CSIPose: Unveiling Human Poses Using Commodity WiFi Devices Through the Wall

Yangyang Gu, Jing Chen, Congrui Chen, Kun He, Ju Jia, Yebo Feng, Ruiying Du, and Cong Wu

Abstract—**The popularity of WiFi devices and the development of WiFi sensing have alerted people to the threat of WiFi sensing-based privacy leakage, especially the privacy of human poses. Existing work on human pose estimation is deployed in indoor scenarios or simple occlusion (e.g., a wooden screen) scenarios, which are less privacy-threatening in attack scenarios. To reveal the risk of leakage of the pose privacy to users from commodity WiFi devices, we propose CSIPose, a privacy-acquisition attack that passively estimates dynamic and static human poses in through-the-wall scenarios. We design a three-branch network based on transfer learning, auto-encoder, and self-attention mechanisms to realize the supervision of video frames over CSI frames to generate human pose skeleton frames. Notably, we design *AveCSI*, a unified framework for preprocessing and feature extraction of CSI data corresponding to dynamic and static poses. This framework uses the average of CSI measurements to generate CSI frames to mitigate the instability of passively collected CSI data, and utilizes a self-attention mechanism to enhance key features. We evaluate the performance of CSIPose across different room layouts, subjects, devices, subject locations, and device locations. Evaluation results emphasize the generalizability of CSIPose. Finally, we discuss measures to mitigate this attack.**

Index Terms—**Channel state information, human pose estimation, human privacy, through the wall.**

I. Introduction

WIFI devices, ubiquitous in modern daily life, continue their growth with global WiFi device shipments projected to increase at a 20% compound annual growth rate from 2023 to 2028 [1]. Despite its convenience and speed, WiFi raises significant privacy concerns [2]. These concerns are amplified by advances in WiFi sensing, particularly through Channel State Information (CSI) [2], [3]. A notable privacy risk is the unauthorized use of WiFi for human sensing. An example is the see-through attack: exploiting WiFi signals for passive human pose estimation in Through-The-Wall (TTW) scenarios.

Researches in body pose estimation have seen varied methods, including the use of custom antennas [4], [5], [6] or multiple receivers [7], [8], [9], [10]. These methods capitalize on how the human body reflects WiFi signals, with some employing deep learning techniques in both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) scenarios [4], [5], [7], [8], [9], [10], [11], [12], [13]. Of particular interest is the NLoS scenario, which poses a potential attack scenario where the transmitter and receiver are situated on opposite sides of an obstruction. Most studies only consider simple obstructions, such as wooden boards [4], [8], [13]. GoPose [5] deals with walls yet places the receiver and the subject on the same side of the wall, which makes CSI loss less information than when the subject and the transmitter are on the same side as detailed in Section II-C. Current works requiring controlled transmitters to emit specific signals (such as transmitting 1000 packets per second) are less applicable to real attack scenarios. Manipulating a transmitter or customizing a receiver is demanding for an attacker. Scenarios with only a wooden screen or co-location of receiver and subject pose limited privacy risks.

A fundamental challenge for the practical body pose exposing attack using CSI is realizing the supervision of stabilized video frames over non-stabilized CSI data during training, and extracting features from attenuated CSI data through walls for both dynamic and static body pose estimation. Using body poses from each frame is feasible due to the fixed video frame rate. However, passively collected CSI data often contains a variable number of measurements within a fixed time interval, especially when the attacker has no control over indoor commercial WiFi devices. The impact of the same pose on CSI varies in static and dynamic settings as explained in Section II-B.

To overcome this challenge, we present *AveCSI*, a novel framework for preprocessing CSI data and extracting essential features for static and dynamic human pose estimation. Specifically, inspired by [14], it first purifies CSI data using a combinatorial method to amplify the impact of the human body on CSI. Next, AveCSI splits the CSI sequence into sub-sequences at a fixed time interval, which is consistent with the frame rate of the supervised video during training and remains constant during testing. Notably, AveCSI adopts the strategy of using

Received 18 August 2024; revised 10 April 2025; accepted 15 May 2025. Date of publication 19 May 2025; date of current version 3 September 2025. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3102100 and Grant 2022YFB3103300, in part by the National Natural Science Foundation of China under Grant 62172303, Grant 62302343, in part by the Key R&D Program of Hubei Province under Grant 2024BAB018, in part by Wuhan Scientific and Technical Key Project under Grant 2023010302020707, in part by Wuhan Knowledge Innovation Project under Grant 2023010201010062, and in part by the Key R&D Program of Shandong Province under Grant 2022CXPT055. Recommended for acceptance by W. Gong. *(Corresponding author: Cong Wu.)*

Yangyang Gu, Jing Chen, Congrui Chen, Kun He, Ruiying Du, and Cong Wu are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: guyangyang@whu.edu.cn; chenjing@whu.edu.cn; congrui_chen@whu.edu.cn; hekun@whu.edu.cn; duraying@whu.edu.cn; cnacwu@whu.edu.cn).

Ju Jia is with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: jiaju@seu.edu.cn).

Yebo Feng is with the College of Computing and Data Science, Nanyang Technological University,, Singapore 639798 (e-mail: yebo.feng@ntu.edu.sg).

Digital Object Identifier 10.1109/TMC.2025.3571469

an average CSI measurement of the sub-sequence as the CSI frame, rather than using the equal-number CSI measurements as the CSI frame [10], [13]. This strategy can avoid leading to large differences in the amplitude frames within the same pose when the difference in the number of CSI measurements within a fixed duration is too large. And it also provides the basis to use the same network framework to train dynamic and static human pose estimation models. Finally, to extract effective features from CSI data that losses more information in TTW scenarios, AveCSI incorporates a three-layer network: 1) a Long Short-Term Memory (LSTM) layer to extract distributional features between subcarriers, 2) a convolutional layer to augment these features, and 3) a Self-Attention (SA) module to extract representative features from CSI data to improve the accuracy of the pose estimation.

In this paper, we introduce CSIPose, a privacy-acquisition attack designed to estimate dynamic and static human poses using the CSI collected via commodity devices in TTW scenarios. To estimate human poses, we propose an innovative three-branch network that fuses the principles of transfer learning [15] and self-attention. Specifically, we first train an Auto-Encoder (AE) network to learn how to represent and reconstruct skeleton frames using Ground-Truth (GT) human pose skeleton frames as input, which are generated by OpenPose [16]. Then, based on the idea of transfer learning, we inherit the parameters of this AE network and combine it with CSI data to train human pose skeleton frame estimation models with CSI as input. The self-attention layer is embedded into the feature extraction components of the network to improve the feature representation.

In the training process, the initial branch employs the GT skeleton frames as input to an AE network. This branch serves to pre-train an encoder (i.e., skeleton encoder) and a decoder (i.e., skeleton reconstruction) to improve the efficiency of subsequent branches. The second branch, building upon skeleton frames and inheriting initial parameters from the first branch, consists of an encoder and a decoder. Simultaneously, the third branch incorporates CSI frames as input, integrating the AveCSI framework and a shared decoder with the second branch. This supervision is executed by designing comprehensive loss functions, which contain constraints between the encoder outputs and constraints between the decoder outputs. In the attack phase, the trained AveCSI framework and decoder are directly employed to estimate skeleton frames based on CSI data.

In summary, the core contributions are as follows.

- We discover a new privacy-acquisition attack, CSIPose, that can passively estimate dynamic and static body poses in TTW scenarios. This attack reveals that commodity WiFi devices can be used for passively exposing the privacy of human poses even in TTW scenarios.
- We design *AveCSI*, a framework for preprocessing CSI data and extracting features for human pose estimation. We also design a three-branch network to train the pose estimation model based on the ideas of transfer learning and self-attention.
- We invite ten volunteers to evaluate seven WiFi devices in six different rooms. Results show the robustness of our attack on unseen devices and subjects in the training dataset. In particular, with a concrete thickness of up to 20 centimeters, our attack is still able to estimate dynamic and static human poses with an average accuracy of 93.27% and 83.22%, respectively. Our dataset and source codes will be available at https://github.com/luojiazhishu/CSIPose-code.

II. Preliminaries and Attack Model

In this section, we introduce the preliminaries of CSI signal and CSI affected by dynamic and static poses and discuss the attack model.

