W-JEPA: WAVELET JOINT-EMBEDDING PREDICTIVE ARCHITECTURE FOR WEARABLE ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

A major challenge in wearable device activity recognition is the heterogeneity across datasets in sensor placement, channel count, sampling rate and subjects. Moreover, the motion patterns that distinguish activities often appear as subtle, scale-specific variations within otherwise similar waveforms. As a result, raw signals are often ambiguous, and models trained on them tend to rely on cues specific to a particular sensor setup or participant, which limits both cross-dataset transfer and generalization to unseen participants. Addressing this requires representations that preserve such activity-relevant, scale-specific variation. We introduce W-JEPA, a self-supervised framework that exploits the multiscale structure of inertial motion and leverages a joint-embedding predictive architecture, by organizing windows into a fixed grid of wavelet tokens indexed by scale, sensor axis, and time. These wavelet tokens help expose activity-specific detail within broadly similar raw signals, and serve as the units of masked latent prediction. W-JEPA predicts masked token representations supplied by an exponential-moving-average (EMA) target encoder. Because recovering a masked token requires relating motion at one scale and axis to motion at other scales and axes, the model learns the cross-scale, cross-axis structure of human movement. Across six wearable activity benchmarks, W-JEPA outperforms state-of-the-art methods under random-window splits and remains the strongest when subjects are held out, improving average macro-F1 by 8.7 percentage points over the strongest published baseline. These results indicate that physically grounded multiscale tokenization, combined together with latent predictive learning, provides a promising foundation for transferable representations of heterogeneous wearable inertial data.

1 INTRODUCTION

Public wearable datasets provide substantial unlabeled recordings for learning representations of human motion, but they differ in sensor placement, channel count, sampling characteristics, participants, and activity vocabularies (Chen et al., 2021). A useful representation must therefore capture which activity is being performed, independent of how and on whom it was recorded. This is difficult because the information that separates activities is often small. Similar physical activities can produce nearly the same rhythmic, periodic pattern in raw sensor recordings and differ mainly in the finer temporal structure within each movement cycle (Kim et al., 2021; Tian et al., 2020). Because this detail is small relative to differences in sensor setup and individual movement style, a model trained on raw signals can fit its training data using cues tied to the setup or to particular participants. Such cues support strong performance on familiar participants but may not carry over to people unseen during training. In our experiments, every method loses at least 14 points of average macro-F1 when evaluation moves from random windows to held-out participants (§4.1). Broad movement patterns and fine activity-specific detail, however, tend to occupy different timescales. A representation that separates motion by timescale can keep this detail from being masked by larger variation in the signal, while still allowing broad patterns to inform recognition.

Figure 1 shows an example of two labels from the DSADS dataset that are easy to confuse. Panel (a) shows horizontal and vertical exercise-bike cycling from the same participant. The two raw waveforms nearly overlap, with a Pearson correlation of 0.992. Their reconstructed D_3 wavelet

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

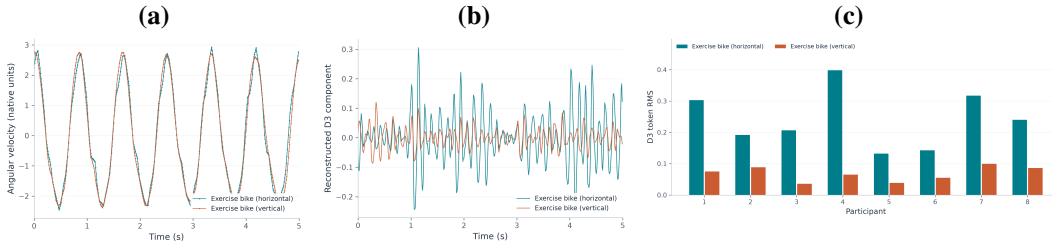


Figure 1: **Wavelet tokens reveal differences between similar activity signals.** **(a)** Sample window of horizontal and vertical exercise-bike cycling from a DSADS participant's left leg gyroscope z-axis exhibit similar raw waveforms. **(b)** The reconstructed db4 D_3 component reveals larger pronounced oscillations during horizontal cycling in this pair. **(c)** Mean RMS of the second D_3 token across 59 windows per activity and participant, with higher values for horizontal cycling in all eight participants. More details appear in Appendix D.1.

components, however, reveal pronounced oscillations during horizontal cycling (panel b). Across all eight participants, the D_3 token has higher mean RMS for horizontal cycling than for vertical cycling (panel c). The information that separates the two activities is present in the signal and becomes visible once this scale is isolated. Self-supervised pretraining could learn such distinctions from unlabeled recordings, but masked-modeling approaches for wearables typically hide and predict raw temporal patches (Ryu et al., 2025; Noor & Baraka, 2026), which mix all timescales within each interval. This raises the question we study: *how should the units of latent prediction be defined to preserve activity-specific detail while learning transferable motion representations from heterogeneous wearable signals?*

We propose W-JEPA, which answers this question by using fixed wavelet tokens as the units of masked latent prediction. A discrete wavelet transform separates each sensor axis into localized components at individual timescales (Daubechies, 1992), and W-JEPA summarizes each resulting band with tokens indexed by scale, sensor axis, and time (Figure 2a). The tokenizer is fixed, meaning no part of it is fit to the pretraining datasets. Because all recordings are resampled to a common rate, each band covers the same nominal frequency range in every dataset, so a token at a given scale carries the same physical meaning across sensor setups. The token grid grows with the number of channels, and only a learned channel embedding depends on the sensor layout. Datasets with different layouts therefore share a common set of prediction units and differ only in how many channels they contribute.

During pretraining, a random subset of tokens is masked. A context encoder processes the visible tokens, and a predictor estimates the latent features that an EMA target encoder assigns to the masked positions (Figure 2b). Predicting in latent space means the model is not asked to reproduce raw sample values, whose amplitude and phase can vary with participant and sensor placement. Masks are drawn independently across channels and bands, so recovering a masked token encourages the model to relate motion at one scale on one axis to motion at other scales on other axes. Because each scale has its own tokens, fine activity-specific detail becomes a prediction target in its own right instead of a small component of a patch dominated by larger variation.

We evaluate W-JEPA on six wearable activity benchmarks. For each target dataset, we pretrain on the remaining five and fine-tune on two to six of its participants. W-JEPA achieves the highest mean accuracy and macro-F1 on every dataset under both subject-disjoint and random-window splits. With participants held out, its average macro-F1 is 80.6, a 12.1% relative improvement over the strongest published baseline at 71.9. The wavelet tokenizer accounts for most of this margin. Replacing it with raw temporal patches lowers average macro-F1 by 7.6 points under participant holdout, compared with 2.9 points under random windows. The tokenizer therefore contributes more than twice as much when test participants are unseen, consistent with wavelet tokens reducing reliance on participant-specific cues.

Our contributions are as follows:

- **Wavelet tokens as prediction units.** A fixed tokenizer that maps each inertial window to tokens indexed by scale, sensor axis, and time. Each band covers the same frequency

range in every dataset, and only a channel embedding depends on the sensor layout. Prior work uses wavelets as classification features, forecasting tokens, or reconstruction targets, W-JEPA uses wavelet tokens as units of masked latent prediction over multi-axis inertial data (§3.1).

- **W-JEPA.** A masked latent-prediction framework over these tokens. Masks are drawn independently across channels and bands, so predicting a masked token requires relating motion across scales and axes without reconstructing raw signal values (§3.2).
- **Cross-dataset evaluation under participant holdout.** With each target excluded from pretraining, W-JEPA outperforms eight baselines on six benchmarks under both subject-disjoint and random-window splits, and it loses the least performance when test participants are unseen (§4.1). Its lead holds when fine-tuning is restricted to wrist sensors (§4.3).
- **Evidence that the wavelets matter.** Replacing wavelet tokens with raw temporal patches lowers subject-disjoint macro-F1 on every dataset, especially when few training participants are available (§4.2). Without wavelet tokens, 41.5% of horizontal exercise-bike windows are classified as vertical cycling, compared with 0.8% with them (§4.2).

2 Related Work

Self-supervised pretraining for wearables has explored many training objectives, but the unit of prediction has received less attention. Existing units are either contiguous intervals of raw samples, which mix all motion scales, or codes learned from the pretraining data, whose meaning depends on that data. W-JEPA uses units that are fixed in advance and separated by scale. We review the work this choice builds on.

