HARBench: A Comprehensive Benchmark for Evaluating Foundation Models in Sensor-based Human Activity Recognition

Kei Tanigaki
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
tanigaki@ist.osaka-u.ac.jp

Takuya Maekawa
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
maekawa.takuya.ist@osaka-u.ac.jp

Takahiro Hara
Graduate School of Information Science and Technology
The University of Osaka
Osaka, Japan
hara@ist.osaka-u.ac.jp

***Abstract*—Foundation models pretrained on large-scale unlabeled data using self-supervised learning (SSL) have demonstrated remarkable capabilities in natural language processing (NLP) and computer vision (CV). Applying this paradigm to sensor-based Human Activity Recognition (HAR) holds promise for robust and generalizable models, but also poses unique challenges due to diverse sensor mounting positions and application domains. Despite this potential, unlike in NLP and CV, there has been a lack of systematic evaluation frameworks in HAR to rigorously assess the generalization and adaptation capabilities of foundation models across these critical dimensions. To address this gap, we introduce HARBench, a comprehensive benchmark designed to evaluate HAR foundation models along key axes, including Domain Robustness, Position Robustness, Few-shot Performance, and Zero-shot Performance. By leveraging a large collection of datasets from multiple domains, totaling over 2.4 million hours of sensor data, HARBench enables evaluation of HAR foundation models from multiple perspectives, providing a thorough and standardized assessment of their capabilities. We conduct large-scale evaluation and analysis of existing HAR approaches, ranging from conventional supervised methods to state-of-the-art SSL models. Our results quantitatively map the capabilities and limitations of current models, thereby providing guidance for future research and the development of next-generation foundation models for HAR. HARBench is publicly released to support the HAR research community, promoting reproducibility and accelerating the development of robust foundation models.**

***Index Terms*—Sensor-based activity recognition, foundation models, benchmark, self-supervised learning**

I. Introduction

A. Background

Sensor-based Human Activity Recognition (HAR) has been actively studied in the PerCom community across diverse domains, such as daily living and industrial activities [1]–[6]. A central challenge in sensor-based HAR is the high cost of collecting large-scale labeled datasets, which are crucial for training high-accuracy models. Self-supervised learning (SSL) has emerged as a promising strategy to address this challenge by enabling the pretraining of versatile feature representations from vast amounts of unlabeled data [7]–[10]. Specifically, the paradigm of building a foundation model on large-scale data and adapting it to diverse downstream tasks with minimal labels offers the potential to significantly improve label efficiency in HAR research.

This approach has already seen remarkable success in the natural language processing (NLP) and computer vision (CV) research communities. Models such as RoBERTa, GPT, and CLIP have demonstrated exceptional performance after fine-tuning with very few labeled data [11]–[13]. A key driver for this progress has been the establishment of standardized benchmarks, such as GLUE and SuperGLUE [14, 15], which have facilitated fair evaluation and fostered healthy competition, accelerating advancements across these fields.

B. Problems

Inspired by the successes in other fields, the HAR community has begun exploring the development of foundation models. However, unlike NLP and CV, the HAR field critically lacks standardized benchmarks to evaluate and compare the performance of these models. This absence of common evaluation frameworks hinders the fair assessment of the generalizability and robustness of new models, creating a significant bottleneck that slows progress across the entire research community. This lack of common benchmarks presents three fundamental barriers to the development of effective foundation models in HAR:

• **Challenges in generalization across various settings:** In contrast to NLP and CV tasks, where the meaning of words or pixels is largely consistent across datasets, sensor-based HAR inputs can vary drastically depending on the sensor's placement and the target domain. For example, accelerometer data collected from a wrist differs significantly from data collected from a foot, even for the same activity. Similarly, sensor data characteristics differ between domains such as daily living and industrial environments. Consequently, a foundation model for HAR must be robust against various sensor positions and domains. However, current evaluation is fragmented, with studies using few different datasets, sensors, and placements,

and no common evaluation framework exists to fairly assess a model's generalization capability across these variations.

• **Lack of standardized evaluation for unseen tasks:** A key advantage of foundation models is their ability to adapt to new, unseen tasks. This capability is especially critical for few-shot and zero-shot learning scenarios, where minimal or no labeled data exist for the target task. However, existing benchmarks are not designed to systematically evaluate models across diverse downstream tasks, leaving a gap in assessing the practical utility of foundation models under realistic, unseen conditions.

• **A dysfunctional development ecosystem:** The lack of standardized benchmarks forces researchers to repeatedly build ad-hoc evaluation setups, impeding fair comparisons and reproducibility. Another concern is that existing foundation model research (e.g., HARNet, SelfPAB) relies on access-restricted datasets (UK Biobank, HUNT4), limiting participation to a few well-funded institutions. This "data oligopoly" discourages diverse perspectives, slows innovation, and hampers the collaborative progress needed for field maturation.

C. Approach and Contributions

To address these fundamental challenges, we introduce HARBench, the first comprehensive benchmark for evaluating the generalization capabilities of HAR foundation models. Analogous to the role that GLUE and SuperGLUE [14, 15] played in NLP, HARBench aims to provide a unified evaluation infrastructure that fosters fair competition and collaboration, thereby accelerating the development of a healthy research ecosystem for HAR. Our contributions are threefold:

• **A comprehensive evaluation framework:** Few studies have systematically evaluated SSL methods across diverse HAR datasets, and most evaluate on fewer than five datasets. By curating and integrating 32 public HAR datasets, we establish a systematic framework that measures generalization across essential axes for foundation models including: (1) Domain Robustness, (2) Position Robustness, (3) Few-shot Performance, and (4) Zero-shot Performance. Importantly, all datasets are publicly available and accessible, enabling researchers and educational institutions with limited resources to participate in cutting-edge foundation model research.

• **An open leaderboard system:** We develop and release a public leaderboard system where researchers can submit their models for automated evaluation against a unified standard. This will promote transparency, reproducibility, and fair comparison within the community.

• **Large-scale evaluation and analysis:** We conduct an extensive evaluation of existing models on HARBench, ranging from traditional supervised methods to the latest SSL models. Through this analysis, we quantitatively map the capabilities and limitations of current approaches and offer new insights and concrete guidelines for the future development of next-generation HAR foundation models.

By standardizing evaluation, HARBench seeks to unify the currently fragmented research landscape and lay the groundwork for collaborative progress. We expect this benchmark to accelerate the development of versatile foundation models and contribute to the sustained growth of the HAR community.

II. Related Work

A. Benchmark Design and Evolution

Evaluation benchmarks play a central role in shaping research directions and advancing machine learning. In NLP, the GLUE benchmark [14] integrated nine language understanding tasks and provided standardized protocols and open leaderboard. This framework paved the way for breakthroughs such as BERT [16]. However, within a year, model performance saturated, necessitating the introduction of the more challenging SuperGLUE benchmark [15]. Building on this trajectory, HELM [17] expanded evaluation dimensions beyond accuracy to include robustness, fairness, efficiency, and other criteria, underscoring the importance of multi-faceted evaluation.

In CV, ImageNet [18] with its unprecedented scale of 1,000 classes and 1.4M images—together with its annual competition, was crucial in driving the deep learning revolution. More recently, CLIP [13] demonstrated zero-shot transfer learning by aligning images with text, enabling models to adapt to novel tasks without explicit training. The CLIP benchmark, spanning 43 datasets, further illustrated that evaluating foundation models requires assessing cross-domain transferability rather than performance on a single dataset.