A. CSI Signal

CSI can characterize how a signal travels multiple paths from the transmitter to the receiver [17], [18]. Specifically, if a signal with carrier frequency f arrives at the receiver through M different paths, then the CSI value $H(f,t)$ can be denoted as:

$$H(f,t) = \sum_{j=1}^{M} \omega_j(t) e^{-i2\pi f \tau_j(t)}, \tag{1}$$

where $\omega_j(t)$ and $\tau_j(t)$ are the amplitude attenuation factor and the propagation delay of the j_{th} path, respectively. When a person is in the physical channel, he or she will have an impact on the CSI even in TTW scenarios [19]. Therefore, we have an opportunity to decode human poses from the CSI measurement. In general, CSI can be directly exported from the network interface card [3], [20], [21], and a CSI measurement $h(t)$ contains N Orthogonal Frequency Division Multiplexing (OFDM) subcarriers and can be denoted as:

$$h(t) = \left[H(f_1,t), H(f_2,t), \ldots, H(f_N,t)\right], \tag{2}$$

where $H(f_N,t)$ can also be denoted as $H(f_N,t) = |H(f_N,t)|e^{i\angle H(f_N,t)}$ with the amplitude $|H(f_N,t)|$ and the phase $\angle H(f_N,t)$ [22], [23]. In our work, we use the commodity mobile device with a single antenna as the detector to demonstrate the strong applicability of our attack. Therefore, we only exploit the CSI amplitude, since it is more reliable than the CSI phase when using a single antenna to collect CSI [14], [17], [24], [25] and the CSI phase does not show better performance in our evaluations as shown in Section VI-B. In what follows, the term CSI measurement generally refers to the amplitude of the CSI measurement if not otherwise specified.

B. CSI Affected by Dynamic and Static Poses

The impact of the same pose on CSI varies distinctly between static and dynamic contexts. Consider, for instance, a static scenario where the subject holds the left arm flat at the side of the body, compared to a dynamic scenario where the subject gradually transitions the left arm from hanging down to holding it flat and then lowering it. Despite both scenarios involving the pose of holding the left arm flat, the individual CSI measurement patterns differ. Fig. 1(a) and (b) visually represent partial CSI measurements in dynamic and static contexts, respectively. The correlation coefficient between the CSI measurements in Fig. 1(a) and (b) reaches a maximum value of only 0.9304. In

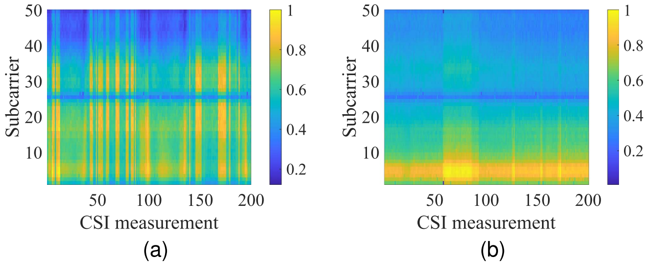

Fig. 1. (a) Partial CSI measurements when a subject gradually moves the left arm from hanging down at the side of the body to holding it flat and then lowering it. (b) Partial CSI measurements when a subject holds the left arm flat at the side of the body.

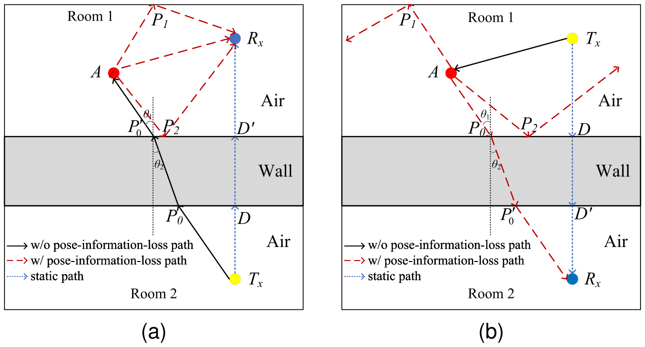

Fig. 2. Signal propagation of the moving subject on (a) the receiver side and on (b) the transmitter side.

contrast, the correlation coefficient between the CSI measurements in Fig. 1(b) has a minimum value of 0.9528, with most coefficients exceeding 0.98.

While (1) and (2) indicate that the CSI measurement comprises values for multiple subcarriers primarily influenced by the current multipath signal, it's essential to note that CSI, designed to enhance communication quality [18], results in the transmit signal at the current moment being influenced by prior CSI. Consequently, in both static and dynamic contexts, the CSI pattern for the same pose exhibits notable differences.

C. CSI Affected by the Relative Location Between the Subject and the Transceiver

When the body and the transmitter are on the same side of the wall, more information related to the human pose is lost. We consider a signal undergoes multiple paths to reach the receiver R_x from the transmitter T_x. When there is a moving subject in the path, we can model CSI as a linear superposition of dynamic and static paths.

$$H(f,t) = H_d(f,t) + H_s(f,t), \tag{3}$$

where $H_d(f,t)$ and $H_s(f,t)$ are signal components corresponding to dynamic paths and static paths. When the transceiver is on both sides of a wall, we can model the signal propagation [26]. As shown in Fig. 2, when the direct propagation path of the transceiver is perpendicular to the wall, we can represent the static path as a blue dashed line. We first consider the case where the moving subject and the receiver are on the same side, as shown in Fig. 2(a). The signal arrives at point P_0 of the wall, undergoes refraction from point P_0', then reaches the moving subject A. Signals reflect from different parts of A and reach the receiver R_x through multiple paths (e.g., $A \rightarrow R_x$ and $A \rightarrow P_2 \rightarrow R_x$). The propagation paths directly affected by A are concentrated in Room 1 where R_x is located. Therefore, R_x can capture all the detailed information affected by the moving subject. However, when we swap the positions of the transceivers, the moving subject and the transmitter are on the same side as shown in Fig. 2(b). Signals pass through the black straight line to reach the moving object A, undergoes reflection to reach the wall. The propagation paths directly affected by A are concentrated in Room 1 where T_x is located. Some paths can refract into the wall and finally penetrate out of the wall to reach the receiver R_x (e.g., $A \rightarrow P_0 \rightarrow P_0' \rightarrow R_x$). Others may not penetrate the wall (e.g., $A \rightarrow P_1$ and $A \rightarrow P_2$). Therefore, some paths carrying information of A may not be captured by R_x. As a result, more information about the human pose is lost when the subject is on the transmitter side.

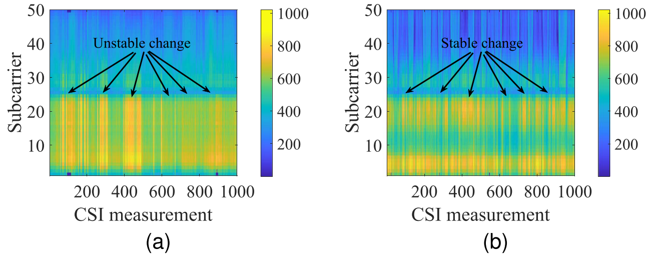

Fig. 3. (a) Partial CSI measurements when a subject is on the transmitter side. (b) Partial CSI measurements when a subject is on the receiver side.

To assess the impact of the relative location between the subject and the transceiver on CSI, we conducted an experiment where a subject performed the *wave* action on both the transmitter and receiver sides. This action involved uniformly raising the right arm from the side of the body and then lowering it. Fig. 3(a) displays partial CSI measurements collected when the subject is on the transmitter side, while Fig. 3(b) illustrates partial CSI measurements collected when the subject is on the receiver side. Observing Fig. 3, it is evident that when the subject is on the transmitter side, the changes in CSI measurements are notably uneven and unstable. Consequently, achieving reliable signal quality poses a more formidable challenge when the person is positioned at the transmitter end for human pose estimation.

D. Attack Model

Consider a typical scenario where an individual, situated in his own room, seeks to surveil his neighbor in the adjoining room. Furthermore, we assume that both rooms share similar layouts and wall types, as is common in commercial buildings with consistent architectural designs between adjacent residences. Given these premises, we assume specific capabilities for the attacker as follows.

- The attacker lacks access to the neighbor's WiFi network and the ability to compromise the neighbor's network devices.

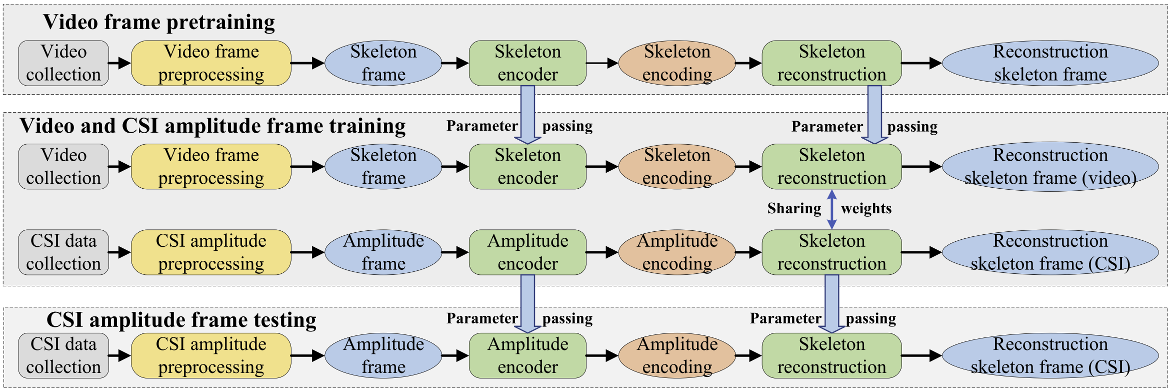

Fig. 4. Workflow of CSIPose .

