TransfHAR: Self-Supervised Wrist Representations for On-Demand Activity Recognition

Aidan Bradshaw
Computer Science
Northwestern University
Evanston, Illinois, USA
aidanbradshaw2025@u.northwestern.edu

Riku Arakawa
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
rarakawa@cs.cmu.edu

Xin Liu
Google
Seattle, Washington, USA
xliu0@cs.washington.edu

Karan Ahuja
Computer Science
Northwestern University
Evanston, Illinois, USA
kahuja@northwestern.edu

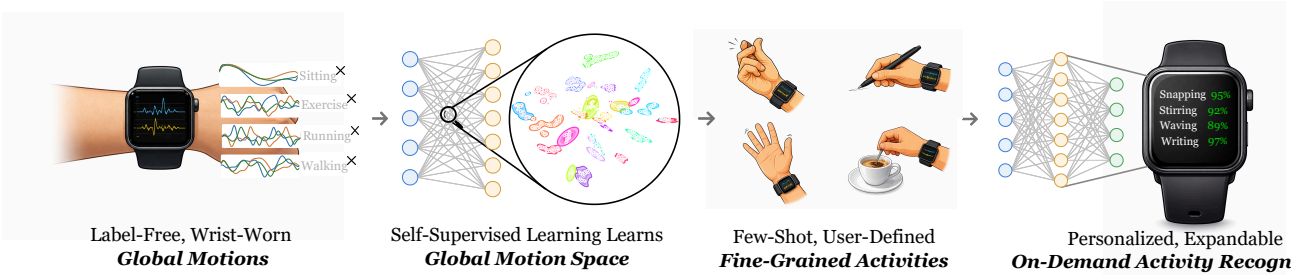

Figure 1: TransfHAR pretrains a ViT-1D encoder with self-supervised masked autoencoding on IMU data of global wrist motions (running, walking, sitting; left). The learned representations transfer to fine-grained, user-defined activities unseen in pretraining (snapping, stirring, writing; center). At deployment, a lightweight linear probe on the frozen encoder learns from a few on-watch demonstrations, enabling personalized, expandable on-demand recognition (right).

Abstract

Fine-grained wrist activity recognition can support applications such as procedural step guidance and context-aware assistance, yet acquiring labeled data for every new task, user, and activity granularity remains a bottleneck. We present TransfHAR, a self-supervised wrist IMU framework for on-demand, fine-grained activity recognition by learning transferable motion priors from global, unlabeled activities. We show that self-supervised pretraining on coarse wrist IMU activities (e.g., sitting, walking, exercise) learns motion structure rich enough to transfer to fine-grained manipulative, gestural, and procedural activities (e.g., snapping, stirring, waving) that are absent from pretraining. We implement TransfHAR as a real-time smartwatch application that lets users define and expand their own activity set for personalized recognition from only a few demonstrations. Across three offline cross-dataset evaluations, TransfHAR matches or exceeds fully supervised baselines that use complete label sets with equal or additional sensor channels, by 6.2 balanced-accuracy points on average. In an in-lab study with 10 participants each performing seven novel wrist activities, TransfHAR reaches 86.7% balanced accuracy across participants with five examples per class and 90.4% when updated from a single one-minute recording per class. These results indicate that broad self-supervised wrist pretraining provides an effective foundation for on-demand fine-grained activity recognition.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; *Empirical studies in ubiquitous and mobile computing*; • **Computing methodologies** → *Machine learning*; Feature representation.

Keywords

Wrist-worn IMU, Self-supervised learning, On-device personalization, Few-shot activity recognition

UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/26/11
https://doi.org/10.1145/3830398.3830553

ACM Reference Format:
Aidan Bradshaw, Riku Arakawa, Xin Liu, and Karan Ahuja. 2026. TransfHAR: Self-Supervised Wrist Representations for On-Demand Activity Recognition. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3830398.3830553

1 Introduction

Everyone has different habits, shortcuts, and mannerisms that shape how they move through their daily lives. This has inspired wrist-worn activity recognition systems that can classify a breadth of human activities directly on a smartwatch, ranging from fine hand gestures [27, 50, 55] to activities of daily living [8, 58], enabling new applications that support habitual, routine, and diverse context-aware interactions [2, 3, 5]. These types of systems are typically designed around pre-defined activity sets, assumptions about persisting activity granularity, and models trained to generalize across users rather than adapt to individual routines [8, 25, 64]. In practice, however, benchmark-driven activity classes rarely match the activities users actually need. Current models are trained on fixed label vocabularies that cannot be extended after deployment, and adding new activities requires collecting fresh labeled data and retraining from scratch [36, 53, 58]. A barista may want to track the steps of their pour-over routine, a rehabilitation patient may need to monitor exercises prescribed by their therapist, or a developer may want gesture shortcuts tailored to their workflow. Most of these personalized activities do not exist in any pre-existing training set, and each is defined by the user's own context, tools, and goals. To maximize utility for end users, the underlying models should be built to optimize for flexibility across motion types [31] and support the self-definition of people's own activities and mannerisms.

The direct approach to this problem, collecting a labeled dataset for each new user-defined activity and training a task-specific classifier, does not scale to the diversity of activities people actually care about, and training from scratch with limited examples is ineffective. Some prior work addresses this through gesture customization or incremental adaptation [29, 55], but these approaches typically extend a fixed gesture vocabulary within a single interaction domain rather than supporting arbitrary new activities across motion and behavior types. The underlying challenge is that fine-grained manipulative, gestural, and procedural activities, the kinds most relevant to personalization, are scarce in public wrist IMU datasets, which are overrepresented by coarse locomotion, posture, and exercise data [11, 14, 38]. However, since these datasets are collected from the same wrist-worn IMU modality, they still capture many of the underlying motion primitives that reappear in fine-grained personalized activities, even when those subtasks are not explicitly labeled. In this case, self-supervised learning could offer a promising path forward, as prior work in wearable sensing and time-series learning has shown that label-free objectives can learn transferable, within-domain motion manifolds from large amounts of unlabeled data while reducing dependence on expensive manual annotation [48, 59, 61, 62]. Despite the many differences in public datasets' hardware, protocols, and labeling schemes, IMU signals are still governed by the same wrist kinematics, which suggests that representations learned from broad global motion may transfer to fine-grained personalized recognition.

In this paper, we present **TransfHAR**, a self-supervised wrist IMU framework for on-demand activity recognition. TransfHAR enables personalized activity recognition from a smartwatch, customized to users' daily mannerisms and preferences. By pooling global motion data from public datasets, we show that training a ViT-1D encoder with self-supervised learning can learn general IMU structure that links different levels of activity types and granularities, in which simple linear probes can adapt effectively to new downstream vocabularies. Our central insight is that global coarse wrist motion provides an effective foundation for fine-grained activities that never appear during pretraining. Although public wrist datasets are dominated by locomotive and postural movements, they still contain recurring motion structure, oscillations, impacts, rotations, and cross-axis coordination patterns that reappear in everyday manipulations, gestures, and task steps.

We evaluate TransfHAR in both offline and interactive settings. Offline, we test transfer ability on three held-out wrist IMU benchmarks spanning manipulative, procedural, and gesture-like activity regimes that are excluded from pretraining, and show that a frozen encoder with a single-layer linear probe trained on each dataset's training split outperforms the supervised baseline on all three datasets, by 8.1 points on SAMoSA, 5.5 on PrISM latte-making, 6.2 points on average across all eight PrISM procedures, and 5.0 on UTD-MHAD, averaging a 6.2-point gain in balanced accuracy. Online, we deploy TransfHAR as a real-time smartwatch system and evaluate it with 10 participants performing seven user-defined activities each, reaching 86.7% balanced accuracy with only five labeled windows per class and 90.4% balanced accuracy from a single one-minute recording per class. We show that self-supervised pretraining on broad, coarse wrist motion can serve as a practical foundation for rapid, user-defined, fine-grained activity recognition in interactive systems.

In this paper, we make the following contributions:

(1) **TransfHAR**, a self-supervised wrist IMU framework that learns reusable motion representations from heterogeneous public datasets and supports rapid on-device adaptation for user-defined activity recognition.
(2) A real-time smartwatch system that enables users to define activity labels, provide a small number of demonstrations, train a lightweight personalized classifier on-device, and receive real-time, on-demand activity predictions, reaching 86.7% balanced accuracy using 12.8 seconds of demonstration per class and 90.4% using one minute per class.
(3) An empirical cross-dataset evaluation showing that frozen self-supervised wrist representations transfer across held-out manipulative, procedural, and gesture-like benchmarks, exceeding supervised baselines by 6.2 balanced-accuracy points on average.

2 Related Work

2.1 Self-Supervised Pretraining for Wearable Sensing

Self-supervised learning has become a standard approach for learning transferable representations from unlabeled data across language [15, 30], vision [10, 22], and time series [60, 63] through objectives such as masked prediction [21] or contrastive learning [13]. Wearable sensing is particularly well suited for this regime since large-scale inertial datasets are expensive to annotate, often heterogeneous in sensor configuration and collection protocol, and typically require substantial preprocessing before downstream use.

Table 1: Public wrist IMU datasets consolidated in TransfHAR and their role in pretraining and probing. All sources are converted to a wrist-only 50 Hz continuous stream in a shared forward-left-up (FLU) coordinate convention.