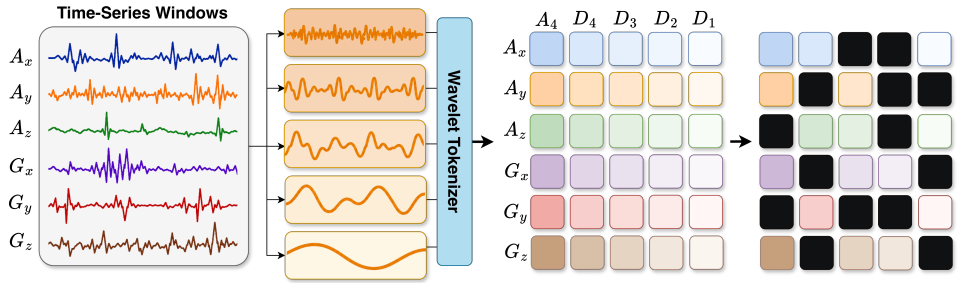
Wavelets for activity recognition. Wavelets separate nonstationary signals into components localized in both time and scale (Daubechies, 1992), whereas Fourier components, used for example in the adaptive spectral blocks of TSLANet (Eldele et al., 2024b), extend over the whole window. Wavelet energy features and wavelet inputs to convolutional networks have long been used for supervised activity recognition (Tian et al., 2020; Vuong et al., 2023). In forecasting, WaveToken quantizes wavelet coefficients into a vocabulary for autoregressive prediction (Masserano et al., 2024). In self-supervised pretraining, Thukral et al. (2026) reconstructs masked wavelet coefficients of PPG signals, which still requires reproducing coefficient values whose amplitudes vary across participants. These methods use wavelets as features, forecasting tokens, or reconstruction targets. To our knowledge, prior work has not used wavelets to define the units of masked latent prediction for multi-axis inertial data. Each token is addressed by scale, sensor axis, and time, so a mask can hide one scale on one axis while others stay visible, and the model predicts learned representations rather than coefficients.

Prediction units and targets in wearable self-supervision. Contrastive and bootstrap methods relate augmented views of a window (Eldele et al., 2021; Grill et al., 2020; Guo et al., 2024; Zhang et al., 2022). Their performance depends strongly on the choice of augmentation (Qian et al., 2022), which can remove the fine detail that separates similar activities. Masked modeling makes the unit explicit. Most wearable methods reconstruct masked sensor values or temporal patches (Zerveas et al., 2021; Haresamudram et al., 2020; Xu et al., 2021; Ryu et al., 2025), which requires reproducing amplitude and phase details that vary with participant and sensor placement. MoPFormer instead predicts codes from a learned vocabulary of motion primitives (Zhang et al., 2025), so its units depend on the pretraining data. Joint-embedding predictive architectures replace reconstruction with the prediction of learned representations (LeCun et al., 2022; Assran et al., 2023; Bardes et al., 2024) and have been extended to time series, skeletal actions, and EEG (Ennadir et al., 2025; Abdelfattah & Alahi, 2024; Guetschel et al., 2024). HAR-JEPA brings this idea to activity recognition with temporal-patch units (Noor & Baraka, 2026). Latent targets remove the burden of signal reconstruction, but temporal patches still mix all motion scales within each interval. W-JEPA keeps the latent target and replaces the unit with wavelet tokens indexed by scale, sensor axis, and time.

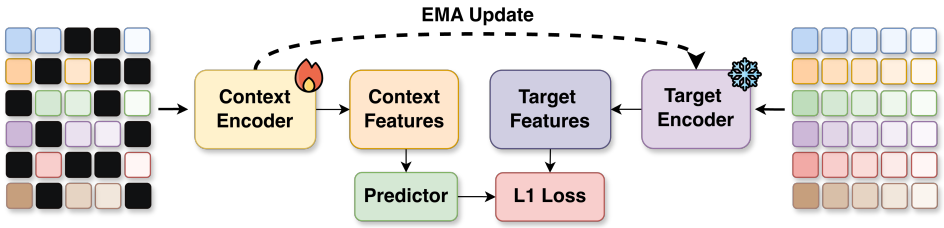
Generalization across datasets and participants. Differences in sensor placement, device, sampling protocol, and participant behavior limit transfer in wearable activity recognition (Chen et al., 2021; Lu et al., 2023). Recent methods address this through physics-informed augmentation (Xu

et al., 2023), hierarchical pretraining with augmentation (Hong et al., 2024), or alignment with language (Zhang et al., 2024; Ding et al., 2026). These methods change the training signal. W-JEPA changes the unit of representation, so that a token at a given scale covers the same frequency band in every dataset, and the two directions are complementary. Evaluation matters as well. Placing windows from the same participant in both training and test sets overstates performance (Hammerla & Plötz, 2015). We therefore treat participant-disjoint results as the primary evaluation, and we also exclude each target dataset from pretraining (§4.1).

3 METHOD



(a) Wavelet tokenization and masking



(b) Masked latent prediction

Figure 2: **W-JEPA.** (a) Wavelet tokenization and masking. Rows are sensor channels and columns are wavelet bands A_4–D_1, ordered from coarse to fine. Each channel and wavelet band combination produces two temporal tokens, shown as a single cell for clarity and masked together; black cells indicate masked token pairs. (b) Masked latent prediction. The predictor estimates representations at masked positions from visible-token context, with targets supplied by an EMA target encoder.

Figure 2 summarizes W-JEPA. Given an inertial window $x \in \mathbb{R}^{C \times T}$ with C sensor channels (one per sensor axis) and T samples, a fixed wavelet tokenizer maps x to a grid of tokens indexed by channel, wavelet band, and time bin (§3.1). During pretraining, a random subset of channel–band token pairs are masked. A context encoder reads the visible tokens, a predictor estimates a latent target at each masked position, and an EMA target encoder computes those targets from the full token sequence. The encoders, predictor, and EMA update follow standard JEPA (LeCun et al., 2022; Assran et al., 2023; Bardes et al., 2024).

The tokenizer supplies the structured prediction space. As mentioned before, Figure 1 shows that detail bands expose fluctuations between hard to distinguish labels. Summarizing each wavelet band separately makes these finer details available to the encoder alongside lower-frequency signal components. Masked prediction can then use relationships among the visible scales, channels, and time bins. This is the motivation for the design; the tokenizer comparison in §4 evaluates its effect on recognition performance. The design choices follow from this view. W-JEPA's changes lie in the tokenizer (§3.1) and in the masking scheme and objective (§3.2).

3.1 WAVELET TOKENIZATION

An inertial window can contain slow variations, rhythmic motion, and brief transients. A discrete wavelet transform separates these timescales while retaining temporal localization (Daubechies,

1992). Tokenizing along those bands means every token arrives with a physically meaningful address: a scale, a sensor axis, and a moment in time.

Channel and band axes. Let $\mathcal{C} = \{1, \ldots, C\}$ index the available sensor channels. Figure 2a illustrates the common six-channel case $\{A_x, A_y, A_z, G_x, G_y, G_z\}$, where A and G denote accelerometer and gyroscope channels. For every $c \in \mathcal{C}$, a four-level Daubechies-4 discrete wavelet transform with reflective boundary extension produces four detail-coefficient sequences $d_{c,1}, \ldots, d_{c,4}$ and one approximation-coefficient sequence $a_{c,4}$ (Wolter et al., 2024). We call these five sequences *wavelet bands* and index them from fine to coarse. The wavelet bands provide a fixed organization of the signal by timescale. Within each band, RMS summarizes coefficient magnitude, while the mean retains signed coefficient information. These statistics provide scale-specific descriptors for token construction.

Token construction. Each of the five coefficient sequences is divided into P contiguous bins of approximately equal length, so that every band yields the same number of bins regardless of how many coefficients it holds. Let q_{ckb} hold the coefficients in bin $b \in \{1, \ldots, P\}$ of band $k \in \{1, \ldots, 5\}$ for channel c. We summarize each bin by two statistics, its mean μ_{ckb} and its RMS magnitude r_{ckb}:

$$\mu_{ckb} = \frac{1}{|q_{ckb}|} \sum_{v \in q_{ckb}} v, \qquad r_{ckb} = \sqrt{\frac{1}{|q_{ckb}|} \sum_{v \in q_{ckb}} v^2 + 10^{-6}}. \tag{1}$$

Together, these summarize scale-specific signal magnitude and preserve signed coefficient information. We use $P = 2$ bins per band. A two-layer MLP $g : \mathbb{R}^2 \to \mathbb{R}^{384}$ maps this pair to a content embedding, and learned embeddings of the channel, band, and bin indices supply the token's address: $u_{ckb} = g([\mu_{ckb}, r_{ckb}]) + e_c^{\text{ch}} + e_k^{\text{band}} + e_b^{\text{bin}}$. Flattening the grid gives a sequence of $N = 5CP$ tokens, in which each position corresponds to exactly one triple (c, k, b). Of the learned components in the token embedding, only e_c^{ch} depends on the sensor layout, which is what makes the adaptation in downstream finetuning cheap.

