Wave2Body: Rethinking mmWave Human Pose Estimation as Radar-to-Body Token Translation

Bo Liang[1] Chen Gong[1] Wei Gao[2] Chenren Xu[1,3,*]

[1]School of Computer Science, Peking University [2]University of Pittsburgh

[3]Key Laboratory of High Confidence Software Technologies, Ministry of Education (PKU)

[*]Corresponding author: chenren@pku.edu.cn

Abstract

Millimeter-wave (mmWave) radar enables privacy-friendly human sensing, but its sparse point clouds are physical measurements of view-dependent electromagnetic reflections and only indirectly characterize body articulation. Recovering a complete 3D pose from such partial, geometry-dependent observations is therefore under-constrained. Existing methods directly regress joint coordinates from paired radar–pose data, relying on the same limited paired supervision to learn radar perception, human-body structure, and their alignment. This coupling can encourage dataset-specific shortcuts under ambiguous radar observations. We propose Wave2Body, a radar-to-body token translation framework that decouples these learning targets using a self-supervised mmWave tokenizer, a pretrained compositional body tokenizer that defines the output space, and a lightweight translator between them. Experiments on M4Human and mmBody show that Wave2Body achieves stronger cross-domain generalization than previous methods while incurring much lower computational costs for training and inference. All the code and experiment results are publicly available at https://github.com/Galaxywalk/Wave2Body.

1. Introduction

Millimeter-wave (mmWave) radar has emerged as a promising sensing modality for human motion understanding in mobile and ubiquitous systems. Unlike cameras, mmWave radar does not record visible appearance, can operate under poor illumination, and remains robust to partial visual occlusions. By reconstructing sparse 3D point clouds from radio reflections, recent mmWave-based human pose estimation (HPE) systems have enabled privacy-friendly 3D body perception for in-home monitoring, clinical rehabilitation, and human–robot interaction [19, 23, 24, 55]. Progress [2, 13, 45, 51] has been made through stronger point-cloud backbones, signal augmentation, diffusion-based pose estimation, and reliability-aware modeling for pose recovery.

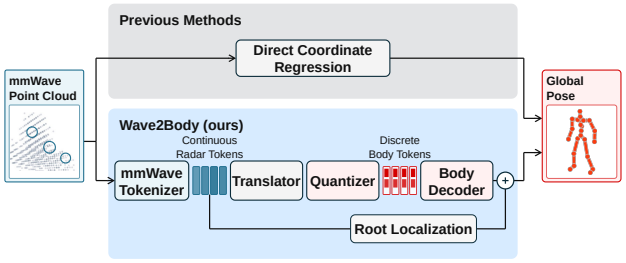

Figure 1. **Wave2Body reformulates mmWave HPE from coordinate regression to radar-to-body token translation.** A self-supervised mmWave tokenizer produces radar tokens, which are translated and quantized into discrete body tokens before decoding; the body encoder is used only to construct training targets.

Despite this progress, mmWave-based HPE differs from vision-based HPE. Radar returns depend on body geometry, orientation, motion, signal propagation, and sensor configuration; they do not provide dense observations of anatomical joints. Some body parts may produce weak or missing returns, while reflections can admit multiple plausible body configurations. Recovering a complete 3D skeleton is therefore an under-constrained inverse problem: the measured returns constrain only part of the pose, and the remaining structure must be inferred from human-body priors.

As illustrated by the upper path in Fig. 1, most existing methods [13, 14, 32, 40, 49] jointly learn radar representations and human-body priors within a single supervised radar-to-coordinate regressor. Even with stronger backbones or pose-aware objectives, they typically predict raw 3D joint coordinates. Consequently, the model must use the same limited paired supervision to encode sparse radar returns, capture valid joint configurations, and align radar with body structure. This coupling is problematic because paired mmWave datasets are costly and cover limited subjects, motions, environments, and sensing conditions.

Under ambiguous observations, a direct regressor may exploit training-specific sensing artifacts and pose frequencies rather than transferable radar representations and anatomical structure. Without reusable radar and body interfaces, radar representation learning, human-body structure modeling, and cross-modal alignment remain entangled within each paired training setup.

As illustrated by the lower path in Fig. 1, Wave2Body addresses this bottleneck by reformulating mmWave HPE as translation from a pretrained radar-token space to a pretrained body-token space. These spaces capture complementary sensor and body knowledge for representing sparse radar returns and valid human configurations. Independent pretraining creates reusable interfaces, allowing the paired training stage to focus on cross-modal mapping rather than learning both spaces from scratch.

Wave2Body realizes this formulation in stages. A self-supervised mmWave tokenizer learns from point clouds without pose supervision, while a compositional VQ-VAE body tokenizer learns from pose-only data. With both tokenizers frozen, a lightweight translator maps radar tokens to embeddings that are quantized and decoded into a pelvis-centered pose. Because body tokens exclude global position, a separate root localization head estimates the pelvis from radar observations. Only the translator and localization head use paired radar–pose supervision for cross-modal mapping and global localization, respectively.

We evaluate Wave2Body on M4Human [14] and mmBody [5] for radar-frame absolute-pose estimation using mean per-joint position error (MPJPE). Wave2Body ranks first across all reported M4Human Cross-Subject action groups, reducing micro-average MPJPE by 9.7%; it also reduces Cross-Action and mmBody MPJPE by 3.8% and 8.2%. These gains support our hypothesis that decoupling radar representation learning, body modeling, and cross-modal alignment improves generalization to unseen subjects and actions. Wave2Body uses $31.00\times$ fewer training and $89.81\times$ fewer inference FLOPs than the matched baseline.

Our contributions are summarized as follows:

- We reformulate mmWave HPE as translation between independently pretrained radar- and body-token spaces, replacing direct radar-to-coordinate regression with reusable interfaces on both sides.
- We instantiate this formulation with a self-supervised mmWave tokenizer pretrained on radar point clouds without pose labels and a compositional body tokenizer pretrained on pose-only data; paired radar–pose supervision is used only for translation and localization.
- We show that this decoupled learning pipeline improves radar-frame absolute-pose generalization, especially under cross-subject and cross-action shifts, while substantially reducing training and inference FLOPs.

