Towards Customizable Foundation Models for Human Activity Recognition with Wearable Devices

MINGHUI QIU, DSA, The Hong Kong University of Science and Technology (Guangzhou), China
CEKAI WENG, DSA, The Hong Kong University of Science and Technology (Guangzhou), China
MINGMING FAN, CMA & IoT, The Hong Kong University of Science and Technology (Guangzhou), China
KAISHUN WU, DSA & IoT, The Hong Kong University of Science and Technology (Guangzhou), China

Foundation models have achieved remarkable success across various domains by learning general representations from raw data, offering a promising paradigm for diverse applications. This concept holds great potential for advancing human activity recognition (HAR), particularly in overcoming challenges associated with collecting large-scale labeled datasets. However, the dynamic nature of HAR tasks, characterized by diverse sensing devices and activity types, results in fragmented datasets that question the feasibility of applying foundation model to this domain. In this work, we propose a novel foundation model training framework that effectively leverages heterogeneous datasets through a two-stage training strategy: (1) self-supervised learning to extract cross-domain sensor patterns, followed by (2) multi-task learning to align representations with semantic contexts. The effectiveness of the trained foundation model is demonstrated through extensive downstream experiments, with the superior fine-tuning performance across various modalities and input configurations—achieving the highest performance metric in 10 out of 12 settings—further validating the robustness and adaptability. While our model shows a performance gap compared to foundation models pre-trained on large-scale or high-quality data in zero- and few-shot scenarios, its competitive results with a more flexible architecture demonstrate the efficiency and potential of our training strategy for HAR foundation models.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Foundation Model, Human Activity Recognition, Wearable Sensing

ACM Reference Format:
Minghui Qiu, Cekai Weng, Mingming Fan, and Kaishun Wu. 2025. Towards Customizable Foundation Models for Human Activity Recognition with Wearable Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 122 (September 2025), 29 pages. https://doi.org/10.1145/3749479

1 Introduction

Foundation Models learn general-purpose representations from large-scale data that can be adapted to diverse downstream tasks [6, 8, 13, 19, 25, 28, 30, 35, 45, 60, 62, 69]. These models enable transfer learning and task-specific adaptation, offering a new approach to solving complex problems. Human Activity Recognition (HAR) systems fall short in generalization and often require extensive task-specific data collection and retraining. The diversity

Authors' Contact Information: Minghui Qiu, DSA, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, mqiu585@connect.hkust-gz.edu.cn; Cekai Weng, DSA, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, cweng368@connect.hkust-gz.edu.cn; Mingming Fan, CMA & IoT, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, mingmingfan@ust.hk; Kaishun Wu, DSA & IoT, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, wuks@hkust-gz.edu.cn.

ACM 2474-9567/2025/9-ART122
https://doi.org/10.1145/3749479

in sensing conditions, including device types [43], placements [5], and user demographics [38] coupled with rapidly evolving sensing technologies and emerging activity types, makes it impractical to collect comprehensive datasets for every scenario. Recent research [30] suggested that foundation models could help in generalization of HAR systems. However, realizing an HAR foundation model that supports full customization faces a fundamental challenge: data fragmentation. This fragmentation stems from varying sensor characteristics across devices (e.g., watches, phones, wristbands), placement-dependent data variations, diverse sensing specifications [38, 41, 43], and inconsistent activity taxonomies [43].

To address these challenges, we propose a two-stage foundation model training framework consisting of representation learning and task alignment training. In the first stage, we employ a self-supervised learning approach to infer robust patterns from fragmented data. We introduce a quantization module to enable the model to extract meaningful embeddings by learning from unlabeled data collected across various sensors and devices. In the second stage, we adopt a multi-task supervised learning approach to refine the learned representations and align them with real-world use cases. The optimization objective of this stage includes three complementary components. First, we introduce a prediction task for quantizing indices, which builds on the quantization module from the first stage and serves as a regularization mechanism to ensure representation consistency across the two stages. Second, we implement downstream activity recognition tasks to facilitate practical applications. Finally, we leverage semantic embeddings from a language model as universal prior knowledge, bridging the gap between sensor data and activity semantics. These components jointly optimize the model to develop robust, practical, and semantically meaningful representations that can generalize effectively to real-world scenarios. To further improve model generalization across downstream tasks, we apply placement-aware training. By accounting for device placement variability during training, placement-aware training ensures that the model performs robustly across different configurations and deployment scenarios. This placement-aware design complements the primary representation learning and task alignment strategies, enabling the model to handle highly fragmented environments effectively.

To evaluate the effectiveness of the proposed foundation model, we trained the HAR foundation model (HAR-FM) on 4 datasets and conducted extensive experiments on 5 other datasets. The results demonstrate that the well-trained HAR-FM advances the field by winning 10 out of 12 cross-user experiments with various input settings (e.g., different modalities and time window configurations) against baselines that cover the state-of-the-art pretraining and/or fine-tuning mechanisms. We also compare the performance in zero and few-shot experiments with other foundation model instances (YuanSSL [62] and UniMTS [65]). Though the performance gap with present foundation model instances persists, the relatively (few times) smaller datasets used for training and more flexible model structure (section 6) of the HAR-FM should further demonstrate the advancement of the proposed training strategy. In general, the contributions of our paper are as follows:

- To the best of our knowledge, we are the first to explore and demonstrate the feasibility of training a multi-sensor foundation model for Human Activity Recognition (HAR) using heterogeneous fractions of open-source sensor datasets.
- We propose a viable foundation model training framework that consists of self-supervised learning with unlabeled data and multi-task supervised learning to align with real-world use cases.
- The pretrained foundation model is made publicly available along with usage tutorials[1], which we hope it could facilitate customized HAR applications across different wearable devices in this domain.

[1] https://github.com/qmhsam/HAR-FM.git

2 Background and Related Works

2.1 Foundation Model For Wearable Devices

Foundation models have revolutionized various domains by providing robust, generalizable frameworks that can be fine-tuned for specific tasks. Foundation model have been pivotal in predicting signals on the different downstream classification tasks,across domains such as energy and finance, emphasizing common signal properties [6, 8, 13, 19, 25, 30, 35, 60, 69]. These foundation models improve performance across tasks, reduce the need for labeled data, and enhance transferability across domains.

Recent research has utilized these models across various sensor types, employing self-supervised pretraining on large datasets to boost accuracy and generalizability [28, 45, 62]. By using contrastive learning to form positive and negative data pairs, these models effectively learn representations. This approach facilitates the transfer of learned insights across different domains, making it easier to develop scalable and efficient solutions in areas like healthcare, smart environments, and industrial monitoring, ultimately enhancing performance across a range of tasks and reducing the reliance on extensive labeled datasets [1, 45, 62].

Inspired by the success of foundation models in sensor data analysis, we aim to explore whether foundation model principles can be effectively adapted for customizable Human Activity Recognition (HAR) tasks. Conventional sensor model training relies on large homogeneous datasets with uniform sensor configurations. Yuan et al. [62] trained different foundation model instances for 30 Hz accelerometer-only data in different window settings with self-supervised learning on a large dataset containing 700,000 person-days of wearable data. Narayanswamy et al. [30] succeeded in pretraining the foundation model, LSM, with multimodal data contained in their self-collected large dataset (6.6 million hours). Zhang et al. [65] used corpus of well annotated motion sequences collected using motion capture camera and trained a foundation model, UniMTS, with synthesis motion sensor patterns.

Beyond extensive data collection efforts, existing training strategies and frameworks seldom consider the diverse nature of both inputs and outputs in HAR tasks. For example, a recent work, CrossHAR [14], proposed a hierarchical self-supervised pretraining framework that generalizes well across datasets, making it a potential foundation model solution. However, it was designed and validated only for the integration of accelerometer and gyroscope data, without addressing task variance across datasets. Other released pretrained models standardize the input configuration (such as input frequency or window size), leaving limited deployment options. For example, Yuan et al.[62] pretrained model instances to support different length of 3 channels sensory inputs (i.e., 5/10/30 seconds) in 30Hz. Zhang et al. [65] released their pretrained model, UniMTS, that supports 10-second 20Hz motion sensor inputs from 22 joints simultaneously. Considering the fact that HAR scenarios often involve heterogeneous data from various wearable devices, the limited options could also limit the broader adoption of the existing foundation/pretrained models.

Through this research, we seek to address these challenges by developing approaches that can handle diverse sensor inputs in customized configurations (such as different sampling rates, different lengths of time windows, and different preferences of sensors) while maintaining adaptability for specific HAR requirements (such as a customized set of activities), reducing the dependency on extensive datasets for each new application.

2.2 Various Human Activity Recognition Datasets

Traditional foundation model training imposes specific requirements on datasets. The landscape of human activity recognition (HAR) datasets is notably diverse, encompassing a broad spectrum of wearable devices, modalities, tasks, and scenarios.

Diversity in Modalities. Human activity recognition (HAR) datasets exhibit significant diversity in the devices and modalities employed. Wearable devices such as smartphones, smartwatches, and earphones are commonly used, equipped with sensors like accelerometers and gyroscopes [17, 24, 27, 38, 40, 41, 43, 57, 59].

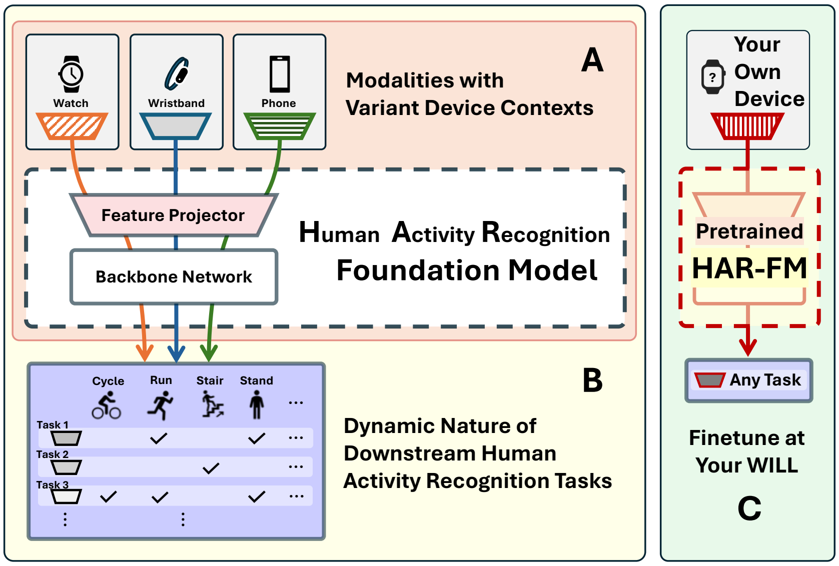


Fig. 1. Framework Overview: The adoption of the foundation model idea consists of pretraining and fine-tuning phases. First, to ensure effective pretraining of the foundation model for human activity recognition (HAR), we employ: (A) Self-supervised learning to capture patterns from modalities across diverse device contexts, and (B) Multi-task learning to enhance the model's capability for downstream recognition tasks. Then, the pretrained foundation model can be fine-tuned with task-specific data for practical applications as illustrated by (C).