3.2 Masked latent pretraining

W-JEPA masks after tokenization and predicts learned representations of the masked tokens. The latent targets allow the training objective to operate on features learned by the target encoder, with no raw-sample reconstruction term. For window i, let $\mathcal{M}_i$ denote the masked token positions and $\mathcal{V}_i$ their visible complement.

Asynchronous paired-token masking. Each channel and wavelet-band combination produces two temporal tokens. We group these tokens into $G = N/2$ disjoint pairs $\{\mathcal{P}_g\}_{g=1}^{G}$, where each pair contains the two tokens from the same channel and wavelet band. Given a mask ratio ρ, we uniformly sample a subset $\mathcal{S}_i$ of $|\mathcal{S}_i| = \lfloor \rho G \rfloor$ pairs, with a fresh draw per window, and mask both tokens in each selected pair, yielding $\mathcal{M}_i = \bigcup_{g \in \mathcal{S}_i} \mathcal{P}_g$. Masking remains asynchronous across channels and bands: selecting one pair does not require masking other channels or bands. This preserves complementary information in the visible context and encourages the predictor to exploit dependencies across sensors, axes, and wavelet scales. MoCA similarly finds that independently masking sensor axes improves reconstruction pretraining relative to synchronized masking (Ryu et al., 2025). Within each pair, masking both temporal tokens prevents the predictor from relying on the other token from that same pair. Figure 2a illustrates our paired-token mask.

Objective. Let $p_{ij} \in \mathbb{R}^D$ be the predictor output and $\bar{z}_{ij} \in \mathbb{R}^D$ the normalized latent target at masked token position j. For batch size B, the prediction loss is

$$\mathcal{L}_{\text{pred}} = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{|\mathcal{M}_i| D} \sum_{j \in \mathcal{M}_i} \left\| p_{ij} - \bar{z}_{ij} \right\|_1 . \tag{2}$$

To discourage representation collapse, we add the variance hinge from VICReg without its covariance term (Bardes et al., 2021). Let σ_{ir} be the sample standard deviation across predictions

$\{p_{ij} : j \in \mathcal{M}_i\}$ in feature coordinate r. We define

$$\mathcal{L}_{\text{var}} = \frac{1}{BD} \sum_{i=1}^{B} \sum_{r=1}^{D} \max(0, 1 - \sigma_{ir}), \qquad \mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{pred}} + 0.0625\, \mathcal{L}_{\text{var}}. \tag{3}$$

The spread is computed within each window over its own masked positions, so the hinge asks that a single window's predictions vary across scales and axes. In a diagnostic comparison, substituting SIGReg (Balestriero & LeCun, 2025) while holding the rest of the pipeline fixed costs accuracy (Appendix Table 6). For downstream fine-tuning, a target dataset with C_{tar} channels, the same tokenizer produces $N = 5C_{\text{tar}}P$ token positions. The target encoder and predictor are discarded, and the context encoder is fine-tuned end-to-end with a classification head on all available channels.

4 Experiments

We ask three questions. Does W-JEPA, pretrained on other datasets, recognize activities of participants unseen during fine-tuning? Does its advantage hold when the target sensor layout is restricted (§4.3)? And what do wavelet tokens contribute beyond the JEPA objective (§4.2)?

4.1 Main Results

Datasets and pretraining pool. We evaluate on PAMAP2 (Reiss, 2012; Reiss & Stricker, 2012), DSADS (Barshan & Altun, 2010; Altun et al., 2010), MHealth (Banos et al., 2014; 2015), RealWorld (Sztyler & Stuckenschmidt, 2016), UCI-HAR (Reyes-Ortiz et al., 2013; Anguita et al., 2013), and USC-HAD (Zhang & Sawchuk, 2012). Appendix A describes each dataset. For each target, the remaining five datasets form the unlabeled pretraining pool, matching MoPFormer's setup (Zhang et al., 2025). Each source contributes the same six-channel view: the x, y, and z axes from one accelerometer and one gyroscope, while downstream fine-tuning uses all configured target channels. All recordings are resampled to 100 Hz and segmented into non-overlapping windows of $T = 500$ samples (5 seconds). We do not apply channel normalization.

Evaluation splits. Our primary evaluation uses subject-disjoint splits: all windows from a participant belong exclusively to either downstream fine-tuning or testing. The target dataset is also excluded from pretraining. For comparison, we evaluate all methods under a random-window protocol that assigns 20% of windows within each activity class to training and 80% to evaluation; this protocol can place the same participant in both partitions. We report accuracy and macro-F1 over five random seeds, with dataset averages weighted equally. Sample and participant counts for the subject-disjoint splits are given in Appendix Table 5.

Baselines. We compare with eight baselines, all re-run with the same preprocessing, splits, and seeds: BYOL (Grill et al., 2020) with the permutation-and-jitter (perm_jit) and low-frequency-component (lfc) augmentations from Qian et al. (2022) and with temporal mixup (Eldele et al., 2024a); ModCL, a contrastive method guided by modality consistency (Guo et al., 2024); MoPFormer, which also pretrains across datasets (Zhang et al., 2025); CALANet, an efficient HAR architecture (Park et al., 2024); TSLANet, a general time-series backbone (Eldele et al., 2024b); and HAR-JEPA, applies JEPA to wearable activity recognition (Noor & Baraka, 2026). W-JEPA (no wavelets), described in §4.2, isolates the tokenizer.

Results. Table 1 evaluates recognition on participants excluded from training and Table 2 reports the random-window evaluation. W-JEPA achieves the highest mean accuracy and macro-F1 on every dataset under both setups, although its margin varies by dataset. Comparing these tables shows how much random-window evaluation overstates performance on new people. TSLANet falls the most, while HAR-JEPA remains the strongest baseline. W-JEPA falls by the least of any method.

4.2 What the wavelet tokens contribute

Replacing wavelet tokens with raw patches. The variant *W-JEPA (no wavelets)* replaces the wavelet tokenizer with a conventional temporal-patch tokenizer while holding the remainder of the

Table 1: Subject-disjoint comparison across six wearable activity recognition datasets. Results are averaged over five random seeds under the unified protocol of §4.

Method	Metric	PAMAP2	DSADS	MHealth	RealWorld	UCI-HAR	USC-HAD
BYOL + perm_jit	Acc.	65.65 ± 3.68	64.66 ± 5.59	57.70 ± 19.68	67.59 ± 0.53	74.49 ± 2.78	57.81 ± 3.15
	F1	65.33 ± 4.65	63.99 ± 6.12	51.32 ± 20.57	62.99 ± 0.45	71.43 ± 2.98	53.71 ± 2.69
BYOL + lfc	Acc.	72.01 ± 3.32	65.13 ± 3.27	73.10 ± 1.25	65.42 ± 1.37	77.36 ± 3.48	59.48 ± 3.70
	F1	72.16 ± 3.56	63.79 ± 4.64	73.46 ± 0.51	59.87 ± 1.51	76.21 ± 3.76	57.63 ± 2.71
BYOL + Mixup	Acc.	72.94 ± 2.00	65.74 ± 5.73	74.13 ± 3.41	64.31 ± 1.74	75.82 ± 4.68	58.09 ± 1.45
	F1	73.30 ± 2.08	64.25 ± 6.91	73.76 ± 3.77	58.92 ± 1.99	72.89 ± 5.86	56.43 ± 1.09
ModCL	Acc.	72.52 ± 2.18	61.00 ± 1.55	69.53 ± 5.47	66.25 ± 2.64	81.35 ± 5.43	55.75 ± 5.69
	F1	72.32 ± 2.79	59.83 ± 2.03	68.16 ± 6.12	64.92 ± 2.04	81.02 ± 5.45	55.06 ± 5.19
TSLANet	Acc.	65.85 ± 1.08	70.03 ± 8.61	77.55 ± 3.00	57.95 ± 1.88	80.19 ± 2.32	60.91 ± 1.53
	F1	66.43 ± 1.27	68.48 ± 9.20	76.82 ± 3.42	54.67 ± 1.95	80.10 ± 2.33	57.40 ± 0.61
CALANet	Acc.	69.40 ± 2.34	63.16 ± 5.53	86.37 ± 2.26	64.51 ± 2.60	81.86 ± 1.00	63.58 ± 2.59
	F1	69.11 ± 2.29	61.78 ± 5.97	85.24 ± 1.90	61.04 ± 2.87	81.38 ± 0.90	59.43 ± 1.96
MoPFormer	Acc.	72.43 ± 1.12	61.10 ± 9.66	81.42 ± 1.16	67.38 ± 2.74	80.93 ± 5.52	60.70 ± 4.62
	F1	69.40 ± 1.09	57.90 ± 10.76	81.23 ± 1.03	68.60 ± 3.41	78.16 ± 6.43	58.76 ± 4.04
HAR-JEPA	Acc.	61.08 ± 0.89	73.16 ± 0.98	87.38 ± 1.89	64.94 ± 0.71	81.57 ± 1.31	61.03 ± 2.75
	F1	63.53 ± 0.76	73.59 ± 0.88	87.64 ± 1.80	64.50 ± 0.89	81.99 ± 0.29	60.22 ± 1.17
W-JEPA (no wavelets)	Acc.	81.34 ± 0.61	70.00 ± 1.95	73.33 ± 2.13	71.06 ± 1.64	82.71 ± 3.40	59.49 ± 0.91
	F1	81.64 ± 0.31	69.98 ± 2.51	73.68 ± 2.68	71.48 ± 1.00	82.00 ± 3.47	59.52 ± 0.53
W-JEPA	Acc.	$\mathbf{86.96 \pm 0.62}$	$\mathbf{82.87 \pm 1.49}$	$\mathbf{89.68 \pm 1.09}$	$\mathbf{80.04 \pm 0.41}$	$\mathbf{84.70 \pm 1.94}$	$\mathbf{64.35 \pm 3.28}$
	F1	$\mathbf{87.19 \pm 0.61}$	$\mathbf{82.76 \pm 1.76}$	$\mathbf{88.89 \pm 1.24}$	$\mathbf{80.53 \pm 1.10}$	$\mathbf{83.97 \pm 2.08}$	$\mathbf{60.35 \pm 1.40}$

