WiFi-JEPA: Self-supervised Learning for WiFi-CSI 3D Human Pose Estimation

Doeon Kim[1⋆], Jungyoon Lee[2⋆], Seongsin Kim[3✉], and Seong-heum Kim[1]

[1] Department of Intelligent Semiconductors, Soongsil University, Republic of Korea
[2] Department of AI Convergence Security, Soongsil University, Republic of Korea
[3] School of AI Software, Soongsil University, Republic of Korea
{ilsin205, jungyoon}@soongsil.ac.kr, {kss0222, seongheum}@ssu.ac.kr

Abstract. WiFi Channel State Information (CSI) enables privacy-preserving human pose sensing in camera-denied environments, but existing WiFi-based pose estimators often fail under environment shifts and rely on costly camera-based annotation pipelines that limit scale. We propose **WiFi-JEPA**, a self-supervised framework that learns CSI-native representations by predicting masked latent embeddings instead of reconstructing raw CSI signals that may contain hardware-specific artifacts. WiFi-JEPA makes three contributions: (i) CSI-specific tokenization and link masking tailored to the CSI tensor over channel, time, and link (C, T, L); masking entire Tx–Rx antenna links forces the model to predict one spatial link view from others, capturing cross-link correlations informative of 3D spatial structure. (ii) A ray-tracing CSI simulation pipeline that generates diverse unlabeled CSI from randomized geometric primitives, providing scalable pre-training data without pose annotations. (iii) State-of-the-art results on Person-in-WiFi-3D: WiFi-JEPA outperforms prior WiFi-CSI baselines on both single- and multi-person 3D pose estimation under the same evaluation protocol. We also show that simulated CSI provides complementary pre-training signal to real CSI, and that four vision-native SSL objectives degrade performance below training from scratch, whereas WiFi-JEPA consistently improves downstream pose estimation.

Keywords: WiFi Sensing · 3D Human Pose Estimation · Synthetic Data · Self-supervised learning

1 Introduction

Human pose estimation (HPE) is a fundamental component of systems for health monitoring, human–computer interaction, and safety-critical perception. Despite rapid progress in camera-based HPE under line-of-sight conditions [20, 21, 27], vision is often limited in camera-denied scenarios: (i) *occlusion* by walls or obstacles, (ii) *low or no illumination*, and (iii) *privacy and regulatory* constraints that prohibit visual capture in sensitive spaces.

⋆ These authors contributed equally to this work.

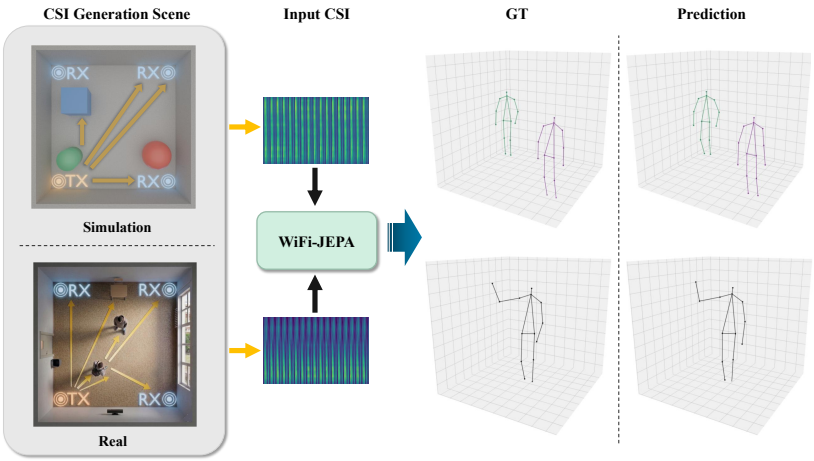

Fig. 1: Overall framework of WiFi-JEPA. *Left*: Pre-training data — sim-object and real CSI from PiW3D. *Center*: Generated CSI input and WiFi-JEPA. *Right*: GT and predicted 3D poses.

WiFi channel state information (CSI) provides a compelling alternative. WiFi signals are ubiquitous indoors and can propagate through many common obstructions. Human motion modulates multipath propagation, which is captured as CSI—a factored, complex-valued measurement over *subcarriers* (frequency), *time*, and multiple *Tx–Rx antenna links*. Prior work has shown that WiFi sensing can recover human-centric outputs without visual imagery, including fine-grained person perception and pose-related estimates [19, 23, 28]. Most notably, PiW3D [28] demonstrates the feasibility of multi-person 3D pose estimation with commodity WiFi even under visual occlusion, reporting a ∼90K-frame dataset collected in multiple indoor areas and 3D joint localization errors on the order of ∼100 mm.

However, robust WiFi-based HPE still faces three bottlenecks. First, *cross-domain generalization* is fragile: performance can degrade when deployment conditions differ from training, such as changes in transceiver placement, surrounding objects and furniture. Second, *label scalability* is limited: CSI 3D pose datasets typically rely on camera-based annotation pipelines and are collected in a small number of rooms with fixed hardware [28]. Third, CSI is *noisy and hardware-dependent*. A common approach is to reshape CSI into image-like 2D grids to leverage ViT/CNN backbones; this axis-mixing can conflate the physical semantics of subcarriers, time, and links and encourage reliance on device-specific artifacts over pose-relevant dynamics.

These challenges motivate CSI-native representation learning that (i) is robust to environment and hardware shifts, (ii) scales without dense camera-derived labels, and (iii) preserves CSI structure while exploiting multi-link spatial diversity.