Additionally, HAR systems incorporate vision-based [32], wireless-based [16, 21, 50, 61, 66, 67], and mobile-based [17, 24, 27, 38, 40, 41, 43, 55, 57, 59] methods.

Diversity in Scenarios. The scenarios covered by HAR datasets are diverse, ranging from general activity monitoring [2, 29, 36, 37, 39, 64] to specialized applications like fine-grained gesture control [26, 68]. This diversity also includes applications such as human activity recognition [17, 24, 40, 57, 59], user authentication [26, 34, 58, 59], and indoor tracking [18, 44].

Diversity in Devices and Body Locations. Data collection for activity recognition often involves using a variety of devices, each offering different sensors and capabilities. These devices, from brands like Samsung, LG, and Apple, are used in diverse settings and worn at different body locations [27, 38, 41, 43].

This diversity in datasets enables comprehensive exploration of Human Activity Recognition (HAR), which is essential for developing robust models. Such diversity is inherently unavoidable due to the complex nature of human activities and varying sensor placements in real-world scenarios. This intrinsic complexity and variability in HAR applications presents a significant challenge, as models trained on one type of dataset often perform suboptimally on others. To address this fundamental challenge, our study introduces an innovative approach using foundation models, enhancing generalization by imposing body placement constraints to ensure consistency across fragmented sensory data.

3 Framework: Foundation Model For Human Activity Recognition

The common pipeline for utilizing a foundation model involves pretraining and fine-tuning phases (as illustrated in Fig. 1). However, applying such an approach to human activity recognition (HAR) with wearable devices presents several challenges, particularly in pretraining with diverse and fragmented datasets. Specifically, existing open-source datasets for HAR rarely share identical data collection settings. Subjects may collect sensory data using various commercial off-the-shelf devices, such as smartwatches [41], smartphones [43], wristbands [7], or sensor arrays [9]. Additionally, researchers often take different perspectives on target recognition scenarios, resulting in variations in the types of human activities represented across individual datasets. The limited size of each dataset further exacerbates the difficulty of extracting transferable knowledge for new applications.

In response to these challenges, we propose an effective training strategy for the HAR foundation model. Particularly, the HAR foundation model consists of a feature projector that helps map the dataset-wise feature vectors to the shared pattern feature space, followed by a backbone network that contains layers of transformer [47] and helps further extracts context related features. As shown in Fig. 1, the training strategy consists of two stages (A and B). In the first stage (A), we implement quantization techniques (detailed in Section 4.1) to enable effective pattern modeling across diverse sources of unlabeled sensory data. In the second stage (B), we address the dynamic nature of downstream tasks by constructing learning objectives specific to each dataset and proposing task alignment solutions to bridge the gaps between datasets (detailed in Section 4.2). The deployment of the foundation model (C) supports full customization with one's own device and scenes related task definition (detailed in Section 4.3).

4 Towards HAR Foundation Model

To enable effective learning of human activity patterns from wearable devices, we propose a training framework that combines representation learning and task alignment. Specifically, we introduce a self-supervised learning solution to learn shared representations across different sources of sensory data. We then apply multi-task supervised learning to align the pretrained model with specific tasks, empowering the model with contextual knowledge.

4.1 Self-supervised Learning For Pattern Modeling

Self-supervised learning has proven successful in learning general data representations from unlabeled samples, particularly in Natural Language Processing and Computer Vision. Drawing inspiration from audio processing works [3, 10], we design a self-supervised learning framework incorporating a Quantizer module to discretize features into finite tokens. As shown in Fig. 2, our framework optimizes three objectives: (A.1) perplexity loss, (A.2) contrastive loss, and (A.3) masked prediction loss.

4.1.1 Quantizer Module. The Quantizer Module employs product quantization [20] to map input vectors to concatenated quantized vectors across multiple groups. As illustrated in Fig. 3, the module (a) maps input vectors to quantized indices through classification, (b) retrieves and concatenates vectors from the codebook using group indices, and (c) transforms the concatenated vectors through a linear projector to obtain the final quantized vector q. To enable end-to-end training, we introduce Gumbel softmax [15] instead of the non-differentiable $\arg\max$ operation for index selection.

The **(A.1) Perplexity Loss** promotes diversity in representations by maximizing the entropy of the averaged distribution $\mathbf{l}$ over codebook entries for each group across a batch:

$$\mathcal{L}_p = 1 - \frac{1}{GV} \sum_{g=1}^{G} \exp(-\sum_{v=1}^{V} p_{gv} \log p_{gv}) \tag{A.1}$$

where G and V denote the number of groups and entries respectively, and p_{gv} represents the probability of selecting the v-th entry for group g.

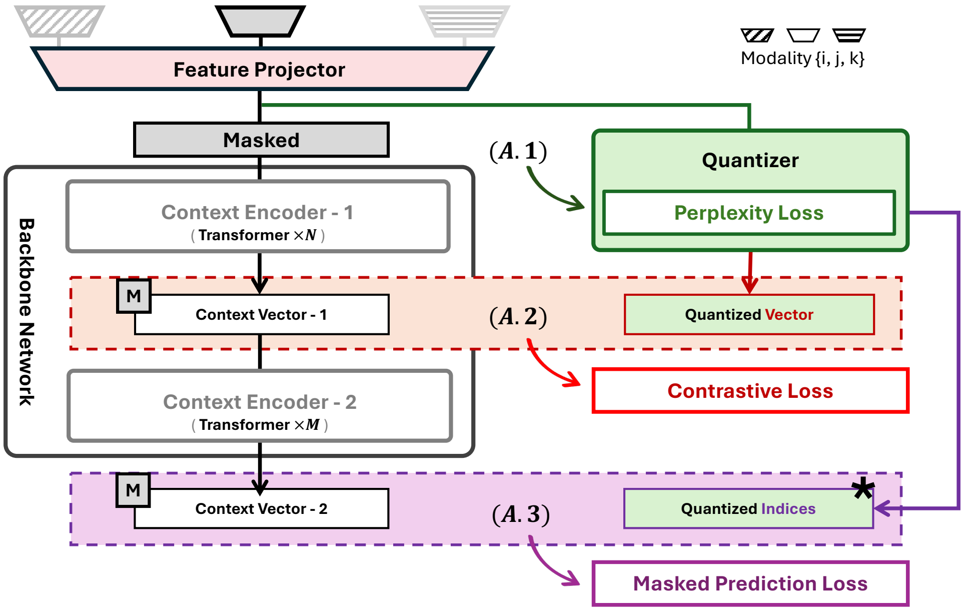
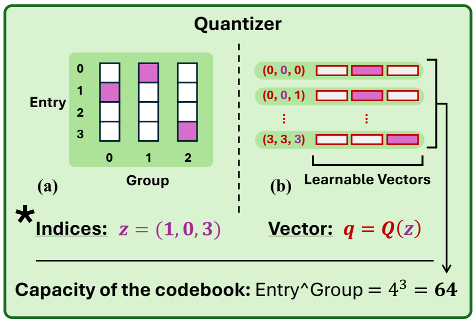


Fig. 2. Overview of the proposed framework's (A) self-supervised learning solutions that leverage enormous amounts of unlabeled data.



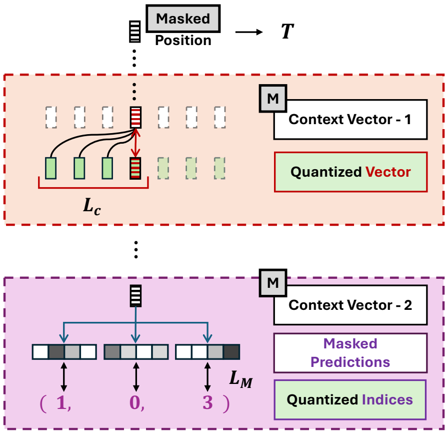


Fig. 3. Quantizer discretizes the input into a finite set of discriminative tokens. The perplexity loss (A.1) is proposed to promote the utilization of each entry in the codebook.

Fig. 4. Illustrations on contrastive loss (A.2, Upper) and masked prediction loss (A.3, Lower) computation.

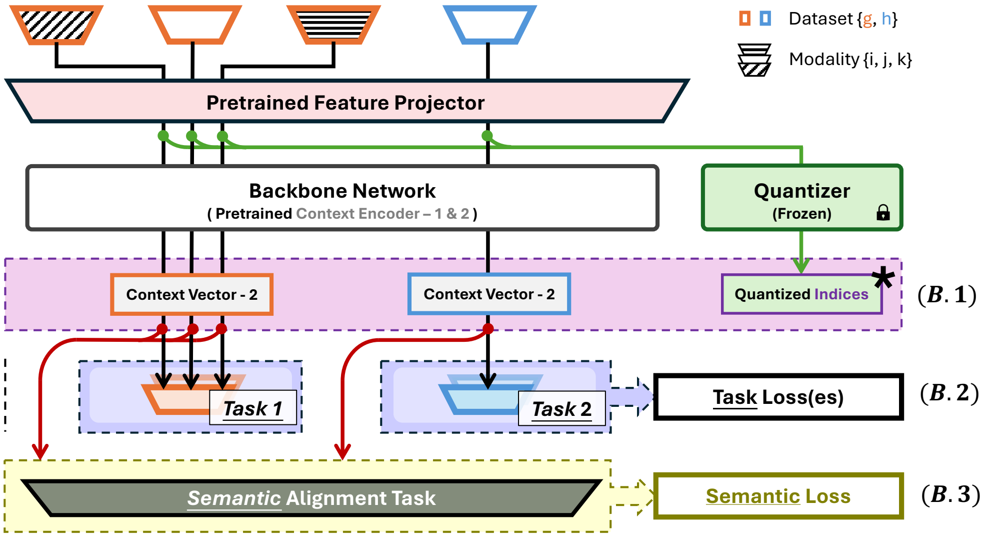


Fig. 5. Overview of (B) supervised learning phase with task alignment.

4.1.2 Self-supervised Learning. The foundation model pretraining involves context vectors from different Backbone Network layers, denoted as *Context Vector-1* and *Context Vector-2* (Fig. 2). Similar to the related works [3, 10, 57], the self-supervised learning relies on task constructing with masked operation.

The **(A.2) Contrastive Loss** requires *Context Vector-1* at masked time step t to identify its quantized vector $\mathbf{q}_t$ among K distractors $\mathbf{Q}^{(K)}$ (Fig. 4, Upper):

$$\mathcal{L}_c = -\log \frac{\exp(sim(\mathbf{c}_t^{(1)}, \mathbf{q}_t)/\kappa}{\sum_{\hat{\mathbf{q}}\sim\mathbf{Q}^{(K)}} \exp(sim(\mathbf{c}_t^{(1)}, \hat{\mathbf{q}}))/\kappa} \tag{A.2}$$

where $sim(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/||\mathbf{u}|| \cdot ||\mathbf{v}||$ computes cosine similarity and κ controls the similarity distribution.

