LEO-Fuse: A Task-General, Modality-Agnostic Framework for Universal Multimodal Human Sensing

Emrecan Aslan[1]

Istanbul Technical University, Istanbul, Turkey
aslane22@itu.edu.tr

Abstract. Robust human-body understanding benefits from fusing heterogeneous sensors—WiFi CSI, depth, LiDAR, mmWave, and RGB—yet each modality carries distinct noise characteristics and may be intermittently unavailable. Existing fusion methods often assume fixed sensor sets and can collapse onto dominant modalities, reducing robustness under missing-modality conditions. We propose *LEO-Fuse*, a task-general, modality-agnostic Local Expert Orchestration (LEO) framework for universal multimodal human sensing. LEO-Fuse freezes unimodal encoders, aligns modality features in a shared token space, fuses available modality tokens with a mask-aware transformer, and applies task-specific local routing for downstream tasks including 3D human pose estimation (HPE) and human activity recognition (HAR). For each target task, a single trained model handles all 31 non-empty subsets of five modalities without retraining. On MM-Fi, LEO-Fuse obtains 75.5 mm mean per-joint position error (MPJPE), averaged over all 31 subsets. For HAR, HPE-trained encoders yield 93.4% average top-1 accuracy, improving to 95.7% with HAR-trained encoders. On XRF55, LEO-Fuse reaches up to 95.5% HAR accuracy. These results demonstrate scalable and robust universal multimodal human sensing under realistic missing-modality settings.

Keywords: 3D human pose estimation · human activity recognition · multimodal fusion · missing modalities · modality invariance

1 Introduction

3D human pose estimation (3D HPE) underpins applications in assisted living, rehabilitation, and human–computer interaction, yet the most accurate approaches largely rely on optical sensing (RGB/RGB-D) [11]. In privacy-sensitive environments, however, cameras are often impractical or restricted. This motivates non-intrusive sensing modalities—WiFi CSI, mmWave radar, LiDAR, and depth—which can perceive human motion and geometry without capturing visual appearance [31].

These modalities exhibit complementary but heterogeneous sensing physics. WiFi CSI provides device-free coverage and can work under occlusions, but its

spatial cues are indirect and environment-dependent, requiring strong learning-based priors [6, 23]. mmWave radar provides motion-sensitive Doppler cues but yields sparse, noisy point clouds; recent radar-only pose estimators therefore emphasise robust feature extraction and priors [7, 33, 39]. In contrast, depth and LiDAR provide more explicit geometry yet remain line-of-sight limited. Multimodal datasets such as MM-Fi [31] enable systematic study of these trade-offs and fusion strategies.

A key challenge is that naive fusion strategies are often unbalanced: optimisation disproportionately relies on the strongest modality and suppresses weaker sensors, leading to modality collapse [1]. Moreover, reliability can be joint-dependent: certain sensors may track torso motion well while failing on extremities, suggesting that global modality weights are insufficient [7, 33].

In practice, most existing non-intrusive 3D HPE methods are tailored to a fixed sensor configuration—typically unimodal [25, 35] or bimodal [32]—and do not support inference across arbitrary modality subsets.

This paper introduces *LEO-Fuse*, whose task-specific models operate on any subset of five sensing modalities—WiFi CSI (W), depth (D), LiDAR (L), mmWave (R), and RGB (I). The architecture combines frozen unimodal encoders, a mask-aware transformer that fuses the available modality tokens, and task-specific Local Expert Orchestration heads—inspired by dense prediction routing in [20, 21]—that perform local modality routing. We evaluate all 31 MM-Fi subsets for HPE and HAR and all seven XRF55 subsets [29] for HAR. We perform ablation studies to isolate the contribution of each component.

Contributions:

- We formulate universal multimodal human sensing as subset-invariant inference over all $2^{|\mathcal{S}|} - 1 = 31$ non-empty modality subsets $\mathcal{M} \subseteq \{W, D, L, R, I\}$, and instantiate the same task-general architecture for 3D HPE and HAR while training a separate subset-universal model for each target task.
- We introduce task-specific Local Expert Orchestration heads that perform dense local modality routing for reliability modelling, while a mask-aware transformer supports fusion over variable modality subsets.
- We introduce a universal training strategy combining weighted subset sampling, soft-target gate supervision, unimodal deep supervision, and mode-aware loss balancing to improve robustness.
- We empirically demonstrate a *cross-modal regularisation* effect: subset-invariant training implicitly transfers knowledge from strong to weak modalities, improving LiDAR singleton HPE by 35.6% and mmWave by 11.5% over dedicated baselines.

2 Background and Related Work

Non-Intrusive Multimodal Datasets. MM-Fi [31] established a large-scale benchmark with synchronised streams from five modalities plus 3D pose labels, enabling systematic evaluation of unimodal and multimodal 3D HPE and HAR under standardised protocols. XRF55 [29] provides a complementary RF-only

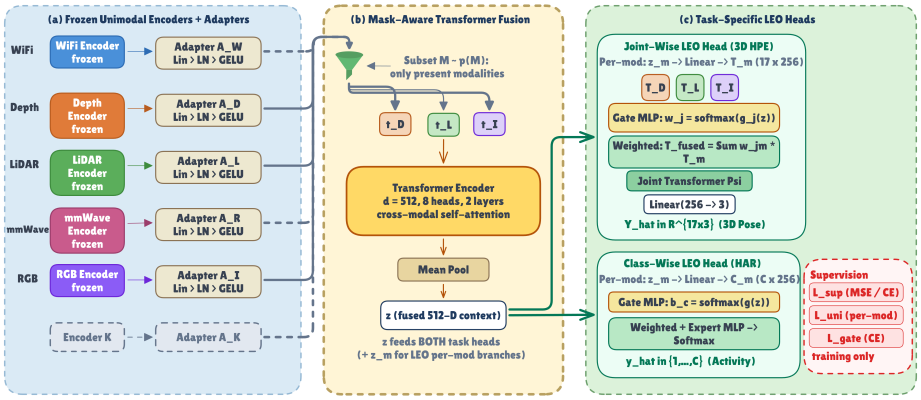


Fig. 1: LEO-Fuse architecture. Frozen unimodal encoders produce feature embeddings, mapped by per-modality adapters into a shared 512-D space. A mask-aware transformer fuses the available modality tokens. The LEO head performs local routing (joint-wise token-level for HPE; class-wise logit-level for HAR), with gate and unimodal supervision during training.

benchmark (WiFi, RFID, mmWave) with 55 action classes and 39 subjects. More recently, CUHK-X [15] offers seven modalities including thermal and infrared with unimodal baselines, M3GYM [30] introduces a large-scale multi-view, multi-person gym video dataset with 2D/3D keypoint, mesh, action, and expert-quality annotations, and XRF V2 [18] extends XRF55 with wearable IMU streams for temporal action summarisation.

mmWave-Based HPE and HAR. mmWave radar provides motion-sensitive, privacy-preserving point clouds but suffers from sparse, noisy returns. mmDiff [7] introduces diffusion-based pose generation with body-part isolation, achieving 65.26 mm MPJPE on MM-Fi. mmSkeleton [19] combines PointNet with Bi-LSTM and dual attention for skeleton estimation from sparse radar points. GF-DecNet [28] decouples geometry-aware features for radar HPE, while mmChainPose [22] exploits temporal chaining with geometric constraints. For HAR, STPM [4] applies Mamba-based spatiotemporal modelling to mmWave point clouds (95.69% on MM-Fi), and mmPoint-Attention [27] proposes a unified attention framework for HPE and HAR. These methods are single-modality; LEO-Fuse subsumes mmWave as one of five modalities.