- The attacker possesses only commodity WiFi devices, devoid of specialized and expensive probing equipment or self-built antenna arrays.
- The attacker is devoid of any prior information regarding the positions of the neighbor's network devices (e.g., the WiFi camera and the router). However, the attacker can strategically position the detector in his own room using publicly available information, such as the Received Signal Strength Indication (RSSI).

Therefore, the attacker can only use his own data to train human pose estimation models.

III. CSIPose Design

In this section, we present the workflow of CSIPose as shown in Fig. 4. Hereafter, we introduce three modules designed to generate human skeleton frames from CSI amplitude frames. The first module is *video frame pretraining*, which pre-trains a relatively stable skeleton encoder network and skeleton reconstruction network for skeleton frame generation using real skeleton frames as input. The second module is *video and CSI amplitude frame training*, which is a two-branch network including a skeleton encoder, an amplitude encoder, and a sharing skeleton reconstruction network, where the skeleton encoder and skeleton reconstruction network inherit the model weights of the same networks in the first module. Based on the idea of transfer learning [27], these two modules form a three-branch training network architecture through pre-training and parameter inheritance. The third module is *CSI amplitude frame testing*, which just uses the trained amplitude encoder and skeleton reconstruction network to generate skeleton frames using CSI amplitude frames.

A. Video Frame Pretraining

To ensure CSI amplitude frame can accurately reconstruct the skeleton frame, we first pre-train the skeleton encoder and the skeleton reconstruction network only using skeleton frames extracted from videos. Specifically, we use a commodity camera to obtain the video and then employ OpenPose [16] to get the Ground Truth (GT) human skeleton frames from the video frames. Two-Dimensional (2D) human skeleton frames are the input for the pretraining model based on AE to reconstruct skeleton frames.

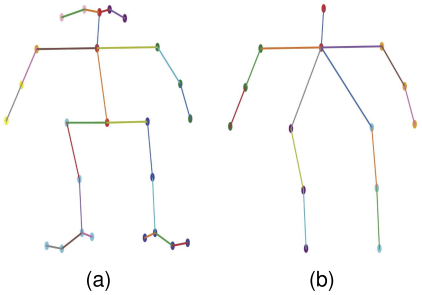

Fig. 5. Skeleton frames of (a) 25 points and (b) 14 points.

Video frame preprocessing. Initially, we use OpenPose to extract 25-point human skeleton frames from the video. Considering the erratic body movements at the beginning of the video and the impact of the CSI receiver's circuitry on the CSI, we remove the first five seconds of video frames. To simplify the calculations and preserve the main information of the human skeleton frame, we design a conversion method to convert the 25-point human skeleton frame to the 14-point frame, as shown in Fig. 5. For subsequent calculations, we normalize the horizontal and vertical coordinates of the skeleton map separately to the interval [0,1] based on the pixel aspect ratio of the video. Next, we reshape this set into a one-dimensional skeleton frame v containing 28 values, denoted as:

$$v = (x_1, y_1, x_2, y_2, \ldots, x_{14}, y_{14}), \tag{4}$$

where x_1 and y_1 represent the horizontal and vertical coordinates of the first point.

Skeleton encoder and skeleton reconstruction: We design a skeleton encoder network and a skeleton reconstruction network based on CNN to process GT skeleton frames. As shown in Fig. 6, the skeleton encoder network contains four $Conv2d$ layers to extract features, and each layer follows a composite layer including a $BatchNorm2d$ layer and the $LeakyReLU$ activation function to reduce the risk of overfitting and introduce nonlinearity. The skeleton frame v can be converted into a

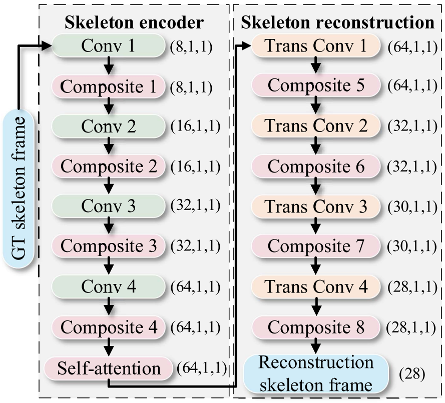

Fig. 6. Architecture of the skeleton encoder network and the skeleton reconstruction network.

64-dimensional skeleton encoding p after the final self-attention layer. The skeleton reconstruction network takes p as the input and outputs the reconstructed skeleton frame v'. This network contains four transposed CNN layers, and each layer follows a composite layer with the same structure as in the skeleton encoder. Finally, the Mean Squared Error (MSE) is calculated between the real skeleton frame v and the reconstructed skeleton frame v' as follows:

$$\mathcal{L}_v = \frac{1}{K}\sum_{i=1}^{K}\left(v_i - v_i'\right)^2, \tag{5}$$

where K is the number of video frames.

B. Video and Amplitude Frame Training

In this section, we design a two-branch network to generate skeleton frames using video frames to supervise CSI amplitude frames. In the video frame processing branch, the skeleton encoder and the skeleton reconstruction inherit model weights from those in the video frame pretraining module based on the idea of transfer learning [28]. The processed video frames serve as direct inputs to the skeleton encoder. Following the generation of the skeleton encoding vector, the skeleton frame reconstruction network utilizes this vector for skeleton frame reconstruction. As for the CSI amplitude frames processing branch, we introduce *AveCSI*, a framework designed to preprocess CSI amplitude and extract salient features. This model comprises two key modules: *CSI amplitude preprocessing* and *amplitude encoder*.

CSI amplitude preprocessing: A WiFi camera is used to acquire supervised video frames in the target room and a commodity mobile device (e.g., a smartphone) is used as the detector in the next room. These two devices are synchronized to collect data by manual time calibration. Specifically, the collect button of the detector and the video recording button of the camera are pressed at the same second while the data is being collected. Since static human poses are time invariant, time calibration does not make much sense. For dynamic human poses, this manual calibration method is sufficient due to the similarity between CSI measurements and the coherence between dynamic actions. Moreover, we had the same person operate the data collection each time and only collected one to two minutes of data at a time, which kept the cumulative error small.

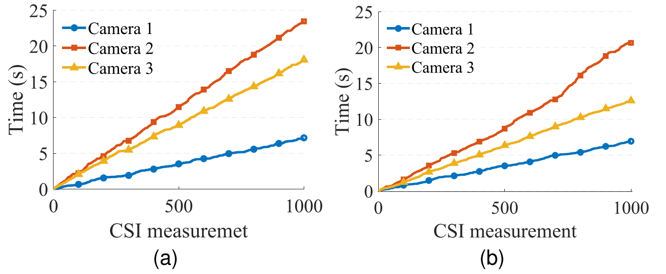

Fig. 7. Time taken by the three cameras to collect 1000 CSI measurements (a) when the left arm is held flat, and (b) when the right arm is swinging.

WiFi transmitters are common fixtures in homes with a high Packet Transmission Rate (PTR) (e.g., cameras and routers). The collected CSI sequence S can be denoted as:

$$S = \left[H_1, H_2, \ldots, H_M\right]^T, \tag{6}$$

where M is the number of collected CSI measurements and H_M just represents the amplitude of the M_{th} CSI measurement. Inspired by [14], we first denoise the sequence based on the Hampel filter, whose window size is equal to the number of collected CSI measurements. Next, we apply the 3-level $Sym4$ wavelet transform with a posterior median threshold rule on the sequence to enhance the information related to human presence.

To ensure that the video frames better supervise the generation of skeleton frames from the amplitude frames, we segment the CSI sequence according to the frame rate of the video. For example, if the video frame rate is 20 FPS/s, we segment the CSI sequence at intervals of 0.05 seconds. However, the camera's PTR always fluctuates due to the variable bit rate encoding, which outputs different numbers of packets depending on the complexity of the image content [29]. Moreover, the number of CSI measurements collected from different devices is significantly different over the same period of time as shown in Fig. 7. When performing CSI amplitude frames synchronized with video frames, even the amplitude frames corresponding to video frames of the same pose may contain different numbers of CSI measurements due to different devices and PTR variations.

To mitigate the effect of this number of CSI measurements on the amplitude frame, classical method to obtain amplitude frames [10], [30] is to use linear interpolation to make the same number of measurement at a fixed time. However, the number of measurements over a fixed period of time varies greatly in our scenario, so using linear interpolation to obtain amplitude frames of the same pose with large differences in the original number may introduce too much error and lead to large differences in the amplitude frames within the same pose. Therefore, we apply the method of taking the average CSI measurement at a fixed time interval as the amplitude frame, which can avoid introducing much error. The amplitude frame a can be denoted as:

$$a = \left[\overline{H}_1, \overline{H}_2, \ldots, \overline{H}_N\right], \tag{7}$$

where $\overline{H}_N$ represents the average amplitude of the N_{th} CSI subcarrier. To facilitate the computation and reduce the influence of walls on the absolute value of CSI amplitude, we scale each

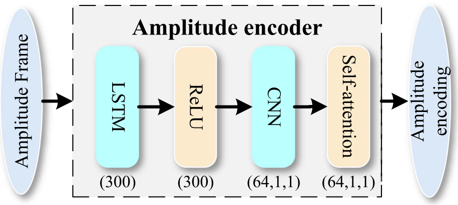

Fig. 8. Architecture of the amplitude encoder network.

frame of a to the interval [0,1] based on the global maximum of this sequence.