The **(A.3) Masked Prediction Loss** $\mathcal{L}_m$ uses *Context Vector-2* to predict the quantized indices corresponding to the masked positions. A linear projector transforms these vectors into $G \times V$ logits, and we compute the average cross-entropy loss between grouped logits and corresponding indices (Fig. 4, Lower).

The final self-supervised learning objective combines these losses:

$$\mathcal{L}_u = \alpha\mathcal{L}_p + \beta\mathcal{L}_c + \gamma\mathcal{L}_m \tag{A.4}$$

where we set the $\alpha = 0.1$, $\beta = 1$ and $\gamma = 1$ empirically.

4.2 Multi-task Learning With Task Alignments

To address the diversity in sensory inputs and task settings, we propose strengthening the foundation model with semantic information through task alignment. As shown in Fig. 5, the model performs three types of tasks during this phase: **(B.1) Quantized Indices Prediction Task**, which preserves patterns learned from unlabeled data, **(B.2) Downstream Tasks**, which provide practical context information, and **(B.3) Semantic Alignment Task**, which learns universal representations shared across different contexts.

4.2.1 Quantized Indices Prediction retains pattern knowledge. The Quantizer Module introduced in self-supervised learning (Section 4.1.1) is retained, as it captures essential knowledge from the general representation learned from

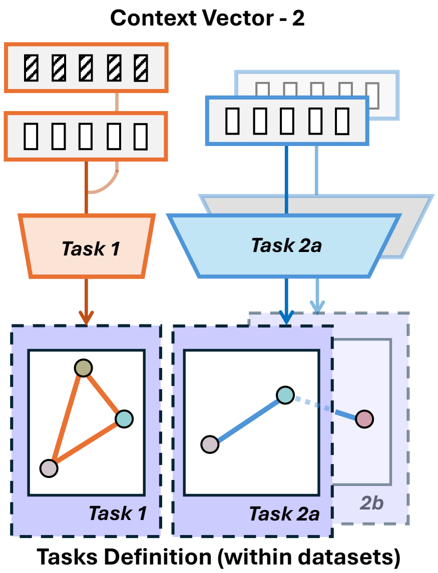


Fig. 6. Downstream Tasks **(B.2)**. The same task projector is shared across modalities (Left). The Context Vector-2 is input to the task projectors to complete the tasks defined on its datasets (Right).

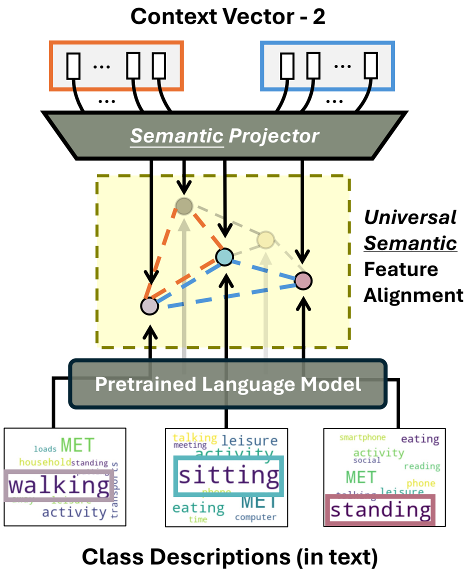


Fig. 7. Semantic Task **(B.3)**. The Context Vector-2 from different datasets is projected to the shared universal semantic feature space to build up the semantic relationship across datasets and tasks.

large-scale unlabeled data. In the supervised phase, we preserve context vector informativeness by performing quantized indices prediction for all time steps (without masking). This task minimizes the **Masked Prediction Loss** (first introduced as (A.3) in Section 4.1.2) on the entire sequence. To differentiate it in this context, we refer to it as the **(B.1) Quantized Indices Prediction Task Loss**, denoted as $\mathcal{L}_m^{(k)}$.

4.2.2 Downstream Task Alignment Reflects Practical Context Usage. In supervised learning, the construction of downstream tasks aligns with each dataset's intended purpose. This alignment ensures precise context information preservation from the dataset contributors' perspective. In practice, we implement a task-specific linear projector (Fig.6) that uses the *Context Vector-2* corresponding to the dataset and task as input to generate predictions. The weighted summary (detailed in Section4.3) of cross-entropy loss values computed between the predicted logits and discrete task labels is referred to as the **(B.2) Downstream Task Loss(es)**, denoted as$\mathcal{L}_d^{(k)}$.

4.2.3 Semantic Alignment bridges gaps across contexts. Downstream tasks often rely on small datasets, each covering only a subset of human activities. To address this limitation, we introduce a semantic alignment task to bridge gaps across different contexts and datasets. First, we collect class descriptions in text from datasets and extract their semantic embeddings using a pretrained language model (e.g., MPNet [42]). A semantic projector is then used to map *Context Vector-2* at each time step to the corresponding semantic embeddings. The learning objective minimizes the Mean Squared Error (MSE) between the projected output and the ground truth semantic embeddings. We refer to this loss as the **(B.3) Semantic Alignment Task Loss**, denoted as $\mathcal{L}_a^{(k)}$.

Table 1. Overview of the datasets used in this study. <$\mathcal{M}$> indicates the number of modalities (up to three: accelerometer, gyroscope, and magnetometer) included in each dataset. <**U**> represents the number of subjects. <**D**> denotes the maximum number of collections for the same modality from different devices, reflecting the diversity of data domains. <**L**> indicates the total time length (in seconds) accumulated from all modalities in the dataset. <*SR*> refers to the sampling rate (in Hertz) used as default for each dataset throughout this work. <$\mathcal{T}$> refers to the number of tasks corresponding to the datasets. Datasets with IDs ranging from 1 to 4 are used to train the foundation model, while the remaining datasets (IDs 5 to 9) are used to evaluate its performance (Neither labeled nor unlabeled data from these datasets was used in the foundation model's training, being used only for fine-tuning, see Section 5.1.2).

Pos.	ID	Dataset	Year	$\mathcal{M}$	**U**	**D**	**L**	*SR*	$\mathcal{T}$
Wrist	1	OPPTY [9]	2012	1	4	4	94K	30	3
	2	realdisp [5]	2012	3	17	2	836K	50	4
	3	WISDM [52]	2019	2	51	1	251K	30	4
	4	capture24 [7]	2020	1	151	1	14M	100	7
	5	PAMAP2 [37]	2012	3	9	2	154K	100	1
	6	mHealth [4]	2014	3	10	1	73K	50	1
	7*	shoaib [41]	2014	3	10	1	50K	50	1
	8	HHAR [43]	2015	2	9	4	119K	100	1
	9	GOTOV [31]	2020	1	35	1	173K	100	1
Upper Arm	1	OPPTY [9]	2012	3	4	6	282K	30	3
	2	realdisp [5]	2012	3	17	2	836K	50	5
	7*	shoaib [41]	2014	3	10	1	50K	50	1
Pocket	1	OPPTY [9]	2012	1	4	1	28K	30	2
	3	WISDM [52]	2019	2	51	1	296K	30	2
	7*	shoaib [41]	2014	3	10	2	101K	50	1

**Multi-task Supervised Learning.* In the supervised learning phase, the foundation model is trained to minimize the total loss, defined as:

$$\mathcal{L}^{(k)} = \mathcal{L}_m^{(k)} + \mathcal{L}_d^{(k)} + \mathcal{L}_a^{(k)} \tag{B.4}$$

4.3 Implementation Details

In this subsection, we outline the foundation model design and training procedure.

4.3.1 Datasets Preparation Based on Placements for Consistent Sensing Objectives. Different sensor modalities provide unique perspectives with distinct physical meanings. For example: Accelerometers measure changes in an object's velocity over time, providing critical data on movement and orientation. Gyroscopes measure angular velocity around an object's axes, indicating rotation in 3D space. Magnetometers detect magnetic fields and are often used to determine orientation relative to the Earth's magnetic poles, aiding navigation and compass functionality. Each sensor captures specific measurements that cannot be derived from others, making them complementary when used together. To ensure consistent measurement objectives and facilitate information sharing across modalities, we focus on sensor placements as prior knowledge. For instance, wearable sensors located on the wrists provide consistent and comparable data across devices.

Based on this principle, we trained foundation model for three placements and evaluated the corresponding performance with samples from these locations separately. The datasets overview is summarized in Table 1 with descriptions attached in Appendix A.

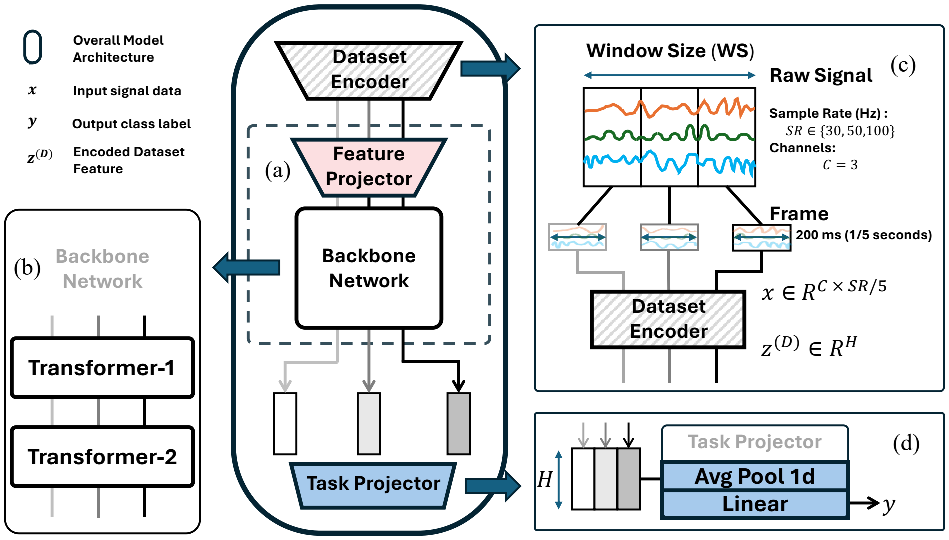


Fig. 8. Overall Model Architecture. (a) The foundation model consists of a feature projector, implemented as a two-layer linear network, and a backbone network. (b) The backbone network comprises two stacked Transformer blocks. (c) The model processes raw signal frames as input, where dimensions are determined by the dataset's sampling rate (SR). The dataset encoder transforms these dataset-specific inputs into feature representations of dimension H. (d) The Task Projector applies average pooling followed by a single-layer linear projection to produce the final output.

4.3.2 Effectively Learning from Imbalanced Data. Imbalanced data is a well-known challenge in the research community, often exacerbated by the sensitivity of sensor data, which limits public sharing, or by the novelty of sensing techniques, which restricts widespread deployment. To mitigate biased learning and improve the generalization capability of the foundation model, we design a compact training solution. Data from multiple sources is combined into a single batch, with weights assigned proportional to the data volume of each source. The overall loss is then computed as a weighted sum of the losses from individual sources, $\mathcal{L} = \sum_i \frac{|D_i|}{\sum_j |D_j|} \mathcal{L}_i$, where $|D_i|$ represents the number of batches provided by source i, and $\mathcal{L}_i$ is the loss value computed using a single batch from source i.

