. Existing baselines exhibit severe degradation in these unseen scenes, with MPJPEs spiking above 600mm in Scene 4 and exceeding 700mm in the more geometrically challenging Scene 5. This failure stems from their implicit overfitting to the training layouts. In contrast, PerceptAlign maintains strong robustness. Despite the significant structural and fps differences, our method achieves an MPJPE of 222.4 mm in Scene 4 and 317.1 mm in Scene 5, outperforming the state-of-the-art by over 54.3%. These results demonstrate that explicitly conditioning on transceiver geometry enables the model to effectively generalize to structurally diverse environments.

Table 5: **Performance comparison on Scene 4 & 5.**

Method	Scene 4 (Workstation)			Scene 5 (Corridor)		
	MPJPE ↓	PCK@20 ↑	PCK@50 ↑	MPJPE ↓	PCK@20 ↑	PCK@50 ↑
Person-in-WiFi 3D	770.5	1.2	6.1	837.7	0.9	4.2
DT-Pose	640.1	2.8	13.5	694.3	1.5	8.7
PerceptAlign	**222.4**	**39.7**	**64.1**	**317.1**	**32.2**	**56.1**

7 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, 23, 31, 38, 56];

and (2) commodity WiFi CSI-based methods, which trade some physical resolution for ubiquity and low cost. Early RF studies demonstrated robust activity recognition and coarse localization using RSSI/CSI features and classical machine learning[40, 47, 52]. As research advanced, deep learning models [25, 27] began to extract richer spatio-temporal patterns from CSI, enabling finer-grained tasks such as pose estimation [9, 11, 21].

(1) Extending from 2D to 3D: Wang et al. [45] 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. [53] 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. [37], Hsu et al. [16], and Yan et al. [53] explored multi-user pose estimation via improved resolution, loss functions, and user separation. (3) Improving accuracy: Huang et al. [19] leveraged the sparsity of joint heat maps and introduced a 3D streaming signal fusion module. Nguyen et al. [35] 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. [62] introduced Vista-Former with dual-stream spatiotemporal attention. Lee et al. [22] 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 [10] 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. [59] employed cross-layer optimization and bilinear time-spectral fusion. More recently, advanced physical modeling and domain adaptation techniques have been proposed. RayLoc [14] introduces a fully differentiable ray-tracing framework to solve localization as an inverse problem. On the other hand, domain adaptation approaches like AdaPose [67] and CrossGR [24] employ alignment losses or source-free adaptation strategies to mitigate distribution shifts. However, these methods typically operate implicitly on feature statistics and often require data from the target domain for adaptation.

However, prior work relies on vision for cross-modal supervision, which causes overfitting between visual and WiFi perspectives and lacks systematic cross-domain evaluation. Existing datasets such as MM-Fi [54] and Person-in-WiFi-3D [53] are limited in scale, subject diversity, scene complexity, and within-scene position/orientation variations, and they omit unified WiFi–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.

8 Discussion

Deployment Prerequisites. Unlike current methods, PerceptAlign relies on transceiver coordinates to disentangle motion from layout. However, existing public datasets do not provide such coordinate information, which makes evaluation on current benchmarks, such as MM-Fi [54], infeasible. Although the deployment environment also requires layout encoding, as detailed in Section 4.1, our approach only relies on two checkerboards and a camera/smartphone. The user simply captures a photo that contains both checkerboards, after which our program automatically performs the coordinate alignment procedure. The entire process completes in under five minutes and introduces minimal user efforts, thereby maintaining high usability.

Geometric Constraints. WiFi sensing captures activity through variations in non-line-of-sight (NLOS) reflections and diffraction effects along the line-of sight-path (LOS). When a transmitter and a receiver are co-located, most of the signal energy propagates directly along the LOS path, while NLOS components carry negligible energy and are only weakly influenced by surrounding motion. Consequently, ambient activities become nearly imperceptible under such conditions. For this reason, existing WiFi sensing studies generally avoid this deployment, and co-located routers and client devices are also uncommon in everyday environments. If multiple receivers are mounted on a single device, our approach treats their spatial encodings as shared and identical.