By contrast, the human activity recognition (HAR) community has historically lacked comprehensive benchmarks. To address this gap, DAGHAR [19] standardized six smartphone-based dataset to enable domain adaptation and generalization studies. DAGHAR was the first to provide a unified benchmark for cross-dataset generalization, harmonizing data formats (units, sampling rates, inclusion of gravity components), activity labels, user splits, and windowing strategies while preserving intrinsic domain differences. Similarly, B-HAR [20] provides an open-source framework for evaluating HAR methodologies across 9 datasets encompassing wearable and ambient sensors. B-HAR standardizes preprocessing pipelines and evaluates traditional ML classifiers (K-NN, Decision Trees, LDA) alongside deep learning models. However, it evaluates each dataset independently without cross-dataset generalization, and does not support foundation model evaluation.

Table I summarizes the key differences between existing HAR benchmarks and HARBench. These benchmarks leave critical gaps for foundation model evaluation: (1) DAGHAR is limited to smartphone sensors; (2) B-HAR evaluates datasets independently as noted above; (3) both focus on daily-living activities with limited industrial coverage; and (4) neither supports few-shot or zero-shot evaluation. These limitations motivate the development of broader and more comprehensive benchmarks.

B. HAR Foundation Models

Recent advances in HAR foundation models are driven by large-scale self-supervised pretraining. HARNet [7] leveraged an unprecedented 700,000 person-days of wearable

TABLE I
COMPARISON OF HAR BENCHMARKS. EVAL AXES REFERS TO THE NUMBER OF EVALUATION DIMENSIONS (OVERALL, DOMAIN, POSITION, FEW-SHOT, ZERO-SHOT). FOUNDATION MODELS INDICATES THE NUMBER OF PRETRAINED HAR MODELS EVALUATED.

	DAGHAR	B-HAR	HARBench
# Datasets	5	9	**32**
Sensors	Phone	Wearable+Ambient	Wearable+Phone
Domains	Daily	Daily, Med.	**Daily, Exer., Ind.**
# Eval Axes	1	1	**5**
# Foundation Models	0	0	**6**

TABLE II
DATASETS USED FOR SELF-SUPERVISED PRETRAINING

Dataset	#Subs	Time[1]	Device[2]	Domain
NHANES [24, 25]	14693	2468k	non dominant W	Daily
ADLRD [26]	16	3	RW	Daily
CHAD [27, 28]	1	13	W, Pocket	Daily
Capture24 [29]	151	3883	W	Daily
DOG [30]	10	12	Trunk, Ankle(3)	Daily
HAR70+ [31]	18	21	Back, T	Daily
HHAR [32]	9	32	Phone, Watch	Daily
IMSB [33, 34]	20	2	W, Knee(3)	Exercise
KDDIKitchen [35]	10	16	Watch	Daily
MotionSense [36, 37]	24	8	Phone	Daily
Opportunity [38]	4	58	RW, RKnee	Daily
SBRHAPT [39]	30	3	Phone	Daily
TMD [40]	13	19	Phone	Daily
WISDM [41]	51	155	RW	Daily

data from the UK Biobank and employed multi-task self-supervised objectives (based on data augmation techniques such as permutation and time warping) on deep CNNs. Evaluations across eight benchmark datasets demonstrated substantial gains, with F1 improvements ranging from 2.5% to 130.9% (median 24.4%), highlighting the generalization capacity across datasets, subjects, and devices. Similarly, SelfPAB [21] pretrained a Transformer encoder on up to 100,000 hours of two-axis accelerometer data from the HUNT4 dataset (35,000 participants, up to seven days each). Using a masked spectrogram reconstruction task, SelfPAB achieved 7.1–14% F1 improvements over purely supervised baselines on downstream datasets including HARTH, HAR70+, and PAMAP2.

While these advances underscore the promise of large-scale pretraining, they also reveal structural challenges: both HARNet and SelfPAB rely on restricted-access datasets (UK Biobank, HUNT4), complicating reproducibility and fair comparison. The lack of open and standardized large-scale pretraining corpora thus represents a fundamental barrier to progress. In addition, the performance of these foundation models has been evaluated on limited labeled datasets for HAR. Recent comprehensive studies [22, 23] have systematically evaluated self-supervised learning methods for HAR, demonstrating the critical need for unified evaluation frameworks that assess models across diverse scenarios including robustness to dataset variations, sensor configurations, and data scarcity conditions. Addressing these problems, our work aggregates 32 public datasets, providing a shared and accessible foundation for pretraining and evaluation.

[1]All time values are in hours.

[2]Abbreviations: R=Right, L=Left, W=Wrist, T=Thigh, FA=Forearm. Parentheses indicate device count. Bold text is explained below.

[3]Numbers in parentheses indicate the actual number of classes used due to factors such as insufficient activity duration.

TABLE III
DATASETS USED AS LABELED DATA

Dataset	#Subs	Time[1]	#Act[3]	Device[2]	Domain
DSADS [42]	8	37	19(19)	**LArm**, **LLeg**(5)	Exercise
ForthTrace [43]	15	19	16(11)	**LWrist**, **LT**(5)	Daily
HARTH [44, 45]	22	62	12(10)	Back, T	Daily
IMWSHA [46, 47]	10	2	11(11)	**W**, **T**(3)	Daily
LARa [48]	8	14	8(6)	**LArm**, **LLeg**(5)	Industry
MEx [49]	30	7	7(7)	**W**, **T**	Exercise
MHEALTH [50]	10	19	12(12)	**RFA**, **LAnkle**(3)	Exercise
OPENPACK [51]	16	141	10(9)	RW, LW(4)	Industry
VTT-ConIoT [52]	13	13	16(16)	Hip, Back(3)	Industry
Exoskeletons [53]	12	8	4(4)	RLeg, LLeg(5)	Industry
PAAL [54]	15	8	24(10)	Phone	Daily
PAMAP2 [55]	9	23	18(12)	**Hand**, **Ankle**	Daily
REALDISP [56]	17	119	33(33)	**LFA**, **LT**	Exercise
RealWorld [57]	15	129	8(8)	**FA**, **T**	Daily
SELFBACK [58]	33	25	9(9)	**Watch**, **T**	Daily
UCA-EHAR [59]	20	6	8(6)	SmartGlasses	Daily
USC-HAD [60]	14	8	12(12)	Phone	Daily
WARD [61]	20	33	13(13)	**LFA**, **LAnkle**	Daily

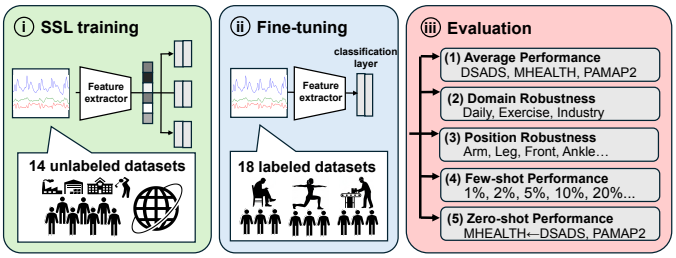

Fig. 1. Overview of HARBench

C. Public HAR Datasets

The HAR domain encompasses a wide range of publicly available datasets; however, their use in standardized evaluation has been limited. To ensure reproducibility and accessibility, we established systematic inclusion criteria for dataset selection: (1) *Public Availability*—freely downloadable without institutional restrictions; (2) *Raw Sensor Data*—raw accelerometer data available rather than preprocessed features; (3) *Documentation*—clear metadata including sampling rate, sensor placement, and activity labels. All 32 selected datasets meet these criteria, enabling broad community participation in foundation model research.

As summarized in Tables II and III, these datasets span diverse domains including daily living activities, sports, and industrial tasks. Among them, large-scale real-world datasets such as NHANES (14,693 participants, 2.5 million hours) are particularly noteworthy, as they are well-suited for self-supervised pretraining of foundation models. The 14 datasets listed in Table II were employed in this study as unlabeled data for pretraining. These datasets are characterized by the absence of activity labels, missing annotations for certain users, or sparsely available labels, making them unsuitable for conventional supervised learning. Leveraging such large-scale unlabeled data allows foundation models to capture diverse motion patterns, thereby enhancing their transferability to downstream tasks.