2. Related Work

mmWave Human Pose Estimation. Early mmWave systems established the feasibility of recovering human skeletons or meshes [28, 40, 49], while datasets and benchmarks such as mRI, HuPR, mmBody, MM-Fi, RT-Pose, and M4Human enabled evaluation under more diverse sensing conditions [3, 5, 14, 22, 30, 53]. Subsequent work explored spatiotemporal learning, radar-tensor representations, multi-view fusion, mesh reconstruction, simulation and data scaling, diffusion models, mmWave point-cloud enhancement and generation, and cross-modal learning [2, 4, 6–9, 12, 13, 21, 22, 25–27, 42, 46–48, 50–52]. Recent studies have highlighted the importance of pose priors and distribution-shift evaluation for generalizable mmWave HPE [20, 38, 43, 45], coverage-aware data scaling [32], and self-supervised or synthetic-data-based wireless learning [17, 18, 57]. Most existing methods, however, still learn task-specific coordinate or heatmap predictors from paired radar–pose data. Wave2Body instead changes the prediction interface, translating radar tokens into a frozen body-token space learned from broader pose data.

Tokenized Representations and Interface Alignment. Masked point-cloud pretraining methods learn local tokens by reconstructing or discriminating masked neighborhoods [33, 37, 54], while VQ-VAE and tokenized human-pose models represent structured continuous signals through discrete, compositional latent spaces [11, 15, 16, 44]. In parallel, multimodal models show that pretrained modality-specific interfaces can be connected through lightweight projection or cross-attention modules [1, 31, 34]. Wave2Body combines these ideas for mmWave HPE: a self-supervised mmWave tokenizer represents local radar returns, a pretrained compositional body tokenizer defines an anatomical output space, and a lightweight radar-to-body translator aligns the two spaces.

3. Method

3.1. Overview

Given a mmWave point cloud $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where each point $\mathbf{x}_i \in \mathbb{R}^{3+C_r}$ contains a 3D coordinate and C_r radar attributes, our goal is to estimate an absolute 3D human pose $\mathbf{Y} \in \mathbb{R}^{J\times 3}$ in the radar coordinate frame, with $J = 22$ joints following the SMPL skeleton [35].

Wave2Body first learns two reusable token interfaces independently. The mmWave tokenizer E_{rad} maps the point cloud to continuous radar tokens $\mathbf{S}$, whereas the body tokenizer $Q_{\text{body}} \circ E_{\text{body}}$, implemented within a VQ-VAE, maps pelvis-centered poses to discrete compositional body tokens. Paired radar–pose data are then used to learn radar-to-body token translation and global radar-frame root local-

ization:

$$\mathbf{S} = E_{\text{rad}}(\mathbf{X}),\ \hat{\mathbf{Z}}_e = T_\theta(\mathbf{S}),\ \hat{\mathbf{B}} = D_{\text{body}}\Big(Q_{\text{body}}(\hat{\mathbf{Z}}_e)\Big), \tag{1}$$

where T_θ maps the radar tokens to pre-quantization body embeddings $\hat{\mathbf{Z}}_e$, and the frozen body quantizer and decoder $Q_{\text{body}}, D_{\text{body}}$ reconstruct a pelvis-centered body configuration $\hat{\mathbf{B}}$. The localization head predicts the pelvis location $\hat{\mathbf{r}}$, yielding the radar-frame pose $\hat{\mathbf{Y}} = \hat{\mathbf{B}} + \mathbf{1}_J \hat{\mathbf{r}}^\top$.

As shown in Fig. 2, the first two stages learn the source and target token interfaces independently. Stage 1 pretrains the mmWave tokenizer without pose labels, and Stage 2 pretrains the body tokenizer from pose-only data without using radar inputs. Stage 3 freezes the pretrained radar and body modules and learns cross-modal translation and global localization from paired radar–pose data. Stage 3 is optimized in two phases: pelvis-centered token translation is learned first, followed by joint refinement of the translator and localization head under absolute-pose supervision.

3.2. Self-Supervised mmWave Tokenizer

The mmWave tokenizer E_{rad} is a Point-MAE-style patch encoder [37] that converts a sparse, unordered point cloud into a fixed-length sequence of continuous radar tokens. We sample G valid patch centers using farthest-point sampling and gather the K nearest support points around each center to form local patches $\{\mathbf{P}_g\}_{g=1}^{G}$. Within each patch, spatial coordinates are expressed relative to its center $\mathbf{c}_g$, while radar attributes such as reflection intensity or velocity are retained. Each patch is embedded, augmented with a positional encoding of $\mathbf{c}_g$, and contextualized by a Transformer. The resulting sequence $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_G] \in \mathbb{R}^{G \times d_s}$ combines centered local structure with radar-frame location.

To learn the radar-token space without pose labels, we pretrain E_{rad} through masked patch reconstruction. A random subset $\mathcal{M}$ of valid patches is masked, and a lightweight reconstruction decoder predicts both their local geometry and radar attributes:

$$\mathcal{L}_{\text{rad}} = d_{\text{Chamfer}}\Big(\hat{\mathbf{P}}_{\mathcal{M}}^{\text{xyz}}, \mathbf{P}_{\mathcal{M}}^{\text{xyz}}\Big) + \gamma \Big\|\hat{\mathbf{P}}_{\mathcal{M}}^{\text{attr}} - \mathbf{P}_{\mathcal{M}}^{\text{attr}}\Big\|_2^2. \tag{2}$$

The Chamfer term measures the bidirectional distance between reconstructed and original masked point sets, while attributes are compared index-wise in KNN-distance order. After pretraining, the decoder is discarded and E_{rad} is frozen. All G patches are then encoded to produce $\mathbf{S}$ as the source sequence for radar-to-body translation.

3.3. Compositional Body Tokenizer

The body tokenizer defines a discrete target space for human articulation independently of global translation. We implement it as a part-structured VQ-VAE [44] comprising a body encoder E_{body}, quantizer Q_{body}, and decoder

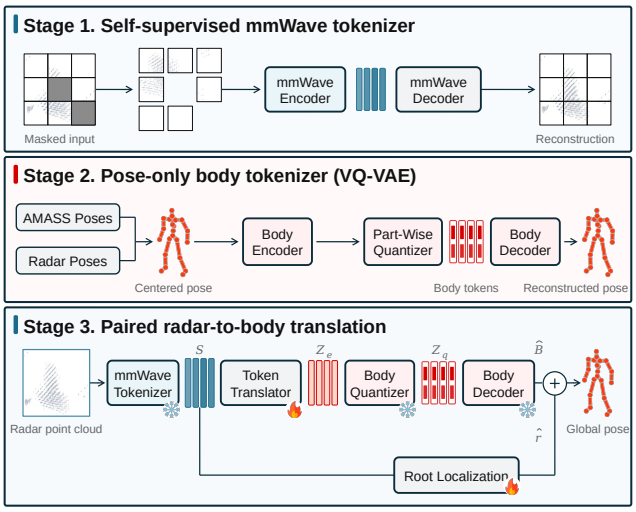

Figure 2. **Three-Stage Training of Wave2Body.** Stage 1 pretrains a continuous mmWave tokenizer through masked radar-patch reconstruction. Stage 2 pretrains a discrete compositional body tokenizer, implemented as a VQ-VAE, using pose-only data from motion-capture and radar datasets. Stage 3 freezes the pretrained radar and body modules and trains the radar-to-body translator and localization head using paired radar–pose data.