Table 2: Random-window comparison across six wearable activity recognition datasets. Results are averaged over five random seeds under the unified protocol of §4.

Method	Metric	PAMAP2	DSADS	MHealth	RealWorld	UCI-HAR	USC-HAD
BYOL + perm_jit	Acc.	84.06 ± 8.26	89.18 ± 5.02	85.49 ± 7.66	95.52 ± 0.27	77.65 ± 3.80	78.45 ± 5.40
	F1	80.53 ± 13.74	88.73 ± 5.45	79.57 ± 11.37	95.60 ± 0.40	75.94 ± 4.48	73.43 ± 6.08
BYOL + lfc	Acc.	90.82 ± 1.20	96.71 ± 0.14	92.28 ± 1.60	96.11 ± 0.34	81.52 ± 4.04	82.85 ± 1.20
	F1	90.66 ± 1.45	96.71 ± 0.14	91.41 ± 1.76	96.14 ± 0.46	80.43 ± 4.76	77.57 ± 1.30
BYOL + Mixup	Acc.	89.76 ± 0.89	97.00 ± 0.52	92.09 ± 1.36	96.00 ± 0.19	81.66 ± 3.67	81.28 ± 2.44
	F1	89.53 ± 1.07	96.97 ± 0.52	91.75 ± 1.52	96.00 ± 0.20	80.54 ± 4.56	75.88 ± 3.34
ModCL	Acc.	90.60 ± 1.17	96.60 ± 2.48	94.51 ± 4.09	95.37 ± 0.87	89.20 ± 3.84	84.54 ± 3.39
	F1	90.47 ± 1.11	96.57 ± 2.54	92.70 ± 5.53	95.42 ± 0.87	88.99 ± 4.19	81.06 ± 3.08
TSLANet	Acc.	92.20 ± 0.58	98.89 ± 0.15	97.12 ± 1.29	97.04 ± 0.16	85.58 ± 8.30	82.93 ± 0.58
	F1	92.16 ± 0.50	98.87 ± 0.16	96.99 ± 1.60	97.11 ± 0.18	84.86 ± 9.32	84.28 ± 0.60
CALANet	Acc.	84.60 ± 0.74	94.91 ± 0.11	92.19 ± 1.56	93.66 ± 0.35	88.10 ± 0.76	77.61 ± 1.49
	F1	81.56 ± 1.00	94.83 ± 0.12	90.66 ± 1.92	91.51 ± 0.64	87.56 ± 0.85	72.19 ± 1.87
MoPFormer	Acc.	90.64 ± 0.52	82.09 ± 5.46	97.15 ± 1.20	91.39 ± 2.81	83.37 ± 7.18	84.23 ± 1.04
	F1	90.24 ± 0.46	79.44 ± 6.58	97.16 ± 1.14	92.13 ± 2.32	79.37 ± 10.47	80.64 ± 1.11
HAR-JEPA	Acc.	91.67 ± 0.19	97.88 ± 0.05	98.43 ± 0.23	95.24 ± 0.10	89.83 ± 0.15	86.56 ± 0.21
	F1	91.96 ± 0.20	97.86 ± 0.05	98.54 ± 0.21	95.46 ± 0.11	90.14 ± 0.14	82.77 ± 0.36
W-JEPA (no wavelets)	Acc.	91.46 ± 0.39	96.94 ± 0.15	96.00 ± 0.38	96.71 ± 0.16	88.64 ± 1.26	85.81 ± 0.36
	F1	91.29 ± 0.34	96.94 ± 0.16	95.55 ± 0.21	96.55 ± 0.13	87.80 ± 1.28	82.66 ± 0.36
W-JEPA	Acc.	$\mathbf{93.75 \pm 0.22}$	$\mathbf{99.22 \pm 0.13}$	$\mathbf{99.28 \pm 0.35}$	$\mathbf{97.13 \pm 0.14}$	$\mathbf{92.91 \pm 1.23}$	$\mathbf{88.54 \pm 0.35}$
	F1	$\mathbf{93.79 \pm 0.21}$	$\mathbf{99.22 \pm 0.13}$	$\mathbf{99.18 \pm 0.19}$	$\mathbf{97.36 \pm 0.12}$	$\mathbf{92.59 \pm 1.26}$	$\mathbf{86.52 \pm 0.33}$

pipeline fixed. Each 500-sample channel is divided into ten consecutive 50-sample patches; the ordered, signed samples of a patch are embedded by an MLP of the same depth and width. With wavelet tokens, subject-disjoint macro-F1 is higher on all six datasets (Table 1). Two JEPAs that predict temporal-patch units, HAR-JEPA and W-JEPA (no wavelets), reach 71.9 and 73.1 average held-out macro-F1; replacing the patches with wavelet tokens raises W-JEPA to 80.6. This comparison supports the tokenizer as a useful design choice within W-JEPA.

Varying participant diversity. Figure 3 compares the tokenizers with 20% of participants held out for testing. This test cohort remains fixed as complete participants from the remaining pool are

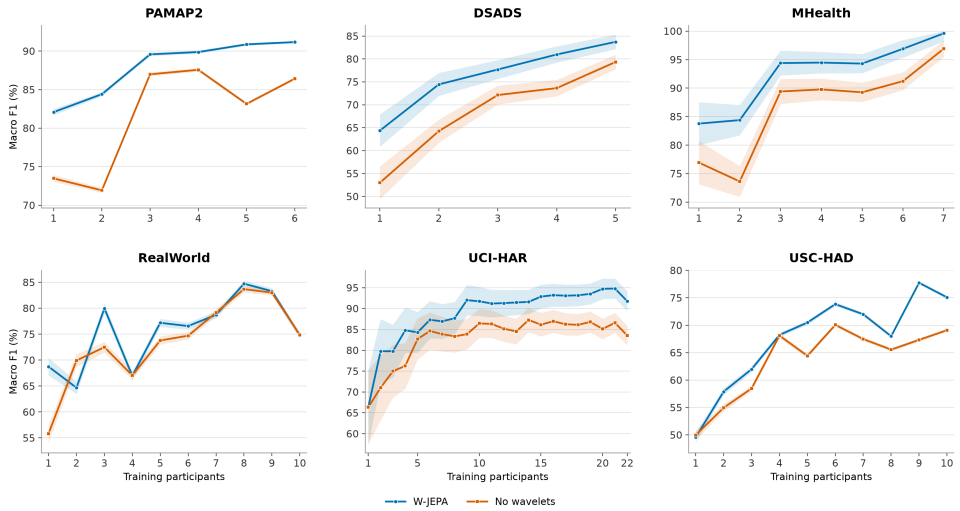
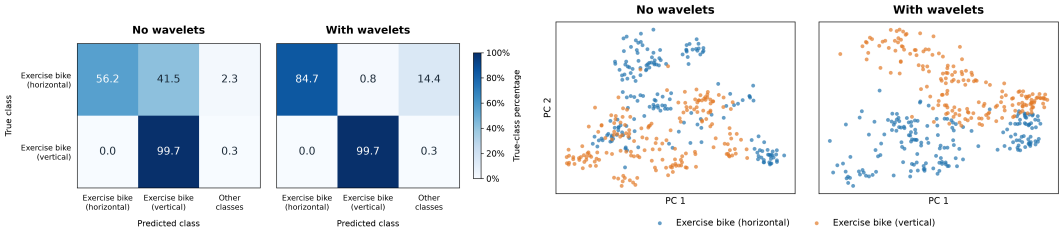


Figure 3: Subject-disjoint macro-F1 with wavelet tokens and learned temporal patches as training participants are added. A fixed 20% participant split is held out for testing, and training participants are drawn from the remaining pool.

progressively added to the fine-tuning set. At matched training set sizes, W-JEPA generally outperforms the *no wavelets* variant, especially when few participants are available. The advantage persists across the displayed training-set sizes on PAMAP2, DSADS, and MHealth, while RealWorld shows smaller and less consistent gains. With the test cohort fixed, the curves compare the benefit of adding training participants on the same evaluation population. They provide evidence that wavelet tokens improve performance under limited participant diversity.