WiFi-Based HPE and HAR. VST-Pose [35] introduces velocity auxiliary targets achieving 169.7 mm MPJPE on MM-Fi, WiLHPE [9] uses lightweight dynamic convolutions reaching 151.75 mm, and graph-based approaches [2] model subcarrier correlations at ∼160.6 mm. Robust WiFi HPE [23] applies denoising autoencoders with dynamic subcarrier attention. For HAR, cross-scenario robustness remains a challenge: He et al. [12] report 30–71% degradation under domain

shift, while Scale What Counts [14] evaluates MAE-based foundation models for zero-shot cross-domain WiFi sensing.

Multimodal Fusion for Human Sensing. Several works combine WiFi or RF signals with vision for HAR. ActionFi [36] fuses restructured CSI with optical flow (99.08% on MM-Fi with 5-fold CV), MaskFi [32] uses self-supervised WiFi–vision pretraining (96.82%), and Wi-Flow [37] combines WiFi CSI dynamic features with optical flow. SkeFi [13] transfers RGB skeleton knowledge to wireless modalities via cross-modal distillation. These methods use fixed two-modality pipelines and cannot handle arbitrary modality subsets at inference.

Foundation Models for Sensing. Recent foundation models target multi-modal human sensing at scale. Babel [5] uses contrastive expandable alignment across six modalities for few-shot HAR. MASTER [40] applies masked pre-training over eight modality types for HAR (89.7% with 2% labels on MM-Fi). TENT [38] connects language models with IoT sensors for zero-shot activity recognition. These models emphasise label efficiency and scalable sensing representation learning, whereas LEO-Fuse targets supervised, subset-universal HPE and HAR over all MM-Fi modality subsets with local routing.

Modality-Invariant and Missing-Modality Methods. X-Fi [3] is the closest prior work, proposing a modality-invariant foundation model with iterative cross-attention (X-Fusion) over six modalities and per-modality Binomial dropout. On MM-Fi, X-Fi achieves 83.7 mm MPJPE and 88.7% HAR accuracy at its best modality combination. Balanced MM [24] uses Shapley-value-based contribution balancing for fixed four-modality HPE, reaching 51.16 mm on Protocol 1; however, their model requires a fixed modality set. RFusion [8] performs dynamic contrastive fusion over three RF modalities for few-shot HAR ($\sim$90.10% on MM-Fi). Ramazanova et al. [26] propose Missing Modality Tokens as explicit placeholders for absent modalities. Kontras et al. [17] combine multimodal and unimodal losses and modulate modality-encoder learning rates from unimodal performance estimates; MIND [10] adds modality-specific prediction heads to a joint-fusion model and uses weighted ensemble distillation from pre-trained unimodal teachers to train a compact multimodal student that also supports unimodal inputs. LEO-Fuse differs by using unimodal prediction errors as soft targets for local modality routing over the currently available modalities, and by reweighting subset losses over availability modes while the main encoders are frozen. Absent modalities are omitted at the fusion input rather than represented by placeholders, while the LEO head models modality reliability. Modality collapse and low-rank entanglement [1, 16] are addressed through local routing and unimodal deep supervision.

3 Method

3.1 Problem Setting

Let $\mathcal{S} = \{W, D, L, R, I\}$ denote the modality set (WiFi CSI, depth, LiDAR, mmWave, and RGB), and let $\mathcal{M} \subseteq \mathcal{S}$ be the set of available modalities for a sample. We observe inputs $\{x_m\}_{m\in\mathcal{M}}$ and a task target $Y_\mathcal{T}$, where $\mathcal{T}$ denotes either 3D HPE or HAR, $Y_{\text{HPE}} \in \mathbb{R}^{J\times 3}$ with J=17, and $Y_{\text{HAR}} \in \{1, \ldots, C\}$ with C action classes. For each task, the goal is a single trained model $f_\mathcal{T}$ that predicts $\hat{Y}_\mathcal{T} = f_\mathcal{T}(\{x_m\}_{m\in\mathcal{M}})$ for any non-empty subset $\mathcal{M}$ without retraining, yielding $2^{|\mathcal{S}|} - 1 = 31$ supported configurations per task.

3.2 Architecture Overview

The full architecture is illustrated in Fig. 1. LEO-Fuse comprises (i) unimodal encoders E_m producing latent vectors $z_m = E_m(x_m) \in \mathbb{R}^{d_m}$, (ii) per-modality adapters A_m that map z_m into a shared fusion dimension via $\tilde{z}_m = A_m(z_m)$, (iii) a transformer fusion module F producing a fused latent $z = F(\{\tilde{z}_m\}_{m\in\mathcal{M}})$, and (iv) a task-specific decision head H. The design is task-general: the same encoder–adapter–fusion–orchestration pattern is used for HPE and HAR, while adapters, fusion, and heads are trained for each target task. For HPE, H is a joint-wise local orchestration head producing $\hat{Y} = H(z, \{\tilde{z}_m\}_{m\in\mathcal{M}})$; for HAR, H is a class-wise local orchestration head producing action logits. Encoders are frozen within each downstream task setting; only adapters, fusion, and the decision head are optimised. To train a universal model, we define a distribution over modality subsets $p(\mathcal{M})$ and sample $\mathcal{M} \sim p(\mathcal{M})$ during training; at test time we evaluate the same task-specific parameters on specified subsets $\mathcal{M}$. The joint-wise head and loss derivations below specialise to HPE; for notational simplicity, we suppress the task subscript and write Y and $\hat{Y}$ for Y_{HPE} and $\hat{Y}_{\text{HPE}}$.

Frozen Unimodal Encoders Each modality encoder is trained as a unimodal baseline for the relevant setting and then frozen for subset-universal fusion training. We use modality-specific architectures producing latents of dimension $d_m \in \{128, 512\}$ (details in Sec. 4.2). Freezing prevents the encoders from drifting towards the dominant modality during fusion training and cleanly separates encoder learning from subset-invariant fusion.