Nonetheless, these datasets have typically been used in isolation, and no comprehensive evaluation framework has been established. To address this gap, we introduce HARBench, the first benchmark that integrates 14 large-scale unlabeled

datasets (Table II) with 18 diverse labeled datasets (Table III), enabling multi-dimensional evaluation of the generalization ability of HAR foundation models.

III. HARBench

HARBench employs a three-stage evaluation protocol, as illustrated in Figure 1: (i) self-supervised pretraining on unlabeled sensor data, (ii) supervised fine-tuning on labeled datasets, and (iii) comprehensive evaluation across five dimensions of model generalizability. This methodology enables systematic assessment of foundation model performance under diverse deployment conditions.

The evaluation framework comprises five complementary metrics designed to capture distinct aspects of model generalizability: (1) Average Performance quantifies overall recognition performance across all evaluation datasets; (2) Domain Robustness measures performance consistency across different application contexts (daily, exercise, industrial domains); (3) Position Robustness evaluates performance consistency across varying sensor placement configurations; (4) Few-shot Performance assesses transfer learning efficiency with limited labeled data; and (5) Zero-shot Performance measures transfer capability to completely unseen data distributions without any prior exposure to target domains.

A. Dataset Integration and Preprocessing

HARBench integrates 14 large-scale unlabeled datasets (Table II) and 18 labeled datasets (Table III), applying unified preprocessing for both pretraining and fine-tuning.

For the pretraining datasets in Table II, we perform data preprocessing tailored to self-supervised learning, which relies on large-scale data. Sensor data are converted to G units, and the sampling frequency is standardized to 30Hz following [7] to ensure compatibility with existing foundation model architectures. For datasets with multiple sensors, data from each sensor are treated independently, and time series are segmented into 5-second windows with 3-second stride. The NHANES dataset, collected using wrist-worn ActiGraph GT3X+ devices with 80Hz triaxial accelerometers, uses non-overlapping windows due to its large volume. Random 3D rotations are applied to promote pose invariance and robustness to diverse sensor mounting angles.

For Table III datasets, we apply the same standardization and windowing as pretraining. However, random 3D rotations are applied only during pretraining for data augmentation; fine-tuning and evaluation use the original sensor orientations without rotation augmentation.

For fine-tuning/evaluation, an activity class label is associated with the window of data as a ground truth label. When multiple activities occur within a single window, the majority label (i.e., the activity with the longest duration) is assigned as the ground truth label for that window. Note that, for Zero-shot Performance evaluation, we do not use original labels of the datasets, and only six common activity classes among the three evaluation datasets are considered.

B. Model Architecture

Following the design principles of existing HAR foundation models such as HARNet [7] and SelfPAB [21], our benchmark adopts a single-sensor input architecture. This design choice is motivated by the practical reality that the number and types of sensors vary significantly across different datasets and deployment scenarios. By processing each sensor independently, the model can handle arbitrary sensor configurations without requiring retraining for each specific setup. Note that while some multimodal foundation models have been evaluated by inputting multiple accelerometer sensors simultaneously [62, 63], the single-sensor approach remains standard for ensuring consistent evaluation across diverse sensor configurations.

Each of HAR foundation models used in this study consists of two main components: a feature extractor and an activity classifier (output classification layer). The feature extractor takes input from a single sensor and is pretrained through SSL on the unlabeled datasets. When multiple sensors are available, a separate extractor processes each sensor's data; however, all extractors share the same parameters. Extracted features from all sensors are then concatenated. The activity classifier receives the concatenated features and is fine-tuned using each of the labeled target datasets (fine-tuning and evaluation datasets). This design allows the model to learn robust, sensor-independent representations during pretraining while adapting effectively to downstream activity recognition tasks through fine-tuning.

Note that, for Zero-shot Performance evaluation, no fine-tuning is performed on the target dataset. Details of the zero-shot evaluation protocol are described in Section III-D5.

C. Basic Evaluation Framework

The evaluation process follows a unified workflow consisting of three stages: pretraining, fine-tuning, and evaluation.

In the pretraining stage, a foundation model is trained on the preprocessed datasets listed in Table II, using data extracted separately from each individual sensor.

During the fine-tuning stage, we perform fine-tuning separately for each labeled dataset used for fine-tuning and evaluation. We adopt a user ID-based 4-fold cross-validation with a standardized user selection strategy. Only datasets with eight or more users are included to ensure sufficient data for robust cross-validation and to prevent evaluation pattern divergence across datasets. The 4-fold validation proceeds as follows: fold 1 uses users 1-2 as the test set, fold 2 uses users 3-4, fold 3 uses users 5-6, and fold 4 uses users 7-8. For each fold, the remaining six users are split into two for validation and four for fine-tuning, with any additional users beyond eight included in the fine-tuning set. This person-independent splitting strategy ensures unbiased evaluation while maintaining reproducibility across all benchmark datasets. Note that this fine-tuning stage is not performed for Zero-shot Performance evaluation, since no labeled data from the target dataset is used in that case.

In the evaluation stage, we use macro F1 score as the primary evaluation metric to mitigate the effects of class

imbalance. We calculate individual quantitative scores for each evaluation axis to comprehensively evaluate the multi-faceted generalizability of foundation models.

D. Multi-dimensional Evaluation Metrics

HARBench defines five core evaluation axes to assess the generalizability of HAR foundation models from multiple perspectives. Each axis targets specific challenges that foundation models encounter in real-world scenarios, quantifying practical performance aspects that cannot be captured through traditional single-dataset evaluations. Unless otherwise noted, evaluations use all available sensors within each dataset (multi-sensor HAR).

1) Average Performance: Average Performance quantifies the fundamental generalization capability of foundation models, equally weighting each dataset.

Let $N_{labeled}$ denote the total number of labeled datasets (18 in our benchmark). The F1 score is computed for each dataset and averaged:

$$\text{Average Performance} = \frac{1}{N_{labeled}} \sum_{i=1}^{N_{labeled}} \text{F1}_{\text{dataset}_i} \quad (1)$$

2) Domain Robustness: Domain Robustness evaluates how well a foundation model maintains balanced performance across different application domains: Daily, Exercise, and Industry.

Let $N_{domains}$ denote the number of evaluation domains, and let $D_{domains} = \{\text{Daily}, \text{Exercise}, \text{Industry}\}$ represent the set of domains (Table III).

Domain Robustness first calculates the average performance within each domain and then computes the arithmetic mean across domains. This approach provides a domain-fair evaluation that mitigates biases due to differing numbers of datasets per domain:

$$\text{Domain Robustness} = \frac{1}{N_{domains}} \sum_{d \in D_{domains}} \text{F1}_d \quad (2)$$

Note that our leaderboard also reports HAR performance of foundation models for each application domain, revealing which models perform best in which domains. This allows practitioners to identify the most suitable foundation model for developing domain-specific HAR applications.

3) Position Robustness: Position Robustness quantifies robustness to sensor mounting position variations using single-sensor HAR. Since identical activities produce different sensor patterns at different body locations (wrist, waist, ankle, etc.), position robustness is essential for practical deployment. Unlike leave-one-position-out evaluation that tests on unseen positions, we measure consistency across 8 placement categories present in training—critical for deployment scenarios where sensor migration occurs within known position types.

Let $N_{positions}$ denote the number of sensor position categories (8 in our framework) and $N_{configs}$ represent the total number of sensor configurations (66 in our benchmark). Here, a sensor configuration corresponds to a single sensor mounted at a specific position within a dataset. For example, if a dataset contains sensors at three different body parts, this would result in three separate configurations, each evaluated independently.