D_{body}. Whereas the mmWave tokenizer produces continuous radar tokens, the body tokenizer $Q_{\text{body}} \circ E_{\text{body}}$ assigns pose embeddings to compositional codebook entries. Because paired mmWave datasets typically contain limited pose diversity, we train the body tokenizer using large-scale motion-capture poses together with pose annotations from the corresponding radar benchmark, without using their paired radar inputs. All poses are converted to the same 22-joint topology and centered at the pelvis:

$$\mathbf{B} = \mathbf{Y} - \mathbf{1}_J \mathbf{r}^\top, \qquad \mathbf{B} \in \mathbb{R}^{J \times 3}, \tag{3}$$

where $\mathbf{r} \in \mathbb{R}^3$ is the pelvis coordinate.

We train the body-token VQ-VAE to define a structured prediction space rather than to serve as a motion generator. To make the tokens compositional, we divide the skeleton into five kinematic regions: left leg (left hip/knee/ankle/foot), right leg (right hip/knee/ankle/foot), spine core (spine1–3, neck, and head), left arm (spine3 and left collar/shoulder/elbow/wrist), and right arm (spine3 and right collar/shoulder/elbow/wrist). Each region has a part-specific encoder, codebook, and decoder, with spine3 shared by the torso and arm regions. Collectively, the encoders produce $T = 24$ pre-quantization embeddings $\mathbf{Z}_e = E_{\text{body}}(\mathbf{B})$, allocated as 5, 5, 4, 5, and 5 slots across the five regions. For embedding $\mathbf{z}_{p,t}$ at slot t of region p, quantization selects the nearest entry in its codebook $\mathcal{C}_p$:

$$Q_p(\mathbf{z}_{p,t}) = \mathbf{e}^*,\ \mathbf{e}^* = \arg\min_{\mathbf{e} \in \mathcal{C}_p} \|\mathbf{z}_{p,t} - \mathbf{e}\|_2^2. \tag{4}$$

The selected codebook entry $\mathbf{e}^*$ is a quantized body to-

ken, and its codebook index is the corresponding discrete token ID. The quantized tokens from all regions are concatenated as $\mathbf{Z}_q = Q_{\text{body}}(\mathbf{Z}_e)$. Part-specific decoders reconstruct their corresponding joint subsets; predictions for shared joints are averaged, and the pelvis remains fixed at the origin. The VQ-VAE is optimized with coordinate reconstruction and vector-quantization objectives:

$$\mathcal{L}_{\text{body}} = \ell_{\text{rec}}(D_{\text{body}}(Q_{\text{body}}(E_{\text{body}}(\mathbf{B}))), \mathbf{B}) + \ell_{\text{vq}}. \quad (5)$$

After pretraining, E_{body}, Q_{body}, and D_{body} are frozen. During translator training, the body encoder supplies target pre-quantization embeddings, while the quantizer and decoder map the translator outputs through the discrete body-token space to a reconstructed pose. The radar model predicts through a body representation learned from broad pose data rather than regressing joint coordinates.

3.4. Radar-to-Body Token Translation and Localization

Given the radar-token sequence $\mathbf{S}$, the query-based Transformer T_θ predicts pre-quantization body embeddings $\hat{\mathbf{Z}}_e \in \mathbb{R}^{T \times d_b}$. It contains T learnable body queries, one for each body-token slot. Each query cross-attends to the radar tokens, after which self-attention models dependencies among body parts. The predicted embeddings are quantized as $\hat{\mathbf{Z}}_q = Q_{\text{body}}(\hat{\mathbf{Z}}_e)$ and decoded according to Eq. 1. The body queries therefore act as anatomical output slots, allowing weakly observed limbs to be inferred jointly from the radar-token sequence and the learned body structure.

Let $\mathbf{Z}_e = E_{\text{body}}(\mathbf{B})$ denote the target pre-quantization body embeddings. The translator is trained with an embedding-alignment loss and a reconstruction loss on the decoded pelvis-centered pose:

$$\mathcal{L}_{\text{trans}} = \lambda_z \left\| \hat{\mathbf{Z}}_e - \text{sg}(\mathbf{Z}_e) \right\|_2^2 + \lambda_b \left\| \hat{\mathbf{B}} - \mathbf{B} \right\|_F^2, \quad (6)$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Gradients through the quantizer are estimated with the straight-through estimator.

Because the body-token space excludes global translation, the lightweight query-attention localization head separately predicts the pelvis location. Its inputs are the radar tokens $\{\mathbf{s}_g\}_{g=1}^G$, their patch centers $\{\mathbf{c}_g\}_{g=1}^G$, and compact valid-point statistics $\rho(\mathbf{X})$. In our implementation, $\rho(\mathbf{X})$ contains global xyz mean, standard deviation, bounding-box extrema, intensity mean, intensity standard deviation, intensity extrema, and the intensity-magnitude-weighted xyz mean. We use the latter as the root prior $\mathbf{r}_0$. Each radar token is augmented with its patch center and its offset from this prior:

$$\mathbf{u}_g = [\mathbf{s}_g, \mathbf{c}_g, \mathbf{c}_g - \mathbf{r}_0]. \quad (7)$$

Let $\mathbf{U} = \{\mathbf{u}_g\}_{g=1}^G$ and $\mathbf{C} = \{\mathbf{c}_g\}_{g=1}^G$. Learnable root queries $\mathbf{Q}_r$ attend to $\mathbf{U}$, and their pooled output is combined with $\rho(\mathbf{X})$ and summaries $\sigma(\mathbf{S}, \mathbf{C})$, comprising mean- and max-pooled radar tokens and the mean, minimum, and maximum patch centers, to predict a residual:

$$\begin{aligned} \mathbf{h}_r &= \text{Pool}(\text{Attn}(\mathbf{Q}_r, \mathbf{U})), \\ \Delta\hat{\mathbf{r}} &= R_\phi([\mathbf{h}_r, \sigma(\mathbf{S}, \mathbf{C}), \rho(\mathbf{X})]), \\ \hat{\mathbf{r}} &= \mathbf{r}_0 + \Delta\hat{\mathbf{r}}. \end{aligned} \quad (8)$$

This design keeps radar-frame localization tied to absolute sensor geometry, while the body-token path predicts pelvis-centered articulation. During the final optimization phase, E_{rad}, E_{body}, Q_{body}, and D_{body} remain frozen, while T_θ and R_ϕ are jointly optimized with

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{trans}} + \lambda_r \left\| \hat{\mathbf{r}} - \mathbf{r} \right\|_2^2 + \lambda_y \left\| \hat{\mathbf{Y}} - \mathbf{Y} \right\|_F^2. \quad (9)$$

Here, $\mathcal{L}_{\text{trans}}$ provides pre-quantization embedding and pelvis-centered pose supervision, while the remaining terms supervise root localization and radar-frame absolute pose.