(a) Confusion between cycling activities

(b) PCA of learned representations

Figure 4: Each panel compares W-JEPA without and with wavelets. (a) Confusion matrices for cycling on a horizontal exercise bike and cycling on a vertical exercise bike, using percentages within each true class; predictions of other activities are "Other classes." (b) 2D PCA projections of learned representations for the same two classes.

Separating near-identical activities. We return to the example of the exercise-bike pair of Figure 1, where we further compare W-JEPA with and without wavelet tokenization on DSADS under the same subject-disjoint evaluation protocol, keeping pretraining, fine-tuning, and all other experimental settings fixed. (Figure 4). Without wavelets, 41.5% of horizontal-bike windows are classified as vertical-bike cycling; with wavelets, this falls to 0.8%, improving horizontal-bike recall while preserving vertical-bike recall; some errors shift to other activities. In the PCA projection of the learned representations, the two classes overlap without wavelets and are more separated with them. This is consistent with wavelet tokens exposing scale-specific differences in Figure 1.

Interpreting attention by timescale. Because every token carries a scale address, the fine-tuned encoder's attention can be read by timescale, which a temporal-patch tokenizer does not allow. We measure the attention that tokens receive in the final encoder layer, averaged over heads and query positions and summed within each band, on held-out DSADS participants (Appendix D.2, Figure 6).

On clean inputs, the coarse A_4 band receives the largest share for most activities, whereas, D_1, receives the largest share for running and jumping. We then add noise at increasing levels, measured against each channel's training RMS. Attention shifts from A_4 toward the detail bands: at 30% noise, A_4 attention falls by 6.1% relative to clean inputs, while D_2 and D_1 rise by 11.1% and 9.4%. Over the same range, held-out accuracy and macro-F1 stay within about one point of their clean values (Appendix Table 7).

4.3 Ablation under restricted sensor settings

We evaluate a restricted setting in which downstream fine-tuning uses only wrist accelerometer and gyroscope channels: nine channels for PAMAP2 and six for MHealth. Pretraining follows the same protocol as in §4, using all datasets in the pretraining pool except the evaluation dataset. All other experimental settings are kept the same. We evaluate both random-window and subject-disjoint splits, using 20% of windows or participants, respectively, for training and the remaining 80% for evaluation. Each experiment uses five training seeds. Tables 3 and 4 report macro-F1 and accuracy under these restrictions. W-JEPA achieves the highest mean score on both metrics for both datasets under both split protocols. Scores are lower with participant holdout for every method, while W-JEPA retains its lead in mean performance when fine-tuning is restricted to wrist inertial signals.

Table 3: Restricted sensor ablation with random-window splits.

Method	PAMAP2		MHealth	
	F1	Acc.	F1	Acc.
BYOL + perm_jit	58.97 ± 6.41	67.32 ± 3.75	40.77 ± 11.21	48.50 ± 6.07
BYOL + lfc	77.85 ± 1.13	79.60 ± 0.71	49.63 ± 3.19	55.08 ± 2.77
BYOL + Mixup	75.16 ± 3.44	78.28 ± 2.03	47.31 ± 3.06	52.76 ± 4.03
ModCL	75.99 ± 5.84	79.10 ± 3.63	72.84 ± 4.67	78.79 ± 2.21
TSLANet	87.63 ± 0.35	86.79 ± 0.56	77.54 ± 2.29	78.02 ± 2.57
CALANet	74.63 ± 1.14	78.11 ± 0.58	75.50 ± 1.17	77.60 ± 1.20
MoPFormer	85.14 ± 0.38	85.52 ± 0.30	88.81 ± 0.67	88.95 ± 0.83
HAR-JEPA	85.96 ± 0.26	86.31 ± 0.28	90.91 ± 0.88	90.67 ± 0.88
W-JEPA (no wavelets)	86.59 ± 0.31	86.54 ± 0.28	89.58 ± 0.55	90.47 ± 0.63
W-JEPA	$\mathbf{88.10 \pm 0.11}$	$\mathbf{88.09 \pm 0.14}$	$\mathbf{93.53 \pm 0.54}$	$\mathbf{94.22 \pm 0.76}$

Table 4: Restricted sensor ablation with subject-disjoint splits.

Method	PAMAP2		MHealth	
	F1	Acc.	F1	Acc.
BYOL + perm_jit	46.88 ± 1.94	50.23 ± 4.10	35.01 ± 6.01	40.87 ± 3.42
BYOL + lfc	48.57 ± 1.54	47.98 ± 2.22	39.08 ± 1.10	43.42 ± 2.30
BYOL + Mixup	51.88 ± 1.64	52.06 ± 2.46	40.23 ± 1.19	44.08 ± 2.77
ModCL	61.71 ± 1.66	62.63 ± 2.44	51.26 ± 6.29	53.89 ± 5.05
TSLANet	63.07 ± 3.29	61.98 ± 3.60	59.24 ± 2.97	58.43 ± 2.74
CALANet	55.90 ± 1.75	55.14 ± 1.63	59.44 ± 1.73	61.63 ± 1.58
MoPFormer	58.71 ± 0.51	59.57 ± 1.16	66.17 ± 3.26	68.19 ± 2.08
HAR-JEPA	60.54 ± 0.26	61.57 ± 0.32	65.29 ± 0.96	66.39 ± 1.09
W-JEPA (no wavelets)	61.93 ± 1.64	62.16 ± 1.77	64.26 ± 1.79	64.77 ± 0.92
W-JEPA	$\mathbf{64.14 \pm 2.12}$	$\mathbf{64.25 \pm 1.16}$	$\mathbf{67.63 \pm 1.13}$	$\mathbf{69.45 \pm 2.62}$

5 Conclusion

We asked how the units of masked latent prediction should be defined for heterogeneous wearable data. W-JEPA organizes inertial windows into wavelet tokens indexed by scale, sensor channel axis, and time, and learns by predicting masked latent features. It outperforms the evaluated baselines, and loses the least when evaluating on unseen participants. Replacing wavelet tokens with temporal patches weakens performance and can increase confusion between similar activities. These results suggest that physically grounded multiscale tokenization and latent prediction together provide a promising foundation for transferable representations of heterogeneous wearable data.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on our research methods and experiments and to assist with implementing methods. We used them to create or edit software code; draft parts of the manuscript; summarize and analyze existing literature; search for information and identify relevant literature; improve the manuscript's readability; and suggest a title and keywords. We did not use generative AI to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims or provide key steps in their proofs; write proofs; propose or refine hypotheses; design research methods or experiments; translate text; clean or reformat datasets; conduct qualitative or thematic analysis; or interpret results. The authors reviewed all AI-assisted work. Two authors verified and tested AI-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide anonymized code at `https://anonymous.4open.science/r/w_jepa_iclr2027-B763/`. It includes the wavelet tokenizer, pretraining, fine-tuning, and evaluation code, a script that downloads and prepares all six datasets, and a pipeline script that runs pretraining followed by fine-tuning under both evaluation protocols. The wavelet tokenizer is specified in §3.1 (Eqs. 1), and the masking scheme and training objective in §3.2 (Eqs. 2–3). All six datasets are publicly available (Burzer et al., 2025); Appendix A describes each one, and §4.1 specifies the preprocessing, the pretraining pools, and both evaluation protocols. Appendix Table 5 lists the sample and participant counts of each subject-disjoint split. Appendix B reports the architecture, optimization settings, learning-rate and EMA schedules, and fine-tuning procedure. All main results are averaged over five random seeds, and every baseline is re-run with the same preprocessing, splits, and seeds. The no-wavelets control is defined in §4.2, the wrist-only setting in §4.3, and the regularizer and sensitivity studies in Appendices C.1 and C.2. Appendices D.1 and D.2 identify the recordings used in Figure 1 and the protocol of the attention analysis.