To further improve the model's robustness, we adopt feature augmentation techniques throughout training. First, we integrate a dense sampling approach [56] that shifts signal data by random offsets in each iteration, ensuring robustness to temporal shifts. Additionally, we generate multiple views of the same input signal by duplicating the feature vectors produced by the dataset-specific encoder and applying different dropout masks to each copy.

4.3.3 Training the Foundation Model. In this section, we clarify the details of the foundation model evaluated in this paper, including its model parameters and training hyperparameters to facilitate reproduction.

Model Architecture. The foundation model is designed to accommodate the customization requirements with variant sensor inputs though, it is necessary to standardize the input sensory signals. Particularly, we assign an encoder for each modality in each dataset, termed dataset-wise feature encoder $f^{(D)}$, which learns from the sliced time-series snippets (as we set a shared window length of around 200 ms) and map them into the shallow feature space $z^{(D)} \in R^{H}$, where the H is the hidden dimension. For an input setting that contains time series longer than 200 ms, the input will be first sliced into small and non-overlapped standard windows (e.g., a sequence of

Table 2. Model Architectures (with hidden dimensions $H = 96$, codebook group number $G = 6$, entry number $V = 4$, and learnable vector dimensions $vq\_dim = 96$)

Component	**Structure**	**Dim. / Params.**
Foundation Model		
(a) Feature Projector	Two-layer Fw Block	$H \rightarrow 144 \rightarrow H$
(b) Backbone Network		
Transformer-1	Number of Layers Attention Heads Feed-Forward Dim.	10 8 256
Transformer-2	Number of Layers Attention Heads Feed-Forward Dim.	24 4 128
Quantizer Module		
Indices Projector	Linear Layer	$H \rightarrow G \times V$
Vector Projector	Two-layer Fw Block	$vq\_dim \rightarrow 64 \rightarrow H$
Semantic Alignment Task Module		
Semantic Embedding Projector	Two-layer Fw Block	$H \rightarrow 1024 \rightarrow 768$
Quantized Indices Prediction Task Module		
Quantized Indices Projector	Two-layer Fw Block	$H \rightarrow 144 \rightarrow G \times V$
(c) Dataset Encoder		
For Training.	Two-layer Fw Block	$R^{3\times(SR/5)} \rightarrow 144 \rightarrow H$
*For Fine-tuning.	CNN Block + Linear Layer	$R^{3\times(SR/5)} \rightarrow ... \rightarrow H$
**It refers to the model architecture we use in the evaluation phase.*		

3 windows for 600 ms input). It allows us to acknowledge the difference among collection settings of different datasets. Particular, we just adopt the default sample rate originate from the datasets.

As depicted in the framework (Fig. 1), the foundation model consists of a general feature projector and a transformer-based backbone network, which enables the implementation of pattern modeling (Self-supervised Learning) and contextual feature extraction (multi-task supervised learning). The quantizer is an additional auxiliary module that gets trained during pattern modeling and helps preserve pattern structure during the multi-task supervised learning. To obtain the downstream task predictions, the output of the Foundation Model (a sequence of context vectors with dimension size of H) is first passed through an average pooling layer and then projected to the task output space with a linear projector. The parameters of these related models in use are as summarized in Table 2.

General Training Configurations. During the training, we primarily adopt the shared input window size setting of 6 seconds, that is 30 frames for each input sequence, for each modality from different datasets. Both the Self-supervised Learning and multi-task supervised learning adopt the Adam optimizer with a linear warmup and exponential decay training strategy (with different settings as summarized in Table 3. All experiments were conducted on an NVIDIA A40 GPU with 48GB of VRAM.

Self-supervised Learning. We adopt a span masking operation with a probability of 0.2, a maximum span length of 3 frames, and a total of 10 masked positions out of 30 frames. This operation randomly selects masked positions and replaces the masked input with a learnable noise vector. The temperature for optimization with Gumbel-Softmax is set to decay exponentially from 2 to 0.5 at a rate of 0.999993 per iteration step. The batch size is set to 1024, with each batch containing samples uniformly from all available datasets.

Table 3. Optimization Details for Self-supervised and Supervised Learning Phases.

Phase	Learning Rate Initial/Min.	 Max.	Warmup Steps	Decay Rate	Total Steps
Self-supervised Learning	1e-5	1e-4	20k	0.99998	200k
Multi-task Supervised Learning	1e-4	5e-4	5k	0.99993	50k $\times R$

R: refers to number of rounds, 2 for Wrist, 1 for Upper Arm, and 1 for POCKET.

Multi-task Supervised Learning. We construct the training data sources with pairs of modality and task. For example, if there are 3 modalities with 4 task definitions from the dataset, the related number of data sources is 12. Same modality from the same dataset share the same encoder, and different modalities of the same task will share the same task projector. We generate the semantic embedding using MPNet [42], an open-source pretrained language model, for each activity with the formatted sentence:

The status or/and activity of the subject is: [ACTIVITY].

where the *[ACTIVITY]* refers to the activity descriptions involved in the datasets. In this training stage, we adopt a multiple-round training strategy. Specifically, we reset the dataset encoder with random initialized parameters to simulate new training tasks for each round. The parameters of the quantizer is frozen throughout the training process. The batch size is set to 2048 for training with wrist, and 1536 for the others (pocket and upper arm).

4.3.4 Deployment with Foundation Model. The easy and effective deployment capability of the foundation model is one of its key characteristics, contributing to its widespread adoption. Our foundation model could be directly deployed with simple fine-tuning pipeline and equipped with zero-shot pipeline listed as follows.

Fine-tuning. We adopt a convolutional neural network (CNN) architecture as the dataset encoder for downstream task deployment and evaluation. Dataset encoders corresponding to data sources with the same sampling rate share the same architecture. The foundation model is jointly trained with the dataset encoder and task projector during the fine-tuning stage, with the batch size set to 64, the maximum number of training epochs set to 1,000, and an early stopping mechanism with a patience of 100 epochs. We use the Adam optimizer with a learning rate of 1e-4 and a weight decay of 1e-6. Dense sampling augmentation [56] is also applied during fine-tuning for all groups of experiments. This augmentation approach is both intuitive and straightforward to implement during the data preprocessing phase. **In this study, we use the same encoder structure for different modalities and different datasets. And therefore the finetuning results should demonstrate the general adaptability of the HAR-FM.**

5 Evaluation

In general, we try to answer the following research questions in the evaluation section, **RQ1**: how the does the foundation model help to the HAR community? **RQ2**: how does the foundation model work? **RQ3**: if there is any generalizability of the proposed training techniques and expirence?

5.1 Experimental Settings

To validate these questions, we first clarify our experimental settings here.

5.1.1 Downstream Datasets Settings. We prepared five datasets (Dataset No. 5–9 in Table 1, which were not used in training the foundation model) to simulate downstream tasks and evaluate the performance of the foundation model. Each dataset includes modalities commonly encountered in practical real-world implementations. To better assess the foundation model's performance in terms of fairness, we simulate the independent deployment of each modality (Accelerometer, Gyroscope, and Magnetometer). Furthermore, we account for the diversity in input customization by preserving the original sampling rate settings of each dataset.

Regarding the downstream task settings, we retain the original definitions based on the purpose of each dataset's construction. This ensures that the evaluation aligns with the customization needs typically required in practice. Detailed information about the datasets is provided in Appendix A.

5.1.2 Cross-User Validation Pipeline. Throughout the experiments, we adopt a 5-fold cross-user validation setting. Specifically, we divide the users in each dataset into five independent folds in advance. For each iteration of the experiments, one fold is selected as the testing dataset, while the remaining four folds are used as the training dataset. The training dataset is further split into 80% for training and 20% for validation to fine-tune the model.

We record the final accuracy on the testing data for each iteration and report the average accuracy across all five iterations. The proposed 5-fold cross-user validation setting closely resembles the practical downstream deployment pipeline of related techniques, making it a reliable reflection of real-world performance.

5.1.3 Baselines. We further clarify the baselines related to the Foundation Model for Human Activity Recognition as follows.

- **HAR-FM (Ours).** It refers to the model gets full trained with the proposed training strategy.
- **TRAIN-FROM-SCRATCH (or T.F.S).** It refers to the fine-tuning the model with the same structure used in the HAR-FM but with random parameters. The fine-tuning of T.F.S shares the same configuration as HAR-FM (Section 4.3.4).

In response to the research questions, we reimplemented several state-of-the-art baselines with different insights.

- **TSFCN [51].** Time-series Full Convolutional Network has the simplest model structure while achieves extraordinary performance in the time-series related tasks. The original paper claims to be a strong baseline with only a few hyper-parameters tuning. It can be considered as one important baseline that does not need pretraining.
- **TinyHAR [71].** It refers to one of the state-of-the-art lightweight deep learning model designs for HAR applications. It is designed following the proposed guidelines for efficient on-device inference which involve careful considerations in temporal as well as modalities dimensions. In our implementation, we set the filter number F as 20 for all settings.
- **MLP-HAR [70].** It refers to the another lightweight deep learning model that adopts purely fully connected layers for efficiency of HAR models on edge devices. We reimplement the MLP-HAR following the settings that produced best performance mentioned in [70] experiments, fixing number of mixer modules N at 3, filter number d at 6, length of intervals τ at 16 and number of frequency bins f at 16 as well.
- **DDLEARN [33].** DDLEARN aims to improve the generalization of the tuned model through adopting Multi-task Learning solutions. Precisely, the authors propose three auxiliary tasks along with several data augmentation methods to help the model learn domain invariant features that improves the model performance on unseen users. In our implementation, we adopt the same group of hyper-parameters for all settings as recommended in the paper.
- **LIMU-BERT [57].** LIMU-BERT is a foundation model training framework that proposes to make use of the ever common unlabeled data with masked language modeling solution. In our paper, we use it as the baseline representing the pretraining methods in the ever critical cross-user setting with limited while diverse data. We preprocessed each modality data following the instructions in the paper. For each iteration, we first pretrained LIMU-BERT backbone with all training data including those without labels (The model is first trained with 3500 epochs for each fold experiments, and the version with minimum validation loss is selected). The pretrained backbone was later fine-tuned along with the GRU classifier for 700 epochs.