Dataset	Subj.	Hours	Original Placement	Hz	Axes	Classes	Role
RecoFit [38]	114	79.8	Right forearm	50	6 (A+G)	28	Pretrain
PAMAP2 [40]	9	~10	Wrist, chest, ankle	100	9 (A+G+M)	18	Pretrain
Opportunity++ [14]	4	19.75	Full body (7 IMU + 12 Acc)	30	17 (A+G+M)	17	Pretrain
WISDM [23]	51	45.9	Watch (dom. wrist) + phone	20	6 (A+G)	18	Pretrain
Shoaib [42]	10	6.5	Wrist (phone) + pocket	50	6 (A+G)	13	Pretrain
WEAR [9]	22	19	Both wrists + ankles	50	3 (A)	18	Pretrain
Capture-24 [11]	151	2,562	Dominant wrist	100	3 (A)	206	Pretrain
UT-Watch [8]	20	~17	Dominant wrist	50	6 (A+G)	23	Pretrain
SAMoSA [36]	20	14.2	Dominant wrist	50	9 (A+G+O)	27	Probe
PrISM-Tracker [6]	14	~2.7	Right wrist	50	9 (A+G+O)	8–19	Probe
UTD-MHAD [12]	8	~1.6	Right wrist (1–21), right thigh (22–27)	50	6 (A+G)	27	Probe

Prior work has explored a range of self-supervised objectives for wearable signals, including multi-task transformation prediction [41], cross-dimensional motion prediction [45], contrastive learning [20, 26, 47], and masked reconstruction [18]. More recent efforts have scaled this direction through broader empirical assessment, larger pretraining corpora, and stronger pipelines [19, 32–34, 59], demonstrating that inertial signals contain substantial latent structure learnable without semantic labels. However, most evaluations still focus on aligned benchmark settings where pretraining and downstream tasks remain similar in activity regime, dataset structure, or label space [25, 46, 59], where the downstream model is fine-tuned end to end rather than used as a fixed feature extractor [46, 59]. This leaves open the question central to our work: whether broad self-supervised pretraining on predominantly coarse wrist motion can produce a frozen representation that remains useful for qualitatively different fine-grained manipulative, gestural, procedural, and user-defined activities.

2.2 Limited-Label Adaptation in HAR

Most HAR systems train supervised task-specific models for a fixed dataset, sensor configuration, and activity vocabulary [17, 37] defined a priori. When models do address transfer learning, the focus is on cross-dataset or cross-user generalization, usually within the same activity granularity with relatively small label spaces of fewer than 10 classes. For example, LIMU-BERT adapts BERT-style masked prediction to unlabeled IMU sequences [53], UniHAR applies physics-informed augmentation grounded in the IMU sensing process [52], and CrossHAR combines sensor augmentation with hierarchical self-supervised pretraining to learn more generalizable representations [25]. These methods focus on improving robustness to domain shift and label efficiency, but the downstream task structure is largely preserved. A separate line of work studies label alignment more directly through few-shot learning and invariant feature methods using limited labeled target data [16, 24, 49, 57]. However, these approaches generally assume that target activities share the granularity and environment of the source data, leaving underexplored the case most relevant to interactive wearable systems, where target activities may be sparse, personalized, defined after deployment, and qualitatively different from the coarse motions seen during pretraining.

2.3 Customization in Wearable Activity Systems

A complementary thread of work in wearable computing asks how end users can define, customize, or extend recognition systems with minimal effort. Prior smartwatch and on-body systems have shown that wrist-worn sensing can support fine-grained recognition of predefined behaviors, including finger motions on commodity smartwatches [50], bio-acoustic hand and object interactions [28], and daily hand activities such as writing and stirring [27, 36]. This literature establishes the feasibility of rich wrist-level interaction, but typically assumes that the activity set is specified during system design and training.

In user customization work, there has been a focus on using demonstration-based learning, template- and distance-based recognition, and lightweight personalized models. uWave, for example, supports user-defined gestures from a few demonstrations using template-based matching [29]. More recent smartwatch work brings this idea closer to on-device adaptation. Xu et al. show that users can extend an existing wrist-worn recognizer with one to five examples by attaching a lightweight user-specific branch to a supervised pretrained model while preserving the original gesture set [55]. Building on these approaches to user customization, TransfHAR supports open, user-defined activity vocabularies drawn from a general self-supervised motion prior, extending beyond predefined gesture sets and single interaction domains.

3 Datasets

3.1 Normalization and Alignment

TransfHAR uses pooled public wrist IMU data, which differ substantially in sensor hardware, sampling rate, placement, channel availability, axis convention, and units. We therefore standardize all sources in Table 1 into a common smartwatch-oriented representation. Each dataset retains a wrist-only IMU stream, resampled at 50 Hz with a shared forward-left-up (FLU) axis convention and consistent physical units. We use the dominant wrist as the canonical reference when available, since it is the most common placement across our sources and best matches smartwatch deployment. In total, the consolidated corpus spans 381 subjects recorded in homes, laboratories, free-living settings, and workshops. All

Table 2: Motor complexity taxonomy by signal structure.

Level	Category	Signal Character	Examples
L1	Quasi-static	Gravity-dominated, slow postural drift, low energy above DC	Sitting, Standing, Lying, Watching TV, Car driving
L2	Locomotor	Quasi-periodic, dominant spectral peak	Walking, Running, Cycling, Repetitive exercise
L3	Manipulative	Aperiodic, variable amplitude, no dominant frequency	Chopping, Brushing teeth, Typing, Drilling
L4	Procedural	L3-like per window, separable only with temporal context	Action steps for cooking, Latte-making

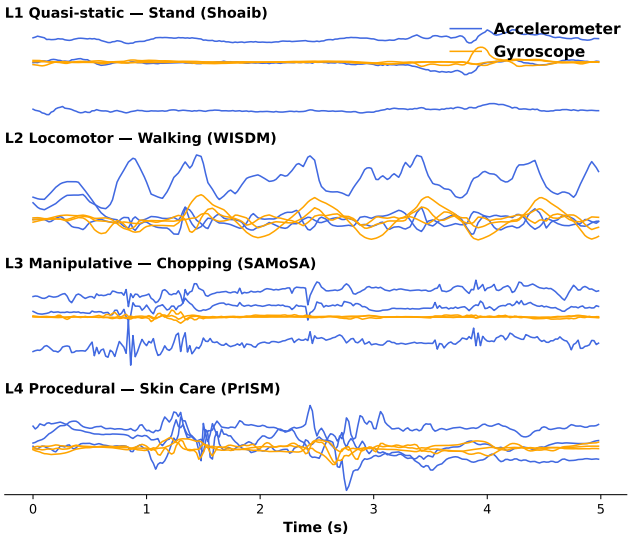

Figure 2: Representative 5-second wrist IMU windows for each motor complexity level from public datasets.

dataset annotations and benchmarks are made publicly available at: https://github.com/Abradshaw1/IMU_LM_Data.

3.2 Activity Taxonomy

Beyond sensor alignment, a second challenge in pooling wrist datasets is inconsistency in how activities are described. HAR papers frequently label datasets as coarse or fine-grained, but these distinctions are often based on class count or naming conventions rather than discriminability from a wrist IMU signal. We therefore define a four-level motor complexity taxonomy grounded in dominant wrist signal characteristics, summarized in Table 2.

L1 activities are gravity-dominated with minimal energy above DC and are often separable with simple statistics. L2 activities exhibit quasi-periodic structure with a dominant spectral peak, and include repetitive exercises whose discriminative signature is periodic regardless of semantic label. L3 activities are aperiodic

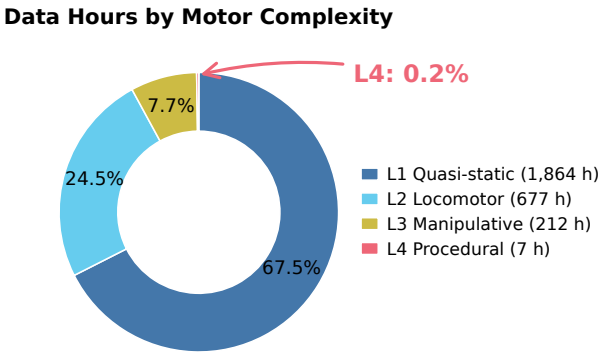

Figure 3: Distribution of data hours by motor-complexity level. The data landscape is dominated by L1/L2 activities, with comparatively limited L3 and negligible L4 coverage.

with variable-amplitude bursts, no stable dominant frequency and typically require learned representations. L4 activities are locally similar to L3 within short windows but become separable only with temporal context across an ordered sequence of steps. Figure 2 visualizes representative windows from each level, showing that the defining distinctions are signal-level rather than semantic. Figure 3 shows the distribution of L types after aggregating all the datasets, exposing a practical imbalance in public wrist IMU data. Most large datasets emphasize posture, locomotion, and exercise at L1/L2, while fine-grained manipulative and procedural motions at L3/L4 remain much scarcer.

4 TransfHAR Framework

We design TransfHAR as a two-stage framework. In Stage A (Figure 4), we pretrain a ViT-1D encoder on pooled public wrist IMU data (3-axis accelerometer and 6-axis accelerometer + gyroscope variants) using a self-supervised masked reconstruction objective over L1/L2 activities, where public data is most abundant. In Stage B (Figure 5), we freeze this encoder and train a lightweight linear probe on top of its representations to classify user-defined L3/L4 activities. We port the frozen encoder from Stage A and an adaptable Stage B into a single framework on a smartwatch, where users can define activities, record a few demonstrations, train the probe, and receive real-time predictions over their personalized label set. In the remainder of this section, we describe the pretraining framework, the downstream adaptation mechanism, and the watch-based personalization workflow.