Per-Modality Adapters We add a per-modality adapter block $A_m : \mathbb{R}^{d_m} \to \mathbb{R}^d$ (with $d = 512$) that maps each modality latent z_m into a shared fusion space:

$$\tilde{z}_m = A_m(z_m) = \text{Dropout}(\text{GELU}(\text{LN}(W_m z_m + b_m))) \in \mathbb{R}^d, \tag{1}$$

where LN denotes layer normalisation and $W_m z_m + b_m$ implements $\text{Linear}(d_m \to d)$. The adapter output $\tilde{z}_m$ forms the modality token input to fusion and is subsequently passed to the pose head.

Mask-Aware Transformer Fusion Given a sampled subset $\mathcal{M} \subseteq \mathcal{S}$, we construct tokens only for modalities $m \in \mathcal{M}$. We instantiate F as a transformer encoder over a variable-length set of modality tokens and mean-pool:

$$\begin{aligned} t_m &= W_{f,m}\tilde{z}_m + e_m, \\ [h_m]_{m\in\mathcal{M}} &= F([t_m]_{m\in\mathcal{M}}), \\ z &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} h_m, \end{aligned} \tag{2}$$

where $e_m \in \mathbb{R}^d$ is a learnable modality embedding. Tokens are concatenated in a fixed canonical modality order, and missingness is represented by omission rather than explicit placeholder tokens [26]; this avoids introducing uninformative tokens that may dilute attention across present modalities.

Joint-Wise Local Expert Orchestration Head The pose head performs joint-wise token-level modality routing. For each modality $m \in \mathcal{M}$, we project $\tilde{z}_m$ into per-joint tokens $T_m \in \mathbb{R}^{J\times d_j}$ and compute joint-wise modality weights:

$$\begin{aligned} T_m &= \mathrm{reshape}(W_m^{(J)}\tilde{z}_m + b_m^{(J)}) + E^{(\mathrm{joint})}, \\ \hat{y}_j^{(m)} &= W_o^{(\mathrm{uni})}T_{m,j} + b_o^{(\mathrm{uni})}, \\ g_j &= \phi([T_{m,j}]_{m\in\mathcal{M}},\, z), \qquad \alpha_j = \mathrm{softmax}(g_j), \\ T_j^{(\mathrm{fused})} &= \sum_{m\in\mathcal{M}} \alpha_{j,m}\, T_{m,j}, \\ U &= \Psi(T^{(\mathrm{fused})}), \qquad \hat{y}_j = W_o U_j + b_o, \end{aligned} \tag{3}$$

where $E^{(\mathrm{joint})}$ are learnable joint embeddings, ϕ is a gating MLP that conditions on the modality joint tokens and the fused context z and outputs logits over the available modalities, and $U \in \mathbb{R}^{J\times d_j}$ is regressed joint-wise to $\hat{Y} = [\hat{y}_j]_{j=1}^{J}$. $W_o^{(\mathrm{uni})}, b_o^{(\mathrm{uni})}$ parameterise the unimodal auxiliary prediction head, while W_o, b_o parameterise the fused prediction head. The per-modality predictions $\hat{y}_j^{(m)}$ are used for unimodal deep supervision and soft-target gate supervision. Unlike global cross-attention approaches, LEO-Fuse routes per-joint, reflecting the observation that modality reliability varies across body parts (e.g., mmWave captures torso geometry accurately but performs poorly on extremities).

3.3 Training Objectives

Weighted Subset Sampling Rather than sampling modality subsets uniformly, we define a structured distribution $p(\mathcal{M})$ over all 31 non-empty subsets of $\mathcal{S}$, parameterised by subset cardinality $K{=}|\mathcal{M}|$. We allocate probability mass

per cardinality level and distribute it uniformly within each level:

$$p_K(K) = \begin{cases} 0.15 & K = 1 \text{ (5 subsets, 0.03 each)}, \\ 0.25 & K = 2 \text{ (10 subsets, 0.025 each)}, \\ 0.25 & K = 3 \text{ (10 subsets, 0.025 each)}, \\ 0.15 & K = 4 \text{ (5 subsets, 0.03 each)}, \\ 0.20 & K = 5 \text{ (1 subset)}. \end{cases} \tag{4}$$

This design assigns the largest probability mass to the mid-cardinality levels, while the full-set mode (K=5) retains 20% of training iterations.

Loss Functions We optimise the expected training objective over sampled modality subsets $\mathcal{M} \sim p(\mathcal{M})$. The primary supervision is coordinate MSE,

$$\mathcal{L}_{\text{sup}} = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{3} \|\hat{y}_j - y_j\|_2^2. \tag{5}$$

To mitigate modality collapse, we define per-joint per-modality mean-squared errors $\epsilon_{j,m} = \frac{1}{3}\|\hat{y}_j^{(m)} - y_j\|_2^2$ and soft gate targets $\tilde{\alpha}_j = \text{softmax}([-\epsilon_{j,m}/\tau]_{m\in\mathcal{M}})$ with τ=0.05. We add two auxiliary losses:

$$\begin{aligned} \mathcal{L}_{\text{gate}} &= \frac{1}{J} \sum_{j=1}^{J} \text{CE}(\tilde{\alpha}_j, \alpha_j), \\ \mathcal{L}_{\text{uni}} &= \frac{\sum_{m\in\mathcal{M}} \lambda_m \left(\frac{1}{J} \sum_{j=1}^{J} \epsilon_{j,m} \right)}{\sum_{m\in\mathcal{M}} \lambda_m}, \end{aligned} \tag{6}$$

where $\hat{y}_j^{(m)}$ is a modality-specific prediction for joint j, and λ_m are per-modality weights (λ_{WiFi}=0.5, all others 1.0). Although every modality branch for joint j is supervised against the same 3D pose target y_j, the error vector $[\epsilon_{j,m}]_{m\in\mathcal{M}}$ defines the soft gate target $\tilde{\alpha}_j$ over the available modalities. Because this error vector can vary across joints, different joints in the same sample can favour different modalities. Gate supervision uses λ_{gate}=0.05 with linear warmup over the first 2 epochs, and unimodal deep supervision uses λ_{uni}=0.02. The total objective is:

$$\mathcal{L} = \mathbb{E}_{\mathcal{M}\sim p(\mathcal{M})} \left[\mathcal{L}_{\text{sup}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}} + \lambda_{\text{uni}} \mathcal{L}_{\text{uni}} \right]. \tag{7}$$

For a sampled mode $\mathcal{M}$, let $\mathcal{L}_{\mathcal{M}}$ denote the bracketed per-mode loss in Eq. (7).