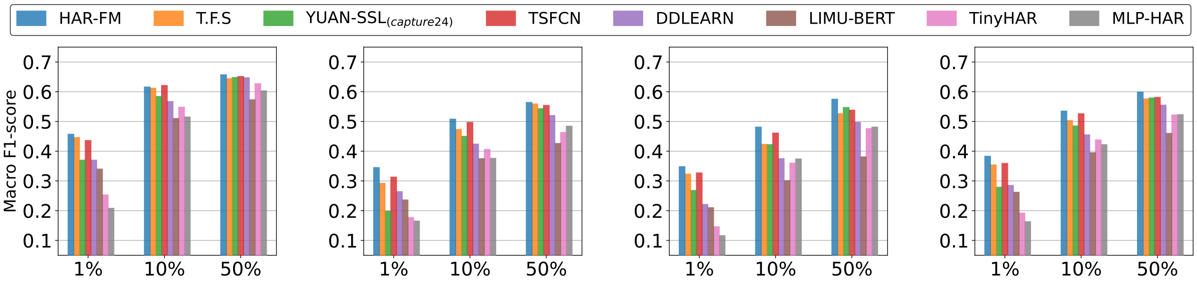
In the evaluation section, we also choose an open-source pretrained model instance to inspect the performance obtained using homogeneous data under comparable dataset conditions[2].

- **YUAN-SSL (capture24) [62].** Yuan et al.[62] contributes to the community with models pretrained with large-scale dataset. In section 5, we primarily compare the performance with one of their released model that were pretrained with the largest wearable dataset, capture24 (Dataset No.4), that is involved in our foundation model's training. The comparison should demonstrate the benefits in making use of multi-modalities as well as the labels. To match the pretrained model's input shape requirement of 10 seconds (at a 20 Hz sampling rate), we zero-pad shorter signal segments (e.g., 1s, 3s, or 5s).

[2] While we compare a YUAN-SSL instance trained on comparable datasets in this section, broader comparisons (e.g., YUAN-SSL[62] at 700k person-days and UniMTS[65]) are reserved for Section 6 due to their larger/higher-quality training data.

Table 4. Overall Performance with Varying Window Sizes (WS). The left table presents detailed performance metrics across different datasets, grouped by modalities (Accelerometer, Gyroscope, and Magnetometer). The right table shows metrics averaged across datasets for each modality (groups A, G, and M) and across all experiments (group ALL). For clearer comparisons in subsequent experiments, we report the averaged results by default.

			Details												Average			
WS	BASELINES	METRICS	Accelerometer					Gyroscope				Magnetometer			A.	G.	M.	ALL
			5	6	7*	8	9	5	6	7*	8	5	6	7*	(5) +	(4) +	(3) =	(12)
1s	TSFCN	Acc	0.533	0.797	**0.921**	0.566	**0.716**	0.479	0.537	0.831	0.573	0.379	0.657	0.769	**0.707**	0.605	0.602	0.638
		F1	**0.418**	0.797	**0.918**	0.538	**0.691**	0.388	0.493	0.822	0.552	0.278	0.642	0.763	**0.672**	0.564	0.561	0.599
	DDLEARN	Acc	**0.544**	0.748	0.901	0.639	0.690	0.408	0.531	0.791	0.524	0.380	0.629	0.735	0.704	0.564	0.581	0.616
		F1	0.405	0.747	0.899	0.604	0.621	0.323	0.489	0.786	0.511	0.271	0.608	0.731	0.655	0.527	0.537	0.573
	LIMU-BERT	Acc	0.456	0.692	0.789	0.462	0.653	0.335	0.369	0.708	0.497	0.256	0.540	0.523	0.610	0.477	0.440	0.509
		F1	0.347	0.684	0.779	0.435	0.613	0.265	0.313	0.700	0.483	0.173	0.533	0.514	0.572	0.440	0.407	0.473
	TinyHAR	Acc	0.533	0.705	0.839	**0.662**	0.679	0.427	0.430	0.736	0.590	0.383	0.581	0.672	0.684	0.546	0.545	0.592
		F1	0.408	0.700	0.834	**0.627**	0.630	0.326	0.376	0.728	**0.583**	0.260	0.552	0.673	0.640	0.503	0.495	0.546
	MLP-HAR	Acc	0.500	0.682	0.840	0.585	0.659	0.413	0.507	0.748	0.556	0.375	0.514	0.774	0.653	0.556	0.554	0.588
		F1	0.388	0.681	0.836	0.563	0.609	0.323	0.458	0.742	0.546	0.272	0.492	0.771	0.615	0.517	0.512	0.548
	TRAIN-FROM-SCRATCH	Acc	0.496	0.795	0.890	0.615	0.712	**0.488**	0.546	**0.839**	0.606	0.365	0.717	0.751	0.702	0.620	0.611	0.644
		F1	0.395	0.797	0.881	0.589	0.686	**0.392**	0.508	**0.830**	**0.583**	0.275	0.718	0.748	0.670	0.578	0.580	0.609
	YUAN-SSL (capture24)	Acc	0.512	0.783	0.891	0.603	0.695	0.431	0.588	0.815	0.590	0.377	0.706	0.725	0.697	0.606	0.603	0.635
		F1	0.403	0.787	0.887	0.563	0.663	0.343	0.547	0.807	0.566	0.283	0.708	0.723	0.661	0.566	0.571	0.599
	HAR-FM (Ours)	Acc	0.488	**0.828**	0.911	0.583	0.712	0.479	**0.630**	0.822	**0.610**	**0.396**	**0.729**	**0.793**	0.704	**0.635**	**0.639**	**0.660**
		F1	0.384	**0.830**	0.909	0.555	0.677	0.383	**0.595**	0.817	0.581	**0.299**	**0.737**	**0.790**	0.671	**0.594**	**0.609**	**0.625**
3s	TSFCN	Acc	0.570	0.837	**0.940**	0.583	0.740	**0.513**	0.573	**0.877**	0.576	0.424	0.727	0.826	0.734	0.635	0.659	0.676
		F1	0.435	0.834	0.933	0.542	**0.725**	**0.419**	0.520	**0.871**	0.540	0.321	0.721	0.821	0.694	0.588	0.621	0.634
	DDLEARN	Acc	0.533	0.771	0.932	0.702	0.692	0.394	0.541	0.781	0.502	0.347	0.624	0.772	0.726	0.554	0.581	0.620
		F1	0.405	0.772	0.931	**0.678**	0.629	0.320	0.490	0.772	0.459	0.253	0.596	0.764	0.683	0.510	0.538	0.577
	LIMU-BERT	Acc	0.500	0.758	0.817	0.545	0.710	0.371	0.431	0.750	0.535	0.267	0.643	0.553	0.666	0.522	0.488	0.558
		F1	0.378	0.753	0.808	0.517	0.670	0.295	0.381	0.737	0.517	0.181	0.638	0.543	0.625	0.483	0.454	0.521
	TinyHAR	Acc	0.562	0.699	0.838	0.690	0.695	0.452	0.518	0.780	0.615	0.394	0.641	0.774	0.697	0.591	0.603	0.630
		F1	0.424	0.687	0.828	0.638	0.638	0.357	0.465	0.773	0.594	0.270	0.615	0.773	0.643	0.547	0.553	0.581
	MLP-HAR	Acc	0.510	0.722	0.836	0.612	0.692	0.369	0.528	0.735	0.549	0.373	0.540	0.763	0.674	0.545	0.559	0.593
		F1	0.388	0.716	0.830	0.555	0.624	0.291	0.481	0.730	0.529	0.271	0.525	0.764	0.623	0.508	0.520	0.550
	TRAIN-FROM-SCRATCH	Acc	0.530	0.854	0.893	0.599	0.744	0.480	0.535	0.834	0.614	0.382	0.743	0.799	0.724	0.616	0.641	0.660
		F1	0.417	0.851	0.885	0.556	0.715	0.393	0.481	0.824	0.578	0.286	0.744	0.794	0.685	0.569	0.608	0.621
	YUAN-SSL (capture24)	Acc	**0.582**	0.846	0.939	**0.713**	0.745	0.472	**0.642**	0.830	0.640	**0.441**	0.769	0.826	**0.765**	0.646	0.679	**0.697**
		F1	**0.456**	0.839	**0.936**	0.677	0.715	0.368	**0.606**	0.819	0.605	**0.336**	**0.770**	0.821	0.725	0.599	0.642	0.655
	HAR-FM (Ours)	Acc	0.569	**0.884**	0.929	0.680	**0.752**	0.465	0.640	0.843	**0.647**	0.426	**0.773**	**0.843**	0.763	**0.649**	**0.681**	**0.697**
		F1	0.445	**0.883**	0.923	0.664	0.721	0.374	0.599	0.835	**0.610**	0.321	**0.770**	**0.841**	**0.727**	**0.604**	**0.644**	**0.659**
5s	TSFCN	Acc	0.567	0.833	**0.962**	0.629	0.751	**0.534**	0.605	**0.872**	0.635	0.415	0.729	0.826	0.748	0.662	0.657	0.689
		F1	0.444	0.835	**0.961**	0.616	0.741	**0.439**	0.565	**0.860**	0.575	0.314	0.721	0.815	0.719	0.610	0.617	0.649
	DDLEARN	Acc	0.491	0.761	0.900	0.696	0.711	0.352	0.543	0.787	0.456	0.373	0.621	0.786	0.712	0.534	0.593	0.613
		F1	0.377	0.750	0.896	0.655	0.632	0.289	0.499	0.775	0.373	0.278	0.589	0.781	0.662	0.484	0.549	0.565
	LIMU-BERT	Acc	0.522	0.788	0.844	0.539	0.705	0.370	0.428	0.771	0.543	0.290	0.653	0.567	0.680	0.528	0.503	0.570
		F1	0.386	0.772	0.842	0.493	0.657	0.298	0.386	0.759	0.500	0.194	0.636	0.563	0.630	0.486	0.464	0.527
	TinyHAR	Acc	0.545	0.543	0.819	0.695	0.711	0.386	0.466	0.783	0.597	0.392	0.616	0.757	0.663	0.558	0.588	0.603
		F1	0.400	0.457	0.810	0.637	0.629	0.281	0.408	0.762	0.533	0.276	0.569	0.755	0.587	0.496	0.533	0.539
	MLP-HAR	Acc	0.505	0.634	0.829	0.660	0.709	0.355	0.492	0.755	0.574	0.335	0.536	0.778	0.667	0.544	0.550	0.587
		F1	0.386	0.618	0.823	0.624	0.630	0.269	0.448	0.749	0.526	0.245	0.489	0.774	0.616	0.498	0.503	0.539
	TRAIN-FROM-SCRATCH	Acc	0.566	0.851	0.946	0.646	0.763	0.525	0.571	0.867	0.590	0.419	0.752	0.780	0.754	0.638	0.650	0.681
		F1	0.467	0.846	0.946	0.616	0.744	0.415	0.527	0.858	0.569	0.319	0.748	0.772	0.724	0.592	0.613	0.643
	YUAN-SSL (capture24)	Acc	**0.611**	0.842	0.921	**0.754**	**0.771**	0.467	0.611	0.834	0.667	**0.471**	0.769	0.862	**0.780**	0.645	0.701	0.708
		F1	**0.484**	0.837	0.919	**0.690**	**0.748**	0.377	0.578	0.832	0.610	**0.350**	0.763	0.855	**0.736**	0.599	0.656	0.664
	HAR-FM (Ours)	Acc	0.585	**0.876**	0.936	0.712	0.761	0.497	**0.646**	0.851	**0.689**	0.467	**0.807**	**0.873**	0.774	**0.671**	**0.716**	**0.720**
		F1	0.459	**0.871**	0.930	0.676	0.734	0.394	**0.606**	0.844	**0.616**	**0.350**	**0.806**	**0.870**	0.734	**0.615**	**0.675**	**0.675**



(a) Acc. (b) Gyro. (c) Magneto. (d) ALL

Fig. 9. Influence by Data Volume (WS=1s).