REFERENCES

Mohamed Abdelfattah and Alexandre Alahi. S-jepa: A joint embedding predictive architecture for skeletal action recognition. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024.

Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.04.019. URL `https://www.sciencedirect.com/science/article/pii/S0031320310001950`.

D. Anguita, Alessandro Ghio, L. Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *The European Symposium on Artificial Neural Networks*, 2013. URL `https://api.semanticscholar.org/CorpusID:6975432`.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Oresti Banos, Rafael Garcia, and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TW22.

Oresti Banos, Claudia Villalonga, Rafael Garcia, Alejandro Saez, Miguel Damas, Juan A Holgado-Terriza, Sungyong Lee, Hector Pomares, and Ignacio Rojas. Design, implementation and validation of a novel open framework for agile development of mobile health applications. *Biomedical engineering online*, 14:1–20, 2015.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Billur Barshan and Kerem Altun. Daily and Sports Activities. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5C59F.

Maximilian Burzer, Tobias King, Till Riedel, Michael Beigl, and Tobias Röddiger. Whar datasets: An open source library for wearable human activity recognition. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1315–1322, 2025.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40, 2021.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Zihan Ding, Liyu Zhang, and Xiaomin Ouyang. Halo: A heterogeneity-aware language-aligned imu foundation model for open-set human activity recognition, 2026. URL https://arxiv.org/abs/2608.27233.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Contrastive domain adaptation for time-series via temporal mixup. *IEEE Transactions on Artificial Intelligence*, 5(3):1185–1194, 2024a. doi: 10.1109/TAI.2023.3293473.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, and Xiaoli Li. TSLANet: rethinking transformers for time series representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embeddings go temporal. *arXiv preprint arXiv:2509.25449*, 2025.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-jepa: towards seamless cross-dataset transfer through dynamic spatial attention. *arXiv preprint arXiv:2403.11772*, 2024.

Changru Guo, Yingwei Zhang, Yiqiang Chen, Chenyang Xu, and Zhong Wang. Modality consistency-guided contrastive learning for wearable-based human activity recognition. *IEEE Internet of Things Journal*, 11(12):21750–21762, 2024. doi: 10.1109/JIOT.2024.3379019.

Nils Y. Hammerla and Thomas Plötz. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 1041–1051, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744. doi: 10.1145/2750858.2807551. URL https://doi.org/10.1145/2750858.2807551.

Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, ISWC '20, pp. 45–49, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380775. doi: 10.1145/3410531.3414306. URL https://doi.org/10.1145/3410531.3414306.

Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2), May 2024. doi: 10.1145/3659597. URL https://doi.org/10.1145/3659597.

Hyejoo Kim, Hyeon-Joo Kim, Jinyoon Park, Jeh-Kwang Ryu, and Seung-Chan Kim. Recognition of fine-grained walking patterns using a smartwatch with deep attentive neural networks. *Sensors*, 21(19), 2021. ISSN 1424-8220. doi: 10.3390/s21196393. URL https://www.mdpi.com/1424-8220/21/19/6393.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification, 2023. URL https://arxiv.org/abs/2209.07027.

Luca Masserano, Abdul Fatir Ansari, Boran Han, Xiyuan Zhang, Christos Faloutsos, Michael W Mahoney, Andrew Gordon Wilson, Youngsuk Park, Syama Rangapuram, Danielle C Maddix, et al. Enhancing foundation models for time series forecasting via wavelet-based tokenization. *arXiv preprint arXiv:2412.05244*, 2024.

Mohd Halim Mohd Noor and Abdulrahman M. A. Baraka. Joint-embedding predictive architecture for sensor-based activity recognition, 2026. URL https://arxiv.org/abs/2607.16350.

Jaegyun Park, Dae-Won Kim, and Jaesung Lee. CALANet: Cheap all-layer aggregation for human activity recognition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 69419–69444. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8078e76f913e31b8467e85b4c0f0d22b-Paper-Conference.pdf.

Hangwei Qian, Tian Tian, and Chunyan Miao. What makes good contrastive learning on small-scale wearable-based tasks? In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 3761–3771, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539134. URL https://doi.org/10.1145/3534678.3539134.

Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Jorge Luis Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C54S4K.

Howon Ryu, Yuliang Chen, Yacun Wang, Andrea Z LaCroix, Chongzhi Di, Loki Natarajan, Yu Wang, and Jingjing Zou. Moca: Multi-modal cross-masked autoencoder for digital health measurements. *arXiv preprint arXiv:2506.02260*, 2025.

Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–9, 2016. doi: 10.1109/PERCOM.2016.7456521.

Megha Thukral, Cyrus Tanade, Simon A. Lee, Juhyeon Lee, Hao Zhou, Keum San Chun, Migyeong Gwak, Viswam Nathan, Md Mahbubur Rahman, Li Zhu, Mehrab Bin Morshed, Subramaniam Venkatraman, and Sharanya Arcot Desai. Wavelet-driven masked multiscale reconstruction for ppg foundation models, 2026. URL https://arxiv.org/abs/2601.12215.

Yiming Tian, Jie Zhang, Jie Wang, Yanli Geng, and Xitai Wang. Robust human activity recognition using single accelerometer via wavelet energy spectrum features and ensemble feature selection. *Systems Science & Control Engineering*, 8(1):83–96, 2020.

Thi Hong Vuong, Tung Doan, and Atsuhiro Takasu. Deep wavelet convolutional neural networks for multimodal human activity recognition using wearable inertial sensors. *Sensors*, 23(24):9721, 2023.

Moritz Wolter, Felix Blanke, Jochen Garcke, and Charles Tapley Hoyt. ptwt-the pytorch wavelet toolbox. *Journal of Machine Learning Research*, 25(80):1–7, 2024.

Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, pp. 220–233, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450390972. doi: 10.1145/3485730.3485937. URL https://doi.org/10.1145/3485730.3485937.

Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. Practically adopting human activity recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399906. doi: 10.1145/3570361.3613299. URL https://doi.org/10.1145/3570361.3613299.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Hao Zhang, Zhan Zhuang, Xuehao Wang, Xiaodong Yang, and Yu Zhang. Mopformer: Motion-primitive transformer for wearable-sensor activity recognition. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 39334–39360. Curran Associates, Inc., 2025. doi: 10.52202/085713-1313. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/37ecb317d11004f77defbca54ef3b928-Paper-Conference.pdf.

Mi Zhang and Alexander A. Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pp. 1036–1043, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312240. doi: 10.1145/2370216.2370438. URL https://doi.org/10.1145/2370216.2370438.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. Unimts: Unified pre-training for motion time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 107469–107493. Curran Associates, Inc., 2024. doi: 10.52202/079017-3414. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c290d4373c495b2cad0625d6288260f0-Paper-Conference.pdf.

A Datasets and Splits

Table 5: Dataset statistics for the subject-disjoint train/test splits. The number of samples corresponds to the number of segmented sensor windows used for training and testing.

Dataset	Train samples	Test samples	Train participants	Test participants
PAMAP2	835	2,998	2	7
DSADS	2,242	6,726	2	6
MHealth	267	1,068	2	8
RealWorld	2,551	10,209	3	12
UCI-HAR	486	1,946	6	24
USC-HAD	1,069	4,284	3	11

PAMAP2. The PAMAP2 Physical Activity Monitoring dataset contains recordings from nine participants wearing inertial measurement units on the dominant wrist, chest, and dominant ankle, together with a heart-rate monitor (Reiss, 2012; Reiss & Stricker, 2012). The inertial units record acceleration, angular velocity, and magnetic field at 100 Hz, while heart rate is sampled at approximately 9 Hz. The collection protocol specifies 12 activities spanning static postures, locomotion, household tasks, and exercise, including sitting, walking, cycling, vacuum cleaning, and rope jumping. Six additional optional activities bring the full release to 18 activity labels. These recordings capture both differences among body locations and variation in how participants perform the same activity.

DSADS. The Daily and Sports Activities dataset contains 19 activities performed by eight participants, with five minutes recorded for each participant–activity pair (Barshan & Altun, 2010; Altun et al., 2010). Five sensor units on the torso, both arms, and both legs each provide three-axis acceleration, angular velocity, and magnetic field measurements at 25 Hz, yielding 45 sensor channels in the original release. Activities include sitting, lying, stair climbing, elevator travel, walking and running on a treadmill, cycling, rowing, jumping, and basketball. Participants perform the activities in their own style, introducing variation in movement speed and amplitude. The release divides each five-minute recording into 60 segments of five seconds each.