4. Experiments

4.1. Experimental Setup

Benchmarks and Protocols. We evaluate Wave2Body on M4Human [14] and mmBody [5]. M4Human is used for generalization evaluation under its official Cross-Subject and Cross-Action protocols, which hold out identities and action classes, respectively. We omit the Random split because it shares both subjects and action classes between training and test data and mainly reflects in-distribution interpolation. Following the benchmark protocol, we report In-Place, Sit-In-Place, Non-In-Place, and a sample-weighted micro-average over all actions. On mmBody, we use the official train/test partition to assess whether the gains persist on a second benchmark, training a model for each benchmark and split.

Input Preprocessing. For Wave2Body, M4Human inputs contain $N = 1024$ points with xyz coordinates and reflection intensity, causally aggregated from the current and previous three frames; mmBody inputs contain $N = 200$ points with xyz, intensity, and velocity over three frames. The reimplemented baselines operate on the same official examples, splits, radar point-cloud modality, coordinate frame, and pose targets.

Pose Representation and Evaluation. All annotations are mapped to a common SMPL-style 22-joint skeleton and evaluated in the radar coordinate frame. The primary metric is absolute MPJPE in millimeters, which measures both body articulation and global pelvis localization. We additionally report root error, root-aligned MPJPE, and

Table 1. **Radar-frame absolute-pose estimation across benchmarks.** Absolute MPJPE (mm; lower is better) on M4Human and mmBody. For M4Human, Micro Average is sample-weighted, and Cross-Action groups contain only held-out test actions. Each learned entry is reported as the mean ± standard deviation over three seeds. Bold and underline indicate the best and second-best results.

Dataset / Split	Action Group	**Wave2Body (ours)** ↓	WiCompass [32]* ↓	RT-Mesh [14] ↓	mmMesh [49]* ↓	mmDiff [13]* ↓
M4Human Cross-Subject	In-Place	**90.62** ±0.74 (**-10.1%**)	<u>100.83 ±0.73</u>	109.60	128.96 ±2.71	131.58 ±5.33
	Sit-In-Place	**99.30** ±2.54 (**-15.4%**)	<u>117.44 ±2.53</u>	130.80	162.65 ±2.67	158.11 ±13.50
	Non-In-Place	**127.62** ±0.64 (**-7.5%**)	<u>137.91 ±4.20</u>	151.80	175.33 ±2.57	182.60 ±8.13
	Micro Average	**102.53** ±0.29 (**-9.7%**)	<u>113.53 ±1.91</u>	120.20	146.10 ±2.46	149.42 ±5.32
M4Human Cross-Action	In-Place	<u>97.30 ±0.66 (+3.6%)</u>	**93.92** ±2.18	107.20	129.32 ±1.09	138.71 ±3.70
	Sit-In-Place	**118.29** ±4.30 (**-1.4%**)	125.77 ±1.69	<u>120.00</u>	122.51 ±1.68	151.82 ±15.10
	Non-In-Place	**137.26** ±0.75 (**-9.2%**)	<u>151.16 ±8.71</u>	152.60	183.84 ±1.64	193.12 ±3.19
	Micro Average	**115.26** ±0.98 (**-3.8%**)	<u>119.81 ±4.36</u>	122.00	150.34 ±1.03	161.64 ±4.55
mmBody	All	**128.17** ±1.87 (**-8.2%**)	<u>139.69 ±4.43</u>	–	145.79 ±2.91	267.37 ±19.36

*Re-evaluated under our radar-frame absolute-pose protocol; original metrics may use different alignments. RT-Mesh uses official M4Human results. Parentheses show relative change from the best competitor; negative values indicate improvement.

Table 2. **M4Human absolute-pose diagnostics.** Errors (mm; lower is better) for Wave2Body and the matched WiCompass baseline under Cross-Subject and Cross-Action shifts. Each entry is reported as the mean ± standard deviation over three seeds; bold indicates the better result.

		Absolute MPJPE ↓		Root error ↓		Root-aligned MPJPE ↓		PA-MPJPE ↓	
Split	Action Group	Wave2Body	WiCompass	Wave2Body	WiCompass	Wave2Body	WiCompass	Wave2Body	WiCompass
Cross-Subject	In-Place	**90.62** ±0.74	100.83 ±0.73	**67.76** ±0.80	78.31 ±0.89	**66.80** ±0.20	72.65 ±1.59	**50.83** ±0.07	53.42 ±0.35
	Sit-In-Place	**99.30** ±2.54	117.44 ±2.53	**85.16** ±3.41	98.94 ±4.30	**64.14** ±0.20	77.57 ±0.89	**48.69** ±0.17	56.51 ±0.89
	Non-In-Place	**127.62** ±0.64	137.91 ±4.20	**88.20** ±0.60	105.57 ±4.46	**95.40** ±0.31	95.54 ±1.54	**68.36** ±0.03	68.44 ±0.49
Cross-Action	In-Place	97.30 ±0.66	**93.92** ±2.18	62.23 ±0.48	**60.31** ±0.77	75.64 ±0.46	**74.59** ±2.02	67.64 ±0.46	**63.83** ±1.42
	Sit-In-Place	**118.29** ±4.30	125.77 ±1.69	**84.41** ±6.58	102.62 ±11.73	89.93 ±2.75	**81.14** ±1.89	73.09 ±0.72	**67.33** ±2.36
	Non-In-Place	**137.26** ±0.75	151.16 ±8.71	**91.10** ±0.79	108.88 ±7.69	**100.84** ±0.41	102.06 ±4.17	**78.68** ±0.09	78.93 ±2.12

Procrustes-aligned MPJPE (PA-MPJPE). Unless noted otherwise, learned entries report the mean and sample standard deviation over seeds 42, 43, and 44; for each seed, the validation-selected checkpoint is evaluated on the test split.