5.2 Overall Performance (RQ1)

In response to the first research question upon the effectiveness of the foundation model, we evaluate the overall performance with variant inputs (i.e., different length of frame as inputs and variant modalities) and variant data settings (i.e., different volume and subjects involved).

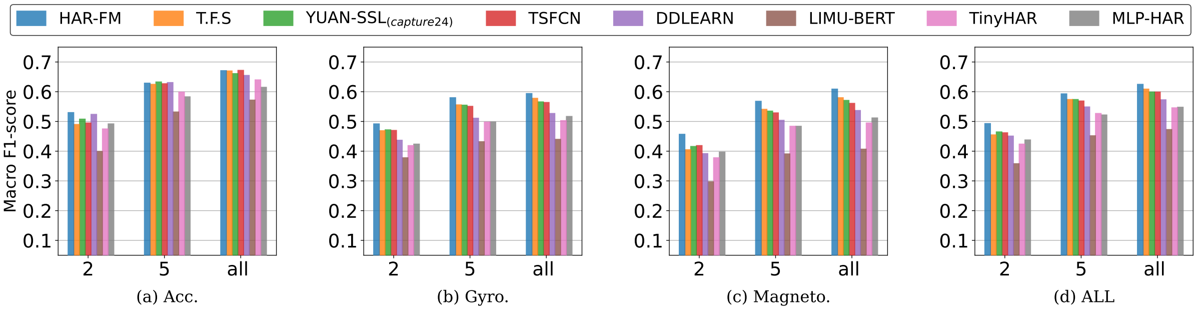
5.2.1 Flexible Input: Variant Window Size and Modalities. Table 4 presents the overall performance across different input settings. In general, **HAR-FM** achieves the highest averaged metrics in the majority of groups (top accuracy on 9 and top Macro F1-score on 10 out of 12 groups in the Average panel). The model instance, **YUAN-SSL(capture24)**, which was pretrained on the accelerometer dataset, present its superiority on accelerometer experiments when the size of window gets larger. It is also interesting to observe the generalizability of **YUAN-SSL(capture24)** on the other two modalities (under 3-second and 5-second window size settings), which indicates the potential information sharing related to the placement.

Among the baseline methods, while **TSFCN** and **TRAIN-FROM-SCRATCH** demonstrate competitive performance, the remaining four baselines exhibit significant limitations due to various shortcomings. **DDLEARN**, which augments the fine-tuning process, achieves accuracy comparable to **HAR-FM** in the accelerometer group under the 1-second window size setting; however, this performance fails to generalize across other settings. The **TinyHAR** and **MLP-HAR** could trade accuracy for computational efficiency, resulting in consistently lower performance. The pretraining solution **LIMU-BERT** fails to achieve high performance in any experimental group, seemingly suffering from over-fitting to the training data that consists of users with limited similarity to the test set.

It is worth noting that comparing absolute metric values across different time window size settings may be inappropriate, as smaller windows can contain several times more samples than larger ones. Nevertheless, within the time window settings, **HAR-FM** demonstrates clear superiority in the magnetic modality across all settings, further highlighting the effectiveness of the foundation model training framework even when faced with extremely imbalanced data distributions (Section 4.3.2).

5.2.2 Reducing Effort in Downstream Data Collection. The fine-tuning performance with a limited amount of labeled data is a key indicator of the practical effectiveness enabled by the Foundation Model. To investigate this, we simulate experiments by controlling two key factors: Data Volume (in percentage) and Number of Subjects, focusing on the 1-second window size setting.

Data Volume (%). In our experiments, the data is trimmed in chronological order to simulate real-world scenarios, thereby reducing the effort required from contributors. Fig. 9 illustrates the impact of data volume (ranging from 1% to 50%). Overall, **HAR-FM** demonstrates superior performance compared to the baselines across all settings. The **TSFCN** was tailored to benefit various time-series datasets, which makes it a strong



(a) Acc.
(b) Gyro.
(c) Magneto.
(d) ALL

Fig. 10. Influence by Number of Subjects (WS=1s).

baseline challenging the effectiveness of the pretrained model in resource-constraint scenarios as well. The steep trend between experiment groups suggests the reliance of the system performance on the large data volume. The substantial trends of **TinyHAR** and **MLP-HAR** suggests their higher reliance on the large data volume for better performance. Notably, in accelerometer modality (Fig. 9(a)), the steeper performance curve of **YUAN-SSL(capture24)** indicates higher customization effort than **HAR-FM**, further suggesting the benefit in training with data from variant devices.

Subjects Number. The number of subjects represents scenarios where only limited participants are available for data collection. We simulate this by trimming the number of subjects in the training dataset for each iteration of validation (see 5-fold cross-user validation setting, Section 5.1.2). Fig. 10 illustrates the impact of reducing the number of subjects. Similarly, **HAR-FM** surpasses other baselines in most experimental settings. Notably, in the accelerometer group, the performance gap among all baselines (except for **LIMU-BERT**, **TinyHAR** and **MLP-HAR**) diminishes as the number of subjects increases. This suggests that the accelerometer data's richer and lower-variance nature, combined with increased subject diversity, enables all models to achieve better performance, thus narrowing the performance gap.

These trends presented in the experiments collectively underscore **HAR-FM**'s advantage in balancing performance with minimal contributor effort, even under extreme data reduction in either volume or number of participants.

5.3 Ablation Study and Visualization (RQ2)

In response to the research question on how the foundation model works, we conduct an ablation study by creating groups in which certain components of the proposed training framework are removed. Additionally, we interpret the effectiveness of the proposed training framework by visualizing changes in data distributions across different modalities and datasets using t-SNE [46].

5.3.1 Ablation Study Settings and Results. The **HAR-FM** is trained using a pipeline consisting of two stages: self-supervised learning for pattern modeling and multi-task supervised learning with task alignments. These stages are referred to as **HAR-SSL** and **HAR-FM**, respectively. To evaluate the effectiveness of the proposed foundation model training, we use the **TRAIN-FROM-SCRATCH** group as a key baseline, representing the lower bound of training effectiveness. In addition to the full-stack version of our framework, we compare groups that exclude individual components, as summarized in Table 5. Groups within the same stage share the same hyperparameters set (e.g., the same number of training iterations, the loss weights if any).

During the Self-supervised Learning (SSL) stage, the version without the masked prediction loss (**-noMP**) fails to effectively model the patterns, leading to performance that is even worse than the baseline **TRAIN-FROM-SCRATCH**.

In the Multi-task Supervised Learning stage, we first compare two groups: one without pretrained parameters (**-noSSL**) and another without the Quantized Indices Prediction Task Loss (**-noQIPT**). Both groups share the same optimization objectives (i.e., Downstream Task Losses and Semantic Alignment Task Loss). The results clearly demonstrate the benefits of pretrained parameters, with 1.2% and 0.8% improvements in accuracy on the ALL groups for the 1-second and 5-second window size settings, respectively.

When comparing the full-stack version (**Ours**) with other ablation groups using pretrained models (i.e., **-noQIPT**, **-noTask**, and **-noSA**), we observe performance degradation in the **-noTask** group, emphasizing the significance of incorporating downstream task alignment, which facilitates the model's adaptation to real-world applications.

Although both the semantic alignment and quantized indices prediction tasks aim to improve frame-level representations—through discrete codebook learning in the SSL stage and pretrained language model knowledge, respectively—the semantic embeddings demonstrate superior performance. This is evidenced by **-noQIPT**

Table 5. Definition of experimental groups used in the ablation study of HAR-SSL and HAR-FM components (Section 5.3.1).

HAR-SSL	
-noMP	Replaces **(A.3) the Masked Prediction Loss**. The new training solution uses the Context Vector-2 to compute the Contrastive Loss instead (described in Section 4.1) in a BERT[57]-like fashion.
HAR-FM	
-noSSL	Training without parameters pretrained during the **HAR-SSL** stage. Considering that the quantizer is only initialized during the self-supervised learning stage, the related **(B.1) Quantized Indices Prediction Task Loss** is excluded as well (described in Section 4.2.1).
-noQIPT	Excludes the **(B.1) Quantized Indices Prediction Task Loss** but uses pretrained parameters (compared to the previous group -noSSL).
-noTask	Excludes the **(B.2) Downstream Task Losses** (described in Section 4.2.2).
-noSA	Excludes the **(B.3) Semantic Alignment Task Loss** (described in Section 4.2.3).

Table 6. Results of ablation study (Section 5.3.1).

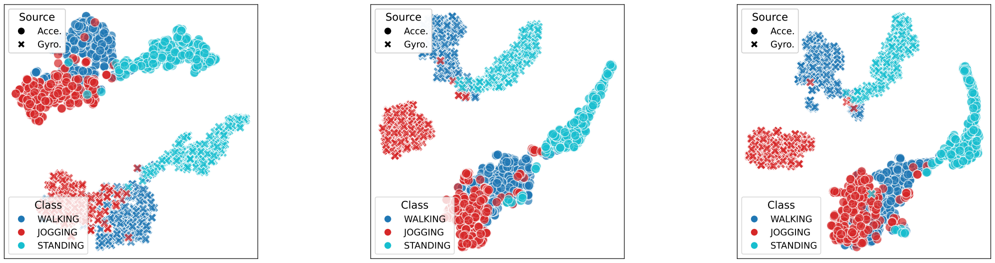
		WS	**1s**				**5s**			
		METRICS	A.(5)	G.(4)	M.(3)	ALL(12)	A.(5)	G.(4)	M.(3)	ALL(12)
TRAIN-FROM-SCRATCH		Acc	0.702	0.620	0.611	0.644	0.755	0.638	0.650	0.681
		F1	0.670	0.578	0.581	0.610	0.724	0.592	0.613	0.643
HAR-SSL	-noMP	Acc	0.675	0.583	0.588	0.615	0.705	0.614	0.636	0.652
		F1	0.635	0.541	0.543	0.573	0.668	0.554	0.587	0.603
	Ours	Acc	0.695	0.621	0.612	0.643	0.736	0.652	0.682	0.690
		F1	0.662	0.580	0.583	0.609	0.705	0.598	0.645	0.649
HAR-FM	-noSSL	Acc	0.699	0.603	0.613	0.638	0.759	0.650	0.688	0.699
		F1	0.659	0.561	0.576	0.599	0.723	0.597	0.646	0.656
	-noQIPT	Acc	0.693	0.619	0.637	0.650	0.768	0.656	0.697	0.707
		F1	0.656	0.576	0.602	0.611	0.731	0.613	0.657	0.667
	-noTask	Acc	**0.709**	0.613	0.603	0.642	0.741	0.648	0.647	0.679
		F1	**0.676**	0.570	0.574	0.607	0.705	0.600	0.609	0.638
	-noSA	Acc	0.698	0.607	0.619	0.642	0.759	0.656	0.699	0.705
		F1	0.664	0.570	0.582	0.605	0.723	0.612	0.655	0.663
	Ours	Acc	0.704	**0.635**	**0.640**	**0.660**	**0.774**	**0.671**	**0.716**	**0.720**
		F1	0.671	**0.594**	**0.609**	**0.625**	**0.734**	**0.615**	**0.675**	**0.675**

achieving higher metrics than **-noSA**, attributable to semantic embeddings' capability to capture broader and more precise global relationships.