Amplitude encoder: As shown in Fig. 8, the amplitude encoder network contains three layers. The first layer is an LSTM model with two LSTM layers to extract the distribution characteristics across subcarriers, where the hidden state of the last time step is used as the output of this layer. A $ReLU$ activation function has followed this layer. The second layer is a CNN layer to be consistent with the output of the skeleton encoder. The third layer contains a self-attention layer to enhance critical features. The skeleton frame a can be converted into a 64-dimensional amplitude encoding q. To make video frames better supervise amplitude frames, the skeleton reconstruction network of the amplitude encoder network and the skeleton encoder network share weights during training.

Loss calculation: To better exploit the supervisory role of video frames, we design loss constraints from two perspectives. The first perspective is the output of the two encoder networks. The skeleton encoder outputs a relatively meaningful encoding p by inheriting the encoder parameters from the video pretraining module. This encoding is used to guide the output of the amplitude encoder for efficient knowledge transfer. We use MSE to compute the loss between the skeleton encoding p and the amplitude encoding q as follows:

$$MSE_{pq} = \frac{1}{K}\sum_{i=1}^{K}\left(p_i - q_i\right)^2, \tag{8}$$

where K is the number of frames. The second perspective is the output of the skeleton reconstruction network. This network generates reconstructed skeleton frames (Video) s and reconstructed skeleton frames (CSI) z with skeleton encoding p and amplitude encoding q as outputs, respectively. We first calculate the loss between the two reconstructed skeleton frames (i.e., s and z) to ensure the stability of the skeleton reconstruction network. Next, we calculate the loss between the two reconstructed skeleton frames (i.e., s and z) and the real skeleton frame v separately to achieve accurate supervision and knowledge transfer from the real skeleton frames to the reconstructed skeleton frames. We use MSE to calculate these three losses. Finally, the objective of the entire video and CSI amplitude frame training module is to minimize the following loss function:

$$\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K}\Big(\alpha\left(p_i - q_i\right)^2 + \beta\left(s_i - z_i\right)^2 + \gamma\left(s_i - v_i\right)^2 + \delta\left(z_i - v_i\right)^2\Big), \tag{9}$$

where α, β, γ, and δ are weights adjusting the impact of each objective to the overall loss function.

TABLE I
Specific Devices Used in Our Experiments

Device Type	Device Model	Mean Packet Per Second
WiFi Camera	360 Cloud Camera 6C (6C)	166
	Camera 83do (83do)	77
	360 D806 Cloud Camera (D806)	66
	EZVIZ H6C (H6C)	194
	XiaoMi Cloud Camera (MI)	102
WiFi Router	TP-LINK C7A4 (C7A4)	60
	CISCO RV100W (Cisco)	90

C. CSI Amplitude Frame Testing

After the model training, only CSI amplitude frames are involved in the testing phase, where the amplitude encoder and the skeleton reconstruction network inherit the corresponding parameters from the trained model. CSI data collection and amplitude frame generation are the same as presented in the previous module. The fixed time interval is consistent with that during training the model. In addition, as mentioned in Section II-B, the same pose in static and dynamic contexts has different impacts on CSI, so we train a skeleton reconstruction model using static and dynamic data, respectively. As shown in Fig. 1, there is a strong distinction between the effects of static and dynamic poses on CSI, which can be recognized simply by some statistical features (e.g., standard deviation [31]). Therefore, we can calculate the standard deviation of the CSI sequence corresponding to the amplitude frame to determine whether the amplitude frame is an input to the static skeleton reconstruction model or the dynamic reconstruction model.

IV. Evaluation

In this section, we report performance evaluation results under different setups.

A. Experimental Setup

To extract CSI measurements, we utilize the capabilities of the CSI extraction tool, nexmon_csi [3], known for its versatility on mobile devices. Our detector is the LG Nexus 5 smartphone. We activate the monitor mode of the WiFi chip to ensure the comprehensive capture of all proximate WiFi packets. In our experimental setup, a CISCO router operating in the 2.4 GHz band, featuring a 20MHz channel bandwidth, is designated as the WiFi access point situated within the target room. Our work is to reveal the privacy risks associated with commodity WiFi devices commonly found in households. Therefore, we select devices that are ubiquitous and have relatively high PTRs. A detailed listing of the specific devices employed in our experiments is provided in Table I. We use six rooms in our experiments and some of them are as shown in Fig. 9. As shown in Fig. 10, we collect data in one pair of adjacent rooms. Room 1 and Room 2 have the same dimensions (i.e., 8.8 m × 5.9 m) and wall types (i.e., three sides of 20cm concrete and one side of 8cm

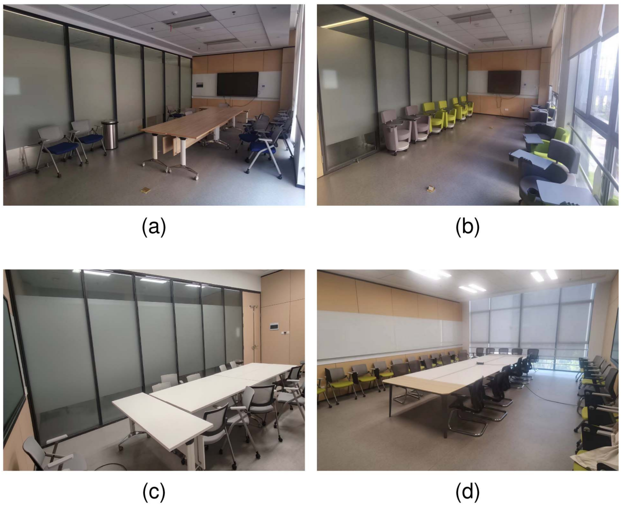

Fig. 9. (a) Room 1, (b) Room 2, (c) Room 4 and (d) Room 6 used in our experiments.

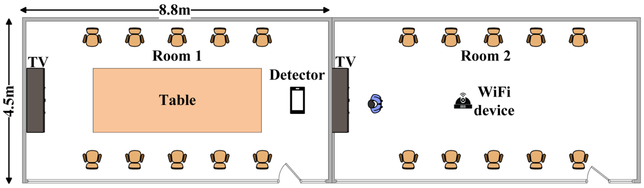

Fig. 10. Floor plan of Room 1 and Room 2.

double-glazed). Compared with Room 1 and Room 2, Rooms 3 and Room 4 have the same wall types and similar dimensions, but different room layouts. As shown in Fig. 9(d), Room 5 and Room 6 are discussion rooms with the size of 14.5 m × 9 m and the wall type of two sides of 20 cm concrete and two sides of 8 cm double-glazed. In addition, Room 1&2, Room 3&4, and Room 5&6 represent the three scenarios of our experiments.

Model settings: In the training phase, two models in the *video frame pretraining* module and *video and CSI amplitude frame training* module are trained for 1300 epochs and 1000 epochs using the $Adam$ optimizer with a learning rate of 0.001 respectively. The **Parameter passing** is implemented through the $torch.load\_state\_dict()$ method. For example, if the $Encoder$ in $model1$ and the $Encoder$ in $model2$ have the same network structure and model1 has been trained, then you can use this code $model1.Encoder.load\_state\_dict(model2.Encoder.state\_di\ ct())$ to implement the $Encoder$ in $model2$ to directly load the weights of the $Encoder$ in $model1$. Moreover, model weights and the $LeakyReLU$ slope are initialized with the default values. For four weights used to adjust the overall loss function $\mathcal{L}$, they are empirically set to 1 for α and γ, to 0.5 for β, and to 0.6 for δ. The entire network is trained and tested on a computer with Intel(R) Core(TM) i7-8700 CPU and 16GB of RAM.

Dataset: In this study, we recruit a cohort of 10 participants who exhibit variations in height, weight, and age to gather data. According to poses in existing works, we define four dynamic poses: left/right arm raised and lowered at the side (*Wave_left/right*) and left/right leg raised and lowered at the side (*Leg_left/right*). We also define four static poses: left/right arm raised flat at the side (*Arm_left/right*), standing with the feet apart (*Stand_apart*), and standing with feet closed (*Stand_closed*). As shown in Fig. 10, the WiFi transmitter is positioned about 2 meters directly in front of the subject, while the detector is positioned about 1 m away from the wall. We collect a total of 165.6K effective CSI measurements.

Baselines. To evaluate the effectiveness of our network, we conduct comprehensive evaluations by benchmarking it against state-of-the-art pose estimation systems. However, existing works [4], [5], [10], [30], [32] achieve the human pose estimation using a multi-transceiver system by actively sending custom WiFi signals in indoor scenarios, which is significantly different from our work on human pose estimation based on the CSI data passively collected from one pair of transceiver antennas in TTW scenarios. Therefore, we can only provide one-channel CSI data from one pair of antennas for other multi-antenna systems, and we use replicated interpolation to preprocess the data to meet the data sampling rate requirement in the comparison works. Moreover, we does not select some schemes [4], [5] based on the angle of arrival due to we can not estimate accurate angles of arrivals using a pair of antennas [31]. Specifically, we select two representative works on human pose estimation: Avola et al. [10] employed a two-branch network for supervised human dynamic pose generation using video information, and Chen et al. [32] estimated the static human pose based on a spatial encoder and the attention mechanism.