4.1 Self-Supervised Pretraining on Global Motion

4.1.1 Pretraining motivation. Public wrist-IMU datasets overrepresent coarse whole-body motion because locomotion, exercise, and daily-living activities are easy to collect at scale. These data come from the same wrist-worn modality as the downstream tasks we target, so the pretraining distribution is shifted mainly in label granularity rather than in hardware or body placement. Fine-grained

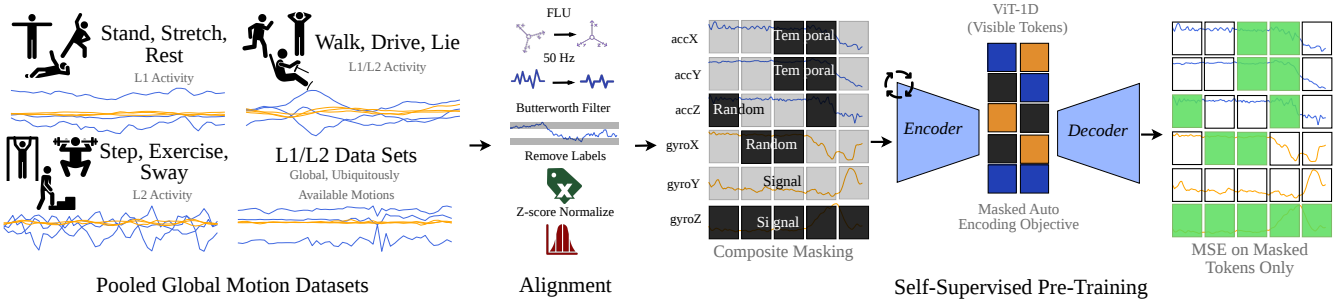

Figure 4: Stage A, self-supervised pretraining on global wrist motion. Public wrist IMU datasets, dominated by coarse L1/L2 activities, are pooled and aligned into a shared representation through a common forward-left-up axis convention, resampling to 50 Hz, low-pass filtering, label removal, and per-channel z-score normalization. The aligned accelerometer and gyroscope channels are tokenized independently, after which a composite masking strategy hides 65% of tokens under random, temporal, or whole-channel patterns. A ViT-1D encoder processes only the visible tokens and a lightweight decoder reconstructs the sequence, with the mean squared error loss computed over masked positions alone.

downstream activities such as manipulations, gestures, and procedural steps are rarely labeled in public corpora, yet they are composed of the same lower-level wrist motion primitives that appear throughout coarser recordings under different or absent category names. A self-supervised objective exploits this overlap by discarding labels and reconstructing masked signal patches [21, 54], learning temporal structure and cross-axis coordination shared across activity types.

4.1.2 Encoder architecture. We adopt a 1D Vision Transformer (ViT-1D) adapted from LSM-2 [54] for wrist IMU signals. The model operates directly on raw time-series inputs using a channel-independent patch tokenization scheme. A shared 1D convolutional kernel (kernel size and stride of 4 samples) is applied independently to each sensor channel, producing 32 patch tokens per channel that are concatenated into a single token sequence. To preserve both temporal and sensor-axis structure, we use a learned 2D positional encoding that decomposes into separate embeddings for time and channel identity. The combined positional embedding is added to each token, allowing the transformer to distinguish temporal position within a channel and the originating sensor axis. The token sequence is processed by a pre-norm transformer encoder with 12 layers, 384-dimensional hidden states, and 6 attention heads, with a 4× expansion in the feed-forward layers and GELU activations. Following the masked autoencoder design, masked tokens are removed prior to the encoder during pretraining so that the transformer operates only on visible tokens. During inference, all tokens are processed and mean-pooled to produce a single 384-dimensional embedding per window.

4.1.3 Input representation and windowing. The continuous IMU signal is segmented into fixed-length windows of 2.56 seconds (128 samples at 50 Hz) with 50% overlap; we validate this choice with a window-length ablation in Appendix A. Each window is normalized per channel using z-score normalization. We apply a 4th-order Butterworth low-pass filter at 24 Hz prior to resampling to reduce aliasing and high-frequency noise from resampling and hardware variability. We consider two input configurations, a 3-axis variant uses tri-axial accelerometer signals, and a 6-axis variant that additionally includes tri-axial gyroscope measurements. In both cases, we use channel-independent patching to produce 32 tokens per channel, resulting in 96 tokens for the 3-axis setting and 192 tokens for the 6-axis setting, all with embeddings having dimension 384. The 3-axis encoder is pretrained on a substantially larger corpus than the 6-axis encoder, since the largest public wrist sources (e.g., Capture-24) are accelerometer-only. Retaining both variants lets us study the downstream value of gyroscope signals alongside the effect of pretraining corpus size. Only corresponding L1/L2 activities were selected from each pretrain dataset (Table 1), with all activity labels discarded during pretraining.

4.1.4 Training details. We pretrain the encoder using a masked autoencoder (MAE) objective [21]. At each training step, 65% of patch tokens are selected for masking and removed from the encoder input, so the transformer processes only the remaining visible tokens. A lightweight decoder (4 layers, 192-dimensional hidden size, 4 attention heads) reconstructs the masked patches after reinserting learnable mask tokens and restoring positional structure. The training loss is mean squared error computed only over masked patch positions, and the decoder is discarded after pretraining. To improve robustness to temporal and channel structure, we use a composite masking strategy that randomly selects one of three modes per sample: (i) random masking over all tokens, (ii) temporal masking across all channels at selected time steps, and (iii) signal masking across all time steps for selected channels. Each mode masks the same fraction of tokens (65%), keeping the visible token count constant. In preliminary experiments, we observed that combining random, temporal, and channel-level masking produced more stable downstream transfer than random token masking alone. This design is also motivated by prior findings that structured masking improves representation quality for IMU data, as explored in LSM-2 [54].

The 3-axis encoder is pretrained on all datasets listed in Table 1, while the 6-axis encoder is trained only on datasets that include gyroscope signals. Datasets used for downstream evaluation are

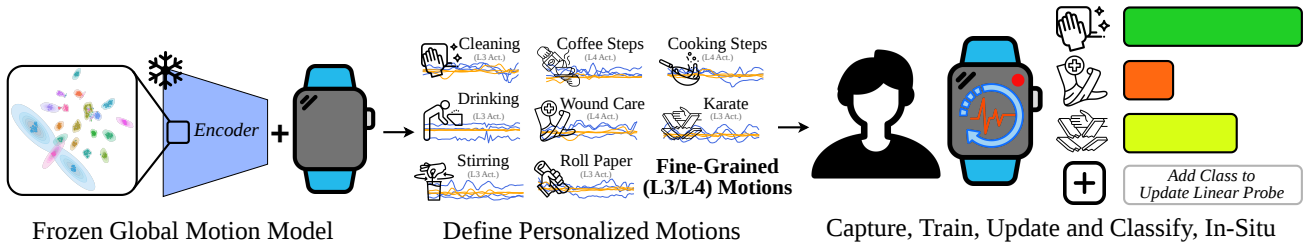

Figure 5: Stage B, on-demand task learning on the watch. The encoder pretrained in Figure 4 is frozen and paired with the smartwatch (left). Users define their own fine-grained L3/L4 activities, spanning manipulations such as stirring and procedural sequences such as coffee-making steps, and record a small number of in-situ demonstrations for each (center). The watch then streams IMU data, encodes each window with the frozen encoder, and classifies it with the on-device linear probe to return real-time confidence over the personalized vocabulary, while an add-class control expands the label set and retrains the probe without modifying the encoder (right).

excluded entirely from pretraining. We optimize using AdamW [35] with a learning rate of 1.5×10^{-4}, $\beta = (0.9, 0.95)$, and weight decay 10^{-4}, with a cosine learning rate schedule and 1,000-step warmup. Training uses mixed precision (FP16), batch size 512, and gradient clipping at norm 1.0 on 4 NVIDIA H100 GPUs. We validate on a 5% subject-held-out split and apply early stopping based on validation reconstruction loss. The encoder checkpoint with the lowest validation loss is used for all downstream experiments.

4.2 On-Demand Task Learning for User-Defined Activities

4.2.1 Participant-specific linear probe and watch workflow. Figure 5 illustrates the transfer and personalization stage (Stage B). After Stage A, the encoder is frozen and a lightweight linear probe is attached on top for transfer learning. Both the 3-axis and 6-axis encoders produce a 384-dimensional mean-pooled embedding per window, which serves as input to the probe. The probe is a single fully connected layer of size $384 \times C + C$, where C is the number of current user-defined activity classes. During adaptation, only the probe parameters are updated while the encoder remains fixed.

On the watch, users define an activity label, record examples in situ, and trigger probe training. During recording, the watch buffers IMU data at 50 Hz, segments it into 2.56-second windows with 50% overlap, and encodes each window with the frozen encoder to produce embedding-label pairs for the current activity set. During live use, incoming windows are encoded and classified by the current probe in real time, and users can iteratively add examples or expand their activity vocabulary without retraining the encoder.

4.2.2 Probe adaptation and training. Because the probe is a single linear layer, retraining is fast and can be repeated whenever the user adds examples or changes the activity vocabulary. Probe training uses AdamW with learning rate 5×10^{-4}, weight decay 10^{-4}, gradient clipping at norm 1.0, mixed precision (FP16), and batch size 256. We use inverse-frequency class-weighted cross-entropy loss, select checkpoints by validation balanced accuracy, and apply early stopping with patience 40.

4.3 Implementation and Deployment

4.3.1 Prototype pipeline. We implement TransfHAR on Apple Watch Series 10 (Figure 5). The application is written in Swift and uses Core Motion for IMU acquisition, local buffering, windowing, on-device encoder inference, storage of embedding-label pairs, probe training, and real-time classification. The ViT-1D encoder is trained offline in PyTorch and converted to Core ML for deployment, allowing the full interaction loop of capture, train, and classify to run locally on the watch without requiring a paired phone or remote server.

4.3.2 Latency and footprint. The system is designed so that expensive representation learning is performed offline, while only lightweight personalization occurs on-device. Predictions are generated from 2.56-second windows with 50% overlap, and once a window is available, encoder inference and probe prediction complete in approximately 23.8 ms on Apple Watch Series 10, yielding a refresh interval of 1.3 s including stride. For a 7-class set, probe training completes in 709, 706, 794, and 786 ms at $K = 1, 5, 10, 20$ respectively, fast enough to run between demonstrations. The deployment footprint remains small because the watch stores only the encoder weights, current probe parameters, label set, and cached embedding-label pairs; the encoder has 21.4M parameters, requires 8.83 GFLOPs per window, and occupies approximately 40.8 MB at FP16, while the probe adds only 385 parameters per additional class. Apple's watchOS does not expose fine-grained thermal or power telemetry, so we report latency and GFLOPs as compute-cost proxies rather than direct battery or thermal measurements.

5 Offline Evaluation

In order to evaluate whether representations pretrained on broad L1/L2 wrist motion transfer to held-out fine-grained L3/L4 activities, we probe the frozen encoder on three public wrist IMU benchmarks spanning manipulative, procedural, and gesture-like activities.