5.3.2 Visualization of Context Feature Evolution. To further analyze the effectiveness of the foundation model, we visualize the evolution of the context features (referred to as Context Vector-2, as illustrated in Fig. 5). We prepare two groups of validation data used during the foundation model training and visualize the distributions of the context feature outputs at different training stages (Step 1, 10000, and 50000) using t-SNE [46].

Crossing Modalities. Fig. 11 illustrates the evolution of features from two modalities, accelerometer and gyroscope, from Dataset No.3 (WISDM). The visualization reveals that as training progresses, the data points form increasingly distinct clusters based on activity classes (walking, jogging, and standing). Moreover, the corresponding clusters from different modalities (indicated by dots and crosses) gradually align and merge, demonstrating the effectiveness of the shared downstream tasks alignment.

Crossing Datasets. Fig. 12 demonstrates the cross-dataset alignment using accelerometer data from Dataset No.2, realdisp, and No.3, WISDM. Despite these datasets having different task definitions, the model learns to organize similar activities into coherent clusters across datasets. This alignment becomes progressively more pronounced through the training steps, from initial scattered distributions to well-defined clusters at Step 50000. This improvement can be attributed to the semantic alignment and quantized indices prediction tasks, which enable the model to learn generalizable activity representations across different datasets.

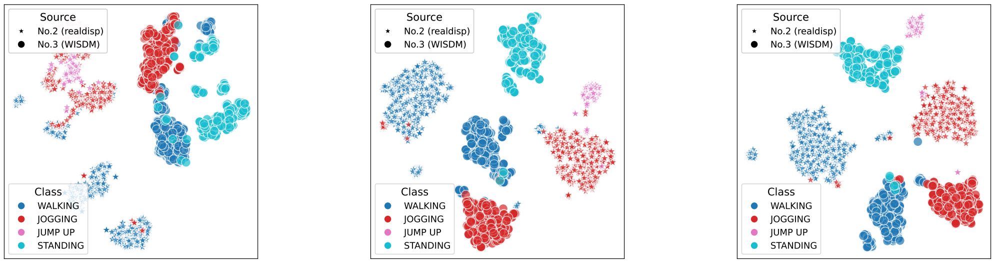


(a) Step 1.

(b) Step 10000.

(c) Step 50000.

Fig. 11. Evolving alignment of the classes (examples with WALKING, JOGGING and STANDING) across modality along with the training process, inspected with accelerometer and gyroscope modalities of training Dataset No.3, WISDM.



(a) Step 1.

(b) Step 10000.

(c) Step 50000.

Fig. 12. Evolution of the alignment across datasets that do not share the same task definitions, inspected with accelerometer on Dataset No.2, realdisp (with WALKING, JOGGING and *JUMP UP*), and No.3, WISDM (with WALKING, JOGGING and *STANDING*).

5.4 Generalization (RQ3): New Wearable Layouts

To explore the generalization capabilities of the proposed foundation model training strategy, we extend the training to datasets collected from other body parts (Upper Arm and Pocket; see Section 4.3) to validate the effectiveness of our approach across different wearable layouts.

For experiments on the generalization of the HAR-FM training strategy, we extract sensory data from two new placements - upper arm and pocket - and train the foundation model for each placement respectively (marked as **HAR-FM-UpARM** and **HAR-FM-POCKET**, with training details summarized in Section 4.3.3). We use data from Dataset No.**7*** (shoaib), which is the only dataset among the five that contains data from these target placements. Notably, for the pocket placement, the test data contains magnetometer modality that was not included in the foundation model training data.

Table 7. The fine-tuned performance of the foundation model pretrained on <Upper Arm> data.

	WS	1s				5s			
No. 7^* - shoaib	METRICS	A.(1)	G.(1)	M.(1)	ALL(3)	A.(1)	G.(1)	M.(1)	ALL(3)
TSFCN	Acc	0.893	0.848	0.659	0.800	0.914	0.901	0.711	0.842
	F1	0.892	0.843	0.644	0.793	0.911	0.899	0.698	0.836
DDLEARN	Acc	**0.900**	0.772	0.731	0.801	0.893	0.817	0.753	0.821
	F1	**0.898**	0.763	0.729	0.797	0.891	0.813	0.749	0.818
LIMU-BERT	Acc	0.742	0.709	0.466	0.639	0.857	0.779	0.491	0.709
	F1	0.735	0.703	0.461	0.633	0.852	0.772	0.482	0.702
TinyHAR	Acc	0.845	0.725	0.626	0.732	0.791	0.739	0.664	0.731
	F1	0.843	0.706	0.619	0.723	0.787	0.699	0.656	0.714
MLP-HAR	Acc	0.803	0.760	0.716	0.760	0.771	0.756	0.756	0.761
	F1	0.799	0.758	0.715	0.757	0.766	0.751	0.753	0.757
TRAIN-FROM-SCRATCH	Acc	0.883	**0.849**	0.682	0.805	0.901	0.891	0.727	0.840
	F1	0.880	**0.845**	0.675	0.800	0.899	0.889	0.718	0.835
HAR-FM-UpARM	Acc	0.894	0.841	**0.744**	**0.826**	**0.936**	**0.927**	**0.851**	**0.905**
	F1	0.893	0.839	**0.739**	**0.824**	**0.934**	**0.926**	**0.849**	**0.903**

Table 8. The fine-tuned performance of the foundation model pretrained on <Pocket> data, where the pretrained data does not contain the magnetometer data (marked as Unseen).

		Seen Modalities						**Unseen**	
	WS	1s			5s			1s	5s
No. 7^* - shoaib	METRICS	A.(1)	G.(1)	ALL.(2)	A.(1)	G.(1)	ALL.(2)	M.(1)	M.(1)
TSFCN	Acc	**0.966**	0.861	0.914	0.990	0.903	0.946	0.801	0.881
	F1	**0.966**	0.857	0.912	0.990	0.899	0.944	0.803	**0.882**
DDLEARN	Acc	0.957	0.833	0.895	0.980	0.853	0.916	**0.838**	0.854
	F1	0.957	0.831	0.894	0.980	0.846	0.913	**0.839**	0.856
LIMU-BERT	Acc	0.889	0.804	0.846	0.905	0.827	0.866	0.519	0.568
	F1	0.889	0.802	0.846	0.902	0.819	0.860	0.528	0.568
TinyHAR	Acc	0.922	0.836	0.879	0.977	0.828	0.902	0.766	**0.882**
	F1	0.922	0.831	0.876	0.977	0.805	0.891	0.767	0.881
MLP-HAR	Acc	0.904	0.824	0.864	0.902	0.813	0.857	0.786	0.812
	F1	0.904	0.821	0.862	0.901	0.808	0.854	0.786	0.812
TRAIN-FROM-SCRATCH	Acc	0.956	0.878	0.917	0.979	0.908	0.944	0.826	0.818
	F1	0.956	0.876	0.916	0.979	0.899	0.939	0.827	0.817
HAR-FM-POCKET	Acc	0.956	**0.887**	**0.922**	**0.993**	**0.914**	**0.954**	0.817	0.866
	F1	0.956	**0.886**	**0.921**	**0.993**	**0.914**	**0.954**	0.818	0.865

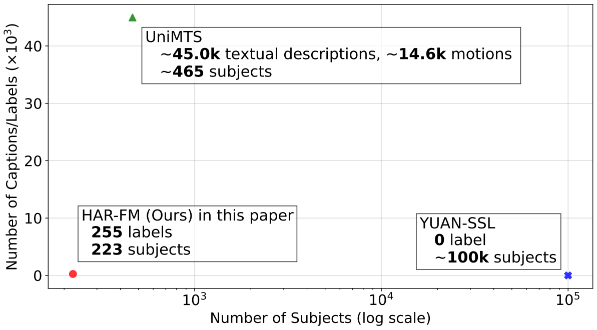


Fig. 13. The proof-of-concept HAR-FM model was trained with modest data settings.

Tables 7 and 8 summarize the fine-tuning results on the test dataset. Overall, **HAR-FM** demonstrates superiority on modalities that were involved in the training dataset. Although it does not achieve the highest performance on the unseen modality in the POCKET experiments, the improvements over the **TRAIN-FROM-SCRATCH** baseline still demonstrate the potential effectiveness of the foundation model training strategy.

6 Discussion, Limitations and Future Work

In our study, we trained an HAR-FM that benefits the downstream deployment. It indicates the feasibility of the proposed training framework to accumulate valuable information from the fragmented open-source datasets, even they do not share the same collection settings nor the purposes.

6.1 Towards Customizable HAR Foundation Model(s)?

In this paper, we mention the concept of a customizable foundation model for Human Activity Recognition (HAR), designed to handle diverse sensor modalities across different deployment scenarios. Our HAR-FM framework implements this vision through a two-stage training strategy, demonstrating effectiveness in extensive experiments while acknowledging that foundation models for HAR remain at an early developmental stage.

6.1.1 Foundation Model Candidates. Recent work has approached foundation model through two primary paradigms: Yuan et al. [62] leveraged large-scale unlabeled data collected from over 100k subjects using wristband accelerometer and pretrained separate models for different window sizes (5/10/30s inputs) with three pretext tasks. Zhang et al. [65] utilized high-quality labeled data. Particularly, they synthesized IMU signals using motion equations from 14,616 motions performed by 465 subjects and approached the IMU sensor foundation model with 44,970 textual descriptions (on the motions) through aligning sensor-text embeddings via contrastive learning. In comparison, our proof-of-concept HAR-FM model (Fig.13) demonstrates the feasibility in FM training with modest data settings, trained on just 4 datasets containing approximately 255 labels and 223 subjects.

6.1.2 Customizable Foundation Model Architecture. Architecturally, while existing foundation model primarily adopt encoder that has fixed input format across datasets, HAR-FM distinguishes itself through its explicitly modular design: (1) dataset-wise encoders that can be customized to accommodate different sensor configurations through adjustable depth and input dimensionality, and (2) shared Transformer-based context encoders that maintain consistent downstream processing. This approach, as quantified in Table 9 for an 18-class single-IMU case study, achieves both parameter and computation efficiency while supporting dynamic input configurations - a critical advantage for real-world deployment scenarios where sensor specifications may vary.