Evaluation metrics: To evaluate the performance of CSIPose, we follow the methodology introduced by Guo et al. [13]. Specifically, we calculate the average pixel distance between the predicted coordinate points in the reconstructed skeleton frame and the real coordinate points in the skeleton frame generated by OpenPose. We consider skeleton frames with average pixel distances below a threshold as correctly predicted skeleton frames. The ranges of the threshold θ include 25, 30, 40, and 50 to gauge the accuracy of the network's output according to the previous work [13]. In our work, we adopt their evaluation metric known as Percentage of Correct Skeletons (PCS). This metric is defined as:

$$PCS(\theta) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{L}\left(\left|\left|p_{i,j}^{n} - g_{i,j}^{n}\right|\right|_2 \leq \theta\right), \tag{10}$$

where N represents the number of test frames, $\mathbb{L}$ is a logical operation that outputs 1 if True and outputs 0 if False. $p_{i,j}^{n}$ and $g_{i,j}^{n}$ are the values for the $(i,j)_{th}$ coordinate points on the prediction coordinate points and corresponding ground-truth coordinate points, respectively, where $i \in 1, 2, \ldots, 14, j \in 1, 2, \ldots, 14$.

B. Overall Performance

We first show the overall performance of our system, and also compare CSIPose with two typical and latest schemes. We only use the data from the device 6 C in Room 1&2 as training data. The test data is from the same device and does not include any

TABLE II
RESULTS OF STATIC AND DYNAMIC POSES FOR DIFFERENT ROOM LAYOUTS

Pose	CSIPose	Avola et al.	Chen et al.	Pose	CSIPose	Avola et al.	Chen et al.
static poses	70.40%[1]	16.79%	66.75%	dynamic poses	81.74%	55.77%	70.85%
	75.38%[2]	19.43%	73.27%		85.97%	60.67%	78.15%
	80.95%[3]	25.10%	78.79%		91.67%	69.39%	87.03%
	85.38%[4]	32.60%	84.06%		94.02%	75.11%	91.33%

[1,2,3,4] represent values of $PCS(25)$, $PCS(30)$, $PCS(40)$, and $PCS(50)$ respectively.

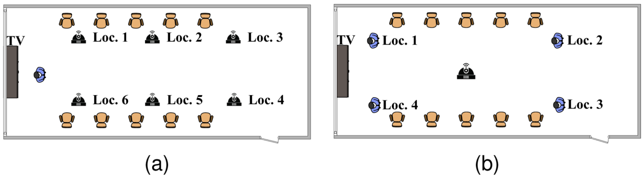

Fig. 11. (a) Different locations of the transmitter (i.e., a camera) and (b) different locations of the subject in Room 1.

training data. The models that we trained are also used for the testing of other impact factors in later subsections.

Table II shows the average accuracy of the three schemes for static and dynamic poses. Specifically, the values of $PCS(25)$ and $PCS(50)$ of CSIPose for static poses are 70.40% and 85.38%, while those of other two schemes are 16.79% and 32.60%, 66.75% and 84.06%, respectively. Under the evaluation metrics of $PCS(25)$ and $PCS(50)$, the values of CSIPose are 3.65% and 1.32% higher than the best scheme, respectively. The values of $PCS(25)$ and $PCS(50)$ of CSIPose for dynamic poses are 81.74% and 94.02%, while those of other two schemes are 55.77% and 75.11%, 70.85% and 91.33%, respectively. Under the evaluation metrics of $PCS(25)$ and $PCS(50)$, the values of CSIPose are 10.89% and 2.69% higher than the best scheme, respectively. These results demonstrate that for static and dynamic poses, CSIPose is superior to the other two schemes. We think the reasons could be twofold. On the one hand, it is that our data comes from a strict through-wall scenario in which the quality of the data is relatively poor. On the other hand, it is that our data comes from a pair of transceiver antennas, which do not provide enough spatial information.

To visualize the performance of the system, we show skeleton frames corresponding to the different poses of the subject. The first line of Fig. 12 shows the skeleton frames after OpenPose processing, and the second line shows the estimated skeleton frames of CSIPose . Specifically, Fig. 12(a) and (c) show four static poses, while Fig. 12(b) and (d) show four dynamic poses. From Table II and Fig. 12, we can observe that the estimation accuracy of the dynamic poses is high, while the estimation accuracy of the static poses is slightly lower. The reason may be that the influence of the static poses on CSI are more likely to be overwhelmed by background noise, while CSI influenced by the dynamic pose has a higher signal-to-noise ratio. However, overall, the estimated skeleton frames of CSIPose are still relatively accurate, which is satisfactory in passive through-wall scenarios.

TABLE III
RESULTS OF STATIC AND DYNAMIC POSES FOR DIFFERENT ROOM LAYOUTS

Pose	Room 1&2	Room 3&4	Room 5&6	Pose	Room 1&2	Room 3&4	Room 5&6
arm_ left	71.85%	71.89%	69.02%	wave_ left	83.61%	88.79%	38.42%
	76.15%	75.29%	72.07%		89.44%	89.77%	46.26%
	80.02%	78.42%	75.07%		91.73%	91.77%	59.63%
	82.33%	79.94%	77.07%		93.24%	92.75%	62.29%
arm_ right	73.56%	71.71%	74.11%	wave_ right	66.99%	84.06%	34.49%
	76.72%	76.37%	76.94%		76.91%	82.48%	40.07%
	80.24%	81.2%	80.56%		85.49%	90.9%	55.86%
	81.86%	83.15%	82.64%		91.98%	93.01%	59.09%
stand_ apart	65.64%	64.23%	60.5%	leg_ left	84.94%	72.52%	28.4%
	69.09%	69.24%	64.36%		89.2%	80.63%	34.61%
	76.22%	79.39%	71.89%		94.60%	90.51%	40.11%
	81.02%	82.74%	77.06%		95.45%	94.48%	54.18%
arm_ closed	76.33%	73.73%	73.75%	leg_ right	69.72%	78.1%	29.67%
	79.5%	77.24%	75.97%		76.93%	82.48%	35.21%
	84.51%	83.67%	79.57%		86.46%	88.65%	35.86%
	87.66%	89.02%	82.64%		92.39%	93.87%	45.86%

C. Impact of Different Room Layouts

To evaluate the impact of the different room layouts on the performance of our proposed system, we conduct experiments on three sets of different room layouts with varying dimensions and structural compositions, where each set contains two adjacent rooms as shown in Fig. 10. We use the model discussed in Section IV-B and the training dataset only includes the data collected from the Room 1&2. For the testing data, we invite a subject to collect it in three sets of room layouts. As shown in Table III, most of the values of $PCS(25)$ exceed 70% for static and dynamic test data collected from Room 1&2 and Room 3&4 respectively, while the average values of $PCS(25)$ of Room 5&6 are 69.35% and 32.75% for static and dynamic test data. The average values of $PCS(50)$ for static and dynamic data from these three set of room layouts are 83.22%, 83.71%, 79.85% and 93.27%, 93.53%, 55.36%. The reason is that the training dataset is collected from Room 1&2, and Room 3&4 has a more similar layout to Room 1&2 than Room 5&6. This result suggests that the impact of room layout on CSIPose performance is significant. However, in our attack model, we assume that the target room has the same or similar room layouts. Therefore, this attack of training a good model in one's room and then going on to estimate the privacy of the neighbor's poses in his own home is feasible.