MHealth. The Mobile Health dataset records ten volunteers performing 12 physical activities using sensors attached to the chest, right wrist, and left ankle (Banos et al., 2014; 2015). Signals are acquired at 50 Hz and include acceleration, angular velocity, and magnetic field measurements; the chest device additionally records two ECG leads. The activity vocabulary covers static postures and movements with different body involvement and intensity, including standing, sitting, lying, walking, cycling, jogging, running, arm elevation, and knee bending. This combination provides examples of both localized limb movements and whole-body exercise. The original multimodal release supplies physiological measurements alongside the motion signals used to characterize these activities.

RealWorld. The RealWorld activity recognition dataset includes 15 participants and eight activities: walking, running or jogging, sitting, standing, lying, jumping, and ascending and descending stairs (Sztyler & Stuckenschmidt, 2016). Devices are worn at seven positions—the chest, forearm, head, shin, thigh, upper arm, and waist—with acceleration recorded at 50 Hz. The release also provides gyroscope, magnetic field, GPS, light, and sound-level measurements. Each participant performs each activity for approximately ten minutes, except jumping, which lasts approximately 1.7 minutes because of the exertion involved. Devices are carried using everyday attachments such as pockets and armbands, providing variation in how activity signals are observed across body locations.

UCI-HAR. The Human Activity Recognition Using Smartphones dataset contains recordings from 30 volunteers aged 19–48 wearing a Samsung Galaxy S II smartphone at the waist (Reyes-Ortiz et al., 2013; Anguita et al., 2013). Its accelerometer and gyroscope record three-axis acceleration and angular velocity at 50 Hz during six activities: walking, ascending stairs, descending stairs, sitting, standing, and lying. The original benchmark distributes filtered inertial signals in 128-sample

windows of 2.56 seconds with 50% overlap, together with engineered features, and assigns 70% of participants to training and 30% to testing. The results in this paper follow the common preprocessing and evaluation protocols described in §4; the original windowing and split describe the source benchmark.

USC-HAD. The University of Southern California Human Activity Dataset contains 14 participants performing 12 daily activities (Zhang & Sawchuk, 2012). A MotionNode device attached to the front of the right hip records three-axis acceleration and angular velocity at 100 Hz, providing six motion channels. The activities are walking forward, walking left, walking right, ascending stairs, descending stairs, running forward, jumping, sitting, standing, sleeping, and taking an elevator up or down. Each participant repeats each activity five times. The activity vocabulary distinguishes movement direction and includes both active motion and relatively stationary behaviors, providing a benchmark for recognizing daily activities from a single body location.

B Implementation Details

Pretraining and fine-tuning settings. We use $P = 2$ temporal bins per band and a mask ratio of $\rho = 0.65$ under the sampling protocol in §3.2. Pretraining runs for 50 epochs with batch size 256; fine-tuning runs for 70 epochs with batch size 64. Both stages use AdamW with weight decay 0.05 (Loshchilov & Hutter, 2019).

Architecture. The encoder uses eight Transformer blocks of width 384; the predictor uses four blocks of width 192. Both have six attention heads, an MLP expansion ratio of four, GELU activations, and layer normalization before each sublayer and after the final block. Linear layers project features between these widths. Both use fixed sinusoidal position embeddings, and the predictor adds a learned mask token.

Learning rates and targets. Pretraining and fine-tuning learning rates are updated each epoch, with linear warmup from 10^{-6} over eight and ten epochs, followed by cosine decay toward 10^{-6}. The target encoder is initialized from the context encoder and kept in evaluation mode with gradients disabled. EMA updates occur after each optimizer step, with momentum increasing linearly from 0.998 toward 1.

Classification and channel adaptation. We flatten all encoder output tokens and apply layer normalization, dropout of 0.1, and a linear classifier. The classification loss is cross-entropy with label smoothing of 0.1. For six-channel targets, pretrained channel embeddings are reused by index. Otherwise, the entire channel-embedding table is reinitialized using a truncated normal distribution ($\sigma = 0.02$) and learned during fine-tuning. All other compatible encoder weights are transferred, and fixed position embeddings are regenerated for the target token count.

C Additional Experiments

C.1 Regularizer ablation

For *W-JEPA (SIGReg)* and *W-JEPA (VICReg)*, we keep the model and pipeline unchanged, except that we replace the original variance regularizer with SIGReg (Balestriero & LeCun, 2025) or a VICReg-style variance–covariance regularizer, respectively. The JEPA prediction loss remains unchanged.

C.2 UCI-HAR sensitivity study

What is varied. The first three panels test sensitivity to pretraining and input-design choices, whereas the fourth tests label efficiency after the baseline pretraining run. The mask sweep covers 5–95% in ten-percentage-point increments. The temporal-bin sweep uses 1, 2, 4, 5, or 10 bins for each of the five wavelet bands, producing 30, 60, 120, 150, or 300 tokens for the six-channel input. Its horizontal coordinate is the nominal ratio of the 500-sample window to the number of bins. The input-length sweep uses 100, 250, 500, 750, or 1000 samples. The label sweep uses 1%

Table 6: W-JEPA regularizer random-window split ablations, summarizes five seeds mean and standard deviation across the six benchmarks. Bold indicates the highest mean per metric, including ties.

Variant	Metric	PAMAP2	DSADS	MHealth	RealWorld	UCI-HAR	USC-HAD
W-JEPA (SIGReg)	Acc.	93.62 ± 0.07	99.02 ± 0.09	99.23 ± 0.27	96.92 ± 0.06	89.74 ± 0.30	87.76 ± 0.45
	F1	93.70 ± 0.04	99.01 ± 0.10	$\mathbf{99.22 \pm 0.24}$	97.13 ± 0.07	89.39 ± 0.29	84.42 ± 0.37
W-JEPA (VICReg)	Acc.	$\mathbf{93.77 \pm 0.20}$	99.11 ± 0.08	98.97 ± 0.57	$\mathbf{97.13 \pm 0.18}$	90.66 ± 0.87	87.97 ± 0.71
	F1	$\mathbf{93.82 \pm 0.22}$	99.10 ± 0.08	99.04 ± 0.53	97.33 ± 0.14	90.45 ± 0.88	84.75 ± 0.62
W-JEPA	Acc.	93.75 ± 0.22	$\mathbf{99.22 \pm 0.13}$	$\mathbf{99.28 \pm 0.35}$	$\mathbf{97.13 \pm 0.14}$	$\mathbf{92.91 \pm 1.23}$	$\mathbf{88.54 \pm 0.35}$
	F1	93.79 ± 0.21	$\mathbf{99.22 \pm 0.13}$	99.18 ± 0.19	$\mathbf{97.36 \pm 0.12}$	$\mathbf{92.59 \pm 1.26}$	$\mathbf{86.52 \pm 0.33}$

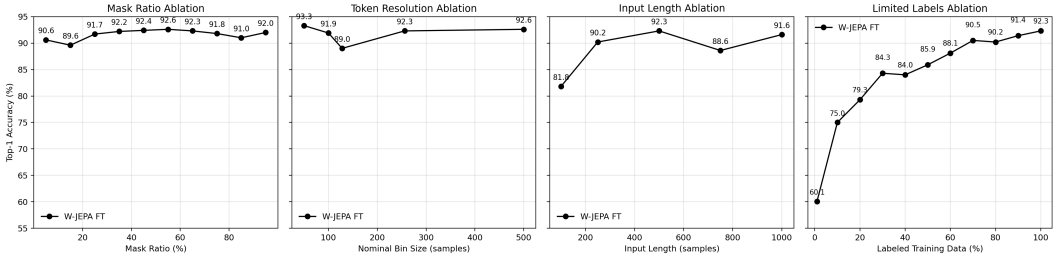


Figure 5: One-factor-at-a-time sensitivity study on UCI-HAR. From left to right, the panels vary the fraction of token positions masked during pretraining, the number of temporal bins used to summarize each wavelet band (shown as the nominal input samples per bin), the input-window length, and the requested fraction of labeled fine-tuning windows. The shared baseline uses a 65% mask ratio, two bins per band, and 500-sample windows. Points report top-1 test accuracy from one run; nominal bin sizes are not literal wavelet receptive fields.

and 10–100% of the training pool in ten-percentage-point increments. These experiments assess hyperparameter and data sensitivity; they do not test whether wavelets or the JEPA objective are necessary.