We propose **WiFi-JEPA**, a self-supervised learning (SSL) framework that learns transferable CSI representations without reconstructing raw CSI. WiFi-JEPA uses a *CSI-specific tokenization* and predicts masked latent representations. We further introduce *link masking*, which masks entire Tx-Rx link observations and forces prediction from the remaining links, explicitly leveraging the multi-link spatial views intrinsic to CSI and central to pose recoverability. To reduce dependence on scarce labels, we also propose a **CSI simulation pipeline** via ray-tracing (NVIDIA Sionna [15]).

Our main contributions are:

- **WiFi-JEPA**: a CSI-native SSL framework with *CSI-specific tokenization* and *link masking* that learns CSI representations by predicting masked latent embeddings, improving *cross-domain robustness*.
- **CSI simulation pipeline** (sim-object): a ray-tracing-based pre-training paradigm providing evidence that dynamics diversity may matter more than geometric realism for transfer; geometric primitives outperform human meshes (sim-human, 100.1 vs. 110.3 mm MPJPE) and ∼90K simulated frames match ∼90K real frames in independent pre-training value.
- **State-of-the-art WiFi-based results**: combining real and simulated pre-training achieves 76.8 mm single-person and 93.5 mm multi-person MPJPE on PiW3D, improving over all prior WiFi HPE baselines under the same evaluation protocol [9, 28, 29].

2 Related Work

2.1 Self-supervised representation learning

Self-supervised learning has evolved through three paradigms: contrastive methods (SimCLR [7], MoCo v3 [8]), momentum-teacher approaches (BYOL [13], DINO [6]), and masked prediction. MAE [14] reconstructs masked inputs in pixel space, retaining any low-level noise, whereas JEPA [3] predicts targets in a learned latent space that need not preserve such artifacts. This latent objective is particularly relevant to WiFi CSI, where raw measurements carry hardware-specific distortions (clock offsets, quantization noise) irrelevant to downstream sensing. Structured masking has proven effective beyond images: V-JEPA [2, 5] showed that structured spatio-temporal block masking outperforms random masking for learning temporal structure in video.

Several recent works apply SSL to WiFi Channel State Information (CSI). SSLCSI [26] systematically benchmarked four categories of SSL algorithms for CSI-based activity recognition. CIG-MAE [18] learns to allocate masking based on per-patch information density but does not differentiate masking strategies across distinct physical axes of CSI. AM-FM [30] scales WiFi pre-training to 9.2 M samples across nine downstream tasks including activity recognition, gesture recognition, and WiFi imaging; however, none addresses multi-person 3D skeletal pose estimation. In the broader wireless domain, WirelessJEPA [10]

applies JEPA to raw multi-antenna in-phase/quadrature streams for communication and RF classification; its inputs and objectives differ fundamentally from estimated CSI for human sensing. Many CSI-SSL methods adapt image-domain augmentation or masking schemes without explicitly modeling the distinct physical semantics of CSI's frequency, spatial, and temporal axes.

2.2 WiFi-based Human Pose Estimation

Human motion modulates WiFi multipath propagation, enabling through-wall sensing without cameras. WiPose [16] first demonstrated single-person 3D WiFi pose estimation with a CNN–LSTM architecture. PiW3D [28] introduced the first multi-person 3D benchmark with a Transformer-based pose decoder. MetaFi++ [29] processed each receiver through a shared CNN and fused features via Transformer self-attention. HPE-Li [12] used dual selective-kernel CNNs with teacher–student distillation. These supervised methods all require synchronized visual supervision (motion capture or camera-based pose annotations), limiting data scalability.

To our knowledge, DT-Pose [9] is the only prior SSL method for WiFi pose estimation. It combines MAE pretraining with temporal contrastive learning and uniformity regularization, and uses a GCN–Transformer decoder that enforces skeleton topology constraints. Its tokenization flattens CSI into a 2D grid with fixed sinusoidal positional encoding, jointly encoding all three axes into a single sequence without preserving their independence. Because DT-Pose further employs a decoder with explicit skeletal-topology constraints (an inductive bias absent from our PETR-based decoder), we control decoder architecture and isolate the effect of the pre-training objective in Sec. 5.4.

2.3 Simulation data for representation learning

Domain randomization [22] showed that training on diverse synthetic data with randomized scene parameters enables sim-to-real transfer without photorealistic rendering. FractalDB [17] and Dead Leaves [4] extended this idea to representation learning, showing that vision backbones pretrained on procedurally generated images can learn transferable features, confirming that structural diversity matters more than geometric realism. This principle is particularly relevant for CSI, where multipath fading causes even simple moving scatterers to produce rich channel variations that encode spatial dynamics—making the diversity of motion patterns more informative than the geometric fidelity of the scatterer itself. In the wireless domain, ray-tracing tools such as Sionna [15] and synthetic channel frameworks like DeepMIMO [1] can generate physically grounded channel data and have been used for supervised tasks (channel estimation, beam prediction), but not for self-supervised pretraining of sensing models. Across these three threads, no prior work combines axis-aware masking with latent-space prediction tailored to the physical structure of CSI, nor has ray-tracing simulation been used for self-supervised pretraining in WiFi sensing.

3 Simulated CSI from Geometric Primitives

Collecting large-scale WiFi CSI datasets is expensive, requiring dedicated indoor environments, synchronized receivers, and co-located motion capture for labeling. We bypass this bottleneck with a simulation pipeline that generates benchmark-compatible CSI from scenes of simple geometric shapes—no human models or external motion data required. We refer to this geometric-primitive dataset as sim-object throughout; a human-mesh variant (sim-human) is evaluated as a comparison in Sec. 5.3.