In the evaluation protocol, the $N_{configs}$ sensor settings are grouped into $N_{positions}$ main mounting position categories. Table IV provides a detailed breakdown of datasets and sensor positions included in each category. Multiple sensor settings within a category are evaluated individually, and the average F1 score for each category is computed.

This allows for analyses indicating, for example, that a foundation model performs well on wrist-mounted sensors or that phone-based recognition remains challenging. Because averaging all $N_{configs}$ settings without regard to category would bias the evaluation toward categories with more configurations, such as Arm and Leg, the Position Robustness metric computes the arithmetic mean across the $N_{positions}$ categories, giving equal weight to each. This approach ensures fair evaluation for categories with fewer settings, such as Phone, Back, and Head, thereby providing a true measure of position robustness:

$$\text{Position Robustness} = \frac{1}{N_{positions}} \sum_{i=1}^{N_{positions}} \text{F1}_{\text{position}_i} \quad (3)$$

4) Few-shot Performance: Few-shot Performance evaluates adaptation efficiency when only limited labeled data is available.

Let N_{shots} denote the number of few-shot data percentage settings (6 in our evaluation), and let $P_{shots} = \{1\%, 2\%, 5\%, 10\%, 20\%, 50\%\}$ represent the set of fine-tuning data percentages prepared for each dataset. In the evaluation protocol, fine-tuning data are progressively sampled from each dataset. Specifically, we perform fine-tuning under the N_{shots} settings using the percentages defined in P_{shots}. To provide concrete context, the median number of samples per class across all datasets at each percentage level is: 1%≈4 samples, 2%≈7 samples, 5%≈18 samples, 10%≈37 samples, 20%≈74 samples, and 50%≈185 samples (see Table X for detailed per-dataset breakdown). The 1% setting simulates extreme few-shot learning, while higher percentages assess scalability with increasing data availability.

In addition to computing F1 scores at each setting, we examine the performance growth curve with respect to the amount of fine-tuning data. The Few-shot Performance score is computed as the arithmetic mean of the F1 scores across all data-percentage settings, equally evaluating generalization capability at each setting:

$$\text{Few-shot Performance} = \frac{1}{N_{shots}} \sum_{p \in P_{shots}} \text{F1}_p \quad (4)$$

5) Zero-shot Performance: Zero-shot Performance measures transfer capability to unseen datasets. While a strict definition of zero-shot learning (ZSL) refers to recognition

TABLE IV
POSITION CATEGORIES AND SENSOR CONFIGURATIONS FOR POSITION ROBUSTNESS EVALUATION

Category	Settings	Dataset-Position Configurations
Arm	15	DSADS (LeftArm, RightArm), LARa (LeftArm, RightArm), OPENPACK (LeftUpperArm, RightUpperArm), REALDISP (LeftLowerArm, LeftUpperArm, RightLowerArm, RightUpperArm), REALWORLD (forearm, upperarm), VTT-ConIoT (UpperArm), WARD (LowerLeftForearm, LowerRightForearm)
Leg	17	DSADS (LeftLeg, RightLeg), Exoskeletons (RightLeg, LeftLeg), FORTHTRACE (RightThigh), HARTH (thigh), IMWSHA (thigh), LARa (LeftLeg, RightLeg), MEx (Thigh), REALDISP (LeftCalf, LeftThigh, RightCalf, RightThigh), REALWORLD (shin, thigh), SELFBACK (thigh)
Front	11	DSADS (Torso), Exoskeletons (Chest), FORTHTRACE (Torso), IMWSHA (chest), LARa (Torsor), MHEALTH (Chest), PAMAP2 (chest), REALWORLD (chest, waist), VTT-ConIoT (Hip), WARD (FrontCenter)
Ankle	5	FORTHTRACE (LeftAnkle), MHEALTH (LeftAnkle), PAMAP2 (ankle), WARD (LeftAnkle, RightAnkle)
Wrist	11	Exoskeletons (RightWrist, LeftWrist), FORTHTRACE (LeftWrist, RightWrist), IMWSHA (wrist), MEx (Wrist), MHEALTH (RightWrist), OPENPACK (LeftWrist, RightWrist), PAMAP2 (hand), SELFBACK (watch)
Phone	2	PAAL (Phone), USCHAD (smartphone)
Back	3	HARTH (back), REALDISP (Back), VTT-ConIoT (Back)
Head	2	REALWORLD (head), UCAEHAR (Glasses)
Total	**66**	**18 unique datasets with specific sensor positions**

of unseen activity classes, our evaluation focuses on generalization where activity class names are known but sensor data distributions are unknown. Recent work [19, 64, 65] has highlighted the importance of this task.

Let $N_{activities}$ denote the number of common activity classes (6 in our framework), and N_{cross} represent the number of datasets used for zero-shot evaluation (3 in our evaluation: DSADS, MHEALTH, PAMAP2), which are the only datasets containing all 6 common activity classes, as shown in Table V. This evaluation employs a comprehensive leave-one-dataset-out cross-validation across the N_{cross} datasets, where each target dataset is completely excluded from the fine-tuning process and the remaining datasets including supporting datasets are used to train a HAR model consisting of a pretrained feature extractor and activity classifier. We utilize arm and leg sensors (indicated in bold in Table III) to ensure consistent sensor placement across all evaluation datasets. To enable fair comparison, we map activity labels to the $N_{activities}$ common classes.

The $N_{activities}$ common activity classes are: (1) Static – stationary postures (standing, sitting, lying merged due to low sensor distinguishability); (2) Walking – various walking variations; (3) Running – fast locomotion including jogging; (4) Stairs – ascending and descending; (5) Jumping – various jumping activities; (6) Cycling – stationary bike and regular cycling.

For each dataset serving as the evaluation target, we calculate the macro F1 score across the available common activity classes in that dataset. The final Zero-shot Performance score is computed as the arithmetic mean across the N_{cross} datasets:

$$\text{Zero-shot Performance} = \frac{1}{N_{cross}} \sum_{i=1}^{N_{cross}} \text{F1}^{\text{zero-shot}}_{\text{dataset}_i} \quad (5)$$

Since the dataset split is fixed in the leave-one-dataset-out protocol, we average all zero-shot evaluations over 4 random seeds to ensure reproducibility.

TABLE V
COVERAGE OF COMMON ACTIVITY CLASSES FOR ZERO-SHOT ACTIVITY RECOGNITION

Dataset	Static	Walk	Stairs	Run/Jog	Jump	Cycle	Total
LODO Evaluation Targets							
DSADS	✓	✓	✓	✓	✓	✓	6/6
MHEALTH	✓	✓	✓	✓	✓	✓	6/6
PAMAP2	✓	✓	✓	✓	✓	✓	6/6
Supporting Datasets							
ForthTrace	✓	✓	✓	-	-	-	3/6
REALDISP	-	✓	-	✓	✓	✓	4/6
RealWorld	✓	✓	✓	✓	✓	-	5/6
SELFBACK	✓	✓	✓	✓	-	-	4/6
WARD	✓	✓	✓	✓	✓	-	5/6

For comparison, we include **Supervised** baseline, which represents the traditional supervised learning scenario where models are trained and evaluated on the same target dataset using 4-fold cross-validation. This baseline shows the performance when labeled data from the target domain is available, providing context for understanding the zero-shot transfer challenge where no target domain labels are accessible during training.

E. Implementation of HARBench

HARBench is implemented as a unified Python-based framework designed to facilitate reproducible and scalable benchmarking of HAR foundation models. Built on PyTorch 2.7.0 and Python 3.10.12, the framework provides a streamlined pipeline for dataset integration, model training, and multi-dimensional evaluation.