Wave2Body Implementation. Wave2Body follows the three-stage training pipeline described in Sec. 3. The mmWave tokenizer uses $G = 96$ patch centers selected by farthest-point sampling and $K = 32$ nearest neighbors per patch, and outputs 128-D radar tokens; it is pretrained for 80 epochs with 60% patch masking. The part-level body tokenizer is trained for 30 epochs on pelvis-centered AMASS [36] poses and the corresponding benchmark training poses, using 24 latent slots, five kinematic regions (left leg, right leg, spine core, left arm, and right arm), 256-D embeddings, and 96-entry codebooks. In Stage 3, we first train the radar-to-body translator for 25 epochs with all tokenizer modules frozen, then add the localization head and jointly train it with the translator for another 25 epochs. The translator uses 24 body queries, four cross-attention blocks, and two self-attention blocks. All stages use AdamW and bf16 mixed precision. No component is trained on test samples; checkpoints are selected using the official M4Human validation split or held-out training data.

Baselines. We compare with WiCompass [32], mmMesh [49], and mmDiff [13] under a matched point-cloud protocol. These baselines use the same paired radar–pose splits, validation and test protocol, radar-frame 22-joint targets, and evaluation code as Wave2Body, but retain architecture-specific input preprocessing. They directly predict absolute joint coordinates. We also report RT-Mesh [14], the official M4Human 4D radar-tensor baseline, but exclude it on mmBody because mmBody provides only point clouds. We do not mix root-relative results from prior reports with our absolute-coordinate evaluation. The only additional supervision used by Wave2Body is AMASS pose-only data for body-tokenizer pretraining.

4.2. Main Results

Pose Accuracy. Table 1 reports absolute MPJPE by split and action group. Wave2Body achieves the best overall result in seven of eight M4Human cross-domain rows: all Cross-Subject groups and the Cross-Action Sit-In-Place, Non-In-Place, and micro-average rows. It reduces the Cross-Subject and Cross-Action micro-averages by 9.7% and 3.8%, respectively, and mmBody MPJPE by 8.2%. Cross-Action results are mixed: it lowers Non-In-Place error by 9.2% but is 3.6% worse than WiCompass on In-Place

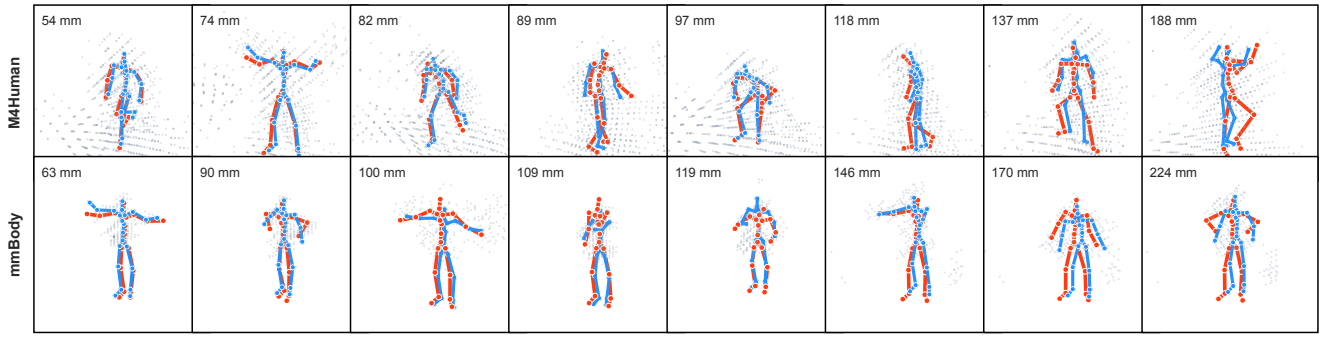

Figure 3. **Qualitative results from easy to challenging cases.** Representative M4Human and mmBody test samples are selected across the absolute-MPJPE spectrum. Gray points denote the input mmWave point cloud, red skeletons denote ground truth, and blue skeletons denote Wave2Body predictions. The value in each panel is the sample-level radar-frame absolute MPJPE.

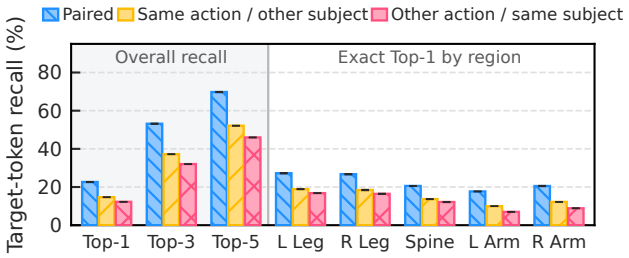

Figure 4. **Body-token accuracy.** M4Human Cross-Subject token-ID accuracy. Error bars show standard deviations across three seeds.

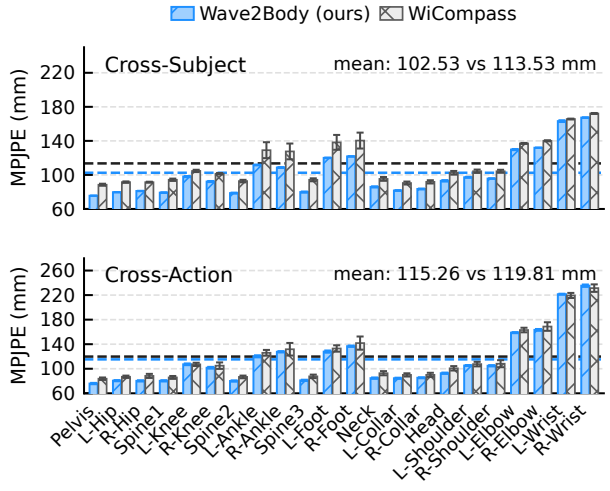

Figure 5. **M4Human per-joint error.** Error bars show three-seed standard deviations; dashes mark the mean.

actions; its 3.8% micro-average gain also trails its 9.7% Cross-Subject gain. Benefits thus concentrate on subject shift and dynamic held-out motions, suggesting a trade-off between a discrete body prior's robustness and direct coordinate regression's fine-grained precision (Sec. 5).

Error Decomposition. Table 2 decomposes absolute-pose error into global localization and body-configuration components. Under Cross-Subject, Wave2Body improves all four metrics in every action group, so its absolute-pose gains extend across both localization and articulation. Under Cross-Action, the pattern is more localized: Wave2Body reduces absolute and root errors for Sit-In-Place and Non-In-Place, while the root-aligned and Procrustes-aligned results are mixed. Only Non-In-Place improves across all four metrics. Thus, the cross-action gains arise primarily from stronger root localization, whereas the articulation benefit is clearest under subject shift and dynamic held-out actions.