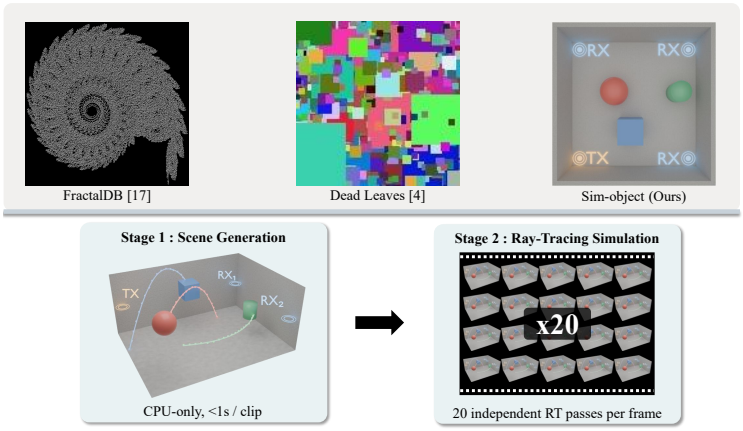

Fig. 2: sim-object pipeline. **Top:** Analogy to FractalDB and Dead Leaves—geometric primitives replace human models. **Bottom:** Stage 1 generates randomized scenes (CPU-only); Stage 2 runs Sionna RT with $\times 20$ independent passes per frame.

Inspired by the success of non-semantic pre-training in vision (Sec. 2.3), we randomize room geometry, object count, trajectory, and wall materials to cover a wide range of channel conditions.

Hypothesis. We hypothesize that SSL pre-training may not require CSI from real human poses; rather, it needs exposure to *how* WiFi signals vary across time, frequency, and space. At 5.64 GHz (wavelength ≈ 5 cm), a sphere and a human body differ substantially in scattering behavior—the body is articulated, non-convex, and has complex dielectric properties. Nevertheless, SSL pre-training may still succeed with geometric primitives, because the learning objective benefits from diverse spatio-temporal channel variation, rather than faithful reproduction of body-specific multipath. Specifically, we test whether *dynamics diversity*—the range of positions, velocities, and directions across training scenes—matters more than geometric fidelity in Sec. 5.3.

Stage 1: Scene generation. A pure-Python generator (CPU-only, <1 s/clip) creates randomized indoor scenes. Room dimensions are sampled uniformly (3–8 m per side, 2.5–4 m height), with walls assigned ITU-standard materials (concrete,

plasterboard, wood, glass). Each scene contains 1–4 geometric primitives (sphere, cube, cylinder, ellipsoid; radius 0.1–0.5 m) that follow physics-based trajectories: initial velocities are sampled up to 3 m/s—exceeding typical indoor speeds per [22]—with elastic wall reflections producing varied spatial coverage. A single transmitter and three receivers are placed at randomized wall positions matching the SIMO configuration of the target benchmark (1 Tx, 3×3 Rx antennas = 9 links). The real PiW3D training set contains ∼90K frames; our 90K simulated frames thus constitute a comparable pre-training corpus.

Stage 2: Ray-tracing simulation. NVIDIA Sionna RT [15] simulates radio propagation in each scene, configured to match the target benchmark. A critical design choice is performing **20 independent** ray-tracing passes per frame: each pass computes the channel for a static scene snapshot, so we sample object positions at 20 sub-frame locations within each measurement window (∼50 ms). Without this, a single pass yields a near-constant time axis, rendering temporal masking ineffective; 20 passes produce realistic temporal variation matching real data. The output $H \in \mathbb{C}^{20\times30\times3\times3}$ (time × 30 subcarrier groups × 3 receivers × 3 antennas per Rx) is decomposed into amplitude and phase, yielding a real-valued tensor of shape $(60, 20, 9)$: 60=2×30 subcarrier channels, 20 time samples, and 9=3 Rx×3 ant links—analogous to a multi-channel spectrogram—for the tokenizer (Sec. 4.1). Generating 90K frames takes ∼10 GPU-hours on one RTX 4090.

4 WiFi-JEPA

4.1 CSI-specific Tokenization

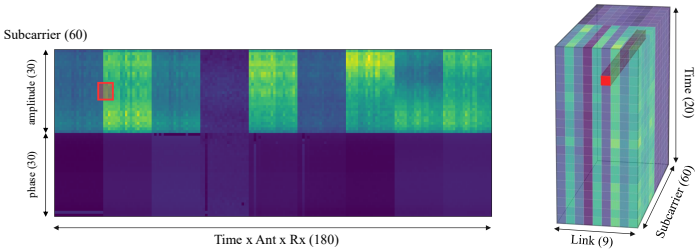

Fig. 3: (a) Flattening mixes temporal and link dimensions, causing patches to cross physical boundaries. (b) Our CSI-specific tokenization keeps (T, L) separate so each token corresponds to a specific spatio-temporal coordinate.

The raw CSI data in the PiW3D dataset [28] is a tensor of shape $N_{rx} \times N_{ant} \times T \times N_c = 3\times3\times20\times30$, where N_{rx} is the number of receivers, N_{ant} is the number of antennas per receiver, T is the number of time steps, and N_c is the number of subcarriers. Following [28], we use amplitude directly and denoise phase values using PhaseFi [24], resulting in a real-valued tensor of shape $3 \times 3 \times 20 \times 60$, where $60 = 2 \times N_c$ (30 amplitude + 30 denoised phase channels).

As illustrated in Fig. 3, PiW3D [28] flattens CSI tensor into a 2D spectrogram of shape $(60, 180)$. Although compatible with standard ViTs, this format mixes temporal sequences with spatially distinct antenna-links, blurring physical boundaries. With 6×6 patch embedding on the flattened grid, patches frequently cross link boundaries and the amplitude-phase boundary, merging physically disparate components within a single token.

CSI-specific tokenization preserves the structural integrity of the signal by reshaping the tensor into $(C, T, L) = (60, 20, 9)$, where the $L = 9$ represents independent spatial links ($3R_x \times 3Ant$), each observing the same physical scene from a different spatial viewpoint. The subcarrier features (C=60) serve as the channel dimension and T denotes time steps. We apply a linear embedding with patch size $(1, 1)$ over the (T, L) dimensions, producing $T \times L = 180$ tokens, each representing a precise spatial-temporal coordinate. A separable 2D sinusoidal positional embedding is added.