The dataset integration pipeline handles the integration of 32 heterogeneous datasets through automated format conversion and standardization. Users initiate the workflow by executing provided shell scripts that download datasets from their

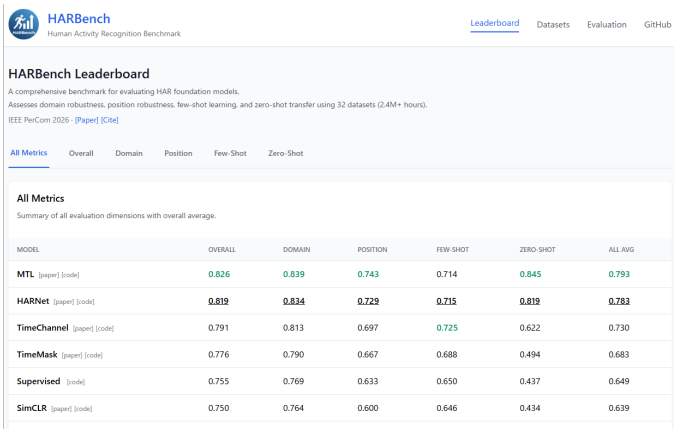

Fig. 2. Web UI of HARBench

respective sources. For large-scale datasets like NHANES, which contains over 30TB of raw data, our framework implements streaming preprocessing that incrementally converts raw signals into windowed segments, applies float32 conversion, and employs efficient compression, reducing the total storage requirement to approximately 1 TB while preserving signal fidelity.

HARBench provides a simple framework for both self-supervised pretraining on the 14 unlabeled datasets and comprehensive evaluation on the 18 labeled datasets. The framework supports various SSL methods (masked reconstruction, contrastive learning, etc.) for pretraining and automatically handles the complete evaluation pipeline across all five dimensions with a few lines of code. Figure 2 shows the web-based leaderboard for benchmark results.

The framework, along with comprehensive documentation, tutorials, and pre-trained model checkpoints, is openly available at https://github.com/litchi7777/harbench under the MIT license. The leaderboard system is publicly available at https://litchi7777.github.io/harbench/. Comprehensive documentation, tutorials, and pre-trained model checkpoints are provided to facilitate adoption and extension by the research community.

IV. Experimental Results and Analysis

A. Baseline Methods

We evaluate a comprehensive set of baseline methods on HARBench, covering conventional supervised learning, existing HAR foundation models, and self-supervised learning approaches.

Supervised Baseline: We implement standard supervised learning using ResNet18 architecture, trained directly on labeled data without any pretraining.

HAR Foundation Models: We include four state-of-the-art HAR foundation models in our evaluation: (1) HARNet [7], pretrained on 700,000 person-days of wrist-worn accelerometer data from the UK Biobank, using three multi-task self-supervised objectives: arrow-of-time prediction, permutation, and time warping tasks; (2) SelfPAB [21], a Transformer-based model pretrained on 100,000 hours of HUNT4 accelerometer data using masked spectrogram reconstruction; (3) LIMU-BERT [66], a BERT-based model for IMU sensing using masked reconstruction (we use the HHAR-pretrained checkpoint); (4) IMU-MAE [67], a masked autoencoder pretrained on video-aligned IMU data.

General Foundation Models: We include PatchTST [68], which uses patch-based tokenization for time series, and MOMENT [69], a large-scale foundation model pretrained on diverse time series datasets.

Self-Supervised Learning Methods: We implement six SSL approaches: (1) Multi-Task Learning (MTL), combining arrow of time, permutation, and time warping predictions; (2) TimeChannel [70], using spatio-temporal masking; (3) TimeMask, focused on temporal masking; (4) CPC [71], using contrastive predictive coding; (5) SimCLR [72] and (6) MoCo [72], both adapted for time series.

All models use ResNet18 as the backbone architecture for fair comparison, consistent with HARNet's design. Models undergo identical preprocessing and evaluation protocols as described in Section III. Table VI summarizes the pretraining data used by each model.

TABLE VI
Pretraining Data for Each Model

Model	Pretraining Data	Access
HARNet	UK Biobank (700k person-days)	Restricted
SelfPAB	HUNT4 (100k hours)	Restricted
LIMU-BERT	HHAR	Open
IMU-MAE	WEAR (video+IMU)	Open
PatchTST	Time series datasets	Open
MOMENT	Large-scale time series	Open
SSL (HARBench)	14 datasets (Table 1)	Open

TABLE VII
Overall Performance and Domain Robustness Results

Model	Overall Avg	Daily (10)	Exercise (4)	Industry (4)	Domain Avg
HARNet	0.819	0.791	0.923	**0.787**	0.834
SelfPAB	0.692	0.660	0.818	0.644	0.708
LIMU-BERT	0.595	0.580	0.651	0.576	0.603
IMU-MAE	0.735	0.671	0.799	0.763	0.744
PatchTST	0.695	0.671	0.802	0.648	0.707
MOMENT	0.650	0.599	0.815	0.611	0.675
MTL	**0.826**	**0.799**	**0.932**	**0.787**	**0.839**
TimeChannel	0.791	0.748	0.914	0.776	0.813
TimeMask	0.776	0.746	0.893	0.732	0.790
CPC	0.738	0.709	0.848	0.702	0.753
SimCLR	0.750	0.722	0.861	0.709	0.764
MoCo	0.737	0.706	0.838	0.714	0.753
Supervised	0.755	0.727	0.866	0.713	0.769

B. Overall Performance and Domain Robustness

Table VII reveals interesting and noteworthy findings about HAR foundation model performance. Most strikingly, MTL significantly outperforms the supervised baseline by approximately 7%, while achieving comparable performance to HARNet despite using only one-tenth of the training data. Notably, HARNet was pretrained solely on wrist-mounted data, whereas MTL leverages diverse data from multiple public datasets. These results highlight that dataset diversity across domains can be more valuable than simply having large volumes of data from a single restricted source. Overall, a clear hierarchy among SSL methods emerges: multi-task learning consistently outperforms masked reconstruction approaches, which in turn surpass contrastive learning methods.

TABLE VIII
POSITION ROBUSTNESS RESULTS ACROSS SENSOR PLACEMENTS

Position	Arm(15)	Leg(17)	Front(11)	Ankle(5)	Wrist(11)	Phone(2)	Back(3)	Head(2)	Avg
HARNet	0.705	0.767	0.682	0.743	0.731	0.780	0.706	0.720	0.729
SelfPAB	0.501	0.630	0.549	0.625	0.577	0.556	0.534	0.551	0.565
LIMU-BERT	0.461	0.619	0.483	0.603	0.505	0.590	0.473	0.623	0.545
IMU-MAE	0.582	0.650	0.599	0.681	0.627	0.544	0.604	0.616	0.618
PatchTST	0.505	0.596	0.554	0.653	0.557	0.547	0.419	0.636	0.558
MOMENT	0.408	0.503	0.483	0.595	0.467	0.433	0.353	0.552	0.474
MTL	**0.709**	**0.771**	**0.702**	**0.770**	**0.732**	**0.788**	**0.729**	**0.743**	**0.743**
TimeChannel	0.675	0.745	0.642	0.701	0.695	0.718	0.694	0.709	0.697
TimeMask	0.634	0.706	0.592	0.681	0.674	0.726	0.649	0.673	0.667
CPC	0.538	0.651	0.550	0.612	0.607	0.687	0.540	0.653	0.605
SimCLR	0.539	0.659	0.540	0.604	0.619	0.651	0.548	0.641	0.600
MoCo	0.577	0.670	0.559	0.627	0.615	0.694	0.594	0.658	0.624
Supervised	0.592	0.692	0.573	0.633	0.637	0.679	0.586	0.676	0.633