D. Impact of Different Devices

To evaluate the applicability of CSIPose for different devices, we invite a subject to collect CSI data from different WiFi devices as depicted in Table I. We use the model discussed in Section IV-B and the training dataset only includes the data collected from the device 6 C. Fig. 13 presents the results of testing the model with the test data of one static pose (arm_left) and one dynamic pose (leg_left). The values of $PCS(25)$ and $PCS(50)$ of most devices for the static pose can achieve more than 70% and 80%, while that for the dynamic pose can achieve more than 70% and 90%. The PTRs of our devices are slower compared to the PTR 1000 in previous works [4], [5], [8], [9],

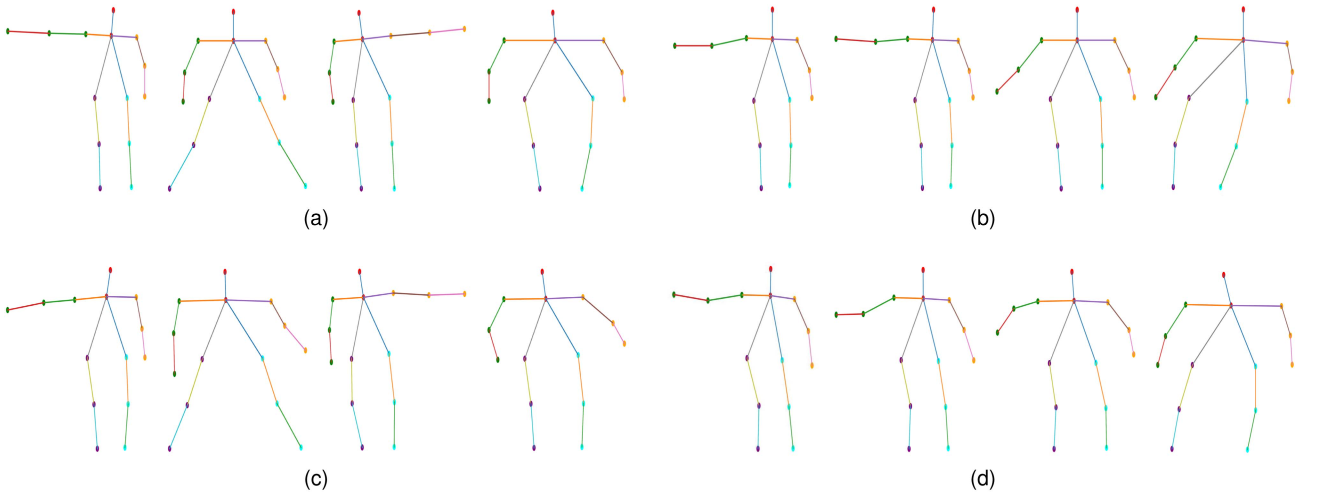

Fig. 12. Examples of the constructed skeletons. (a) Ground truth static pose skeletons. (b) Ground truth dynamic pose skeletons. (c) Generated static pose skeletons. (d) Generated dynamic pose skeletons.

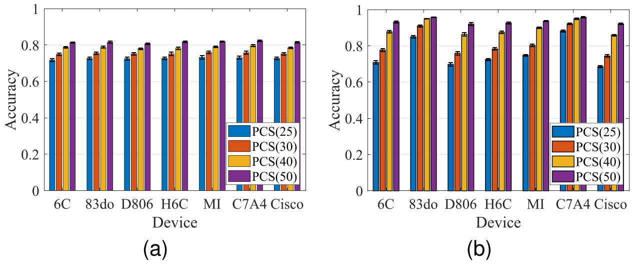

Fig. 13. Accuracy of (a) the arm_left pose and (b) the leg_left pose under different devices.

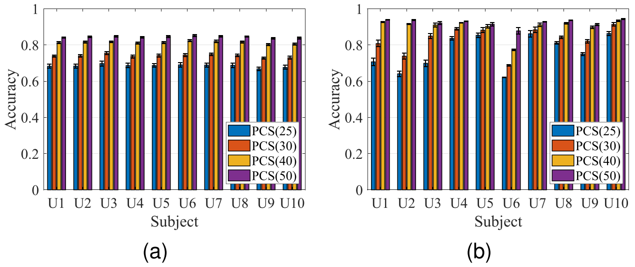

Fig. 14. Accuracy of (a) the arm_left pose and (b) the wave_left pose under different subjects.

[32], but with the processing framework we designed, CSIPose still achieves a high recognition accuracy. This result indicates that CSIPose is applicable for different devices and also effective for unseen devices in the training dataset. The attacker can train the attack model with the device different from the device in the target room.

E. Impact of Different Subjects

To verify the effectiveness of the model trained by an attacker with his own data against an unknown subject, we meticulously select 10 subjects spanning diverse characteristics, including age, gender, height, and weight. We use the model discussed in Section IV-B and the training dataset only includes the data collected from one subject. For the testing data, we ask 10 subjects to perform static and dynamic poses to collect it. Fig. 14 presents the results of testing the training model with the test data of one static pose (arm_left) and one dynamic pose (wave_left). CSIPose is implemented on the basis that the effect of the same pose on CSI is similar in the same context, so that when body size differences are large enough to dramatically change the effect of the same posture on CSI, attack performance declines. However, the attack performance of CSIPose is acceptable since the average values of $PCS(25)$ for static and dynamic poses still reach 68.49% and 77.07%. This result indicates that the model is also effective for unseen subjects in the training dataset. The attacker can train the model just using his own training data, which increases the viability and stealth of the attack.

F. Impact of Different Transmitter Locations and Subject Locations

To evaluate the impact of different transmitter locations on the performance of CSIPose, we invite a subject to collect CSI data at different transmitter locations as shown in Fig. 11(a). We just ask the subject to perform a static pose (arm_right) and a dynamic pose (wave_right), since the difference in accuracy between different poses is little as reported in previous evaluations. Fig. 15 illustrates the accuracy of the arm_right pose and the wave_right pose at different transmitter locations. We can observe that the accuracy of $PCS(25)$ and $PCS(50)$ for arm_right remains about 75% and 82%, while that for wave_right remains over 60% and 90%. The accuracy for the static pose is not significantly affected by the different transmitter locations, while the values of $PCS(25)$ for the dynamic pose is significantly affected by that. The reason may be that Loc. 1 and Loc. 6 are closer to location of the training data. This result indicates that CSIPose is robust to different transmitter locations, and the attacker can directly implement the see-through attack even without knowing the transmitter location.

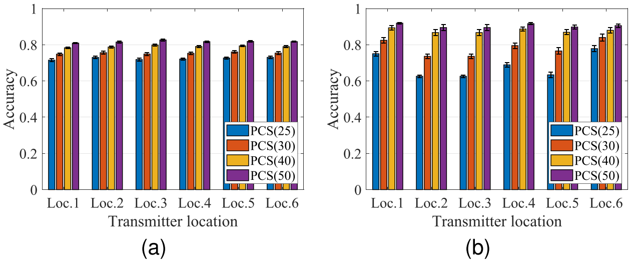

Fig. 15. Accuracy of (a) the arm_right pose and (b) the wave_right pose under different transmitter locations.

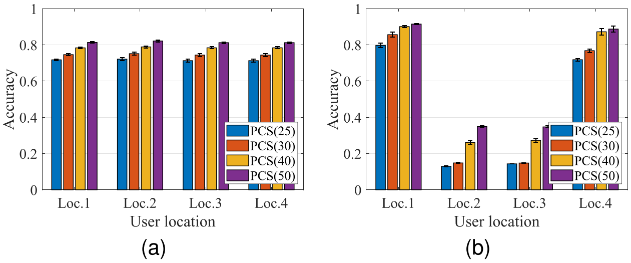

Fig. 16. Accuracy of (a) the arm_right pose and (b) the wave_right pose under different user locations.

To investigate the impact of subject locations on the performance of our proposed system, we invite a subject to collect CSI data at four different locations as shown in Fig. 11(b), respectively. We also ask the subject to perform a static pose (arm_right) and a dynamic pose (wave_right). The transmitter is placed at the room's center as the setup in Fig. 10. From Fig. 16, we can observe that the performance of four locations has little difference for the arm_right pose, while for the wave_right pose, the performance of Loc. 1 and Loc. 4 is significantly better than the performance of other two locations. Specifically, the accuracy of $PCS(50)$ for Loc.1 and Loc. 4 can reach 91.50% and 88.75%, while that for Loc. 2 and Loc. 3 are only 35.01% and 34.73%. The results show that changes in the user's position significantly affect accuracy, with closer proximity generally resulting in higher accuracy. This result indicates that CSIPose does not have generalization with different user locations, and the attacker may need to first estimate the transmitter location, which can be achieved by collecting RSSI as mentioned in [31]. Therefore, the attacker still can estimate the subject's body pose in some areas.

G. Ablation Study

To evaluate the effectiveness of our design framework *AveCSI* in CSI processing and feature extraction, we first compare two methods to convert the CSI frame. We design a method to compute the mean measurement of a period of CSI measurements as the CSI frame, while classical method is to make the fixed-number CSI measurement sequence as the CSI frame. We utilize the training dataset and the testing dataset used for the impact of different devices in Section IV-D. Using mentioned two methods, we train two models and present the average accuracy of different devices for each pose. As shown in Table IV, the average accuracies $PCS(25)$ and $PCS(50)$ of our method are 76.05% and 89.46%, while those of the classical method are 64.51% and 83.64%. This result demonstrates that our method to convert the CSI frame is superior to classical method. The self-attention layer helps to extract significant features for human pose estimation. To quantify the gain of the self-attention layer, we delete the self-attention layer from the model and train a model. As shown in Table IV, the test results show that the average accuracies $PCS(25)$ and $PCS(50)$ of the method without SA are 61.83% and 80.68%, which are 14.22% and 8.78% lower than our accuracies, respectively. This result demonstrates that SA is very helpful for feature extraction. In summary, These results indicate that AveCSI can improve the accuracy of the pose estimation.