D Additional Analyses

D.1 Exercise-bike example details

Signal and example selection. Figure 1 uses the left-leg gyroscope z-axis (`LL_zgyro`) from DSADS. Panels (a) and (b) show horizontal and vertical exercise-bike cycling from WHAR subject 7, displayed as participant 8 in panel (c). The windows begin at 264.25 s in session 119 and 37.27 s in session 127, respectively. The raw Pearson correlation is 0.992. Signals are resampled from 25 to 100 Hz and divided into 500-sample windows that remain within session boundaries. No amplitude normalization or post-extraction time shifting is applied.

D.2 Attention by timescale under input noise

This section details the attention analysis summarized in §4.2. We examine how the fine-tuned encoder allocates attention across wavelet bands on held-out DSADS participants (Figure 6). Attention received by tokens in the final encoder layer (layer 8) is averaged over heads and query positions, then summed within each band. The middle heatmap divides each band's attention by the uniform allocation of $1/5$, so a value of 1 denotes uniform attention.

Added noise is measured against a fixed, training-derived per-channel dynamic RMS reference, and attention is averaged over three different randomly generated noises per window. Both change panels show $100(a_{\text{noise}} - a_0)/a_0$, calculated after aggregating attention within each band. Shaded area shows 95% bootstrap intervals for the mean attention change, estimated by resampling held-out participants while keeping their clean and noisy measurements paired.

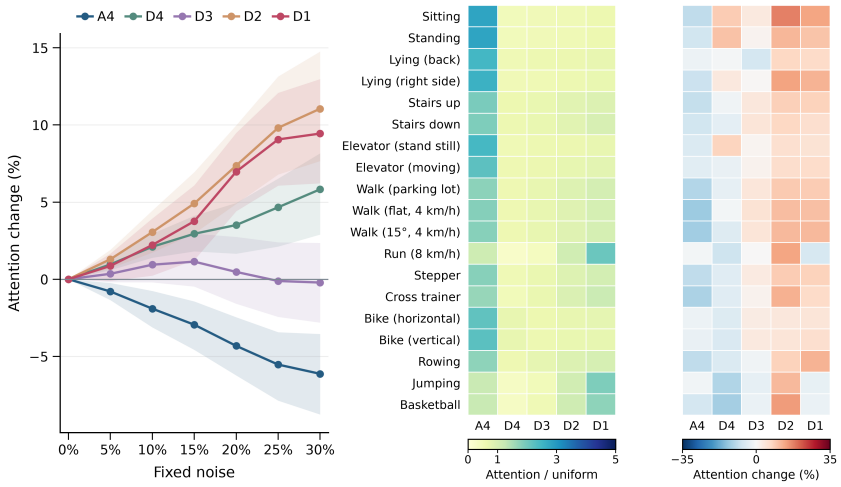


Figure 6: **Wavelet-scale attention on held-out DSADS participants.** Left: relative attention change across noise doses across all activities; Middle: clean attention by activity relative to uniform allocation. Right: relative attention change by activity at 30% noise. All panels use the final encoder layer.

Table 7: DSADS full-test performance under fixed input noise on a single seed. Nonzero doses report mean ± sample standard deviation across three noise realizations; clean inputs were evaluated once.

	Held-out participants		Random windows	
Noise (%)	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)
0	81.73	81.77	99.28	99.28
5	81.65 ± 0.02	81.58 ± 0.02	99.22 ± 0.01	99.21 ± 0.01
10	81.87 ± 0.03	81.67 ± 0.02	99.05 ± 0.04	99.04 ± 0.04
15	81.14 ± 0.07	80.77 ± 0.11	98.90 ± 0.02	98.88 ± 0.02
20	81.00 ± 0.05	80.73 ± 0.08	98.66 ± 0.04	98.63 ± 0.04
25	81.72 ± 0.11	81.59 ± 0.15	98.03 ± 0.03	97.98 ± 0.03
30	81.77 ± 0.21	81.60 ± 0.26	97.14 ± 0.09	97.05 ± 0.09

On clean inputs, A_4 receives the largest attention share for the activities, whereas D_1 receives the largest share for running and jumping. At 30% added noise, pooled A_4 attention falls by 6.1% relative to clean inputs, while D_2 and D_1 rise by 11.1% and 9.4%.

Full-test performance remains largely intact across the tested noise levels (Table 7). At 30% noise, held-out accuracy rise slightly, while random-window accuracy and macro-F1 fall by about two points. Thus, the attention shift toward detail bands accompanies preserved classification performance under added noise. This is consistent with greater use of fine-scale tokens when classifying noisy inputs, although attention weights alone cannot establish that the shift causes the observed robustness.

D.3 VISUALIZATION OF LEARNED REPRESENTATIONS

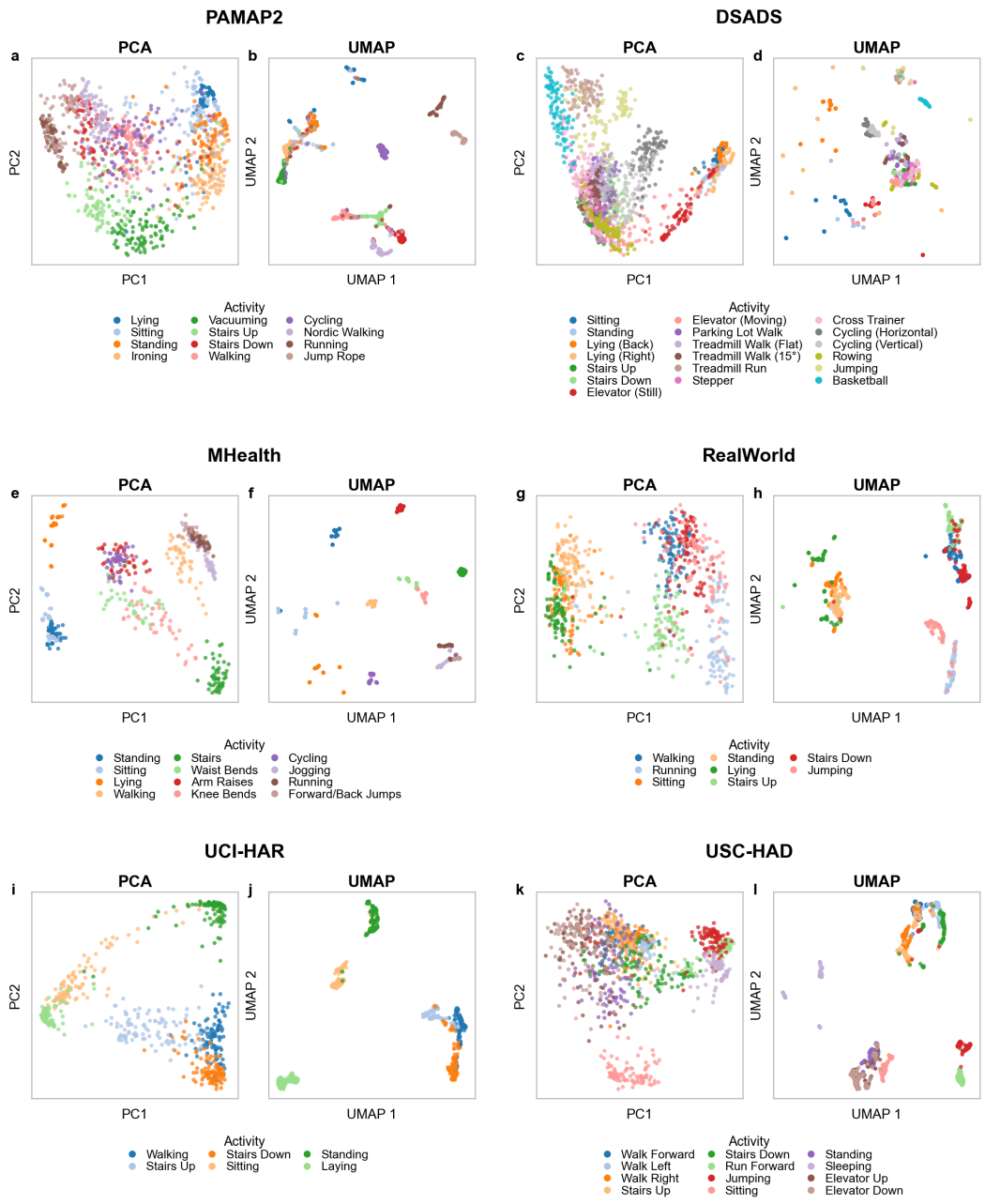


Figure 7: PCA and UMAP projections of W-JEPA's final-block representations across six human activity recognition datasets. Each point represents a test window, colored by activity class within each dataset.