TABLE IX
FEW-SHOT LEARNING PERFORMANCE ACROSS DIFFERENT DATA PERCENTAGES

Model	1%	2%	5%	10%	20%	50%	Avg
HARNet	0.564	0.636	0.713	0.762	0.798	**0.818**	0.715
SelfPAB	0.483	0.542	0.610	0.663	0.675	0.685	0.610
LIMU-BERT	0.377	0.440	0.514	0.541	0.572	0.586	0.505
IMU-MAE	0.471	0.522	0.605	0.654	0.693	0.725	0.612
PatchTST	0.344	0.424	0.516	0.590	0.628	0.652	0.526
MOMENT	0.513	0.557	0.605	0.624	0.637	0.648	0.597
MTL	0.559	0.631	0.716	0.764	**0.800**	0.814	0.714
TimeChannel	**0.608**	**0.658**	**0.730**	**0.769**	0.789	0.798	**0.725**
TimeMask	0.544	0.630	0.695	0.735	0.754	0.771	0.688
CPC	0.490	0.566	0.644	0.697	0.716	0.728	0.640
SimCLR	0.508	0.575	0.644	0.684	0.726	0.742	0.647
MoCo	0.454	0.531	0.624	0.673	0.707	0.725	0.619
Supervised	0.517	0.582	0.641	0.706	0.720	0.736	0.650

TABLE X
NUMBER OF TRAINING SAMPLES USED AT EACH FEW-SHOT RATIO ACROSS DATASETS (FOLD 0)

Dataset	Train	#Class	1%	2%	5%	10%	20%	50%
DSADS	4402	19	44	88	220	440	880	2201
ForthTrace	3145	11	31	62	157	314	629	1572
HARTH	23136	10	231	462	1156	2313	4627	11568
IMWSHA	422	11	11	11	21	42	84	211
LARa	1399	6	13	27	69	139	279	699
MEx	3577	7	35	71	178	357	715	1788
MHEALTH	1240	12	12	24	62	124	248	620
OPENPACK	32517	9	325	650	1625	3251	6503	16258
VTT-ConIoT	2717	16	27	54	135	271	543	1358
Exoskeletons	3741	4	37	74	187	374	748	1870
PAAL	7183	10	71	143	359	718	1436	3591
PAMAP2	3314	12	33	66	165	331	662	1657
REALDISP	3167	33	33	63	158	316	633	1583
RealWorld	15608	8	156	312	780	1560	3121	7804
SELFBACK	13569	9	135	271	678	1356	2713	6784
UCA-EHAR	5275	6	52	105	263	527	1055	2637
USC-HAD	6567	12	65	131	328	656	1313	3283
WARD	6210	13	62	124	310	621	1242	3105
Median	3960	11	40	81	203	407	814	2035

Domain-wise analysis reveals that the HAR foundation models substantially outperform the supervised baseline across all domains. This result indicates that learning from large-scale unlabeled data is consistently beneficial, regardless of the target domain. Notably, although the unlabeled pretraining data do not include any industrial-domain recordings, MTL still achieves the best performance in the industrial domain. This finding is particularly interesting, as it suggests that pretraining on a large number of datasets with diverse sensor placements enables effective cross-domain knowledge transfer, allowing representations learned from daily-life activities to generalize well to industrial scenarios.

C. Position Robustness Evaluation

Table VIII reveals notable position-specific patterns in foundation model performance. MTL achieves the highest accuracy across all sensor positions, while HARNet attains comparable performance to MTL on wrist and arm placements. This behavior can be attributed to the fact that HARNet was pretrained on large-scale unlabeled data predominantly collected from wrist-worn sensors. As a result, HARNet appears to focus primarily on hand-related motion patterns. In contrast, MTL excels at rarer sensor placements, i.e., the head, indicating that pretraining on diverse datasets with heterogeneous sensor locations is more effective for capturing motion patterns associated with such less common positions.

D. Few-shot Learning Evaluation

Table IX reveals a critical transition in optimal SSL strategies around 10–20% data availability. In extremely limited data scenarios (1–10%), TimeChannel achieves the highest performance, while MTL dominates from 20% onward. This transition suggests that when labeled samples are scarce, predictive learning strategies that reconstruct masked portions generalize better than discriminative approaches. In MTL, discriminative learning corresponds to defining binary classification boundaries, such as detecting whether a specific data augmentation has been applied. When only a few examples per class are available, predicting missing temporal patterns is more valuable than relying on these classification boundaries. As data availability increases beyond this threshold, MTL's approach of combining multiple self-supervised tasks (arrow-of-time, permutation, and time warping) becomes increasingly effective, leveraging complementary signals from these different pretext tasks to build robust representations.

E. Zero-shot Generalization Evaluation

Table XI demonstrates the transformative impact of self-supervised learning on cross-dataset generalization. Interestingly, dataset difficulty varies significantly. MHEALTH proves easiest for zero-shot transfer, whereas PAMAP2 poses the greatest challenge. To understand this performance gap, we analyzed the sensor data distributions across datasets. The standard deviation of accelerometer measurements reveals that while most datasets show similar variability (σ=0.565-0.666 for SELFBACK, ForthTrace, RealWorld, WARD, REALDISP,

TABLE XI
ZERO-SHOT PERFORMANCE ON UNSEEN DATASETS

Model	DSADS	MHEALTH	PAMAP2	Avg
HARNet	0.725	0.849	0.693	0.756
SelfPAB	0.683	0.825	0.541	0.683
LIMU-BERT	0.326	0.298	0.361	0.328
IMU-MAE	**0.747**	0.760	0.704	0.737
PatchTST	0.664	0.584	0.466	0.571
MOMENT	0.689	0.711	0.621	0.674
MTL	0.746	**0.966**	**0.709**	**0.807**
TimeChannel	0.594	0.819	0.495	0.636
TimeMask	0.514	0.572	0.595	0.560
CPC	0.512	0.564	0.448	0.506
SimCLR	0.484	0.475	0.455	0.471
MoCo	0.524	0.451	0.439	0.471
Supervised	0.924	0.976	0.745	0.882

DSADS, MHEALTH), PAMAP2 exhibits substantially higher variability (σ=0.746). This distinct sensor data distribution in PAMAP2 creates a significant domain shift that challenges zero-shot transfer. The distribution mismatch, rather than activity complexity per se, appears to be the primary factor affecting zero-shot performance. Importantly, such distribution shifts frequently occur in real-world deployments when models encounter different sensor manufacturers, hardware specifications, or data collection protocols. Our zero-shot evaluation metric is therefore highly valuable for assessing model resilience to these inevitable deployment challenges, providing practitioners with crucial insights into model robustness before real-world implementation.

V. DISCUSSION

The comprehensive experiments using HARBench on 32 diverse datasets provided reliable insights that could not be obtained from previous studies relying on only a few datasets.

A. Data Diversity vs. Quantity

Our key finding challenges the common assumption that larger datasets automatically lead to better foundation models. MTL achieves performance comparable to HARNet while using only *one-tenth* of the training data, highlighting that diversity of training data is more critical than total quantity for HAR foundation models. This insight indicates that carefully curated datasets encompassing multiple domains and sensor configurations can outperform massive single-source datasets. By leveraging publicly available, diverse datasets, institutions without access to restricted large-scale data can still develop competitive HAR foundation models, thereby democratizing research in this field.

B. SSL Method Hierarchy and Few-shot Strategy Shift

Our evaluation establishes a clear hierarchy among SSL approaches: multi-task learning (MTL) generally achieves the strongest performance across most settings, followed by masked reconstruction methods and then contrastive learning approaches. However, the relative effectiveness of each SSL strategy depends on the availability of training data and deployment conditions.