5.1 Datasets

5.1.1 SAMoSA. SAMoSA [36] contains 27 everyday activity classes collected from 20 participants wearing a smartwatch in home and

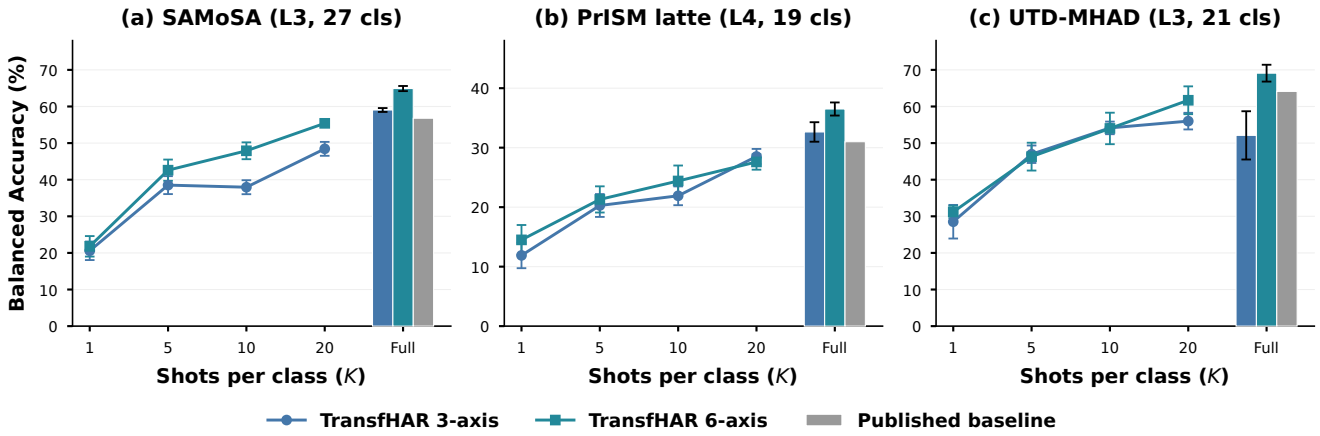

Figure 6: Balanced accuracy as a function of shots per class (K = 1 shot = 2.56s) on three held-out probe targets, with the full-split setting as the bar chart. Supervised baselines are shown as gray bars for each dataset.

workshop environments. The dataset includes synchronized 50 Hz wrist IMU and audio; we use only the motion stream. Activities span kitchen, bathroom, and workshop manipulations, placing it primarily in the L3 regime.

5.1.2 PrISM-Tracker. PrISM-Tracker [6] is a procedural activity dataset collected from smartwatch motion and audio during multi-step daily tasks. The full dataset contains eight procedures, each composed of multiple labeled task steps. We evaluate TransfHAR on PrISM in two ways. First, we use the latte-making procedure, which contains 19 discrete steps performed across 23 sessions from 14 unique participants, for both few-shot and full-split comparison. Second, we compare against the PrISM motion-only baselines across all procedures in the dataset. These tasks are locally similar within short windows and must be distinguished within flexible multi-step sequences, placing PrISM primarily in the L4 regime.

5.1.3 UTD-MHAD. UTD-MHAD [12] is a multimodal human action dataset containing depth, RGB, skeleton, and inertial recordings for 27 actions performed by 8 subjects. We use only the inertial modality, restricted to the 21-class wrist subset (UTD1) and compare with the published inertial baseline from Yang et al. [56]. This subset contains short, burst-like gesture and exercise motions such as arm swings, drawing motions, knocks, and throws, representative of the L3 regime.

5.2 Protocol

For probe training and evaluation, we use each dataset's evaluation protocol. Since PrISM does not report all-procedure performance, we pull their original code and run their motion-only model using their data and protocol. For SAMoSA we also pull their code and run their model under the same protocol to produce baselines. We evaluate both 3-axis and 6-axis encoders, each paired with its corresponding pretrained checkpoint.

5.2.1 Few-shot and full-split conditions. We evaluate two probe training regimes. In the *few-shot* regime, the probe is trained on $K \in \{1, 5, 10, 20\}$ labeled windows per class, sampled uniformly at random from the training split rather than taken contiguously from its start. One shot is one 2.56-second window, so $K = 1, 5, 10, 20$ correspond to 2.56, 12.8, 25.6, and 51.2 seconds of labeled data per class. Only the training split is subsampled; validation and test sets are unchanged across all K. In the *full-split* regime, the probe is trained on all windows in the dataset's training partition. In both regimes the encoder is frozen and only the probe is optimized, and evaluation follows each published baseline's protocol. All results are reported as the mean over five random seeds with the standard deviation in parentheses.

5.2.2 Metrics. We report balanced accuracy to account for class imbalance across benchmarks, defined as: Balanced Accuracy = $\frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FN}_c}$. Balanced accuracy is used as the primary model-selection metric during probe training. For supervised comparison, we benchmark against the corresponding motion-only baseline for each dataset. The SAMoSA and PrISM baselines use 9-axis input (accelerometer, gyroscope, and magnetometer/orientation), while the UTD-MHAD baseline uses 6-axis accelerometer and gyroscope input; all are fully supervised.

5.3 Offline Results

5.3.1 Few-shot scaling. Figure 6 shows balanced accuracy as a function of shots per class on the three held-out probe targets, with the full-split setting included as the final point for reference. Across all three datasets, 6-axis performance rises sharply between K=1 and K=5, then improves more gradually through K=20 before reaching its highest value in the full-split setting. For example, on SAMoSA the 6-axis probe increases from 21.8% (SD = 2.8) at K=1 to 42.6% (SD = 2.9) at K=5, and on UTD-MHAD it rises from 31.2% (SD = 1.7) to 46.3% (SD = 3.8) over the same range. This early jump suggests the frozen encoder already provides class boundaries that additional data refines rather than creates.

As K increases, performance continues to improve but with diminishing returns, reaching 55.4% (SD = 0.9) on SAMoSA and

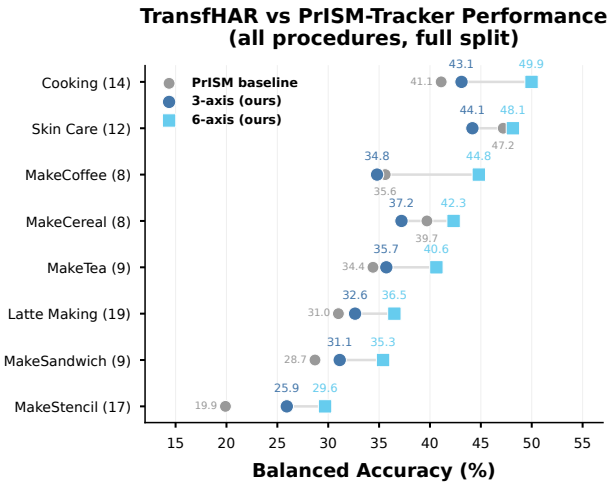

Figure 7: Full-split balanced accuracy across all eight PrISM-Tracker procedures for 3-axis and 6-axis encoders. Number of procedure steps is shown in parentheses.

61.7% (SD = 3.8) on UTD-MHAD at K=20, before rising to 64.9% (SD = 0.7) and 69.1% (SD = 2.3) in the full-split setting. In contrast, gains on PrISM latte-making are smaller, increasing from 14.5% (SD = 2.5) to 27.6% (SD = 1.3) for 6-axis between K=1 and K=20 and reaching 36.5% (SD = 1.1) with the full training split. This reflects the underlying task structure. While L3 activities are often separable from short wrist windows, L4 procedural steps are more locally similar and depend more strongly on temporal context beyond a single 2.56-second segment, consistent with our window-length ablation (Appendix A), where PrISM latte-making alone benefits from longer 5.12 s windows. Across datasets, the 6-axis encoder outperforms the 3-axis variant at nearly every supervision level, with the two variants at parity on UTD-MHAD at intermediate shot counts, and the magnitude of the gain depends on the activity type. The gap is larger on SAMoSA, e.g. 55.4% (SD = 0.9) vs. 48.4% (SD = 1.9) at K=20 and 64.9% (SD = 0.7) vs. 59.0% (SD = 0.5) in the full-split setting, where rotational dynamics are important, and smaller on UTD-MHAD at intermediate shot counts, where burst-like motions are already well captured by acceleration alone.

5.3.2 Full-split baseline comparison. In the full-split setting (Figure 6), the 6-axis TransfHAR probe exceeds the corresponding supervised baseline on all three datasets: 64.9% (SD = 0.7) vs. 56.8% on SAMoSA, 36.5% (SD = 1.1) vs. 31.0% on PrISM latte, and 69.1% (SD = 2.3) vs. 64.1% on UTD-MHAD. On UTD-MHAD, this full-split advantage is also substantial relative to the 3-axis variant, with 69.1% (SD = 2.3) for 6-axis versus 52.1% (SD = 6.6) for 3-axis. Across these three benchmarks, the frozen TransfHAR probe exceeds the corresponding published supervised task-specific pipelines. Notably, the gains on SAMoSA and PrISM require only accelerometer and gyroscope inputs against 9-axis baselines, suggesting that self-supervised pretraining recovers useful rotational structure without explicit orientation features. Per-class confusion structure for all three benchmarks is given in Figure 14.

Table 3: Architecture-matched comparison on the full training split, all using the same 6-axis ViT-1D. Balanced accuracy, mean (SD) over five seeds.