4.2 Link Masking

In the JEPA framework, the masking strategy defines the pretext task and thus dictates the semantic quality of the acquired representations.

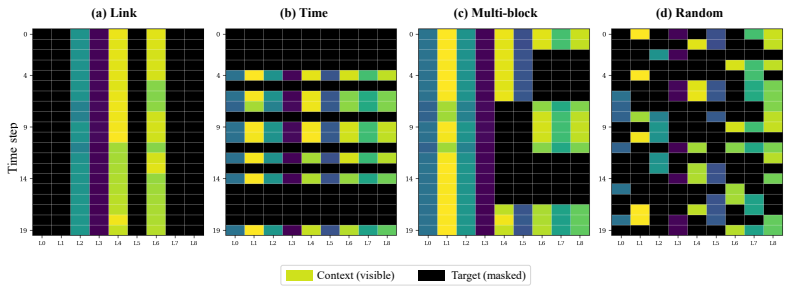

Fig. 4: Four masking strategies on the (T, L) token grid. The vertical axis corresponds to the 20 time steps, and the horizontal axis represents the 9 antenna links.

We propose *link masking*, which masks entire columns of the (T, L) token grid, as shown in Fig. 4(a). Given a masking ratio r=0.6, we randomly select $\text{round}(r \times L) = 5$ of the 9 links as targets and use the remaining 4 links as context. All time steps of each masked link are removed simultaneously, compelling the model to predict the complete temporal signal of unseen antenna-receiver pairs from the remaining links. This aligns with the intuition of multi-view learning: as the 9 links observe the same physical scene from different viewpoints, the model must exploit cross-link spatial redundancy to reconstruct the missing signals.

To validate this hypothesis, we compare link masking against three alternatives: (i) *time masking*, which masks entire rows to test whether temporal interpolation alone suffices; (ii) *random block masking*, which masks contiguous rectangular regions of the (T, L) grid following I-JEPA [3]; and (iii) *random*

masking, which independently selects individual tokens regardless of grid position. Ablation results in Sec. 5.4 substantiate our approach: link masking consistently outperforms all alternatives, confirming that cross-link spatial correlation provides the strongest pretext signal.

4.3 Pre-training Objective

As illustrated in Phase 1 in Fig. 5, our architecture follows the JEPA framework [3] with three components: a context encoder f_θ, a predictor p_ϕ, and a target encoder $f_{\bar{\theta}}$. After tokenization and link masking, the context encoder takes the masked sequence as input, processing only the visible tokens $\mathbf{x}_{\mathcal{C}_i}$. The predictor receives these context embeddings concatenated with learnable mask tokens $\mathbf{p}$ at positions $\mathcal{M}_i$ and predicts the target representations. The target encoder processes all tokens, and only the representations at $\mathcal{M}$ are used as prediction targets.

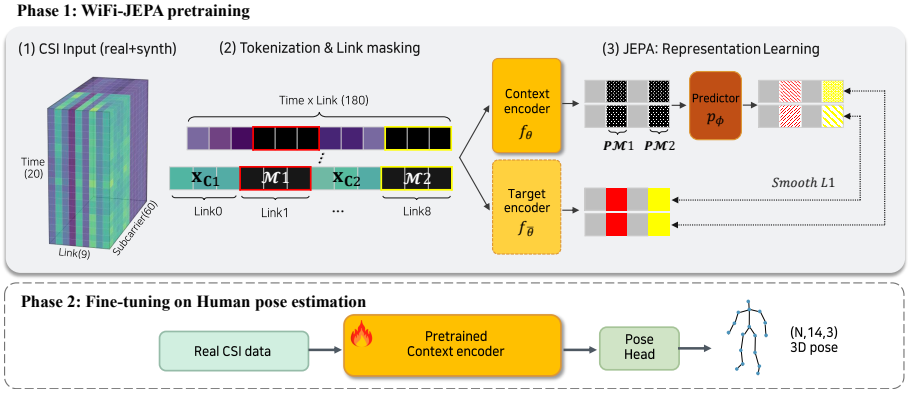

Fig. 5: WiFi-JEPA architecture. **Phase 1** (top): self-supervised pre-training with link masking on the (T, L) token grid. **Phase 2** (bottom): supervised fine-tuning with a PETR decoder for multi-person 3D pose estimation.

We minimize the Smooth L1 loss between predicted and target embeddings:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{SmoothL1}(\hat{z}_i,\ \text{sg}[\text{LN}(f_{\bar{\theta}}(\mathbf{x})_i)]) \,, \quad \hat{z}_i = p_\phi(f_\theta(\mathbf{x}_\mathcal{C}),\ \mathbf{p}_\mathcal{M})_i \tag{1}$$

where $\mathcal{M}$ and $\mathcal{C}$ denote the masked and context token index sets, $\mathbf{p}_\mathcal{M}$ are the positional embeddings at masked locations, $\text{sg}[\cdot]$ denotes stop-gradient, and LN is layer normalization applied to the target embeddings. The target encoder is updated via EMA with cosine momentum schedule from 0.996 to 1.0.

Both the context and target encoders share the same ViT backbone. We use a shallower predictor to limit its capacity and encourage the context encoder to learn more transferable, pose-relevant features. Architecture specifications and training details are provided in Sec. 5.1.