Modality Alignment Analysis. We test whether the translator maps mmWave representations into the body-token space [10]. Low end-to-end MPJPE along the radar-to-token-to-pose path shows accurate recovery but cannot rule out frequent-pose shortcuts. We therefore evaluate the translator-only checkpoint before localization. On each M4Human test split, Top-k recall [29] asks whether the paired ground-truth token at each of 24 slots is among the k pose codes nearest the predicted embedding; Top-1 is exact token agreement. To test sample specificity, we use two deterministic one-to-one controls: same action/other subject and other action/same subject. Across three seeds on Cross-Subject, paired Top-1/3/5 recall is 22.63/53.16/69.77%, exceeding the 14.68/37.20/52.12% and 12.25/32.02/45.98% controls (Fig. 4); the advantage also holds in all five regions, supporting sample-specific recovery. On Cross-Action, the paired and control Top-1 recalls are 21.60%, 12.49%, and 10.80%, respectively. We then test whether token alignment tracks articulation error. For each sample, we correlate the Top-k miss rate (fraction of 24 targets absent from Top-k) with root-aligned MPJPE using Spearman's rank correlation ρ [41]. On Cross-Subject, Top-1/3/5 mean correlations are $\rho = 0.595$, $\rho = 0.742$, and $\rho = 0.774$; Cross-Action Top-1 is $\rho = 0.615$. Under Schober et al.'s scale [39],

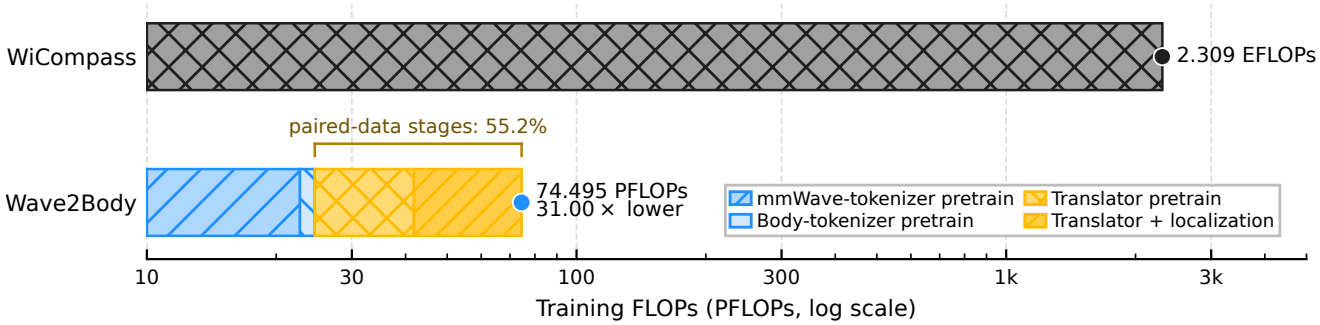

Figure 6. **Cold-start training computation on M4Human.** Wave2Body is decomposed into two tokenizer-pretraining phases and two paired-data phases. FLOPs include supported profiler operations in forward, loss, backward, and optimizer computation.

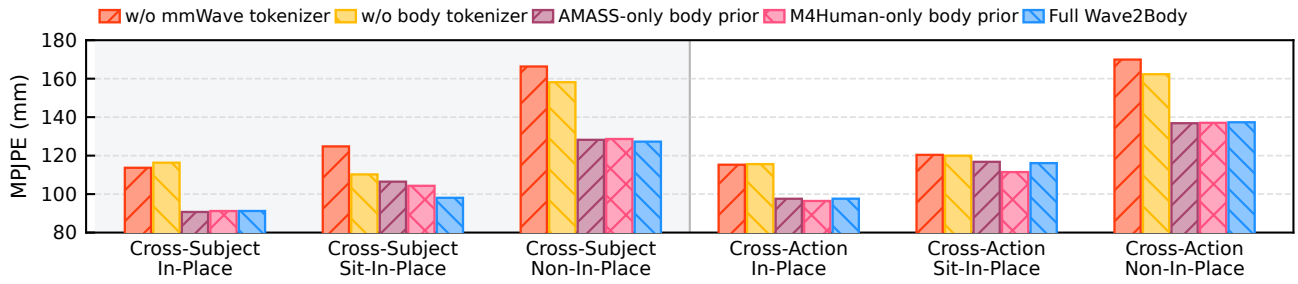

Figure 7. **Component ablations studies.** Lower MPJPE is better.

the Top-1 correlation is moderate, while the Top-3/5 correlations are strong; token-alignment mismatch correlates with predicted-pose error. Together, low end-to-end error, stronger paired-than-control recovery, and this sample-level error link support alignment beyond pose-frequency, action, or identity shortcuts.

Qualitative Results. Figure 3 complements aggregate results with representative M4Human and mmBody predictions selected from low- to high-error buckets. At low and moderate error levels, predictions closely follow ground-truth torso and limb configurations despite sparse, partial radar coverage. In the hardest cases, Wave2Body still tends to produce human-like skeletons, but errors manifest as global root shifts or ambiguous distal-limb articulation. Such failures are not caused only by sparse points: dynamic motion, unusual poses, and difficult scenarios can make short-window radar observations compatible with multiple plausible poses. This supports the decomposition in Table 2: decoupling improves average articulation and localization under shift, but sparse, cluttered, or ambiguous returns can still yield plausible but incorrect body tokens.

Per-Joint Error Analysis. We analyze absolute-pose error across body joints on M4Human against WiCompass, the strongest matched baseline (Table 1). Dashed lines in Fig. 5 show the unweighted joint mean. Under Cross-Subject, Wave2Body improves every joint and lowers mean error by 9.7%, with the largest gains at ankles, feet, spine, pelvis, and hips, indicating better global placement and lower-body localization under subject shift. Under Cross-Action, it lowers the mean by 3.8% and improves 19 of 22 joints, most clearly at the pelvis, hips, spine, neck, and collar. In both splits, wrists, elbows, ankles, and feet have the highest errors, consistent with distal joints being less constrained by mmWave reflections than the pelvis and torso.

Table 3. **Inference efficiency on M4Human.** Measurements are CUDA-synchronized and exclude data loading and transfer.

Method	Params	GFLOPs/sample	Throughput (samples/s, B=192)	Peak CUDA Mem. (MiB, B=192)
WiCompass	26.91M	62.381	536.6	7556.4
Wave2Body	16.93M	0.6946	4328.5	558.9

4.3. Efficiency Analysis

Profiling Protocol. We profile Wave2Body and WiCompass, the strongest matched baseline, on the M4Human Cross-Subject full-action setting using a single NVIDIA GeForce RTX 4090 with bf16 mixed precision, TF32, PyTorch eager execution, and `torch.compile` disabled. For each training phase, we average supported-operation FLOPs over 128 batches, covering forward propagation, loss computation, backward propagation, and optimizer updates, and scale the per-sample average by the training samples per epoch and configured epochs. Training batch sizes are 192 for Wave2Body and 48 for WiCompass due to GPU memory limits. Data loading, validation, and checkpointing

are excluded.