Mode-Aware Loss Balancing Different modality subsets converge at different rates; subsets dominated by weak modalities (e.g., WiFi-only) exhibit persistently higher loss and receive insufficient weight under uniform weighting. We address this with an exponential moving average (EMA)-based per-mode loss balancer. For each mode $\mathcal{M}$, we maintain a running mean $\bar{\mathcal{L}}_{\mathcal{M}}$:

$$\bar{\mathcal{L}}_{\mathcal{M}} \leftarrow (1-\beta)\, \bar{\mathcal{L}}_{\mathcal{M}} + \beta\, \mathcal{L}_{\mathcal{M}}, \tag{8}$$

where β=0.1 is the EMA decay rate. We compute the global average $\bar{\mathcal{L}} = \frac{1}{N_{\text{modes}}} \sum_{\mathcal{M}'} \bar{\mathcal{L}}_{\mathcal{M}'}$, where N_{modes} is the number of modes with tracked EMA statistics, and rescale each batch loss by:

$$s_{\mathcal{M}} = \text{clamp}\left(\left(\frac{\bar{\mathcal{L}}_{\mathcal{M}}}{\bar{\mathcal{L}}} \right)^{\gamma},\ s_{\min},\ s_{\max} \right), \quad \mathcal{L}_{\mathcal{M}} \leftarrow s_{\mathcal{M}} \cdot \mathcal{L}_{\mathcal{M}}, \tag{9}$$

with γ=0.25, $s_{\min}$=0.5, $s_{\max}$=2.0. This gently up-weights high-loss modes without extreme reweighting, improving robustness across all 31 subsets. In contrast, X-Fi [3] manually tunes per-modality Binomial occurrence probabilities; our EMA-based scheme achieves analogous rebalancing automatically and adapts over the course of training.

3.4 Extension to Human Activity Recognition

For HAR, LEO-Fuse replaces the joint-wise HPE head with a class-wise local orchestration head. Each modality branch predicts class logits, and a class-wise gate combines the available modality logits into the final prediction. The same subset-invariant training principle applies over the available HAR modality set. Encoders are frozen after task-appropriate unimodal training, while adapters, fusion, and the HAR head are trained for subset-universal classification.

4 Experiments

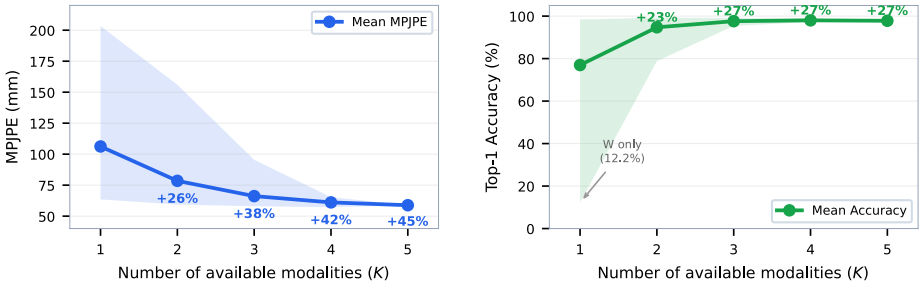
4.1 Dataset and Protocol

We evaluate on the MM-Fi dataset [31], a large-scale non-intrusive sensing benchmark with 40 subjects performing 27 action categories across 4 environments (1,080 sequences, 320k+ synchronised frames at 10 Hz, synchronisation error <25 ms). Each frame provides all five modalities. 3D pose labels (17 joints) are obtained by triangulating stereo infra-red keypoints with temporal and bone-length refinement. We use protocol P3 (all 27 action categories) with the S1 random split; all LEO-Fuse vs. X-Fi comparisons use the same protocol and metric definitions. We report MPJPE and Procrustes-aligned MPJPE (PA-MPJPE), both in millimetres. Additional S2 (cross-subject) and S3 (cross-environment) results are provided in the supplementary material.

For XRF55 HAR [29], we evaluate WiFi, RFID, and mmWave under the official all-scenes first-14/last-6 trial split with 55 classes and 12,870 test samples.

4.2 Implementation Details

Our universal W+D+L+R+I model uses fusion dimension $d = 512$ and joint token dimension $d_j = 256$. Depth and RGB use ResNet-18 encoders with input shapes $1 \times 224 \times 224$ and $3 \times 224 \times 224$, respectively; both produce 512-D latents. LiDAR uses a PointTransformer (1024×3, 128-D latent), mmWave uses a DGCNN-based point encoder (128-D latent), and WiFi uses a CSI 1D CNN



(a) HPE (MPJPE, lower is better). **(b)** HAR (accuracy, higher is better).

Fig. 2: Sensor scaling vs. number of available modalities K. Shaded bands show min–max across subsets of each cardinality. (a) Cardinality-averaged MPJPE decreases monotonically; (b) mean HAR accuracy saturates at $K \geq 3$, with the large $K=1$ band driven by WiFi-only (12.2%).

(128-D latent). Per-modality adapters map every latent to 512-D before fusion. Training uses AdamW (learning rate 6×10^{-4}, weight decay 10^{-4}, batch size 64) for 30 epochs on an RTX 3050 Laptop GPU. We employ unimodal feature caching to accelerate training since encoders are frozen. During universal training we sample modality subsets $\mathcal{M} \sim p(\mathcal{M})$ from the K-based distribution described in Sec. 3.3; at test time we evaluate on each target subset.

4.3 Universal HPE Results

Effect of Subset Cardinality. Cardinality-averaged MPJPE decreases monotonically as subset cardinality K increases (Fig. 2). Singletons ($K=1$) average 106.21 mm MPJPE, decreasing to 78.46 mm at $K=2$, 66.17 mm at $K=3$, and 61.08 mm at $K=4$. The best-performing subset is D+L+R+I at 57.22 mm, slightly outperforming the full five-modality set W+D+L+R+I (58.82 mm); the 1.60 mm gap reflects the difficulty of integrating WiFi's indirect spatial cues without degrading the fusion of stronger geometric modalities.

Table 1 compares LEO-Fuse against X-Fi [3] on all 31 subsets (X-Fi values from Table 1 in [3]). LEO-Fuse wins on **30 of 31** subsets in MPJPE (avg. 75.5 vs. 103.7 mm, +27.2%) and **29 of 31** in PA-MPJPE (avg. 45.7 vs. 58.2 mm, +21.3%). The exceptions are W+R, where WiFi–mmWave fusion without geometric anchors is harder for our gating mechanism, and WiFi-only PA-MPJPE, where X-Fi's Binomial sampling allocates more capacity to WiFi. The largest gains appear for LiDAR-including subsets (+53.4% on singleton L), confirming that frozen encoders and per-joint gating exploit spatial structure more effectively than iterative cross-attention.

Modality-Specific Observations. Among singletons, RGB (63.49 mm) and depth (66.26 mm) perform best, followed by LiDAR (77.93 mm) and mmWave (119.97 mm); WiFi-only is weakest (203.38 mm). WiFi-including subsets average 83.66 mm versus 66.87 mm for WiFi-excluding ones. Figure 3 reveals that

Table 1: LEO-Fuse vs. X-Fi [3]: MPJPE and PA-MPJPE (mm) on all 31 MM-Fi subsets. Δ_{rel}: relative improvement in percent (positive = better). **Bold** = winner per metric. LEO-Fuse wins 30/31 (MPJPE) and 29/31 (PA-MPJPE).