4.4 Fine-tuning for 3D Pose Estimation

We integrate the pre-trained JEPA encoder into PETR [20], with a linear projection layer mapping the encoder output from 512 to 256 dimensions to match the decoder operating dimension. During fine-tuning (Phase 2 in Fig. 5), we train the model end-to-end on real CSI data with differential learning rates. The pre-trained JEPA encoder is fine-tuned with $0.1\times$ the base learning rate to retain pretrained features while adapting to the task, whereas the PETR decoder and regression heads use the $1.0\times$ base learning rate.

The model predicts 14 body keypoints in 3D coordinates for N detected persons ($N \times 14 \times 3$). Hungarian matching assigns queries to instances; the decoder refines predictions via cross-attention over encoded CSI features. Loss functions and hyperparameter settings are detailed in Sec. 5.1.

5 Experiments

We evaluate WiFi-JEPA on the PiW3D dataset [28], the only public WiFi-CSI benchmark that includes multi-person 3D pose annotations. Our experiments address three questions: (i) How much does WiFi-JEPA advance the state of the art in WiFi-based 3D pose estimation? (ii) Why JEPA over alternative SSL objectives, and do the proposed structure-aware tokenization and link masking contribute to downstream performance? (iii) Can sim-object complement or even replace real data for pre-training?

5.1 Experimental Setup

Dataset. PiW3D [28] provides synchronized WiFi-CSI and 3D pose labels (14 joints) captured in three indoor environments—an office, a classroom, and a corridor—each measuring approximately $4\,\mathrm{m} \times 3.5\,\mathrm{m}$, using one transmitter and three Intel 5300 receivers (three antennas each) at 5.64 GHz with 30 subcarriers. Seven volunteers performed eight daily actions (e.g., walking, stretching, bending over, sitting down) across these environments, yielding diverse multipath conditions. Each CSI sample has raw dimensions $1\times3\times3\times30\times20$ (#Tx, #Rx, #Ant, #Subcarrier, #Time). The training set contains 89,946 frames; the test set contains 7,824 frames spanning single-person (2,586), two-person (3,184), and three-person (2,054) scenarios.

Metrics. We report Mean Per-Joint Position Error (**MPJPE**, mm) as the primary metric—the mean Euclidean distance between predicted and ground-truth 3D joint positions. We additionally report Procrustes-aligned MPJPE (**PA-MPJPE**), which removes global translation, rotation, and scale to isolate articulated pose accuracy, and Percentage of Correct Keypoints (**PCK@**τ) at τ=20, 50 mm. For finer-grained analysis, we also report per-dimension absolute errors along the horizontal (x), vertical (y), and depth (z) axes, following [28].

Table 1: Comparison with prior WiFi-CSI pose estimation methods on PiW3D dataset. Best in **bold**, second-best underlined. "–" (gray) indicates metrics not reported in the original paper. † Values taken directly from the original publications.

Method	Venue	Single-person				Multi-person			
		MPJPE↓	PA↓	PCK@20↑	PCK@50↑	MPJPE↓	PA↓	PCK@20↑	PCK@50↑
WiPose [16]	MobiCom'20	101.8†	–	–	–	–	–	–	–
MetaFi++ [29]	IoT-J'23	132.0†	75.8†	62.0	89.3	–	–	–	–
HPE-Li [12]	ECCV'24	120.2†	69.5†	59.1	87.5	–	–	–	–
DT-Pose [9]	arXiv'25	90.0†	58.7†	72.1	90.6	–	–	–	–
PiW3D (baseline) [28]	CVPR'24	91.7†	55.1	69.3	95.2	107.2†	65.6	58.1	91.8
WiFi-JEPA (real)	–	78.2	**53.9**	74.5	96.6	97.1	67.2	59.3	93.0
WiFi-JEPA (real+sim)	–	**76.8**	54.0	**75.9**	**96.7**	**93.5**	**65.1**	**61.5**	**93.7**

Implementation details. All training is conducted on a single NVIDIA RTX 4090 GPU. **Pre-training:** The context and target encoders are (12 layers, embedding dimension 512, 8 heads), the predictor is a ViT (8 layers, 512 dimensions, 16 heads). The target encoder is updated via EMA with momentum scheduled from 0.996 to 1.0. We pre-train for 100 epochs with AdamW (batch size 64, learning rate 5×10^{-4}, weight decay 0.05), using Smooth L1 loss with layer-normalized targets. Link masking randomly drops 60% of antenna links per sample. Pre-training data comprises ∼90K real frames from PiW3D and ∼90K synthetic frames from sim-object (Sec. 3); the real-only variant uses PiW3D frames alone. **Fine-tuning:** We attach the pre-trained WiFi-JEPA context encoder to a PETR-style Transformer decoder (5 layers, 256 dimensions, 8 heads) and fine-tune for 100 epochs with AdamW (base learning rate 2×10^{-5}, weight decay 10^{-4}). The encoder learning rate is scaled to $0.1\times$ the base rate.

5.2 Main Results

Comparison with State-of-the-Art We compare against five WiFi-CSI pose estimation methods (see Sec. 2.2 for architectural details): WiPose [16], HPE-Li [12], MetaFi++ [29] (single-person), DT-Pose [9] (the only prior SSL method), and PiW3D [28] (the first multi-person 3D method).

Table 1 summarizes the results. We note that WiPose, MetaFi++, HPE-Li, and DT-Pose do not report multi-person (MP) results; we compare them only in the single-person (SP) scene. WiFi-JEPA (real+sim) achieves 76.8 mm single-person MPJPE, outperforming the previous best method DT-Pose (90.0 mm) by 13.2 mm (−14.7%) and the PiW3D baseline (91.7 mm) by 14.9 mm (−16.2%). PCK@20 also increases from 72.1% (DT-Pose) to 75.9%. In multi-person scenes—where only PiW3D reports prior results—WiFi-JEPA reduces MPJPE from 107.2 mm to 93.5 mm (−12.8%) and raises PCK@20 from 58.1% to 61.5%. Notably, the real-only pre-training variant already surpasses all baselines (78.2 mm SP, 97.1 mm MP), and adding simulation data yields a further 1.4 mm and 3.6 mm reduction in SP and MP respectively, demonstrating that sim-object provides complementary pre-training signal.