Training Computation. Figure 6 compares total training computation and decomposes Wave2Body by phase. Under the same accounting convention, Wave2Body uses $31.00\times$ fewer cold-start training FLOPs than WiCompass. Its two paired-data stages account for 55.2% of the cold-start computation, while the remaining 44.8% is spent on pretraining the two tokenizers.

Inference Efficiency. Table 3 reports inference-time parameters and supported-operation FLOPs together with CUDA-synchronized throughput and peak allocated GPU memory. Compared with WiCompass, Wave2Body uses 37% fewer parameters and $89.81\times$ fewer forward FLOPs; at batch size 192, it achieves $8.07\times$ higher throughput and reduces peak memory by 92.6%. These results demonstrate substantially lower inference computation and resource use for high-throughput deployment scenarios.

4.4. Ablation Studies

We ablate three design factors behind Wave2Body: the mmWave tokenizer, the body tokenizer, and the pose data used to train the body tokenizer. Figure 7 compares the full model with four single-seed variants on M4Human Cross-Subject and Cross-Action, which evaluate held-out identities and held-out action classes, respectively. All tokenizer pretraining remains restricted to the training split.

Dual-Tokenizer Ablation. The variant without the body tokenizer retains the frozen mmWave tokenizer and localization head, but replaces body-token translation and decoding with a pooled regressor for pelvis-centered joint coordinates. It increases MPJPE in every action group by 3.3–27.5%. Conversely, the variant without the mmWave tokenizer replaces masked-pretrained radar patches with a supervised raw-point encoder while retaining the body-token target and localization head; it increases MPJPE by 3.7–30.7%, with the largest penalty on Cross-Subject Non-In-Place. The comparable degradation from either removal supports the dual-interface formulation: the mmWave tokenizer contributes transferable sensor representations, while the body tokenizer supplies an anatomical target space.

Body-Prior Training Data. We further train the body tokenizer using only M4Human poses, only AMASS poses, or their combination. All three variants remain stronger than removing either tokenizer. The mixed AMASS+M4Human prior performs best on the Cross-Subject micro-average and several Cross-Subject groups, whereas the M4Human-only and AMASS-only variants are each slightly better on individual Cross-Action groups.

5. Discussion

Wave2Body does not eliminate the sensing gap between mmWave radar and vision-, depth-, or LiDAR-based sensors: radar point clouds remain sparse, specular, and viewpoint-dependent, so many body parts may be weakly observed or missing. Instead, it handles this ambiguity across three components: the mmWave tokenizer learns sensor-side regularities from unlabeled radar, the body-token space provides a structured target for articulation from pose-only data, and the localization head recovers global position. This decomposition allows each data source to contribute without requiring all of this knowledge to be learned from paired radar–pose samples [32, 38]. The results, however, show that its benefits are not uniform. Wave2Body improves consistently under Cross-Subject shift, whereas its Cross-Action gains are smaller, arise primarily from root localization, and do not extend to In-Place actions. This suggests a potential trade-off from the discrete body-token space: it constrains ambiguous predictions to plausible body configurations, but quantization may limit fine-grained accuracy when radar observations already provide sufficient pose cues. This may explain why Wave2Body is strongest under subject shift and dynamic held-out motions but weaker on Cross-Action In-Place.

Limitations. Wave2Body still depends on the quality of radar observations. When point clouds are extremely sparse, biased, or cluttered, the translated embeddings may be quantized into plausible but incorrect body tokens, and the localization head may fail when global localization cues are weak. Future work can improve robustness through explicit sequence modeling beyond fixed short-window aggregation, together with motion smoothness and kinematic consistency constraints, so that weakly observed body parts and root trajectories can be inferred over longer sequences. Another important direction is multi-person mmWave HPE [56], where inter-person occlusion and identity association make both body-token selection and radar-frame localization more challenging.

6. Conclusion

Wave2Body reformulates mmWave human pose estimation as radar-to-body token translation, connecting continuous tokens from a self-supervised mmWave tokenizer to the discrete space of a pretrained compositional body tokenizer. This decoupled learning pipeline gives each data source a clear role: unlabeled radar data for sensor representation learning, pose-only motion data for human-body priors, and paired radar–pose samples for cross-modal alignment. This allocation improves radar-frame pose generalization and suggests that structured prediction spaces can bridge radar observations and transferable human knowledge.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Sizhe An and Umit Y Ogras. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pages 889–894, 2022. 1, 2

[3] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in neural information processing systems*, 35:27414–27426, 2022. 2

[4] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. *arXiv preprint arXiv:2210.01346*, 2022. 2

[5] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022. 2, 4

[6] Lin Chen and Guoli Wang. Cpformer: End-to-end multi-person human pose estimation from raw radar cubes with transformers. *IEEE Sensors Journal*, 2025. 2

[7] Lin Chen and Guoli Wang. Dilated point spatio-temporal mesh transformer for mmwave radar-based human mesh reconstruction. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025.

[8] Lin Chen, Cong Li, Shuxin Zhong, Jun Chen, Yufei Wen, Haotian Song, and Kaishun Wu. Sensing life in stillness: Unified dynamic and static human mesh reconstruction with mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(1):1–25, 2026.

[9] Jaeho Choi, Soheil Hor, Shubo Yang, and Amin Arbabian. Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27750–27759, 2025. 2

[10] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022. 6

[11] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1323–1333, 2024. 2

[12] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021. 2

[13] Junqiao Fan, Jianfei Yang, Yuecong Xu, and Lihua Xie. Diffusion model is a good pose estimator from 3d rf-vision. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 2, 5

[14] Junqiao Fan, Yunjiao Zhou, Yizhuo Yang, Xinyuan Cui, Jiarui Zhang, Lihua Xie, Jianfei Yang, Chris Xiaoxuan Lu, and Fangqiang Ding. M4human: A large-scale multimodal mmwave radar benchmark for human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 42836–42846, 2026. 1, 2, 4, 5

[15] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. Vq-hps: Human pose and shape estimation in a vector-quantized latent space. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024. 2

[16] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023. 2

[17] Chen Gong, Bo Liang, Wei Gao, and Chenren Xu. Data can speak for itself: Quality-guided utilization of wireless synthetic data. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, pages 209–222, 2025. 2