Training regime	SAMoSA	PrISM latte	UTD-MHAD
From scratch (no SSL)	28.8 (3.4)	16.2 (3.1)	46.4 (2.2)
Pretrained, fine-tuned	61.4 (0.8)	34.4 (0.8)	65.4 (3.6)
Pretrained, frozen + probe	**64.9 (0.7)**	**36.5 (1.1)**	**69.1 (2.3)**

5.3.3 Transfer across all PrISM procedures. While our main PrISM results focus on the latte-making benchmark, we additionally evaluate full-split transfer across all eight procedures in the dataset to assess how broadly the learned representation generalizes within the L4 procedural regime. Figure 7 shows that the 6-axis encoder matches or exceeds the fully supervised PrISM 9-axis baseline on every procedure, improving balanced accuracy by 6.2 points on average, while the 3-axis encoder is close to parity at 0.9 points and falls below the baseline on three procedures. The largest gains appear on *MakeStencil* (+9.7 pp) and *MakeCoffee* (+9.2 pp), suggesting that gyroscope information is especially valuable for procedures whose steps remain difficult to separate from short wrist windows alone.

At the same time, the figure shows that transfer difficulty varies substantially across procedures. Performance is strongest on *Cooking* (49.9%) and *Skin Care* (48.1%), but remains much lower on *MakeStencil* (29.6%) and *MakeSandwich* (35.3%), where individual steps are likely more locally similar within short windows. This pattern highlights the limitation of window-level classification for L4 activities, where procedural steps are often distinguished more by temporal context than by instantaneous wrist motion.

5.3.4 Architecture-matched comparison. Published baselines differ from our encoder in both architecture and input modality, so we additionally train two supervised variants of the same 6-axis ViT-1D under the identical full-split protocol (Table 3). Training from scratch falls far below both pretrained variants despite identical capacity, indicating that transfer stems from the learned representation rather than the architecture. Fine-tuning the encoder offers no advantage over freezing it, suggesting that a linear probe may be better suited for model adaptation.

6 On-Demand Personalization Study

The offline evaluations in Section 5.3 use fixed public benchmarks. We now evaluate TransfHAR in the intended setting where users define their own wrist activities on a smartwatch and obtain recognition from a few demonstrations using the frozen encoder and a per-participant linear probe.

6.1 Study Design

6.1.1 Participants and apparatus. We recruited 10 participants (7 right-handed, 3 left-handed; mean age 26.1, SD = 2.0) from our university community. The study was approved by the Institutional Review Board, and participants received $10 compensation. Each session lasted approximately 30 minutes. Participants were

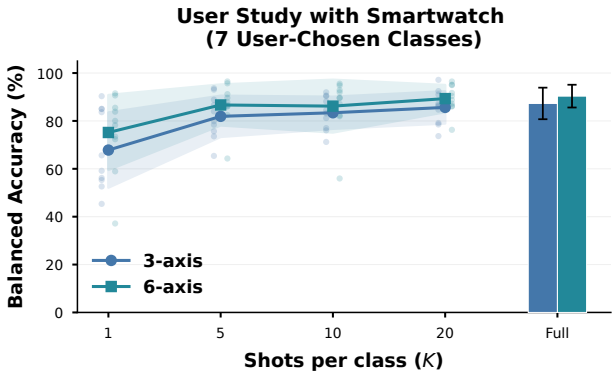

Figure 8: Personalization performance across participants as a function of shots per class (K = 1 shot = 2.56s), with the full-session setting included as the bar chart. Points show per-participant balanced accuracy and lines show participant means for the 3-axis and 6-axis encoders.

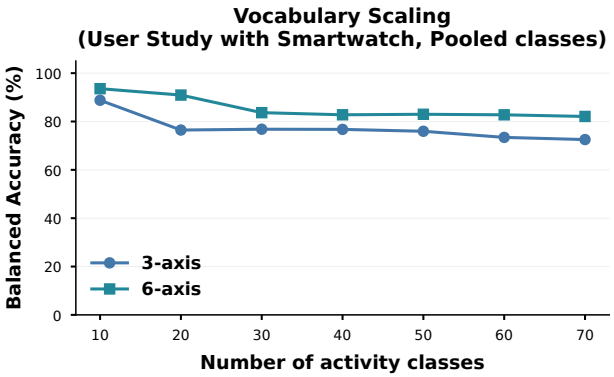

Figure 9: Vocabulary scaling in the user study with all participants' activities pooled into a single non-personalized probe, and balanced accuracy measured as the number of classes the probe must distinguish from 10 to 70. Accuracy declines gradually, with most of the drop occurring below 30 classes.

instructed to wear the Apple Watch on their dominant wrist. Accordingly, 7 participants wore the watch on the right wrist and 3 on the left wrist.

6.1.2 Activity selection and data collection procedure. Participants were shown a bank of 55 candidate activities (Appendix Table 5) spanning hand and air gestures (e.g., drawing a heart in the air, waving goodbye), object manipulation (e.g., stirring coffee with a spoon, opening a laptop lid), workspace actions (e.g., typing on a keyboard, writing with a pen), and personal motions (e.g., brushing hair, filing nails). The bank was designed to encourage choosing L3/L4 granularity and discourage selection of L1/L2 coarse activities like walking or jogging. Participants selected 7 activities from this list or defined their own, after being told: *Below is a list of example daily wrist activities to illustrate the level of granularity we are targeting. Please choose activities from this list, or define your own, based on what would be most useful in your daily life.*

For each activity, we recorded 3 separate 1-minute sessions of continuous IMU data while the participant repeatedly performed that activity. Between sessions, participants removed the watch and took a 30-second break before re-wearing it, introducing natural variation in watch placement and fit. During recording, the participant selected the activity label in the interface, pressed *Start*, performed the activity continuously, then stopped the recording.

6.2 Evaluation Protocol

6.2.1 Segmentation and splits. All recordings were segmented into 2.56-second windows (128 samples at 50 Hz) with 50% overlap, matching the setup used throughout the paper. Evaluation was performed separately for each participant because the class vocabulary differed across users. The study yields 210 one-minute trials (10 participants × 7 activities × 3 sessions), each producing approximately 45 windows. Evaluation is cross-session round robin: for each participant and activity, one session trains the probe and the two held-out sessions form the test set, rotating over all three assignments. Reported values are cross-session and post re-wearing.

6.2.2 Few-shot and full-session probing. We evaluate two probing regimes on top of the frozen encoder. In the *full-session* setting, the linear probe is trained using all windows from the designated training session for each of the participant's 7 activities. Since each training session is one minute long, this condition reflects the performance achievable from a single full recording per class. In the *K-shot* setting, we subsample K training windows per class from that same session, sampled uniformly at random across the session rather than from its beginning. We evaluate $K \in \{1, 5, 10, 20\}$ with the test set fixed as the two held-out sessions.

6.3 Results

We report user-study results on few-shot adaptation, vocabulary scaling, and error structure. Because each participant defined a different set of 7-class activities, the results are aggregated across participants, with participant-specific activity labels provided in the Appendix Table 6.

6.3.1 Personalization and few-shot performance. In the full-session setting, where the probe is trained on one 1-minute recording per class, the 6-axis encoder reaches 90.4% balanced accuracy (SD = 4.8) and 89.9% macro-F1 (SD = 5.3), while the 3-axis encoder reaches 87.3% balanced accuracy (SD = 6.6) and 86.6% macro-F1 (SD = 7.2). Figure 8 shows that personalization remains effective even with very limited supervision. With one labeled window per class (2.56 seconds of user data), the 6-axis encoder reaches 75.2% balanced accuracy on average across participants (SD = 15.6), compared with 67.8% for 3-axis (SD = 15.7). At five shots, performance rises sharply to 86.7% (SD = 8.5) for 6-axis and 81.9% (SD = 8.6) for 3-axis. Gains continue through K = 20, reaching 89.3% and 85.7%, but the largest improvement occurs between one and five shots. By K = 20, performance is already within 1.1 pp of the full-session setting for the 6-axis encoder, suggesting that most of the available personalization benefit can be recovered without collecting a full minute of data per class. Splitting by handedness, right-handed participants (n=7) average 88.7 (5.0) and left-handed participants

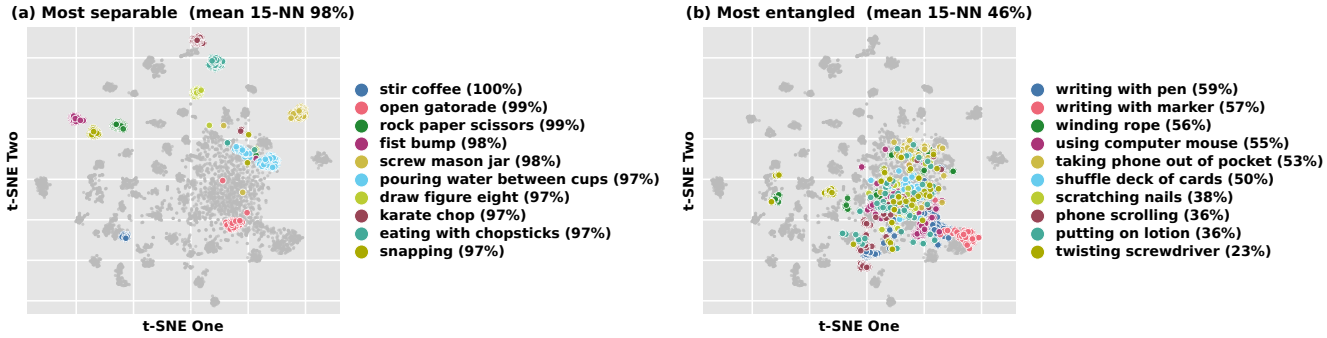

Figure 10: t-SNE of user-study embeddings from the frozen 6-axis encoder, showing the (a) ten most separable and (b) ten most entangled activities against all other windows in gray. Both panels share one projection. Percentages are neighborhood purity, the fraction of each window's 15 nearest neighbors in the 384-dimensional embedding sharing its label.