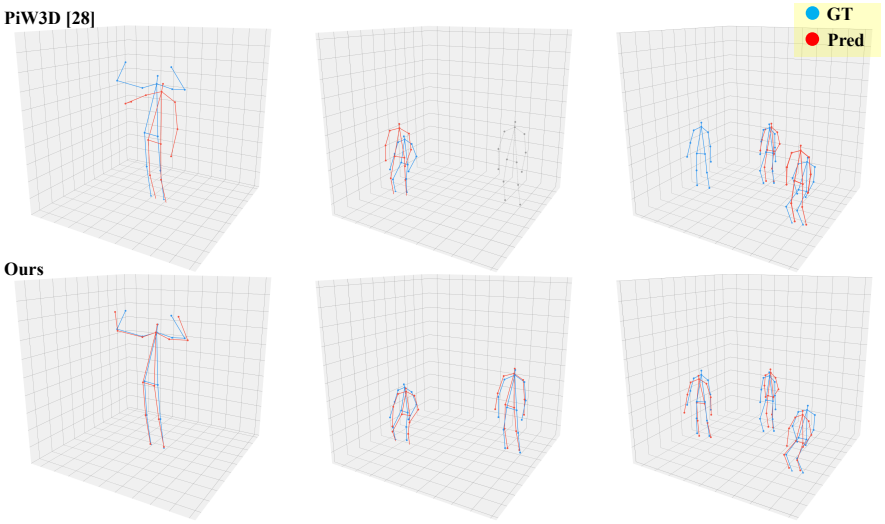

Fig. 6: Qualitative comparison of 3D pose estimation. Top row: PiW3D baseline [28]. Bottom row: WiFi-JEPA (ours). From left to right: 1-person, 2-person, and 3-person scenes.

Challenging Scenarios

Cross-Domain Generalization WiFi CSI amplitude profiles vary across environments due to multipath propagation differences. For instance, the mean amplitude on a single receiver–antenna-link ranges from 5.7 in Office to 20.7 in Corridor (Fig. 7), with similar variation across all nine links. To test whether WiFi-JEPA representations transfer across environments, we conduct a leave-one-environment-out (LOO) evaluation: both pre-training and fine-tuning use data from two environments only, and the model is tested on the held-out environment (Table 2). WiFi-JEPA achieves 248.4 mm on Office, 428.2 mm on Class-

Table 2: Leave-one-environment-out cross-domain evaluation (MPJPE, mm). Off./Class./Corr.: held-out test environment (Office/Classroom/Corridor). Both pre-training and fine-tuning exclude the held-out environment.

Method	Metric	Off.	Class.	Corr.
WiFi-JEPA (ours)	MPJPE	248.4	428.2	296.0
	Mean		**324.2**	
PiW3D (baseline)	Mean		**626.4**	

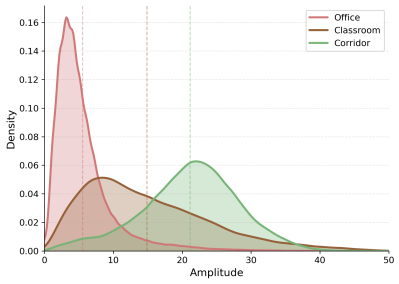

Fig. 7: Amplitude distribution shift across 3 environments.

room, and 296.0 mm on Corridor, yielding a mean MPJPE of 324.2 mm—a 48.2% reduction over the PiW3D baseline (626.4 mm). While absolute errors remain high, WiFi-JEPA halves the baseline error. This is consistent with the masking ablation in Section 5.4, where link masking consistently outperforms alternative strategies, indicating that cross-link correlations are key to environment-robust representations. Furthermore, we evaluate WiFi-JEPA against a standard supervised Domain Adaptation (DA) baseline, DANN [11], in a few-shot cross-environment setup utilizing only 10% of target environment labels. WiFi-JEPA consistently outperforms Supervised+DANN across all held-out environments, reducing the macro mean MPJPE from 208.6 mm to 171.3 mm (a 17.9% error reduction). This demonstrates that our SSL pre-training yields substantial generalization benefits that go beyond standard adversarial DA techniques.

Table 3: Per-person-count MPJPE (mm). Performance degrades gracefully as the number of simultaneous persons increases.

Metric	Method	1P	2P	3P
MPJPE↓	PiW3D	91.7	108.1	125.3
	Ours	**76.8**	**96.0**	**110.7**

Table 4: Extremity error breakdown. Mean L/R elbow and hand errors with directional decomposition (mm). PiW3D: Person-in-WiFi-3D baseline [28].

Metric	Elbows		Hands	
	PiW3D	Ours	PiW3D	Ours
MPJPE↓	128.5	**60.6**	192.5	**68.7**
Horiz↓	68.2	**31.1**	94.4	**34.4**
Vert↓	56.8	**35.0**	104.8	**40.0**
Depth↓	73.9	**23.2**	92.9	**26.4**
Improv.	**↓52.8%**		**↓64.3%**	

Multi-Person Scenes and Per-joint Analysis A known limitation of WiFi-CSI pose estimation is performance degradation in multi-person scenarios, where overlapping signals make it difficult to disentangle individual poses. WiFi-JEPA improves across all person counts: single-person MPJPE drops from 91.7 mm to 76.8 mm (−16.2%), two-person from 108.1 mm to 96.0 mm (−11.2%), and three-person from 125.3 mm to 110.7 mm (−11.7%) (Table 3).