[18] Chen Gong, Bo Liang, Qihao Zhu, Wei Gao, Yin Chen, Jin Nakazawa, and Chenren Xu. Towards generalizable wireless sensing models via pre-training on multi-source datasets. In *Proceedings of the 2026 ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems*, pages 488–502, 2026. 2

[19] Zhanzhong Gu, Xiangjian He, Gengfa Fang, Chengpei Xu, Feng Xia, and Wenjing Jia. Millimeter wave radar-based human activity recognition for healthcare monitoring robot. *arXiv preprint arXiv:2405.01882*, 2024. 1

[20] Rongxiao Guo and Qingchao Chen. Dghmesh: A large-scale dual-radar mmwave dataset and generalization-focused benchmark for human mesh reconstruction. *arXiv preprint arXiv:2604.22827*, 2026. 2

[21] Hoang Hai Pham, Shuntian Zheng, Jiaqi Li, and Yu Guan. A two-stage motion-aware framework for mmwave-based human mesh recovery. *arXiv e-prints*, pages arXiv–2605, 2026. 2

[22] Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, Wenhao Chai, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark. In *European Conference on Computer Vision*, pages 107–125. Springer, 2024. 2

[23] Shuting Hu, Peggy Ackun, Xiang Zhang, Siyang Cao, Jennifer Barton, Melvin G Hector, Mindy J Fain, and Nima Toosizadeh. mmwave radar for sit-to-stand analysis: A comparative study with wearables and kinect. *IEEE Transactions on Biomedical Engineering*, 72(9):2623–2634, 2025. 1

[24] Yuxuan Hu, Kuangji Zuo, Boyu Ma, Shihao Li, Zhaoyang Xia, Feng Xu, and Jianfei Yang. Waveman: mmwave-

based room-scale human interaction perception for humanoid robots. *arXiv preprint arXiv:2601.07454*, 2026. 1

[25] Niraj Prakash Kini, Ruey-Horng Shiue, Ryan Chandra, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Transhupr: Cross-view fusion transformer for human pose estimation using mmwave radar. In *BMVC*, 2024. 2

[26] Niraj Prakash Kini, Shiau-Rung Tsai, Guan-Hsun Lin, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. millimamba: Specular-aware human pose estimation via dual mmwave radar with multi-frame mamba fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1481–1490, 2026.

[27] Ailing Kong, Hengyu Gu, Xueya Yan, and Jian Guo. Twrpose3d: Through-wall radar based 3d human pose estimation with dual-stage temporal and view fusion. In *2026 38th Chinese Control and Decision Conference (CCDC)*, pages 4617–4622. IEEE, 2026. 2

[28] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 491–503, 2022. 2

[29] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 6

[30] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023. 2

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[32] Bo Liang, Chen Gong, Haobo Wang, Qirui Liu, Rungui Zhou, Fengzhi Shao, Yubo Wang, Wei Gao, Kaichen Zhou, Guolong Cui, et al. Wicompass: Oracle-driven data scaling for mmwave human pose estimation. *arXiv preprint arXiv:2602.18726*, 2026. 1, 2, 5, 8

[33] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5

[37] Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1:2440001, 2023. 2, 3

[38] Zhuoxuan Peng, Boan Zhu, Xingjian Zhang, Wenying Li, and S-H Gary Chan. Expanding mmwave datasets for human pose estimation with unlabeled data and lidar datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21221–21230, 2026. 2, 8

[39] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018. 6

[40] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE sensors journal*, 20 (17):10032–10044, 2020. 1, 2

[41] Charles Spearman. The proof and measurement of association between two things. 1961. 6

[42] Chang Su, Beihong Jin, Qiwen Shi, and Zhi Wang. mmwaveflow: Unified enhancement and generation of mmwave human point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31366–31376, 2026. 2

[43] Yuanzhi Su, Huiying Cynthia Hou, and Chun Zhao. Posegraphnet: Pose prior and graph structure for 3d human pose estimation using mmwave radar. *Measurement*, page 118851, 2025. 2

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[45] Zhenyu Wang, Mahathir Monjur, and Shahriar Nirjon. mmjoints: Expanding joint representations beyond (x, y, z) in mmwave-based 3d pose estimation. *arXiv preprint arXiv:2510.08970*, 2025. 1, 2

[46] Xijia Wei, Yuan Fang, Kevin Chetty, Youngjun Cho, and Nadia Bianchi-Berthouze. Maepose: Self-supervised spatiotemporal learning for human pose estimation on mmwave video. *arXiv preprint arXiv:2605.00242*, 2026. 2

[47] Yingxiao Wu, Zhongmin Jiang, Haocheng Ni, Changlin Mao, Zhiyuan Zhou, Wenxiang Wang, and Jianping Han. mmhpe: Robust multiscale 3-d human pose estimation using a single mmwave radar. *IEEE Internet of Things Journal*, 12 (1):1032–1046, 2024.

[48] Qian Xie, Qianyi Deng, Ta Ying Cheng, Peijun Zhao, Amir Patel, Niki Trigoni, and Andrew Markham. mmpoint: Dense human point cloud generation from mmwave. In *BMVC*, pages 194–196, 2023. 2

[49] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th annual international conference on mobile systems, applications, and services*, pages 269–282, 2021. 1, 2, 5

[50] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Pro-*

ceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, pages 391–406, 2022. 2

[51] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023. 1

[52] Hongliu Yang, Zizhou Fan, Yueyang Wang, Jie Xiong, Duo Zhang, Xusheng Zhang, Junzhe Wang, Fusang Zhang, and Daqing Zhang. Bridging the resolution gap: Cost-effective human point cloud generation via low-bandwidth mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(2):1–26, 2026. 2

[53] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023. 2

[54] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 2

[55] Duo Zhang, Xusheng Zhang, Shengjie Li, Yaxiong Xie, Yang Li, Xuanzhi Wang, and Daqing Zhang. Lt-fall: The design and implementation of a life-threatening fall detection and alarming system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7 (1):1–24, 2023. 1

[56] Duo Zhang, Zhehui Yin, Xusheng Zhang, Junzhe Wang, Hongliu Yang, Zhiyun Yao, Zizhou Fan, Wenwei Li, and Daqing Zhang. Mupose: Breaking the scalability barrier of mmwave multi-user pose estimation in the wild. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, pages 622–636, 2026. 8

[57] Haoyu Zhang, Dongheng Zhang, Ruiyuan Song, Zhi Wu, Jinbo Chen, Liang Fang, Zhi Lu, Yang Hu, Hui Lin, and Yan Chen. Umimo: Universal unsupervised learning for mmwave radar sensing with mimo array synthesis. *IEEE Transactions on Mobile Computing*, 2025. 2