A critical shift occurs around 10-20% labeled data availability: TimeChannel excels in extremely low-data settings, while MTL and HARNet becomes more effective as data increases. This provides practical guidance for deployment—TimeChannel for rapid prototyping with minimal data, MTL for general applications. However, the relatively gradual improvement across data regimes shows that current foundation models still struggle to achieve substantial gains when labeled data is extremely scarce. Future research should aim to develop models that improve more sharply in few-shot scenarios, potentially by better integrating masked reconstruction and multi-task learning strategies.

C. Zero-shot Transfer and Foundation Model Robustness

Our Zero-shot Activity Recognition results reveal two critical insights. First, MTL achieves superior zero-shot performance (0.807), demonstrating that multi-task learning on diverse datasets inherently develops domain-invariant representations—a key advantage for real-world deployment. Second, while existing complex foundation models (SelfPAB: 0.683, LIMU-BERT: 0.328, IMU-MAE: 0.737, PatchTST: 0.571, MOMENT: 0.674) show moderate zero-shot performance, they are all outperformed by MTL trained on diverse public datasets. This suggests that training data diversity matters more than model scale or architecture complexity for cross-dataset generalization. These findings indicate that future HAR models should prioritize multi-dataset training strategies that emphasize domain-invariant feature learning over dataset-specific optimization.

VI. CONCLUSION

This paper introduced HARBench, the first comprehensive benchmark for evaluating HAR foundation models across five dimensions. Our experiments revealed that data diversity outweighs quantity—public datasets can match proprietary approaches—and establish clear SSL method hierarchies with context-dependent optimal selections.

The results also highlighted future directions for HAR foundation model research: (1) fostering industry-academia collaboration to systematically collect diverse HAR datasets, (2) exploring hybrid SSL approaches (e.g., masked reconstruction with multi-task learning) to achieve robust performance across all data regimes, particularly in extreme few-shot scenarios, and (3) enhancing zero-shot transfer through multi-dataset training that prioritizes domain-invariant features.

HARBench addresses the critical gap in standardized HAR foundation model evaluation, providing unified protocols and practical deployment guidance. By showing that public datasets can achieve performance comparable to restricted large-scale datasets, this work democratizes HAR foundation model research and is expected to catalyze advances in domain-adaptive pretraining and extreme few-shot learning.

ACKNOWLEDGEMENT

This study is partially supported by JST K Program Grant Number JPMJKP25V2, and JSPS KAKENHI Grant Number JP21H05299 and JP25K22801, Japan.

REFERENCES

[1] J. Morales, N. Yoshimura, Q. Xia, A. Wada, Y. Namioka, and T. Maekawa, "Acceleration-based human activity recognition of packaging tasks using motif-guided attention networks," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2022, pp. 1–12.

[2] H. Haresamudram, I. Essa, and T. Plötz, "Investigating enhancements to contrastive predictive coding for human activity recognition," in *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2023, pp. 232–241.

[3] Q. Xia, J. Morales, Y. Huang, T. Hara, K. Wu, H. Oshima, M. Fukuda, Y. Namioka, and T. Maekawa, "Self-supervised learning for complex activity recognition through motif identification learning," *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 3779–3793, 2025.

[4] J. Morales, Q. Xia, N. Yoshimura, H. Oshima, M. Fukuda, Y. Namioka, and T. Maekawa, "Multilevel Transfer Learning for Complex Work Activity Recognition in Logistic Domain," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2025, pp. 159–170. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/PerCom64205.2025.00036

[5] Y. Zhang and T. Maekawa, "InterHandNet: Capturing two-hand interaction for robust hand-washing activity recognition," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2025, pp. 13–24.

[6] W. Ge, G. Mou, E. O. Agu, and K. Lee, "Semantically encoding activity labels for context-aware human activity recognition," in *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2025, pp. 190–196.

[7] H. Yuan, S. Chan, A. P. Creagh, C. Tong, A. Acquah, D. A. Clifton, and A. Doherty, "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data," *NPJ Digital Medicine*, vol. 7, no. 1, p. 91, 2024.

[8] S. Ek, R. Presotto, G. Civitarese, F. Portet, P. Lalanda, and C. Bettini, "Comparing self-supervised learning techniques for wearable human activity recognition," *CCF Transactions on Pervasive Computing and Interaction*, pp. 1–18, 2025.

[9] Y. Cai, B. Guo, F. Salim, and Z. Hong, "Towards generalizable human activity recognition: A survey," *arXiv preprint arXiv:2508.12213*, 2025.

[10] H. Chen, C. Gouin-Vallerand, K. Bouchard, S. Gaboury, M. Couture, N. Bier, and S. Giroux, "Contrastive self-supervised learning for sensor-based human activity recognition: A review," *IEEE Access*, 2024.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[14] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[15] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[17] R. Bommasani, P. Liang, and T. Lee, "Holistic evaluation of language models," *Annals of the New York Academy of Sciences*, vol. 1525, no. 1, pp. 140–146, 2023.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[19] O. Napoli, D. Duarte, P. Alves, D. H. P. Soto, H. E. de Oliveira, A. Rocha, L. Boccato, and E. Borin, "A benchmark for domain adaptation and generalization in smartphone-based human activity recognition," *Scientific Data*, vol. 11, no. 1, p. 1192, 2024.

[20] C. Turetta, F. Demrozi, and G. Pravadelli, "B-har: an open-source baseline framework for in-depth study of human activity recognition datasets and workflows," *IEEE Access*, 2024.

[21] A. Logacjov, S. Herland, A. Ustad, and K. Bach, "Selfpab: large-scale pre-training on accelerometer data for human activity recognition," *Applied Intelligence*, vol. 54, no. 6, pp. 4545–4563, 2024.

[22] H. Haresamudram, I. Essa, and T. Plötz, "Assessing the state of self-supervised human activity recognition using wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–47, 2022.

[23] H. Qian, T. Tian, and C. 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*, 2022, pp. 3761–3771.

[24] C. for Disease Control, P. (CDC), and N. C. for Health Statistics (NCHS), "National health and nutrition examination survey data," 2022, hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Physical Activity Monitor - Raw Data 80Hz (PAX80_G). [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/2011-2012/pax80_g.htm

[25] ——, "National health and nutrition examination survey data," 2022, hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Physical Activity Monitor - Raw Data 80Hz (PAX80_H). [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/2013-2014/pax80_h.htm

[26] M. F. Bruno, Barbara and A. Sgorbissa, "Dataset for ADL Recognition with Wrist-worn Accelerometer," UCI Machine Learning Repository, 2012, DOI: https://doi.org/10.24432/C5PC99.

[27] M. Shoaib, S. Bosch, H. Scholten, P. J. Havinga, and O. D. Incel, "Towards detection of bad habits by fusing smartphone and smartwatch sensors," in *2015 IEEE International Conference on Pervasive Computing and Communication workshops (PerCom Workshops)*. IEEE, 2015, pp. 591–596.

[28] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors," *Sensors*, vol. 16, no. 4, p. 426, 2016.

[29] S. Chan Chang, R. Walmsley, J. Gershuny, T. Harms, E. Thomas, K. Milton, P. Kelly, C. Foster, A. Wong, N. Gray *et al.*, "Capture-24: Activity tracker dataset for human activity recognition," 2021.

[30] P. M. Roggen, Daniel and J. Hausdorff, "Daphnet Freezing of Gait," UCI Machine Learning Repository, 2010, DOI: https://doi.org/10.24432/C56K78.