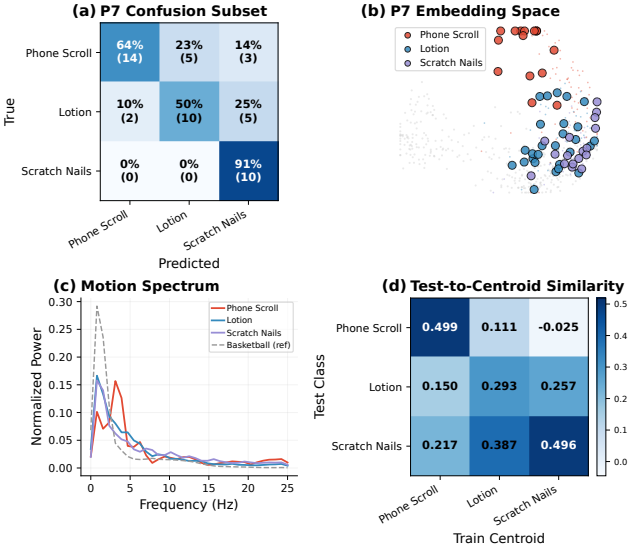

Figure 11: P7 failure analysis in the 6-axis full-session setting. Top left shows the confusion subset for low-performing activities. Top right shows PCA projection of P7 embeddings, with the highlighted classes shown relative to the participant's remaining activities in gray. Bottom left shows normalized acceleration-magnitude spectra. Bottom right shows mean similarity between test samples and train centroids.

(n=3) 94.2 (2.3), though the left-handed group is underpowered (Figure 15). The 6-axis encoder outperforms the 3-axis encoder at every supervision level, with the largest advantage in the lowest-data regime: 7.4 pp at $K = 1$, narrowing to 3.1 pp in the full-session setting. Participant variability is highest in the one-shot regime and decreases as more examples are added.

6.3.2 Vocabulary scaling. Personalization also depends on how many activities the system must distinguish at once. To evaluate this, we pool all participants' data into a single non-personalized model, without participant identity as input, and measure how recognition scales as the class count grows. This pools 70 trials for probing against 140 held-out trials for test, under the same cross-session round robin scheme. Figure 9 shows that balanced accuracy degrades gradually as the class count increases. The 6-axis model drops from 93.6% at 10 classes to 82.1% at 70 classes; the 3-axis model drops from 88.8% to 72.5%. Most of the decline occurs between 10 and 30 classes, with performance largely plateauing beyond that (the 6-axis model varies by less than 1 point from 30 to 60 classes). Users can therefore expand their label vocabulary over time without retraining the encoder.

6.3.3 Error structure and failure analysis. Although overall performance is high, errors are not uniformly distributed. Most participant-activity pairs are recognized near ceiling, while classes below 70% accuracy cluster within P7, with isolated cases in P1, P4, and P8; the per-class accuracy map and per-participant confusion matrices are given in Appendix B (Figures 13 and 15). Figure 10 ranks every user-defined activity by neighborhood purity in the embedding space. The most separable classes form compact, isolated clusters and are dominated by distinct burst or rotational motions such as stirring, snapping, and screwing a jar lid. The most entangled classes collapse into a shared central region and are predominantly sustained, low-amplitude motions such as writing, scrolling, and applying lotion, whose wrist dynamics differ far less than their semantic labels suggest.

We additionally examine P7 as a representative case (Figure 11). The confusion matrix (a) confirms that errors are concentrated among *Phone Scroll*, *Lotion*, and *Scratch Nails* rather than spread across the full 7-class vocabulary, with *Lotion* most often misclassified as *Scratch Nails* (25%). The same pattern appears for P4, whose *winding rope* is predicted as *tying a shoelace* 25% of the time. In PCA space (b), the confused classes occupy a shared local region relative to P7's other activities, indicating overlap within the participant-specific manifold rather than global collapse. Their normalized motion spectra (c) are similar, especially at low frequencies, suggesting that these activities generate closely related wrist dynamics. The test-to-centroid similarity matrix (d) shows weak diagonal margins between confused classes, confirming that errors arise from semantically different activities that produce similar wrist motion which remain difficult to separate from short windows alone.

Figure 12: Example on-demand activity scenarios captured with the watch-based system: espresso-making steps, skin care, surface cleaning, and cooking.

7 Discussion

7.1 Motion Priors from Broad Wrist Data

Our main finding is that broad wrist motion can act as an anchor for fine-grained recognition. Although Stage A sees only coarse L1 and L2 activities, the resulting encoder transfers to manipulative, gestural, and procedural classes that never appear during pretraining. This suggests that many downstream wrist labels do not require entirely new motion structure to be learned from scratch. Instead, they refine distinctions that are already latent in a representation trained on diverse global motion. This is especially important for wrist IMU, where useful structure must be recovered from temporal dynamics rather than rich spatial context.

7.2 Application Scenarios for On-Demand Activity Vocabularies

Our results suggest that users can introduce or revise activity classes at deployment time by providing only a few examples, rather than relying on a fixed, predefined vocabulary. This supports wearable applications where the activity set cannot be fixed in advance. For example, context-aware assistants can adapt to user-specific routines [4], factory systems can track evolving assembly workflows [44], and medical applications can monitor personalized or procedure-specific actions [7], including passive motion-based measurement of symptoms such as hyperactivity [1] and tremor [43], without requiring large-scale recollection and retraining. Figure 12 illustrates four such scenarios captured with our system, spanning espresso preparation, skin care, surface cleaning, and cooking.

At the same time, our results highlight a boundary condition for these scenarios. On-demand recognition is most effective for kinematically distinct L3 activities with different burst structure or rotational dynamics, so a barista's pour-over steps separate well because each involves a distinct tool motion, whereas a desk worker's writing and mouse use do not, since both are small sustained hand movements and fall among the most entangled classes in Figure 10. L4 procedural steps remain harder and only partially solved, since they are locally similar within a single window and separable mainly through temporal context (Appendix A). We found these failures to be modality-specific rather than demonstration-specific, with additional examples not helping when two activities produce genuinely overlapping wrist motion. Practical systems should therefore warn users at definition time when a new class falls too close to an existing one in embedding space and offer to merge them, as explored in prior work on representation-driven feedback [39, 51, 55].

7.3 Limitations and Future Work

TransfHAR has several limitations. First, the current formulation is window-based and therefore cannot model longer temporal structure, which limits performance on procedural L4 tasks whose steps are locally similar but sequentially distinct. A natural next step is to augment the frozen encoder with a lightweight temporal aggregation layer, such as a small recurrent module or state-based model over window embeddings, to capture richer sequential context. Our comparisons also isolate pretraining rather than architecture, so how the frozen representation compares to contemporary self-supervised backbones remains open to future work. Second, our personalization study is short-horizon and controlled. Although we include watch removal and re-wearing, we do not evaluate longer-term behavioral drift or repeated re-personalization over time. Future work should therefore study longitudinal use in more natural settings, including how users revise activity vocabularies and when re-adaptation is needed. Third, while the watch workflow is lightweight, broader adoption questions remain around hardware variation, battery and thermal cost, and interface support to collect better examples. The transformer encoder is also heavier than edge-optimized CNN or LSTM backbones, so model compression and edge-computing optimization remain open for future work. Fourth, our pretraining corpus pools public datasets whose demographic metadata are sparse and inconsistent, so we cannot characterize how well it represents the broader population, and whether the learned representation transfers equally across body types, ages, and motor abilities remains untested.

8 Conclusion

We presented TransfHAR, a framework for on-demand wrist activity recognition built on self-supervised pretraining. By learning broad motion priors from heterogeneous public wrist IMU data and adapting them with a lightweight linear probe, TransfHAR supports rapid personalization to new fine-grained activities without retraining a full model. Our offline experiments across three held-out benchmarks showed that the frozen representation transfers strongly to manipulative, gestural, and procedural tasks, often matching or exceeding supervised baselines. Our user study with a smartwatch further showed that users can define their own activity sets and obtain strong recognition from only a few demonstrations.

Acknowledgments

We thank VESSL AI for providing compute credits that supported this work.

References

[1] Riku Arakawa, Karan Ahuja, Kristie Mak, Gwendolyn Thompson, Sam Shaaban, Oliver Lindhiem, and Mayank Goel. 2023. LemurDx: Using Unconstrained Passive Sensing for an Objective Measurement of Hyperactivity in Children with no Parent Input. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2 (2023), 46:1–46:23. doi:10.1145/3596244

[2] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. 2024. PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch Enabled by Multimodal Procedure Tracking and Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (2024), 180:1–180:26. doi:10.1145/3699759

[3] Riku Arakawa, Manami Nakagawa, and Hiromu Yakura. 2025. UbiLearn: Supporting English-as-a-Foreign-Language Learners in Reflecting on Conversations Using a Smartwatch. *Proc. ACM Hum. Comput. Interact.* 9, 5 (2025), 1–29. doi:10.1145/3743735

[4] Riku Arakawa, Prasoon Patidar, Will Page, Jill Lehman, and Mayank Goel. 2025. Scaling Context-Aware Task Assistants that Learn from Demonstration and Adapt through Mixed-Initiative Dialogue. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, UIST 2025, Busan, Korea, 28 September 2025 - 1 October 2025*. ACM, New York, NY, USA, 145:1–145:19. doi:10.1145/3746059.3747700

[5] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024*. ACM, New York, NY, USA, 15:1–15:16. doi:10.1145/3654777.3676350

[6] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2022. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (2022), 156:1–156:27. doi:10.1145/3569504

[7] Edgar A. Bernal, Xitong Yang, Qun Li, Jayant Kumar, Sriganesh Madhvanath, Palghat Ramesh, and Raja Bala. 2018. Deep Temporal Multimodal Fusion for Medical Procedure Monitoring Using Wearable Sensors. *IEEE Trans. Multim.* 20, 1 (2018), 107–118. doi:10.1109/TMM.2017.2726187

[8] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 42:1–42:28. doi:10.1145/3534582

[9] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Möller. 2024. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (2024), 175:1–175:21. doi:10.1145/3699776

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, Piscataway, NJ, USA, 9630–9640. doi:10.1109/ICCV48922.2021.00951

[11] Shing Chan, Hang Yuan, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. 2024. CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data* 11, 1 (2024), 1135. doi:10.1038/s41597-024-03960-3

[12] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*. IEEE, Piscataway, NJ, USA, 168–172. doi:10.1109/ICIP.2015.7350781

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[14] Mathias C. Ciliberto, Vitor F. Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. 2021. Opportunity++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-Based Human Activity Recognition. *Frontiers in Computer Science* 3 (2021), 792065. doi:10.3389/fcomp.2021.792065