TABLE IV
EVALUATION RESULTS OF ABLATION EXPERIMENTS

Method	arm_ left	arm_ right	stand_ apart	stand_ closed	wave_ left	wave_ right	leg_ left	leg_ right
Calssical method	54.74%	61.55%	42.65%	57.95%	83.11%	74.83%	71.14%	70.10%
	62.52%	65.47%	51.04%	65.36%	85.74%	82.41%	76.82%	75.86%
	65.53%	71.47%	62.70%	74.61%	89.43%	88.42%	86.34%	84.75%
	67.69%	77.57%	79.12%	80.43%	91.66%	90.92%	91.69%	90.03%
w/o SA	67.37%	77.52%	64.78%	72.93%	59.02%	50.03%	53.01%	49.95%
	70.33%	79.07%	73.47%	83.47%	64.23%	59.43%	60.04%	55.87%
	77.39%	82.28%	77.73%	89.02%	71.30%	71.37%	70.97%	64.93%
	79.53%	83.63%	86.95%	91.81%	77.09%	76.84%	77.64%	71.95%
Ours	**74.14%**	76.49%	**67.46%**	**78.37%**	**84.93%**	**78.21%**	**79.28%**	69.52%
	75.87%	78.07%	72.26%	84.18%	87.24%	85.30%	84.32%	76.93%
	79.83%	80.26%	77.09%	89.72%	91.46%	90.33%	91.80%	86.26%
	84.09%	81.74%	86.92%	91.16%	**93.17%**	**92.22%**	**94.92%**	**91.49%**

V. DEFENSE

We now discuss defenses against CSIPose . Defending against wireless sensing in passive scenarios has been a challenging and topical research matter [33]. Since detecting passive sensing attacks is difficult, we can only defend against them on the transmitter side. Specifically, we can design some encryption or obfuscation measures in terms of channels and signals on the transmitter side.

From the channel aspect, we can add additional devices in the target room to change the channel characteristics or emit a synthesized signal to confuse the characteristics of the target signal [31], [34], [35]. For example, Staat et al. [35] designed a configuration algorithm based on smart reflective surfaces to confuse the wireless channel. While this work can be carefully designed to not affect the signal, it is generally more expensive to synthesize. Therefore, this defense can defend users against our attack, but it is not universal.

Considering the signaling aspect, we can use suitable obfuscation or encryption techniques to modify the signal [36], [37], [38]. Meng et al. [37] designed an encryption method based on antenna switching to interfere with passive wireless sensing at the source. However, this work is not applicable to IoT devices with only one antenna. Qiao et al. [38] used full-duplex radios to randomly modify and retransmit physical signals in the environment, which can obfuscate sensitive physical information.

However, this work faces the same weakness of high cost. The defense in [31] is relatively low cost but has the potential to interfere with normal communication. Therefore, defenses from a signaling perspective can stop our attacks accompanied by the risk of affecting the communication performance.

In conclusion, although CSIPose can be defended, defenses are either costly or have the risk of increasing power consumption or affecting communication performance. An applicable, low-cost method for defending against passive sensing that does not affect normal communication is urgently needed.

VI. Discussion

We discuss the limitations of CSIPose and our future work in this section.

A. Limitations

CSIPose still exhibits certain limitations that require further improvement.

Platform: CSIPose necessitates acquiring root access to the phone and relies on nexmon_csi for CSI collection, thereby constraining the diversity of applicable attack platforms. While alternative embedded platforms like the Raspberry Pi, offer the potential for more practical attacks. We are also exploring implementing CSIPose on some embedded platforms. Therefore, the platform is vulnerable to the limitations of CSIPose.

Number of human poses: In our experimental setup, we just design four static poses and four dynamic poses. The number of poses is small compared to some other work, and our scheme is not currently capable of estimating arbitrary movements. This is mainly due to the fact that our scheme is pattern-based and in a TTW scenario, whereas some schemes model the human body in an indoor environment and with multiple antenna arrangements. However, our work fills the gap in human pose recognition based on passively collected CSI in TTW scenarios, which is a huge guidance for future work. In the future, we will also go further to explore pattern-based schemes that include more poses as well as model-based schemes in our attack scenarios.

Subject location: As shown in Section IV-F, the location of the subject has a significant impact on the performance of CSIPose . The nature of our system is pattern recognition of human poses, but CSI is sensitive to location [24], [39]. Therefore, our system does not have a very robust performance. However, with simple passive position estimation [31], an attacker can simulate the corresponding positions for targeted training. Therefore, we believe that the impact of this weakness on our system is limited.

B. Estimation Using CSI Phase

CSI phase is widely used for WiFi sensing since it can be used to obtain some advanced phase features (e.g., Arrival-of-Angle). However, our receiver only have one antenna and is difficult to obtain advanced features. To evaluate the feasibility of estimating poses with CSI phase, we extract CSI phase data from the training dataset and the testing dataset used in Section IV-D and train a model. The evaluation results show that the average estimation accuracy $PCS(25)$ and $PCS(50)$ of this model for human poses are 74.20% and 89.12%, respectively. As shown in Table IV, the average estimation accuracy $PCS(25)$ and $PCS(50)$ of our method are 76.05% and 89.46%, respectively. The CSI phase does not show a better performance than the CSI amplitude. And the CSI amplitude seems to perform better under a more stringent evaluation metric (i.e., $PCS(25)$). Therefore, we choose the CSI amplitude to estimate human poses. And we will further explore the applicability of the CSI phase to human pose estimation.

C. Multi-Subject Pose Estimation

Multi-subject estimation becomes imperative, considering the frequent presence of more than one person in a domestic room. Nevertheless, the coexistence of multiple subjects introduces complexities as their actions superimpose on each other, causing the CSI patterns to become perplexing and indistinguishable. This interference is particularly pronounced in our TTW attack scenario, exacerbating the inherent challenges in deciphering the CSI data. To address this issue, we posit that modeling the propagation of CSI in TTW scenarios holds promise as a viable solution. While Zhang et al. [26] have aimed at developing a model for estimating CSI in through-wall scenarios, its performance experiences a significant degradation, especially in scenarios featuring reinforced concrete. Hence, our subsequent focus centers on further exploration and refinement of CSI estimation models tailored for TTW scenarios, thereby enhancing the accuracy of the multi-subject pose estimation in TTW scenarios.

VII. Related Work

Human pose estimation has been more researched in the fields of smart healthcare, virtual reality, and wireless security. RF-based methods [40], [41], [42], [43], [44], [45] can achieve accurate estimation of position, posture, and motion characteristics due to the superior performance of specialized equipment. However, our work focuses on revealing the privacy threats regarding human poses posed by commodity WiFi devices. Therefore, this section only explores WiFi-based human pose estimation methods, which can be categorized into methods in LOS scenarios and methods in NLOS scenarios according to the location of the transceiver.

Methods in LOS scenarios: The LOS scenario allows the receiver to obtain more accurate information about the human body pose from the CSI as discussed in Section II-C. Therefore, many methods [10], [11], [12] use deep learning networks to generate human pose maps using CSI as input annotated with real video frames. Wang et al. [11] proposed a U-Nets-based deep neural network to map a CSI tensor to a human pose using a MIMO system. Avola et al. [10] designed a novel two-branch generative WiFi sensing framework that inherently considered motion information to synthesize coherent human silhouette and skeleton videos from CSI measurements. Zhou et al. [12] designed a domain-independent neural network to extract subject-independent features and convert them into fine-grained human pose images. Wang et al. [46] designed a self-encoder to estimate human poses and utilized a diffusion model to fit the human poses into avatar poses. However, these methods do not consider the influence of blocks or walls, and require the transmitter to actively connect to the receiver to obtain a stable sampling rate.

Methods in NLOS scenarios: Methods in NLOS scenarios can be categorized into loosely constrained methods and tightly constrained methods depending on the type of obstacle. Loosely constrained methods [4], [7], [8], [9], [13] only experimentally evaluate the system performance in the presence of simple obstacles (e.g., wooden boards and screens). Jiang et al. [5] modeled the human skeleton as a tree and used CSI data from multiple receivers in space to achieve 3D human pose estimation using a deep learning network. Li et al. [9] was based on the 2D Fast Fourier Transform of CSI using a conditional adversarial generative network to segment the boundaries of the object and the human body. Ren et al. [4] realized 3D pose estimation of the human body based on spectrograms of CSI and models of joint movements. Strictly constrained method [5] achieves the estimation of the human pose under the barrier of the wall.

However, since the subject and the receiver were in the same room in GoPose [5], the wall only attenuated the strength of the signal and hardly interfered with the effect of the human body on the signal. Therefore, GoPose is not suitable for our attack scenario where the subject and the transmitter are in the same room. Owing to the spatial resolution provided by the MIMO system, these methods can achieve human pose estimation in the presence of occlusions without having to cope with the effects of occlusions.

Unlike these existing methods, we first use GT skeleton frames to train a dependable skeleton reconstruction network. Second, we implement CSI amplitude-based human pose estimation in a novel deep neural network employing well-designed bounded loss. In particular, we only passively collect WiFi signals with low and unstable sampling rates instead of signals with stable and high sampling rates as in other schemes. Our scheme overcomes the negative impact of low-quality data and emphasizes the significant threat to human pose privacy posed by commodity WiFi devices.