	MPJPE			PA-MPJPE		
Subset	Ours	X-Fi	Δ_{rel}	Ours	X-Fi	Δ_{rel}
W	**203.4**	225.6	+9.8	108.4	**105.3**	−2.9
D	**66.3**	101.8	+34.9	**43.0**	48.4	+11.2
L	**77.9**	167.1	+53.4	**47.2**	103.2	+54.3
R	**120.0**	127.4	+5.8	**62.9**	69.8	+9.9
I	**63.5**	93.9	+32.4	**41.4**	60.3	+31.3
W+D	**72.7**	101.8	+28.6	**43.6**	48.1	+9.4
W+L	**105.6**	159.5	+33.8	**57.5**	102.7	+44.0
W+R	156.0	**117.2**	−33.1	81.7	**62.7**	−30.3
W+I	**68.7**	93.0	+26.1	**41.8**	59.5	+29.7
D+L	**62.4**	102.5	+39.1	**41.2**	48.4	+14.9
D+R	**64.0**	98.0	+34.7	**41.9**	47.3	+11.4
D+I	**59.4**	86.1	+31.0	**39.3**	48.1	+18.3
L+R	**73.7**	109.8	+32.9	**43.0**	63.4	+32.2
L+I	**60.5**	93.0	+34.9	**39.8**	59.7	+33.3
R+I	**61.7**	88.8	+30.5	**40.4**	57.3	+29.5
W+D+L	**66.9**	102.0	+34.4	**41.4**	48.1	+13.9
W+D+R	**70.5**	97.0	+27.3	**42.4**	47.1	+10.0
W+D+I	**61.2**	85.3	+28.3	**39.4**	48.1	+18.1
W+L+R	**95.4**	107.4	+11.2	**50.9**	63.1	+19.3
W+L+I	**64.1**	93.0	+31.1	**39.9**	59.7	+33.2
W+R+I	**67.1**	88.5	+24.2	**40.8**	57.1	+28.5
D+L+R	**60.9**	96.0	+36.6	**40.1**	47.3	+15.2
D+L+I	**58.0**	84.8	+31.6	**38.7**	48.2	+19.7
D+R+I	**58.4**	83.4	+30.0	**38.8**	47.3	+18.0
L+R+I	**59.3**	88.4	+32.9	**38.8**	57.2	+32.2
W+D+L+R	**65.4**	97.6	+33.0	**40.4**	47.4	+14.8
W+D+L+I	**59.5**	86.0	+30.8	**38.8**	48.2	+19.5
W+D+R+I	**60.3**	84.0	+28.2	**39.0**	47.6	+18.1
W+L+R+I	**63.0**	88.6	+28.9	**39.1**	57.1	+31.5
D+L+R+I	**57.2**	83.5	+31.5	**38.2**	47.6	+19.7
W+D+L+R+I	**58.8**	83.7	+29.7	**38.3**	47.6	+19.5
Average	**75.5**	103.7	+27.2	**45.7**	58.2	+21.3

modality gaps are joint-dependent: depth and RGB excel on proximal joints (spine, hips ∼43–55 mm) while all modalities degrade on distal extremities (wrists ∼85–170 mm), motivating per-joint routing.

Singleton Comparison with Dedicated Baselines. Table 2 compares universal singletons against dedicated unimodal baselines using the same encoders. For WiFi, depth, and RGB, the universal model incurs a small penalty (+0.48 to +4.29 mm) due to shared capacity, but *improves* over dedicated baselines for LiDAR (35.6%) and mmWave (11.5%).

4.4 Comparison with Prior Methods

Broader Baseline Context. Table 3 situates LEO-Fuse among published methods on MM-Fi. Direct comparison requires caution: mmDiff [7] uses a random split without specifying Protocols 1–3; Balanced MM [24] reports on Protocol 1 (14 daily actions only); ActionFi [36] uses 5-fold cross-validation; and STPM [4] does

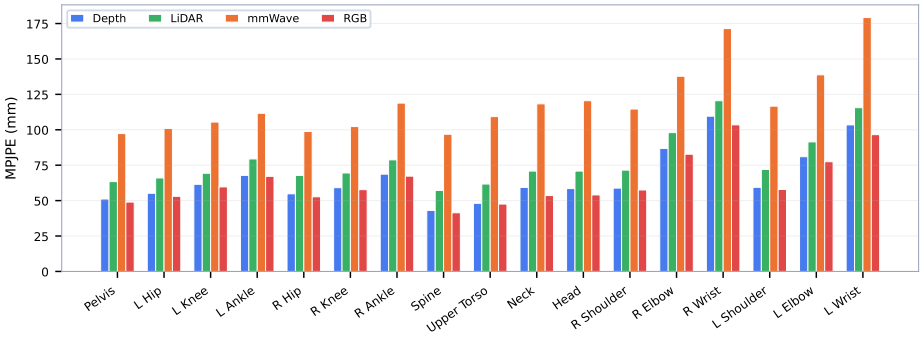


Fig. 3: Per-joint MPJPE (mm) for each singleton modality (WiFi excluded—off-scale at 200+ mm). Proximal joints (spine, pelvis) are easiest across all sensors; distal joints (wrists, elbows) exhibit the largest inter-modality variance.

Table 2: Dedicated unimodal baselines vs. LEO-Fuse universal singletons. MPJPE / PA-MPJPE in mm. Δ_{rel}: relative reduction in percent by the universal model (positive = improved).

	Dedicated		Universal		Δ_{rel}	
Modality	MPJPE	PA	MPJPE	PA	MPJPE	PA
WiFi (W)	199.09	106.12	203.38	108.35	−2.2	−2.1
Depth (D)	65.58	43.33	66.26	43.03	−1.0	+0.7
LiDAR (L)	121.02	56.00	77.93	47.18	**+35.6**	**+15.8**
mmWave (R)	135.51	67.11	119.97	62.86	**+11.5**	**+6.3**
RGB (I)	63.01	42.29	63.49	41.39	−0.8	+2.1

not specify the exact split. For HPE, LEO-Fuse and X-Fi are the only methods averaging over all 31 subsets; the remaining baselines use fixed modality sets and different protocols.

4.5 Ablation Study

Table 4 isolates each component's contribution. The local routing head is the most impactful (+16.9% MPJPE when removed), confirming that joint-wise modality routing is the primary driver of robustness. Reducing pose-head parameters by 68.9% (14.46 M to 4.49 M) and total trainable parameters by 48.1% (20.71 M to 10.74 M) increases average MPJPE by only 0.49 mm (75.53 to 76.02 mm; +0.65%). Unimodal deep supervision has the second-largest ablation effect (+10.5%), while mode-aware loss balancing (+4.3%) and gate supervision (+1.8%) provide complementary gains. Replacing K-based subset sampling with uniform sampling increases MPJPE by 5.3%, confirming that structured cardinality-based exposure matters. The supplementary material

Table 3: Published methods on MM-Fi (MPJPE in mm, HAR accuracy in %). Protocol differences exist (see text); numbers are not directly comparable across rows.