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). Third, our current design primarily focuses on layout shifts and does not explicitly address individual differences or other domain factors. Attributes such as body shape, user position, and motion style also modulate CSI signals, yet they are physically uncorrelated with transceiver geometry. As a result, a residual performance gap remains in cross domain settings, for example an error

of approximately 145 mm in cross subject scenarios. 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. In addition, we plan to extend PerceptAlign by integrating it with existing generalization techniques, such as adversarial domain generalization, to address individual specific statistical shifts that are independent of the physical layout. This direction is expected to further improve the robustness and generalization capability of the system and to narrow the remaining performance gaps.

9 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 12.3% 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. Pretraining 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] Long Fan, Yinghui He, Lei Xie, Serene Zhang, and Jun Luo. 2025. Sense with Polyface Mirror: Enhancing Wi-Fi Sensing Diversity via Programmable Metasurfaces. (2025).
[9] Jiaqi Geng, Dong Huang, and Fernando De la Torre. 2022. Densepose from wifi. *arXiv preprint arXiv:2301.00250* (2022).
[10] 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.
[11] 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).
[12] 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.
[13] 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.
[14] Xueqiang Han, Tianyue Zheng, Tony Xiao Han, and Jun Luo. 2025. RayLoc: Wireless Indoor Localization via Fully Differentiable Raytracing. *arXiv preprint arXiv:2501.17881* (2025).
[15] Yinghui He, Mingming Xu, Zhe Chen, Fu Xiao, and Jun Luo. 2025. Beam-Fi: Integrated Sensing and Communication via MU-MIMO upon Commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–22.

[16] 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.

[17] Jingzhi Hu, Xin Li, Jin Gan, and Jun Luo. 2025. Poison to Cure: Privacy-preserving Wi-Fi Multi-User Sensing via Data Poisoning. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 47–62.

[18] Jingzhi Hu, Tianyue Zheng, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. MUSE-Fi: Contactless muti-person sensing exploiting near-field Wi-Fi channel variation. In *Proceedings of the 29th annual international conference on mobile computing and networking*. 1–15.

[19] 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).

[20] 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.

[21] 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.

[22] 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).

[23] 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.

[24] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojiang Chen, Zhanyong Tang, and Zheng Wang. 2021. CrossGR: Accurate and low-cost cross-target gesture recognition using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–23.

[25] Xin Li, Yinghui He, and Jun Luo. 2025. μCeiver-Fi: Exploiting Spectrum Resources of Multi-Link Receiver for Fine-Granularity Wi-Fi Sensing. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 1045–1059.

[26] 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).

[27] Xin Li, Hongbo Wang, Zhe Chen, Zhiping Jiang, and Jun Luo. 2024. Uwb-fi: Pushing wi-fi towards ultra-wideband for fine-granularity sensing. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 42–55.

[28] 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).

[29] 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.

[30] 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.

[31] 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.

[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[33] 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.

[34] 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.

[35] 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).

[36] 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.

[37] 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).

[38] 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.

[39] 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.

[40] 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.

[41] 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).

[42] 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.

[43] 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.

[44] 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).

[45] 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.

[46] 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).

[47] 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.

[48] 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.

[49] 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).

[50] 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.

[51] 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).

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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).

[60] 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.

[61] 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).

[62] 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).

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

A Appendix A: More details of our dataset

The hardware specifications are listed in Table2, 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 Easy-Mocap.

Table 6: Detailed per-subject MPJPE (mm) for cross-subject evaluation.

Scene	Method	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Sub11	Sub12	Sub13	Avg.
	PiW 3D	226.7	240.2	244.8	216.8	231.0	246.3	233.9	239.4	220.6	266.3	234.0	243.3	243.0	**237.4**
Scene 1	DT-Pose	248.8	255.9	222.5	221.7	238.1	229.5	236.7	203.0	228.4	230.3	228.6	229.5	215.0	**231.4**
	Ours	**131.8**	**128.7**	**132.2**	**122.1**	**129.0**	**124.6**	**133.4**	**121.6**	**135.7**	**128.6**	**131.0**	**134.1**	**132.1**	**129.6**
	PiW 3D	275.1	277.3	289.5	271.1	294.5	275.4	266.9	254.8	267.8	-	-	-	-	**274.7**
Scene 2	DT-Pose	280.4	248.7	292.2	275.9	261.6	261.7	255.5	272.6	262.0	-	-	-	-	**267.8**
	Ours	**138.9**	**151.7**	**139.1**	**148.7**	**151.5**	**144.2**	**144.4**	**148.2**	**154.8**	-	-	-	-	**146.8**
	PiW 3D	334.7	307.7	309.9	299.6	321.2	335.8	-	-	-	-	-	-	-	**318.1**
Scene 3	DT-Pose	300.6	358.5	313.3	316.8	298.0	308.3	-	-	-	-	-	-	-	**315.9**
	Ours	**177.4**	**168.8**	**176.5**	**182.7**	**184.3**	**172.2**	-	-	-	-	-	-	-	**177.0**
												Global Weighted Average:			**145.3**