VIII. Conclusion

This paper presents CSIPose, a privacy-acquisition attack designed to clandestinely estimate dynamic and static human privacy poses based on CSI in TTW scenarios. It can directly and passively collect CSI data from WiFi devices to estimate the dynamic and static poses of the human body indoors. We design a three-branch network model to realize the supervision of ground truth video frames on CSI-generated skeleton frames. Additionally, AveCSI uses the average of CSI sequences as CSI frames to address the problem of data instability, which is caused by the fact that packets emitted by uncontrolled devices during normal operation are low-rate and unstable. For feature extraction, AveCSI consists of a three-layer feature extraction network which contains the LSTM layer, the CNN layer and the self-attention layer. Finally, we evaluate the performance of the system in different room layouts, devices, subjects, device locations and user locations. The evaluation results emphasize the strong generalization of the system and demonstrate the privacy risk posed by commodity WiFi devices.

References

[1] M. Intelligence, "Wi-Fi market," 2023. [Online]. Available: https://www.mordorintelligence.com/zh-CN/industry-reports/wi-fi-market

[2] F. Qi et al., "Unauthorized and privacy-intrusive human activity watching through Wi-Fi signals: An emerging cybersecurity threat," *Concurrency Comput. Pract. Experience*, vol. 35, no. 19, 2023, Art. no. e7313.

[3] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your CSI: A channel state information extraction platform for modern Wi-Fi chipsets," in *Proc. 13th Int. Workshop Wireless Netw. Testbeds, Exp. Eval. Characterization*, 2019, pp. 21–28.

[4] Y. Ren, Z. Wang, S. Tan, Y. Chen, and J. Yang, "Winect: 3D human pose tracking for free-form activity using commodity WiFi," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1–29, 2022.

[5] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "GoPose: 3D human pose estimation using WiFi," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–25, 2022.

[6] M. Xu, Z. Guo, L. Gui, B. Sheng, and F. Xiao, "Wispe: A cots Wi-Fi-based 2-D static human pose estimation," *IEEE Syst. J.*, vol. 17, no. 3, pp. 3560–3571, Sep. 2023.

[7] Y. Wang, L. Guo, Z. Lu, X. Wen, S. Zhou, and W. Meng, "From point to space: 3D moving human pose estimation using commodity WiFi," *IEEE Commun. Lett.*, vol. 25, no. 7, pp. 2235–2239, Jul. 2021.

[8] W. Jiang et al., "Towards 3D human pose construction using WiFi," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–14.

[9] C. Li, Z. Liu, Y. Yao, Z. Cao, M. Zhang, and Y. Liu, "Wi-Fi see it all: Generative adversarial network-augmented versatile Wi-Fi imaging," in *Proc. 18th Conf. Embedded Networked Sensor Syst.*, 2020, pp. 436–448.

[10] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and G. L. Foresti, "Human silhouette and skeleton video synthesis through Wi-Fi signals," *Int. J. Neural Syst.*, vol. 32, no. 05, 2022, Art. no. 2250015.

[11] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-WiFi: Fine-grained person perception using WiFi," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5451–5460.

[12] S. Zhou, L. Guo, Z. Lu, X. Wen, W. Zheng, and Y. Wang, "Subject-independent human pose image construction with commodity Wi-Fi," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[13] L. Guo, Z. Lu, X. Wen, S. Zhou, and Z. Han, "From signal to image: Capturing fine-grained human poses with commodity Wi-Fi," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 802–806, 2019.

[14] Y. Gu, J. Chen, K. He, C. Wu, Z. Zhao, and R. Du, "WiFiLeaks: Exposing stationary human presence through a wall with commodity mobile devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 6997–7011, Jun. 2024.

[15] Z. He et al., "HCR-auth: Reliable bone conduction earphone authentication with head contact response," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 4, pp. 1–27, 2024.

[16] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[17] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 65–76.

[18] Y. Ma and G. Z. S. Wang, "WiFi sensing with channel state information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, pp. 46: 1–46: 36, 2019.

[19] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, "TW-See: Human activity recognition through the wall with commodity Wi-Fi devices," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 306–319, Jan. 2019.

[20] M. Gast, *802.11 Wireless Networks - The Definitive Guide: Creating and Administering Wireless Networks: Covers 802.11a, g, n and i*, 2nd ed. Sebastopol, CA, USA: O'Reilly, 2005.

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

[22] H. Zhu, F. Xiao, L. Sun, R. Wang, and P. Yang, "R-TTWD: Robust device-free through-the-wall detection of moving human with WiFi," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1090–1103, May 2017.

[23] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, “Non-invasive detection of moving and stationary human with WiFi,” *IEEE J. Sel. Areas Commun.*, vol. 33, no. 11, pp. 2329–2342, Nov. 2015.

[24] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, “Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi,” in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2017, pp. 6: 1–6: 10.

[25] Y. Gu, J. Chen, C. Wu, K. He, Z. Zhao, and R. Du, “Loccams: An efficient and robust approachfor detecting and localizing hidden wireless cameras via commodity devices,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 1–24, 2024.

[26] H. Zhang et al., “Understanding the mechanism of through-wall wireless sensing: A model-based perspective,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–28, 2023.

[27] C. Wu et al., “Rethinking membership inference attacks against transfer learning,” *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6441–6454, 2024.

[28] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, “EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices,” in *Proc. 2022 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2022, pp. 2931–2945.

[29] X. Ji, Y. Cheng, W. Xu, and X. Zhou, “User presence inference via encrypted traffic of wireless camera in smart homes,” *Secur. Commun. Netw.*, vol. 2018, pp. 1–10, 2018.

[30] Y. Zhou, H. Huang, S. Yuan, H. Zou, L. Xie, and J. Yang, “MetaFi: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation,” *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14: 128–14: 136, Aug. 2023.

[31] Y. Zhu et al., “Et tu alexa? when commodity wifi devices turn into adversarial motion sensors,” in *Proc. 27th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2020, pp. 1–15.

[32] Y.-C. Chen et al., “Seeing the unseen: WiFi-based 2D human pose estimation via an evolving attentive spatial-frequency network,” *Pattern Recognit. Lett.*, vol. 171, pp. 21–27, 2023.

[33] W. Sun, T. Chen, and N. Gong, “SoK: Inference attacks and defenses in human-centered wireless sensing,” 2022, *arXiv:2211.12087*.

[34] M. Cominelli, F. Gringoli, and R. L. Cigno, “AntiSense: Standard-compliant CSI obfuscation against unauthorized Wi-Fi sensing,” *Comput. Commun.*, vol. 185, pp. 92–103, 2022.

[35] P. Staat et al., “IRSHield: A countermeasure against adversarial physical-layer wireless sensing,” in *Proc. 2022 IEEE Symp. Secur. Privacy*, 2022, pp. 1705–1721.

[36] J. Luo, H. Cao, H. Jiang, Y. Yang, and Z. Chen, “mimoCrypt: Multi-user privacy-preserving Wi-Fi sensing via MIMO encryption,” in *Proc. 2024 IEEE Symp. Secur. Privacy*, 2024, pp. 2812–2830.

[37] X. Meng, J. Zhou, X. Liu, X. Tong, W. Qu, and J. Wang, “Secur-Fi: A secure wireless sensing system based on commercial Wi-Fi devices,” in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[38] Y. Qiao, O. Zhang, W. Zhou, K. Srinivasan, and A. Arora, “PhyCloak: Obfuscating sensing from communication signals,” in *Proc. 13th USENIX Symp. Netw. Syst. Des. Implementation*, 2016, pp. 685–699.

[39] L. Guo, J. Ma, and Y. Xu, “Passive indoor human tracking using commodity Wi-Fi,” in *Proc. Adv. Comput. Sci. Ubiquitous Comput.*, 2023, pp. 337–346.

[40] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, “Making the invisible visible: Action recognition through walls and occlusions,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 872–881.

[41] W. Ding, Z. Cao, J. Zhang, R. Chen, X. Guo, and G. Wang, “Radar-based 3D human skeleton estimation by kinematic constrained learning,” *IEEE Sensors J.*, vol. 21, no. 20, pp. 23: 174–23: 184, Oct. 2021.

[42] M. Zhao et al., “RF-based 3D skeletons,” in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2018, pp. 267–281.

[43] Z. Wu et al., “RFMask: A simple baseline for human silhouette segmentation with radio signals,” *IEEE Trans. Multimedia*, vol. 25, pp. 4730–4741, 2022.

[44] T. Li, L. Fan, Y. Yuan, and D. Katabi, “Unsupervised learning for human sensing using radio signals,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 3288–3297.

[45] C. Yu et al., “Mobirfpose: Portable RF-based 3D human pose camera,” *IEEE Trans. Multimedia*, vol. 26, pp. 3715–3727, 2024.

[46] J. Wang et al., “A unified framework for guiding generative AI with wireless perception in resource constrained mobile edge networks,” *IEEE Trans. Mobile Comput.*, pp. vol. 23, no. 11, pp. 10344–10360, Nov. 2024.