Method	Modality	Scope	MPJPE	HAR Acc
mmDiff [7]	mmWave	HPE	65.3	—
mmSkeleton [19]	mmWave	HPE	71.0	—
WiLHPE [9]	WiFi	HPE	151.8	—
VST-Pose [35]	WiFi	HPE	169.7	—
Balanced MM [24]	4-mod fixed	HPE	51.2†	—
STPM [4]	mmWave	HAR	—	95.69
ActionFi [36]	WiFi+RGB	HAR	—	99.08‡
RFusion [8]	3-RF inv.	HAR	—	∼90.1
X-Fi [3]	5-mod inv.	Both	103.7	62.5
LEO-Fuse	**5-mod inv.**	**Both**	**75.5**	**98.4**

†P1 (14 actions). ‡5-fold CV.
MPJPE: 31-subset average.
HAR: 15-subset comparable average; LEO-Fuse uses HAR-trained encoders.

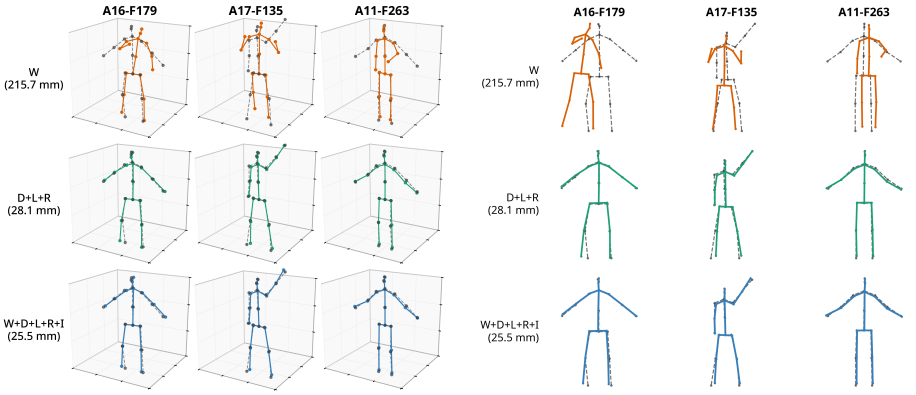
Table 4: Component ablations and head-capacity control (MPJPE / PA-MPJPE in mm, averaged over 31 subsets). Δ_{rel}: relative degradation in percent vs. full (positive = worse).

	Average		Δ_{rel}	
Variant	MPJPE	PA-MPJPE	MPJPE	PA
Full model	75.53	45.75	—	—
Small LEO head	76.02	46.42	+0.65	+1.5
w/o local routing head	88.31	51.65	+16.9	+12.9
w/o unimodal supervision	83.44	49.95	+10.5	+9.2
w/o weighted subset sampling	79.54	46.58	+5.3	+1.8
w/o mode loss balancing	78.75	46.51	+4.3	+1.7
w/o gate supervision	76.86	47.10	+1.8	+3.0

provides full per-mode breakdowns and controlled 10-epoch sensitivity checks for the auxiliary-loss and mode-balancing hyperparameters.

4.6 Qualitative Analysis

Figure 4 visualises three representative MM-Fi frames (A16-F179, A17-F135, A11-F263) under three modality configurations. The trend is consistent across both 3D and 2D views: WiFi-only predictions show large structural errors (row MPJPE 215.7 mm), including unstable torso orientation and pronounced limb drift. Adding geometric modalities (D+L+R) restores global body structure and sharply reduces error (28.1 mm). Using all five modalities further improves alignment to the dashed ground truth (25.5 mm), most visibly on distal joints.



(a) 3D view. **(b)** 2D projection.

Fig. 4: Qualitative 3D pose predictions on MM-Fi test frames. Rows: WiFi-only (W), D+L+R, and all five modalities (W+D+L+R+I). Grey dashed skeletons denote ground truth. Per-configuration MPJPE (mm) is shown for each row.

The raised-arm sample (A17-F135) is particularly diagnostic. In WiFi-only mode, the shoulder–elbow–wrist chain is fragmented and the pose is under-constrained. D+L+R already recovers coherent kinematics, while W+D+L+R+I refines the arm trajectory and leg symmetry. This behaviour is consistent with joint-wise token-level routing: the model can rely more on geometric sensors where WiFi evidence is noisy, rather than enforcing a single global fusion weight.

Residual errors remain in fast or high-articulation extremities, especially wrists and ankles, even in the full-modality row. This matches the quantitative per-joint pattern in Fig. 3, where distal joints exhibit the largest cross-modality variance. Qualitatively, LEO-Fuse improves robustness primarily by fixing global structure first and confining remaining error to local extremities.

4.7 Human Activity Recognition

To isolate the effect of encoder training, we compare an HPE-trained setting, which reuses frozen HPE encoders, with a control whose MM-Fi encoders are trained unimodally using only HAR labels and then frozen before subset-universal training. The HAR-trained control improves the 31-subset average from 93.4% to 95.7% and the full WDLRI mode from 97.8% to 99.7%, with improvements on 28 of 31 subsets relative to the HPE-trained setting. As with HPE, MM-Fi HAR accuracy scales with modality count (Fig. 2b), approaching 98% for $K \geq 3$; the wide $K=1$ band reflects WiFi-only's 12.2%. For XRF55, encoders are trained as unimodal HAR baselines on XRF55 data.

Table 5 reports HAR results on all 31 MM-Fi subsets. Both LEO-Fuse encoder settings outperform X-Fi on **all 15** comparable subsets; the remaining 16 WiFi-including subsets are only evaluated by LEO-Fuse. X-Fi [3] excludes

Table 5: MM-Fi HAR: top-1 accuracy (%) on all 31 subsets. HPE-trained reuses frozen encoders trained for HPE; HAR-trained uses frozen encoders trained unimodally with HAR labels. — indicates not reported by X-Fi [3].

Subset	HPE-trained	HAR-trained	X-Fi	Subset	HPE-trained	HAR-trained	X-Fi
W	12.2	14.8	—	W+D+R	96.1	98.3	—
D	93.0	92.9	48.1	W+D+I	95.4	99.1	—
L	98.4	98.2	52.7	W+L+R	99.2	99.5	—
R	87.3	94.9	85.7	W+L+I	97.2	99.5	—
I	93.9	98.3	26.5	W+R+I	97.4	99.2	—
W+D	92.7	93.4	—	D+L+R	99.0	99.5	80.5
W+L	98.1	97.9	—	D+L+I	97.5	99.7	48.7
W+R	78.8	93.2	—	D+R+I	97.2	99.5	70.7
W+I	93.3	98.2	—	L+R+I	99.1	99.6	77.8
D+L	97.9	98.8	51.6	W+D+L+R	98.7	99.6	—
D+R	96.5	98.4	79.8	W+D+L+I	97.2	99.7	—
D+I	95.4	99.0	45.3	W+D+R+I	97.3	99.5	—
L+R	99.2	99.5	88.7	W+L+R+I	98.9	99.6	—
L+I	97.5	99.5	35.2	D+L+R+I	97.9	99.7	72.2
R+I	97.9	99.2	73.4	W+D+L+R+I	97.8	99.7	—
W+D+L	98.0	98.9	—				
Average	93.4	95.7	—	Average (comparable)	96.5	98.4	62.5

WiFi from MM-Fi HAR because MM-Fi synchronises all modalities at 10 Hz, so each WiFi CSI frame spans only 100 ms ($3\times114\times10$ tensor) [31]—too little temporal context to distinguish the 27 action categories. We deliberately retain WiFi to stress-test robustness to low-quality modalities; WiFi HAR on XRF55 is substantially better (82.2%), confirming that this limitation is dataset-specific rather than fundamental. Table 6 shows XRF55 results, where LEO-Fuse wins **6 of 7** subsets, losing only on RF+R by −0.1 percentage points.