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4171–4186. doi:10.18653/V1/N19-1423

[16] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (2017), 11:1–11:28. doi:10.1145/3090076

[17] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. AAAI Press, 1533–1540. http://www.ijcai.org/Abstract/16/220

[18] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan A. Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked reconstruction based self-supervision for human activity recognition. In *ISWC '20: 2020 ACM International Symposium on Wearable Computers, Virtual Event, Mexico, September 12-17, 2020*. ACM, New York, NY, USA, 45–49. doi:10.1145/3410531.3414306

[19] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2022. Assessing the State of Self-Supervised Human Activity Recognition Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022), 116:1–116:47. doi:10.1145/3550299

[20] Harish Haresamudram, Irfan A. Essa, and Thomas Plötz. 2021. Contrastive Predictive Coding for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2 (2021), 65:1–65:26. doi:10.1145/3463506

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, Piscataway, NJ, USA, 15979–15988. doi:10.1109/CVPR52688.2022.01553

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, Piscataway, NJ, USA, 9726–9735. doi:10.1109/CVPR42600.2020.00975

[23] Mohammadreza Heydarian and Thomas E. Doyle. 2023. rWISDM: Repaired WISDM, a Public Dataset for Human Activity Recognition. *CoRR* abs/2305.10222 (2023), 13 pages. doi:10.48550/ARXIV.2305.10222

[24] Alexander Hoelzemann and Kristof Van Laerhoven. 2020. Digging deeper: towards a better understanding of transfer learning for human activity recognition. In *ISWC '20: 2020 ACM International Symposium on Wearable Computers, Virtual Event, Mexico, September 12-17, 2020*. ACM, New York, NY, USA, 50–54. doi:10.1145/3410531.3414311

[25] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. 2024. CrossHAR: Generalizing Cross-dataset Human Activity Recognition via Hierarchical Self-Supervised Pretraining. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2 (2024), 64:1–64:26. doi:10.1145/3659597

[26] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1 (2022), 17:1–17:28. doi:10.1145/3517246

[27] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300568

[28] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST 2016, Tokyo, Japan, October 16-19, 2016*. ACM, New York, NY, USA, 321–333. doi:10.1145/2984511.2984582

[29] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive Mob. Comput.* 5, 6 (2009), 657–675. doi:10.1016/J.PMCJ.2009.07.007

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019), 13 pages. arXiv:1907.11692

[31] Jeffrey W. Lockhart and Gary M. Weiss. 2014. Limitations with activity recognition methodology & data sets. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14 Adjunct Publication, Seattle, WA, USA - September 13 - 17, 2014*. ACM, New York, NY, USA, 747–756. doi:10.1145/2638728.2641306

[32] Aleksej Logacjov. 2024. Self-supervised Learning for Accelerometer-based Human Activity Recognition: A Survey. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (2024), 149:1–149:42. doi:10.1145/3699767

[33] Aleksej Logacjov and Kerstin Bach. 2024. Self-supervised learning with randomized cross-sensor masked reconstruction for human activity recognition. *Eng. Appl. Artif. Intell.* 128 (2024), 107478. doi:10.1016/J.ENGAPPAI.2023.107478

[34] Aleksej Logacjov, Sverre Herland, Astrid Ustad, and Kerstin Bach. 2024. SelfPAB: large-scale pre-training on accelerometer data for human activity recognition. *Appl. Intell.* 54, 6 (2024), 4545–4563. doi:10.1007/S10489-024-05322-3

[35] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 18 pages. https://openreview.net/forum?id=Bkg6RiCqY7

[36] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022), 132:1–132:19. doi:10.1145/3550284

[37] Francisco Javier Ordóñez Morales and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016), 115. doi:10.3390/S16010115

[38] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises. In *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI)*. ACM, New York, NY, USA, 3225–3234. doi:10.1145/2556288.2557116

[39] Prasoon Patidar, Riku Arakawa, Ricardo Graça, Ruben Moutinho, Adriano Soares, Ana Vasconcelos, Filippo Talami, Joana Couto da Silva, Inês Silva, Cristina Mendes-Santos, Mayank Goel, and Yuvraj Agarwal. 2025. OrganicHAR: Towards Activity Discovery in Organic Settings for Privacy Preserving Sensors Using Efficient Video Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4 (2025), 203:1–203:32. doi:10.1145/3770674

[40] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *16th International Symposium on Wearable Computers, ISWC 2012, Newcastle, United Kingdom, June 18-22, 2012*. IEEE Computer Society, Piscataway, NJ, USA, 108–109. doi:10.1109/ISWC.2012.13

[41] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task Self-Supervised Learning for Human Activity Detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2 (2019), 61:1–61:30. doi:10.1145/3328932

[42] Muhammad Shoaib, Stephan Bosch, Özlem Durmaz Incel, Hans Scholten, and Paul J. M. Havinga. 2016. Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors. *Sensors* 16, 4 (2016), 426. doi:10.3390/S16040426

[43] Luis Sigcha, Ignacio Pavón, Nélson Costa, Susana Costa, Miguel F. Gago, Pedro M. Arezes, Juan-Manuel López-Navarro, and Guillermo de Arcas. 2021. Automatic Resting Tremor Assessment in Parkinson's Disease Using Smartwatches and Multitask Convolutional Neural Networks. *Sensors* 21, 1 (2021), 291. doi:10.3390/S21010291

[44] Georgios Sopidis, Michael Haslgrübler, Behrooz Azadi, Bernhard Anzengruber-Tanase, Abdelrahman Ahmad, Alois Ferscha, and Martin Baresch. 2022. Micro-activity recognition in industrial assembly process with IMU data and deep learning. In *PETRA '22: The 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 29 June 2022 - 1 July 2022*. ACM, New York, NY, USA, 103–112. doi:10.1145/3529190.3529204

[45] Setareh Rahimi Taghanaki, Michael J. Rainbow, and Ali Etemad. 2021. Self-supervised Human Activity Recognition by Learning to Predict Cross-Dimensional Motion. In *ISWC 2021: Proceedings of the 2021 ACM International Symposium on Wearable Computers, Virtual Event, September 21-26, 2021*. ACM, New York, NY, USA, 23–27. doi:10.1145/3460421.3480417

[46] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Søren Brage, Nicholas J. Wareham, and Cecilia Mascolo. 2021. SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 36:1–36:30. doi:10.1145/3448112

[47] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. 2020. Exploring Contrastive Learning in Human Activity Recognition for Healthcare. *CoRR* abs/2011.11542 (2020), 6 pages. arXiv:2011.11542 https://arxiv.org/abs/2011.11542

[48] Daoyu Wang, Mingyue Cheng, Zhiding Liu, and Qi Liu. 2025. TimeDART: a diffusion autoregressive transformer for self-supervised time series representation. In *Proceedings of the 42nd International Conference on Machine Learning*, Vol. 267. JMLR.org, 25 pages. https://proceedings.mlr.press/v267/wang25r.html

[49] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* 119 (2019), 3–11. doi:10.1016/J.PATREC.2018.02.010

[50] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. Serendipity: Finger Gesture Recognition using an Off-the-Shelf Smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. ACM, New York, NY, USA, 3847–3851. doi:10.1145/2858036.2858466

[51] Jason Wu, Chris Harrison, Jeffrey P. Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, New York, NY, USA, 1–14. doi:10.1145/3313831.3376875

[52] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2023, Madrid, Spain, October 2-6, 2023*. ACM, New York, NY, USA, 85:1–85:15. doi:10.1145/3570361.3613299

[53] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *SenSys '21: The 19th ACM Conference on Embedded Networked Sensor Systems, Coimbra, Portugal, November 15 - 17, 2021*. ACM, New York, NY, USA, 220–233. doi:10.1145/3485730.3485937

[54] Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob E. Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak N. Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. 2025. LSM-2: Learning from Incomplete Wearable Sensor Data. *CoRR* abs/2506.05321 (2025), 32 pages. doi:10.48550/ARXIV.2506.05321

[55] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E. Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 496, 19 pages. doi:10.1145/3491102.3501904

[56] Po Yang, Congmin Yang, Vitaveska Lanfranchi, and Fabio Ciravegna. 2022. Activity Graph Based Convolutional Neural Network for Human Activity Recognition Using Acceleration and Gyroscope Data. *IEEE Trans. Ind. Informatics* 18, 10 (2022), 6619–6630. doi:10.1109/TII.2022.3142315

[57] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek F. Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, New York, NY, USA, 351–360. doi:10.1145/3038912.3052577

[58] Taeyoung Yeon, Vasco Xu, Henry Hoffmann, and Karan Ahuja. 2025. WatchHAR: Real-time On-device Human Activity Recognition System for Smartwatches. In *Proceedings of the 27th International Conference on Multimodal Interaction, ICMI 2025, Canberra, Australia, October 13-17, 2025*. ACM, New York, NY, USA, 387–394. doi:10.1145/3716553.3750775

[59] Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden R. Doherty. 2024. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digit. Medicine* 7, 1 (2024), 91. doi:10.1038/S41746-024-01062-3

[60] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. TS2Vec: Towards Universal Representation of Time Series. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 8980–8987. doi:10.1609/AAAI.V36I8.20881

[61] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y. Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, and Shirui Pan. 2024. Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 10 (2024), 6775–6794. doi:10.1109/TPAMI.2024.3387317

[62] Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. 2024. Masked Video and Body-Worn IMU Autoencoder for Egocentric Action Recognition. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVIII*, Vol. 15076. Springer-Verlag, Berlin, Heidelberg, 312–330. doi:10.1007/978-3-031-72649-1_18

[63] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Curran Associates, Inc., Red Hook, NY, USA, 16 pages. http://papers.nips.cc/paper_files/paper/2022/hash/194b8dac525581c346e30a2cebe9a369-Abstract-Conference.html

[64] Haozhe Zhou, Riku Arakawa, Yuvraj Agarwal, and Mayank Goel. 2025. IMUCoCo: Enabling Flexible On-Body IMU Placement for Human Pose Estimation and Activity Recognition. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, UIST 2025, Busan, Korea, 28 September 2025 - 1 October 2025*. ACM, New York, NY, USA, 91:1–91:16. doi:10.1145/3746059.3747695

A Window Length Ablation

Table 4 compares three window lengths under the full-split protocol. The 2.56 s window used throughout the paper is best on both L3 benchmarks, while the L4 procedural task benefits from the longer 5.12 s window.