The PiW3D baseline struggles most with elbows and hands (160.5 mm), because arms exhibit faster, more varied motion and occupy a smaller spatial cross-section [28]. WiFi-JEPA reduces this to 64.7 mm, a 59.7% improvement (Table 4), with depth error dropping from 92.9 mm to 26.4 mm for hands.

5.3 Effect of Simulated CSI

We now examine the hypothesis from Sec. 3. Table 5 presents a dataset-level comparison: all rows use the same WiFi-JEPA encoder, PETR decoder, and fine-tuning protocol; only the pre-training data varies. We additionally compare

against sim-human, a human-mesh variant that uses Blender-animated characters with motion-captured animation clips instead of geometric primitives. Both simulated datasets use the same number of clips and are sampled to 90K frames, so the comparison isolates the effect of scene content: geometric primitives with randomized physics trajectories (sim-object) versus animated human meshes with motion-capture sequences (sim-human).

Table 5: Effect of pre-training data on downstream MPJPE (mm). All rows share the same encoder, decoder, and fine-tuning protocol. Δ: improvement over no pre-training.

Pre-train data	Frames		MPJPE↓	Δ
	Real	Sim		
None (scratch)	—	—	102.4	—
sim-object	—	90K	100.1	−2.3
sim-human	—	90K	110.3	+7.9
Real only	90K	—	97.1	−5.3
Real + sim-object	45K	45K	97.2	−5.2
Real + sim-object	90K	90K	**93.5**	**−8.9**

sim-object suffices for pre-training. Pre-training on sim-object (100.1 mm) outperforms sim-human (110.3 mm), with the latter causing *negative transfer* (+7.9 mm worse than no pre-training), likely because the fixed motion-capture sequences lack the trajectory variation needed to learn generalizable channel features. Because sim-human uses anatomically realistic scatterers yet performs worse, the result directly supports the hypothesis that trajectory diversity—randomized physics with varied velocities and elastic reflections—is more important than scatterer realism for CSI pre-training. Furthermore, sim-object alone (100.1 mm) is close to real-data-only pre-training (97.1 mm), with a gap of only 3.0 mm (2.3 vs. 5.3 mm reduction from scratch). That is, ∼90K simulated frames provide comparable pre-training value to ∼90K real frames.

Complementarity of Simulated and Real Data. Combining 90K simulated and 90K real frames reduces MPJPE by 8.9 mm over training from scratch, compared with a 5.3 mm reduction from real-only pre-training. This suggests that the simulation pipeline provides channel variation patterns absent from the real dataset, such as diverse room geometries and wall materials. The simulated data is generated at a cost of ∼10 GPU-hours on one RTX 4090 (Sec. 3), while the real frames simply reuse the existing training set as unlabeled data.

5.4 Effect of WiFi-JEPA Pre-training

Table 6: All methods use the same ViT backbone and PETR-style decoder, pre-trained on ∼90K real frames for 100 epochs, then fine-tuned identically. Best in **bold**, second-best underlined.

Method	SSL Type	MPJPE↓	PA↓	PCK@20↑	PCK@50↑
SimMIM [25]	MIM	145.6	84.0	41.5%	84.3%
MAE [14]	MIM	130.3	77.4	47.6%	87.4%
BYOL [13]	Self-distill	144.4	83.2	42.5%	84.1%
MoCoV3 [8]	Contrastive	141.5	82.6	43.4%	84.9%
WiFi-JEPA (w/o pretrain)	–	102.4	**65.1**	57.8%	91.8%
WiFi-JEPA (w/ pretrain)	JEPA	**97.1**	67.2	**59.3%**	**93.0%**

Comparison with SSL Baselines WiFi-JEPA outperforms all baselines in MPJPE and PCK by a wide margin (Table 6). The gap to the strongest competitor, MAE, is 33.2 mm MPJPE (130.3 → 97.1) and +11.7 pp in PCK@20 (47.6% → 59.3%). PA-MPJPE remains comparable to the no-pre-training baseline, indicating that gains concentrate in global localization rather than relative joint proportions. All four baselines actually *degrade* performance relative to no pre-training (102.4 mm). MIM methods must reproduce raw CSI values including hardware artifacts, which may bias the encoder toward device-specific patterns. Self-distillation and contrastive methods require augmentations that preserve channel structure, however standard image augmentations lack physical grounding for CSI. These negative results may partly reflect the difficulty of adapting vision-native methods to a new modality. WiFi-JEPA avoids these pitfalls by predicting in a learned latent space and exploiting the factored (C, T, L) structure of CSI.

Ablation Studies *Masking strategies:* We compare four masking strategies described in Sec. 4.2: random (individual patches), multi-block [3] (contiguous rectangles), temporal (entire time rows), and link (entire receiver–antenna columns). All variants are pre-trained on ∼90K real frames for 100 epochs with the same hyperparameters, then fine-tuned (Table 7). Link masking achieves the

Table 7: Masking strategy ablation. All variants use the same backbone and training schedule, pre-trained on ∼90K real frames for 100 epochs.

Method	MPJPE↓	PA-MPJPE↓	PCK@50↑
Random	105.14	69.30	88.48%
Multi-block	102.06	67.71	90.54%
Time	104.24	68.25	88.22%
Link	**97.10**	**67.20**	**93.00%**

Table 8: Effect of structure-aware tokenization. w/o: 2D spectrogram (1×60×180); w/: CSI-specific tokenization (60×20×9).

Metric	w/o csi-tok.	w/csi-tok.
PA-MPJPE↓	70.98	**67.20**
MPJPE↓	111.85	**97.10**

lowest MPJPE (97.1 mm) and highest PCK@50 (93.0%), outperforming random masking by 8.04 mm and 4.52 pp respectively. *CSI-specific tokenization:* Table 8 compares our CSI-specific tokenization against a 2D spectrogram baseline that flattens the CSI tensor into a single-channel 60×180 image with 6×6 patches. Both models use the same WiFi-JEPA and link masking. CSI-specific tokenization reduces MPJPE by 14.75 mm (111.85→97.1) and PA-MPJPE by 3.78 mm, confirming that preserving the physical axis semantics of subcarrier, time, and link enables more effective representation learning.