4.8 Discussion and Limitations

Beyond the cross-modal regularisation effect reported in Tab. 2, several limitations remain.

Weak-Modality Pair Degradation. WiFi and mmWave are the two weakest singletons (203.4 and 120.0 mm MPJPE). When paired without a geometric anchor such as depth or LiDAR, the W+R subset degrades to 156.0 mm—the only configuration where X-Fi outperforms LEO-Fuse. The local routing head, designed to assign greater weight to more reliable modalities for each joint, receives two uniformly noisy inputs and cannot recover the geometric structure that stronger modalities would provide.

Table 6: XRF55 HAR: top-1 accuracy (%) on all 7 subsets under the official all-scenes first-14/last-6 trial split. Δ_{pp}: percentage-point difference (Ours minus X-Fi; positive values favour Ours). LEO-Fuse wins 6/7.

Subset	Ours	X-Fi	Δ_{pp}
W (WiFi)	**82.2**	55.7	+26.5
RF (RFID)	**46.8**	42.5	+4.3
R (mmWave)	**85.1**	83.9	+1.2
W+RF	**87.1**	58.1	+29.0
RF+R	86.4	**86.5**	−0.1
W+R	**95.0**	88.2	+6.8
W+RF+R	**95.5**	89.8	+5.7

No Zero-Shot or Continual Learning. TENT [38] supports zero-shot activity recognition, Babel [5] targets scalable multimodal sensing representation learning, and CAREC [34] enables continual learning for wireless HAR. LEO-Fuse requires fully supervised training on each target dataset.

5 Conclusion

We presented LEO-Fuse, a task-general, modality-agnostic framework for universal multimodal human sensing that supports all 31 subsets of five sensing modalities with one trained model per target task. Three key findings emerge: (1) joint-wise local routing is the dominant robustness driver; (2) subset-invariant training produces implicit cross-modal regularisation that improves weak-modality singletons beyond dedicated baselines; (3) the same design pattern—frozen unimodal encoders, subset-universal fusion, and local orchestration—applies to both HPE and HAR. On MM-Fi, LEO-Fuse achieves 75.5 mm average MPJPE and 95.7% average HAR accuracy with HAR-trained encoders; on XRF55, up to 95.5%. Extending to zero-shot or continual-learning settings is a promising direction.

Acknowledgements

This work originated as a project for the *Introduction to Machine Learning for Electrical Engineers* course at Istanbul Technical University. I am grateful for the course's rigorous academic environment and open-ended format, which inspired me to develop the initial project into the present research.

References

1. Chaudhuri, A., Dutta, A., Bui, T., Georgescu, S.: A closer look at multimodal representation collapse. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 7555–7577. PMLR (13–19 Jul 2025)

2. Chen, J., Qu, Y., Tang, R., Slock, D.: Graph-based 3d human pose estimation using wifi signals. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 19992–19996 (May 2026). https://doi.org/10.1109/ICASSP55912.2026.11463884
3. Chen, X., Yang, J.: X-fi: A modality-invariant foundation model for multimodal human sensing. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 97378–97397 (2025)
4. Chen, Y., Guo, Z., Zhang, H., Xu, M.: Stpm: Spatial-temporal point mamba for activity recognition using mmwave radar point clouds. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (June 2025). https://doi.org/10.1109/ICME59968.2025.11208900
5. Dai, S., Jiang, S., Yang, Y., Cao, T., Li, M., Banerjee, S., Qiu, L.: Babel: A scalable pre-trained model for multi-modal sensing via expandable modality alignment. In: Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems. p. 240–253. SenSys '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3715014.3722068
6. Deng, J., Chen, K., Jing, P., Dong, G., Yang, M., Zhu, A., Li, Y.: Csi-channel spatial decomposition for wifi-based human pose estimation. Electronics **14**(4), 756 (2025). https://doi.org/10.3390/electronics14040756
7. Fan, J., Yang, J., Xu, Y., Xie, L.: Diffusion model is a good pose estimator from 3d rf-vision. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 1–18. Springer Nature Switzerland, Cham (2025)
8. Feng, C., Chen, J., Liang, S., Peng, X., Yang, B., Wang, X., Huang, Z., Meng, X., Chen, X.: Rfusion: Dynamic multimodal rf fusion for few-shot human activity recognition. IEEE Transactions on Mobile Computing **25**(7), 10045–10060 (July 2026). https://doi.org/10.1109/TMC.2026.3656490
9. Gian, T.D., Nguyen, T.H., Nguyen, N.T., Nguyen, V.D.: Wilhpe: Wifi-enabled lightweight channel frequency dynamic convolution for hpe tasks. In: 2024 Tenth International Conference on Communications and Electronics (ICCE). pp. 516–521 (July 2024). https://doi.org/10.1109/ICCE62051.2024.10634628
10. Guerra-Manzanares, A., Shamout, F.: MIND: Modality-informed knowledge distillation framework for multimodal clinical prediction tasks. Transactions on Machine Learning Research (2025)
11. Guo, Y., Gao, T., Dong, A., Jiang, X., Zhu, Z., Wang, F.: A survey of the state of the art in monocular 3d human pose estimation: Methods, benchmarks, and challenges. Sensors **25**(8), 2409 (2025). https://doi.org/10.3390/s25082409
12. He, Z., Bouazizi, M., Yin, Y., Gui, G., Ohtsuki, T.: Robust cross-scenario wifi wireless sensing using incremental learning and elastic weight consolidation loss. IEEE Internet of Things Journal **12**(13), 24288–24299 (July 2025). https://doi.org/10.1109/JIOT.2025.3555267
13. Huang, S., Zhou, Y., Yang, J.: Skefi: Cross-modal knowledge transfer for wireless skeleton-based action recognition. IEEE Internet of Things Journal **13**(7), 13888–13898 (April 2026). https://doi.org/10.1109/JIOT.2025.3647603
14. Jiang, C., Yan, Y., Wang, Y., Chou, C.T., Hu, W.: Scale what counts, mask what matters: Evaluating foundation models for zero-shot cross-domain wi-fi sensing. arXiv preprint arXiv:2511.18792 (2025)
15. Jiang, S., Yuan, M., Ji, X., Yang, B., Liu, Z., Xu, L., Li, Y., He, Y., Dong, L., Lu, W., Yan, Z., Jiang, X., Gao, W., Chen, H., Xing, G.: A large-scale multimodal dataset and benchmarks for human activity scene understanding and reasoning.