B Confusion Matrices and Error Maps

Figure 13 maps per-class accuracy for every participant-activity pair in the personalization study, Figure 14 gives per-class confusion structure on the three held-out benchmarks, and Figure 15 gives the per-participant matrices.

Table 4: Window length ablation, frozen 6-axis encoder with a linear probe, full split. Balanced accuracy, mean (SD) over five seeds.

Dataset	1.28 s	2.56 s	5.12 s
SAMoSA (L3)	50.8 (0.6)	**64.9 (0.7)**	61.4 (1.4)
UTD-MHAD (L3)	46.7 (0.3)	**69.1 (2.3)**	57.1 (3.5)
PrISM latte (L4)	29.2 (0.3)	36.5 (1.1)	**38.1 (2.9)**

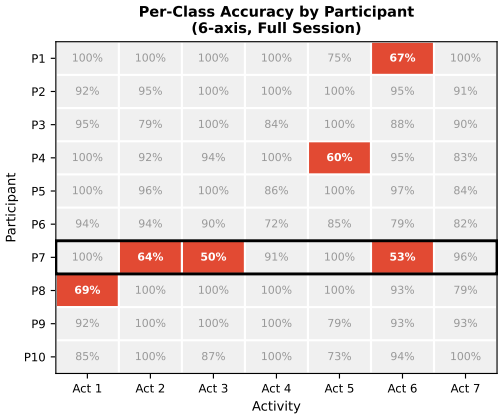

Figure 13: Per-class accuracy for each participant-activity pair, 6-axis full-session setting, with classes below 70% highlighted.

C Activity Bank and Participant Details

Table 5 lists the 55 candidate activities shown to participants, and Table 6 gives each participant's chosen set. Of the 70 activities selected, 43 were taken directly from the bank, 12 were participant-specific variants of bank items, and 15 were new activities with no bank equivalent.

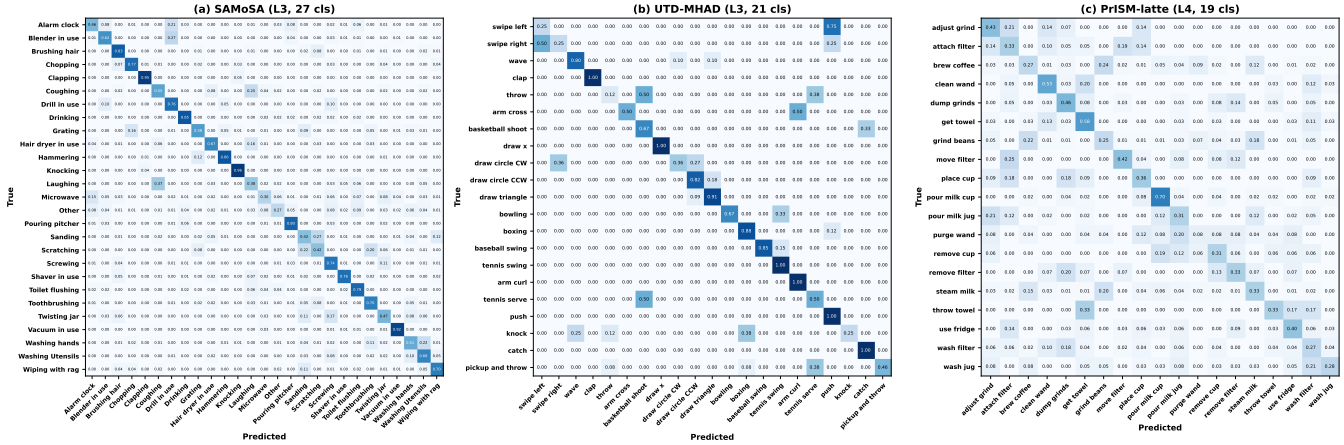

Figure 14: Row-normalized confusion matrices for the frozen 6-axis encoder with a linear probe, full split, on the three held-out benchmarks. Errors concentrate among kinematically similar classes, with PrISM latte showing the broadest off-diagonal mass.

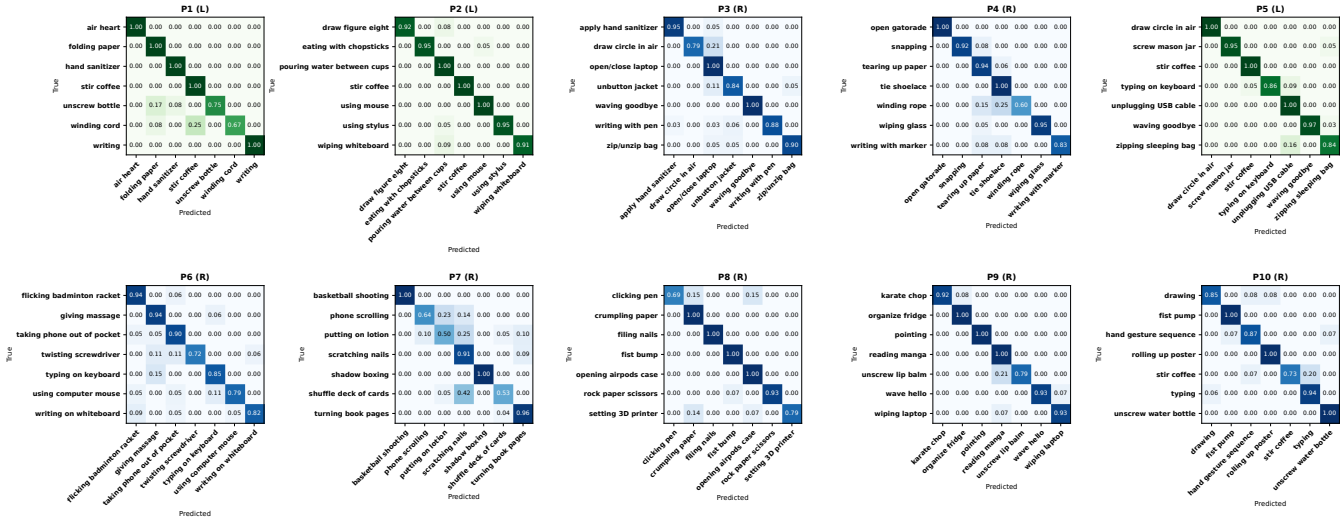

Figure 15: Per-participant confusion matrices for the 6-axis encoder in the full-session setting, colored by which wrist the watch was worn on.

Table 5: Activity bank shown to participants to illustrate the target granularity of on-demand wrist activity recognition. Participants could select from this list or define their own activities.

Category	Candidate activities shown to participants
Hand / air gestures	drawing a heart in the air; drawing a circle in the air; drawing a figure-eight in the air; waving goodbye; conducting an orchestra; air drumming; shadow boxing; finger snapping sequence; thumbs up/down toggle; pointing at corners of the room; tracing the alphabet in the air; karate chop motion; fist pump; jazz hands; air guitar strumming
Object manipulation (common lab items)	stirring coffee with a spoon; opening/closing a laptop lid; turning a doorknob; flipping pages of a book; stacking/unstacking cups; shuffling a deck of cards; peeling tape off a roll; crumpling paper into a ball; folding a sheet of paper in half; uncapping and recapping a marker; rolling a ball of clay; winding a cord around your hand; zipping/unzipping a bag; tearing paper into strips; clicking a retractable pen repeatedly
Tool-like / workspace	typing on a keyboard; using a computer mouse (clicking and dragging); writing with a pen on paper; erasing a whiteboard; cutting paper with scissors; stapling papers; using a screwdriver; tightening a jar lid; pouring water between cups; wiping a table with a cloth; peeling a sticky note off a pad; pressing buttons on a calculator; turning a dial/knob; plugging/unplugging a USB cable; sharpening a pencil
Personal / body	brushing hair; applying hand sanitizer (rubbing hands); putting on / taking off a wristwatch; buttoning / unbuttoning a shirt cuff; tying a shoelace; rolling up a sleeve; stretching a rubber band between fingers; cracking knuckles (hand flex); filing nails with an emery board; putting on and removing a glove

Table 6: Participant-specific activity sets used in the on-demand personalization study. To preserve readability in the main figures, activities are shown as Act 1–Act 7 in participant order; this table provides the corresponding semantic labels.

Participant	Chosen activities
P1	air heart; folding paper; hand sanitizer; stirring coffee; unscrewing bottle; winding cord; writing
P2	drawing figure eight in the air; eating with chopsticks; pouring water between cups; stirring coffee; using mouse; using stylus; wiping whiteboard
P3	applying hand sanitizer; drawing circle in the air; opening and closing laptop; unbuttoning jacket; waving goodbye; writing with pen; zipping and unzipping bag
P4	opening Gatorade bottle; snapping; tearing paper; tying shoelace; winding rope; wiping glass; writing with marker
P5	drawing circle in the air; screwing mason jar lid; stirring coffee; typing on keyboard; unplugging USB cable; waving goodbye; zipping sleeping bag
P6	flicking badminton racket; massage giving; taking phone out of pocket; twisting screwdriver; typing on keyboard; using computer mouse; writing on whiteboard
P7	basketball shooting; phone scrolling; putting on lotion; scratching nails; shadow boxing; shuffling deck of cards; turning book pages
P8	clicking pen; crumpling paper; filing nails; fist bump; opening AirPods case; rock paper scissors; setting 3D printer
P9	karate chop; organizing fridge; pointing; reading Japanese manga; unscrewing lip balm; waving hello; wiping laptop
P10	drawing; fist pump; hand gesture sequence; rolling up poster; stirring coffee; typing; unscrewing water bottle