Table 7: Detailed per-subject PCK@20 / PCK@50 (%) for cross-subject evaluation.

Scene	Method	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Sub11	Sub12	Sub13	Avg.
	PiW 3D	15.0/47.0	13.8/43.9	13.4/42.9	16.0/49.6	14.6/46.0	13.3/42.6	14.3/45.3	13.8/44.1	15.6/48.6	11.8/38.8	14.3/45.3	13.5/43.2	13.5/43.3	**14.1/44.7**
Scene 1	DT-Pose	16.8/48.0	16.1/46.4	19.8/54.8	19.9/55.1	17.9/50.6	18.9/52.8	18.1/50.9	22.8/61.2	19.1/53.1	18.8/52.6	19.0/53.1	18.9/52.8	20.9/57.1	**19.0/53.0**
	Ours	**54.2/88.1**	**55.8/89.5**	**54.0/87.9**	**57.2/91.0**	**55.6/89.3**	**56.8/90.2**	**53.8/87.5**	**57.5/91.2**	**52.9/86.8**	**55.9/89.6**	**54.8/88.5**	**53.5/87.2**	**54.1/87.8**	**55.1/88.8**
	PiW 3D	11.2/37.3	11.1/36.9	10.4/35.1	11.5/38.0	10.1/34.4	11.2/37.2	11.7/38.7	12.6/40.9	11.7/38.5	-	-	-	-	**11.3/37.4**
Scene 2	DT-Pose	14.0/41.5	16.8/48.0	13.2/39.5	14.4/42.4	15.6/45.2	15.5/45.1	16.1/46.5	14.6/43.0	15.5/45.1	-	-	-	-	**15.1/44.0**
	Ours	**53.3/88.1**	**46.7/79.3**	**53.2/88.0**	**48.1/81.2**	**46.8/79.4**	**50.4/84.2**	**50.3/84.1**	**48.4/81.5**	**45.3/77.4**	-	-	-	-	**49.2/82.6**
	PiW 3D	8.4/29.5	9.5/32.6	9.4/32.3	9.9/33.7	8.9/31.0	8.3/29.4	-	-	-	-	-	-	-	**9.1/31.4**
Scene 3	DT-Pose	12.6/38.2	9.7/30.9	11.9/36.4	11.7/35.9	12.8/38.6	12.2/37.1	-	-	-	-	-	-	-	**11.8/36.2**
	Ours	**36.9/65.7**	**39.8/69.7**	**37.2/66.1**	**35.4/63.4**	**34.9/62.8**	**38.6/68.1**	-	-	-	-	-	-	-	**37.1/66.0**
												Global Weighted Average:			**51.3/84.8**

- 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.

B Appendix B: Detailed Cross-Subject Evaluation

This section provides additional details on the cross subject evaluation presented in Section 6.2. Table 6 and Table 7 reports the MPJPE and PCK@20/50 for each subject across all scenes. As shown, all methods achieve their best performance in Scene 1, where the training data exhibits the greatest diversity. Performance degrades in Scene 2 and reaches its lowest level in Scene 3. The baseline methods suffer a pronounced drop in Scene 3, with errors exceeding 300 mm in several cases, indicating that models struggle to generalize when training data is limited. Although PerceptAlign also exhibits increased error in Scene 3, it consistently maintains substantially lower errors than the baselines. This result demonstrates that geometry aware spatial embeddings that leverage layout priors enable the model to learn more effective and transferable features.