In: Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services. pp. 352–370. MobiSys '26, Association for Computing Machinery, New York, NY, USA (2026). https://doi.org/10.1145/3745756.3809209
16. Kim, S., Kokilepersaud, K., Prabhushankar, M., AlRegib, G.: Countering multimodal representation collapse through rank-targeted fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4744–4754 (March 2026)
17. Kontras, K., Chatzichristos, C., Blaschko, M.B., Vos, M.D.: Improving multimodal learning with multi-loss gradient modulation. In: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024. BMVA (2024)
18. Lan, B., Li, P., Yin, J., Song, Y., Wang, G., Ding, H., Han, J., Wang, F.: Xrf v2: A dataset for action summarization with wi-fi signals, and imus in phones, watches, earbuds, and glasses. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **9**(3), 98 (Sep 2025). https://doi.org/10.1145/3749521
19. Li, W., Lei, W., Shi, K., Shi, Z., Wang, Y., Zhou, J.: mmskeleton: 3d human skeleton estimation using millimeter wave radar sparse point clouds. In: 2024 IEEE/CIC International Conference on Communications in China (ICCC). pp. 307–312 (Aug 2024). https://doi.org/10.1109/ICCC62479.2024.10681946
20. Li, Y., Hou, Y., Wei, Y., Zhu, X., Ma, Y., Shao, W., Guo, Y.: Moe3d: Mixture of experts meets multi-modal 3d understanding. arXiv preprint arXiv:2511.22103 (2025)
21. Liu, M., Liu, J., Zhang, J., Li, W., Yuan, J.: Posemoe: Mixture-of-experts network for monocular 3d human pose estimation. IEEE Transactions on Image Processing **34**, 8537–8551 (2025). https://doi.org/10.1109/TIP.2025.3644785
22. Lu, W., Chen, Y., Huang, X.: mmchainpose: Geometry-aware temporal chaining for robust human pose estimation from mmwave point clouds. Neurocomputing **663**, 132019 (2026). https://doi.org/10.1016/j.neucom.2025.132019
23. Nguyen, X.H., Nguyen, V., Luu, Q., Gian, T.D., Shin, O.: Robust wifi sensing-based human pose estimation using denoising autoencoder and CNN with dynamic subcarrier attention. IEEE Internet Things J. **12**(11), 17066–17079 (2025). https://doi.org/10.1109/JIOT.2025.3535156
24. Qi, M., Peng, J., Zhang, X., Ma, H.: Towards balanced multi-modal learning in 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21231–21241 (June 2026)
25. Qu, Y., Ma, H., Xiong, W.: Multiformer: A multiperson pose estimation system based on csi and attention mechanism. IEEE Internet of Things Journal **13**(10), 21160–21172 (May 2026). https://doi.org/10.1109/JIOT.2026.3665246
26. Ramazanova, M., Pardo, A., Alwassel, H., Ghanem, B.: Exploring missing modality in multimodal egocentric datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 75–85 (June 2025)
27. Shi, X., Bouazizi, M., Ohtsuki, T.: mmpoint-attention: A unified attention framework for human pose and activity recognition from mmwave radar point clouds. IEEE Sensors Journal **25**(14), 26794–26805 (July 2025). https://doi.org/10.1109/JSEN.2025.3577772
28. Tang, H., Chen, W., Yan, Y., Tang, Y., Yang, Y., Feng, Y., Jiang, X., Lu, J., Xu, Y., Fang, Q.: Gf-decnet: Geometry-aware feature decoupling network for millimeter-wave radar pose estimation. IEEE Internet of Things Journal **12**(21), 45596–45609 (Nov 2025). https://doi.org/10.1109/JIOT.2025.3601012

29. Wang, F., Lv, Y., Zhu, M., Ding, H., Han, J.: Xrf55: A radio frequency dataset for human indoor action analysis. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **8**(1), 21 (Mar 2024). https://doi.org/10.1145/3643543
30. Xu, Q., Cao, R., Shen, X., Du, H., Wang, S., Yu, X.: M3gym: A large-scale multimodal multi-view multi-person pose dataset for fitness activity understanding in real-world settings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12289–12300 (June 2025)
31. Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C.X., Xie, L.: Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 18756–18768. Curran Associates, Inc. (2023)
32. Yang, J., Tang, S., Xu, Y., Zhou, Y., Xie, L.: Maskfi: Unsupervised learning of wifi and vision representations for multimodal human activity recognition. arXiv preprint arXiv:2402.19258 (2024)
33. Zhang, B., Zhou, Z., Jiang, B., Zheng, R.: Super: Seated upper body pose estimation using mmwave radars. In: 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI). pp. 181–191 (May 2024). https://doi.org/10.1109/IoTDI61053.2024.00020
34. Zhang, T., Fu, Q., Ding, H., Wang, G., Wang, F.: Carec: Continual wireless action recognition with expansion–compression coordination. Sensors **25**(15), 4706 (2025). https://doi.org/10.3390/s25154706
35. Zhang, X., Ye, Z., Zhang, J., Tian, X., Liang, Z., Yu, S.: Vst-pose: A velocity-integrated spatiotem-poral attention network for human wifi pose estimation. arXiv preprint arXiv:2507.09672 (2025)
36. Zhang, Z., Zhang, J.: Actionfi: Human action recognition via multimodal feature fusion of restructured csi and optical flow. In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 6824–6829 (Oct 2025). https://doi.org/10.1109/SMC58881.2025.11343712
37. Zhang, Z., Zhang, J.: Wi-flow: Multimodal human action recognition based on dynamic features in wifi csi and optical flow. In: 2025 IEEE Smart World Congress (SWC). pp. 1155–1162 (Aug 2025). https://doi.org/10.1109/SWC65939.2025.00184
38. Zhou, Y., Yang, J., Zou, H., Xie, L.: Tent: Connect language models with iot sensors for zero-shot activity recognition. IEEE Transactions on Mobile Computing **25**(6), 8314–8326 (June 2026). https://doi.org/10.1109/TMC.2025.3650710
39. Zhu, B., He, Z., Xiong, W., Ding, G., Huang, T., Xiang, W.: Probradarm3f: mmwave radar-based human skeletal pose estimation with probability map-guided multiformat feature fusion. IEEE Transactions on Aerospace and Electronic Systems **61**(6), 15832–15842 (Dec 2025). https://doi.org/10.1109/TAES.2025.3594328
40. Zhu, G., Zhao, D., Li, C., Zhao, M., Zhang, Z., Quan, H., Ma, H.: Master: A multimodal foundation model for human activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **9**(3), 157 (Sep 2025). https://doi.org/10.1145/3749511