[31] A. Ustad, A. Logacjov, S. Ø. Trollebø, P. Thingstad, B. Vereijken, K. Bach, and N. S. Maroni, "Validation of an Activity Type Recognition Model Classifying Daily Physical Behavior in Older Adults: The HAR70+ Model," *Sensors*, vol. 23, no. 5, p. 2368, Jan. 2023.

[32] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015, pp. 127–140.

[33] A. Jalal, M. A. K. Quaid, and A. S. Hasan, "Wearable sensor-based human behavior understanding and recognition in daily life for smart environments," in *2018 International Conference on Frontiers of Information Technology (FIT)*, 2018, pp. 105–110.

[34] A. Jalal, M. A. Quaid, and M. Sidduqi, "A triaxial acceleration-based human motion detection for ambient smart home system," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 353–358.

[35] Y. Mohammad, K. Matsumoto, and K. Hoashi, "A dataset for activity recognition in an unmodified kitchen using smart-watch accelerometers," in *the 16th International Conference on Mobile and Ubiquitous Multimedia*, 2017, pp. 63–68.

[36] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *the International Conference on Internet of Things Design and Implementation*, ser. IoTDI '19. New York, NY, USA: ACM, 2019, pp. 49–58. [Online]. Available: http://doi.acm.org/10.1145/3302505.3310068

[37] ——, "Protecting sensory data against sensitive inferences," in *the 1st Workshop on Privacy by Design in Distributed Systems*, ser. W-P2DS'18. New York, NY, USA: ACM, 2018, pp. 2:1–2:6. [Online]. Available: http://doi.acm.org/10.1145/3195258.3195260

[38] D. Roggen, A. Calatroni, L.-V. Nguyen-Dinh, R. Chavarriaga, and H. Sagha, "Opportunity activity recognition," https://doi.org/10.24432/C5M027, 2012, uCI Machine Learning Repository.

[39] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[40] C. Carpineti, V. Lomonaco, L. Bedogni, M. Di Felice, and L. Bononi, "Custom dual transportation mode detection by smartphone devices exploiting sensor diversity," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2018, pp. 367–372.

[41] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[42] B. Barshan and M. C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *The Computer Journal*, vol. 57, no. 11, pp. 1649–1667, 2014.

[43] K. Karagiannaki, A. Panousopoulou, and P. Tsakalides, "The FORTH-TRACE dataset for human activity recognition of simple activities and postural transitions using a body area network," Jul. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.841301

[44] A. Logacjov, K. Bach, A. Kongsvold, H. B. Bårdstu, and P. J. Mork, "HARTH: A Human Activity Recognition Dataset for Machine Learning," *Sensors*, vol. 21, no. 23, p. 7853, Nov. 2021.

[45] K. Bach, A. Kongsvold, H. Bårdstu, E. M. Bardal, H. S. Kjærnli, S. Herland, A. Logacjov, and P. J. Mork, "A Machine Learning Classifier for Detection of Physical Activity Types and Postures During Free-Living," *Journal for the Measurement of Physical Behaviour*, pp. 1–8, Dec. 2021.

[46] S. B. U. D. Tahir, A. Jalal, and K. Kim, "Wearable inertial sensors for daily activity analysis based on adam optimization and the maximum entropy markov model," *Entropy*, vol. 22, no. 5, p. 579, 2020.

[47] S. B. ud din Tahir, A. Jalal, and M. Batool, "Wearable sensors for activity analysis using smo-based random forest over smart home and sports datasets," in *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*, 2020, pp. 1–6.

[48] F. Niemann, C. Reining, F. Moya Rueda, N. R. Nair, J. A. Steffens, G. A. Fink, and M. Ten Hompel, "Lara: Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, vol. 20, no. 15, p. 4083, 2020.

[49] A. Wijekoon, N. Wiratunga, and K. Cooper, "Mex: Multi-modal exercises dataset for human activity recognition," *arXiv preprint arXiv:1908.08992*, 2019.

[50] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomedical Engineering Online*, vol. 14, no. 2, pp. 1–20, 2015.

[51] N. Yoshimura, J. Morales, T. Maekawa, and T. Hara, "Openpack: A large-scale dataset for recognizing packaging works in iot-enabled logistic environments," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2024, pp. 90–97.

[52] S.-M. Mäkelä, A. Lämsä, J. S. Keränen, J. Liikka, J. Ronkainen, J. Peltola, J. Häikiö, S. Järvinen, and M. Bordallo López, "Introducing vtt-coniot: A realistic dataset for activity recognition of construction workers using imu devices," *Sustainability*, vol. 14, no. 1, p. 220, 2021.

[53] M. Pesenti, G. Invernizzi, J. Mazzella, M. Bocciolone, A. Pedrocchi, and M. Gandolla, "Imu-based human activity recognition and payload classification for low-back exoskeletons," *Scientific Reports*, vol. 13, no. 1, p. 1184, 2023.

[54] P. Climent-Pérez, Ángela María Muñoz-Antón, A. Poli, S. Spinsante, and F. Florez-Revuelta, "Paal adl accelerometry dataset v2.0," 2021.

[55] A. Reiss, "PAMAP2 Physical Activity Monitoring," UCI Machine Learning Repository, 2012, DOI: https://doi.org/10.24432/C5NW2H.

[56] O. Baños, M. Damas, H. Pomares, I. Rojas, M. A. Tóth, and O. Amft, "A benchmark dataset to evaluate sensor displacement in activity recognition," in *the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 1026–1035.

[57] T. Sztyler, H. Stuckenschmidt, and W. Petrich, "Position-aware activity recognition with wearable devices," *Pervasive and Mobile Computing*, vol. 38, pp. 281–295, 2017.

[58] "selfBACK," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C5SK65.

[59] P.-E. Novac, A. Pegatoquet, B. Miramond, and C. Caquineau, "Ucaehar: A dataset for human activity recognition with embedded ai on smart glasses," *Applied Sciences*, vol. 12, no. 8, p. 3849, 2022.

[60] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 1036–1043.

[61] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, pp. 103–115, 2009.

[62] G. Zhu, D. Zhao, C. Li, M. Zhao, Z. Zhang, H. Quan, and H. Ma, "Master: A multi-modal foundation model for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 3, pp. 1–26, 2025.

[63] K. Matsuishi, K. Ukita, and T. Okita, "Multimodal foundation model for cross-modal retrieval and activity recognition tasks," *arXiv preprint arXiv:2506.03174*, 2025.

[64] L. Ban, T. Zhu, X. Lu, Q. Qiu, W. Han, S. Li, L. Chen, K. I.-K. Wang, M. Nie, and Y. Wan, "Har-doremi: Optimizing data mixture for self-supervised human activity recognition across heterogeneous imu datasets," *arXiv preprint arXiv:2503.13542*, 2025.

[65] Z. Hong, Z. Li, S. Zhong, W. Lyu, H. Wang, Y. Ding, T. He, and D. Zhang, "Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–26, 2024.

[66] H. Xu, P. Zhou, R. Tan, M. Li, and G. 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*, 2021, pp. 220–233.

[67] M. Zhang, Y. Huang, R. Liu, and Y. Sato, "Masked video and body-worn imu autoencoder for egocentric action recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 312–330.

[68] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[69] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "Moment: A family of open time-series foundation models," *arXiv preprint arXiv:2402.03885*, 2024.

[70] J. Wang, W. Cui, T. Zhu, H. Ning, and Z. Liu, "An improved masking strategy for self-supervised masked reconstruction in human activity recognition," *IEEE Sensors Journal*, 2024.

[71] H. Haresamudram, I. Essa, and T. Plötz, "Contrastive predictive coding for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–26, 2021.

[72] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo, "Exploring contrastive learning in human activity recognition for healthcare," *arXiv preprint arXiv:2011.11542*, 2020.