6 Conclusion

We presented WiFi-JEPA, a self-supervised framework for WiFi-CSI-based 3D human pose estimation that predicts masked latent embeddings on the factored (C, T, L) CSI tensor. Structure-aware tokenization and link masking capture cross-link spatial correlations central to body localization, while ray-tracing simulation provides scalable pre-training data without annotation. Across single- and multi-person 3D pose estimation on PiW3D, WiFi-JEPA sets a new state of the art and nearly halves cross-environment error, and it consistently improves over training from scratch, whereas four vision-native SSL objectives degrade it. Notably, simulated CSI from simple moving primitives offers pre-training value comparable to real data, indicating that the diversity of channel dynamics matters more than the geometric realism of the scatterer.

Limitations. WiFi-JEPA improves global joint localization more than relative joint configuration (PA-MPJPE stays close to the from-scratch baseline). Cross-environment error, though substantially reduced, remains far higher than same-environment performance. Our evaluation is confined to a single dataset with fixed hardware, so generalization across antenna configurations and frequency bands remains open.

More broadly, CSI-native self-supervised learning, which respects the physical structure of the wireless channel rather than treating CSI as a generic image, is a promising direction for WiFi sensing beyond pose estimation.

Acknowledgements This work was supported by the G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317); the Ministry of Science and ICT (MSIT) and the National IT Industry Promotion Agency (NIPA) through the Advanced GPU Utilization Support Program (02-26-01-0499); the NRF grant funded by the MSIT (RS-2025-16071992); the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea government(MOTIE) (RS-2026-25530975, HRD Program for Industrial Innovation); the MSIT under the Convergence security core talent training business support program(IITP-2024 2024-RS-2024-00426853) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation); and the IITP grant funded by the MSIT (No. 2022-0-00986, Development of artificial intelligence-based base station electromagnetic wave human exposure prediction algorithm).

References

1. Alkhateeb, A.: DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. arXiv preprint arXiv:1902.06435 (2019)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15619–15629 (2023)
4. Baradad, M., Wulff, J., Wang, T., Isola, P., Torralba, A.: Learning to see by looking at noise. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
5. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)
8. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9640–9649 (2021)
9. Chen, Y., Guo, J., Guo, S., Zhou, J., Tao, D.: Towards robust and realistic human pose estimation via WiFi signals. arXiv preprint arXiv:2501.09411 (2025)
10. Chu, V., Mashaal, O., Abou-Zeid, H.: WirelessJEPA: A multi-antenna foundation model using spatio-temporal wireless latent predictions. arXiv preprint arXiv:2601.20190 (2026)
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016)
12. Gian, T.D., Lai, T.D., Luong, T.V., Wong, K.S., Nguyen, V.D.: HPE-Li: WiFi-enabled lightweight dual selective kernel convolution for human pose estimation. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15089, pp. 93–111. Springer (2024)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent – a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 21271–21284 (2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)

15. Hoydis, J., Ait Aoudia, F., Cammerer, S., Nimier-David, M., Binder, N., Marcus, G., Keller, A.: Sionna RT: Differentiable ray tracing for radio propagation modeling. arXiv preprint arXiv:2303.11103 (2023)
16. Jiang, W., Xue, H., Miao, C., Wang, S., Lin, S., Tian, C., Murali, S., Hu, H., Sun, Z., Su, L.: Towards 3D human pose construction using WiFi. In: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom). pp. 1–14. ACM (2020)
17. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. In: Asian Conference on Computer Vision (ACCV). pp. 583–600 (2020)
18. Liu, G., Hao, Y., Zou, Y.: CIG-MAE: Cross-modal information-guided masked autoencoder for self-supervised WiFi sensing. arXiv preprint arXiv:2512.04723 (2025)
19. Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y., Yang, J.: GoPose: 3D human pose estimation using WiFi. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **6**(2), 1–25 (2022)
20. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11069–11078 (2022)
21. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5693–5703 (2019)
22. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 23–30 (2017)
23. Wang, F., Zhou, S., Panev, S., Han, J., Huang, D.: Person-in-WiFi: Fine-grained person perception using WiFi. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5452–5461 (2019)
24. Wang, X., Gao, L., Mao, S.: PhaseFi: Phase fingerprinting for indoor localization with a deep learning approach. In: IEEE Global Communications Conference (GLOBECOM). pp. 1–6 (2015)
25. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMIM: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9653–9663 (2022)
26. Xu, K., Wang, J., Zhu, H., Zheng, D.: Evaluating self-supervised learning for WiFi CSI-based human activity recognition. ACM Transactions on Sensor Networks **21**, 21:1–21:38 (2025). https://doi.org/10.1145/3715130
27. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 38571–38584 (2022)
28. Yan, K., Wang, F., Qian, B., Ding, H., Han, J., Wei, X.: Person-in-WiFi 3D: End-to-end multi-person 3D pose estimation with Wi-Fi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 969–978 (2024)
29. Zhou, Y., Huang, H., Yuan, S., Zou, H., Xie, L., Yang, J.: MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation. IEEE Internet of Things Journal **10**(16), 14128–14136 (2023)
30. Zhu, G., Hu, Y., Jayaweera, S., Gao, W., Wang, W.H., Zhang, J., Wang, B., Wu, C., Liu, K.J.R.: AM-FM: A foundation model for ambient intelligence through WiFi. arXiv preprint arXiv:2602.11200 (2026)