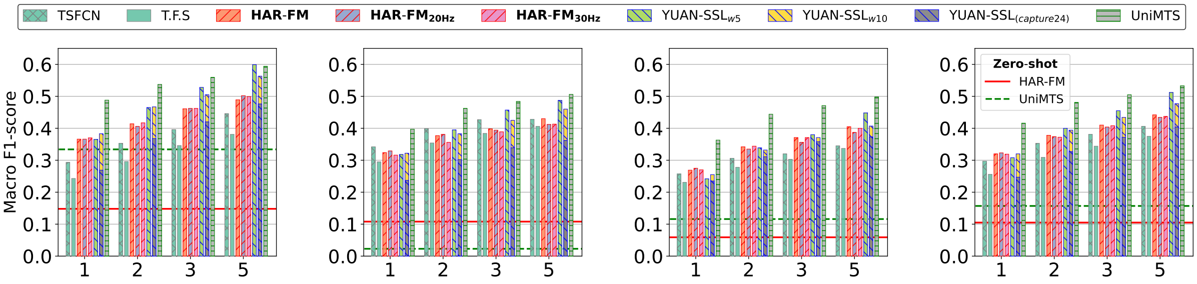
Table 9. Comparison of model parameters and Multiply-Accumulate Operations (MACs) for an 18-class HAR case study. All configurations of HAR-FM (Ours) use dataset-wise encoders with 3-layer convolutional neural network, varying only in input shapes.

Model	Configuration		Input Shape	Params (M)	MACs (M)
	Freq. (Hz)	Window Size (sec)			
HAR-FM (Ours)*	20	1	(5, 3, 4)	2.571	14.543
		3	(15, 3, 4)		43.627
		5	(25, 3, 4)		72.711
	30	5	(25, 3, 6)	2.599	79.413
	50	5	(25, 3, 10)	2.654	92.818
YUAN-SSL	30	5	(3, 150)	4.500	117.835
		10	(3, 300)	10.991	288.941
UniMTS	20	10	(200, 3, 22, 1)	5.175	7,098.916

*The main configuration settings are listed to demonstrate the customizability of our learning framework.

Table 10. Zero-shot performance on <Wrist> data.

		WS	1s				5s			
		METRICS	A.(5)	G.(4)	M.(3)	ALL(12)	A.(5)	G.(4)	M.(3)	ALL(12)
TSFCN (**Rand-init**)		Acc	0.072	0.090	0.089	0.084	0.069	0.089	0.087	0.082
		F1	0.026	0.029	0.033	0.029	0.023	0.028	0.031	0.028
HAR-FM (Ours)	**Rand-init**	Acc	0.079	0.093	0.066	0.079	0.079	0.093	0.068	0.080
		F1	0.033	0.042	0.028	0.035	0.034	0.041	0.030	0.035
	Zero-shot	Acc	0.197	**0.156**	0.126	0.160	0.255	**0.184**	0.115	0.185
		F1	0.119	**0.095**	0.063	0.092	0.148	**0.108**	0.059	0.105
UniMTS	**Rand-init**	Acc	0.096	0.109	0.092	0.099	0.096	0.109	0.092	0.099
		F1	0.022	0.025	0.025	0.024	0.022	0.026	0.025	0.024
	Zero-shot	Acc	**0.358**	0.080	**0.170**	**0.203**	**0.405**	0.070	**0.184**	**0.220**
		F1	**0.266**	0.027	**0.093**	**0.129**	**0.334**	0.023	**0.116**	**0.157**



(a) Acc. (b) Gyro. (c) Magneto. (d) ALL

Fig. 14. Few shot experiment results (WS=5s).

6.2 Towards Training-Free Deployment

Another expectation on the Foundation Model is the superior understanding upon the context of the sensory signals, which is key ability to realize training-free deployment in some future. In this paper, we investigate such ability of existing foundation model candidates with zero- and few-shot experiments constructed on WRIST datasets. The evaluation adopts the 5-fold experimental settings (see section 5.1.2) to avoid ambiguity. Similarly, the averaged classification accuracy and Macro F1-score are reported.

6.2.1 Zero-shot Experiment. We conduct zero-shot evaluation comparing the **UniMTS** and our **HAR-FM**, both of which employ similar idea of aligning the text and sensor embeddings. The zero-shot classification is performed based on the similarity computed between the input sensor embeddings and the semantic embeddings (generated by text encoder). For **HAR-FM**, we initialize dataset-wise encoders with weights obtained during stage-two multi-task learning. Particularly, for accelerometer data, datasets No.5, 8, and 9 use the encoder pretrained on dataset No.4, while datasets No.6 and 7, along with all gyroscope and magnetometer data, use the encoder pretrained on dataset No.2. The gyroscope and magnetometer data in dataset No.5, PAMAP2 and No.8, HHAR, are downsampled to 50 Hz to match the available encoders (from dataset No.2). Baseline comparisons include **Rand-init**, which refers to the model constructed for downstream finetuning with all parameters randomly initialized and serves as the rigorous boundary for learning efficacy. In addition, we introduce TSFCN[51] that has a different model architecture and demonstrates competitive performance in previous experiments.

Results. Table 10 summarizes zero-shot performance across modalities. Both **HAR-FM** and **UniMTS**, the models surpass their **Rand-init** counterparts on modalities seen during the training phase (i.e., in all modalities for **HAR-FM** and accelerometer data for **UniMTS**). Notably, **UniMTS** achieves partial transfer to magnetometer data but fails on gyroscoped data, suggesting that while some unintended cross-modal benefits may occur, explicit training on diverse modalities remains crucial - particularly for gyroscope's unique motion characteristics.

6.2.2 Few-shot Experiment. In this experiment, we investigate 1, 2, 3 and 5-shot data settings with window length at 5 seconds to inspect the foundation model's transferability under variant configurations. In particular, the **HAR-FM** series adopt same model structure varying the input shape to support available maximum, 20 Hz and 30 Hz frequency input respectively. In addition to **UniMTS** (which accepts maximum 10-second inputs), we include two variant pretrained models of YUAN-SSL that were trained in different input settings (5-second and 10-second) as **YUAN-SSL**$_{\mathbf{w5}}$ and **YUAN-SSL**$_{\mathbf{w10}}$ respectively. Similarly, other than pretrained model, we introduce **TSFCN** and **TRAIN-FROM-SCRATCH** as the baselines to demonstrate the benefits in using the pretrained models regarding such extremely limited resource scenario. To investigate the potential benefits from training with larger dataset, we also propose to compare the performance with **YUAN-SSL**$_{\mathbf{(capture24)}}$ baseline that was a model instance pretrained on dataset, capture24[7], using 10-second window size setting. All baselines are finetuned for 200 epochs without any augmentation to ensure the fairness.

Results. As shown in Figure 14, **HAR-FM** consistently outperforms the **TRAIN-FROM-SCRATCH** and **YUAN-SSL**$_{\mathbf{(capture24)}}$ baseline, demonstrating the effectiveness of the two-stage training strategy in extracting transferable patterns. However, the gap between **HAR-FM** and the other foundation model candidates like **UniMTS** and **YUAN-SSL** suggests that incorporating larger datasets (evidenced by the gaps between baseline **YUAN-SSL**$_{\mathbf{(capture24)}}$ and **YUAN-SSL**$_{\mathbf{w10}}$) and more fine-grained activity descriptions (evidenced by the significant performance of **UniMTS**) could lead to significant performance improvements. Interestingly, **TSFCN**'s unexpected competitiveness on gyroscope data (outperforming all but **UniMTS** in 1-/2-shot) highlights the modality-specific challenges that the current pretraining paradigms may not fully address.

6.3 Limitations and Future Works

In this paper, we verify the feasibility of applying the foundation model concept—learning general representations—to Human Activity Recognition (HAR) using fragmented datasets. Although our proposed foundation model demonstrates promising progress and provides valuable insights, certain limitations remain. We discuss these limitations and potential future research directions in this section.

Customizable Foundation with More Tasks, Modalities, and Practical Configurations. The proposed training strategy for a HAR foundation model presents a base that can effectively learn from fragmented datasets. The customizable modular design (a dataset-wise encoder upon the pretrained context encoder) supports variant modality input format (e.g., channels and frequency configuration) and demonstrates its flexibility to realize more resource-efficient real-world deployment (Table 9). However, the lag in the performance of the proof-of-concept HAR-FM comparing to the other foundation model candidates in the one- and few-shot experiments should be acknowledged. To address the performance gap, training with more data could be the most direct viable solution, though the extensive experiments in this paper shall also reveal several insights. Firstly, the ablation study (Table 6) and generalization study with unseen modalities (Table 8) highlight the importance of downstream task and modality diversity in improving foundation model performance during early stages. We observe frequent failures with the accelerometer modality under 1-second window settings, indicating potential challenges in scenarios with rarer configurations. Secondly, in the zero- and few-shot experiments, Yuan et al.[62] and Zhang et al.[65] demonstrates either larger volume or more precise annotations could contribute to the model performance, both of which could be covered by our two-stage training strategy (i.e., self-supervised learning and multi-task supervised learning). Notably, the failures observed in zero- and few-shot experiment on gyroscope data suggest the importance in incorporating new modality data explicitly which is also part of the essential design principals of HAR-FM training framework. Based on these analysis, scaling the foundation model to incorporate more diverse tasks (such as motion tracking [23]), modalities (such as electromyogram [11]), and practical configurations becomes crucial. Nevertheless, training with more datasets could introduce more engineering efforts and require addressing new challenges, including training fairness and data imbalance. What is also an interesting topic to be explored could be related to the model capacity. Given the continuous expansion of tasks and sensing modalities, the challenge of catastrophic forgetting [49] (as analyzed in Appendix B) must be carefully considered in system design. In summary, we envision a future HAR foundation model capable of perceiving human activities across diverse configurations (such as available sensors, potential scenarios or even new wearable devices, as mentioned in Section 2.2).

Integrating Placements. Currently, we train the HAR-FM in a placement-wise manner to extract consistent patterns across different modalities (Section 4.3.1). Other than better explanation (that the modalities might share underlying representation of motion), the placement-wise training further reduces the requirements upon the sensory data collection, especially for uncommon modalities (e.g., electromyogram [11]). In comparison, UniMTS[65] is trained with the data that contains synthesized IMU readings from 22 joint nodes of the human body graph. Its superior performance against HAR-FM and YUAN-SSL[62] that are pretrained with data from wrist only, suggests the benefits in integrating all placement-wise models. Future work for HAR-FM could explore a framework that leverages the placement-specific models as experts to develop a unified model. Such a model could learn to map and transfer knowledge across different sensor placements, potentially through techniques like knowledge distillation or ensemble learning. This advancement would be particularly valuable as dataset diversity continues to expand and could ultimately lead to a more generalized solution for the placement invariance challenge in HAR.

Closer Connection with Other Foundation Model. Our current implementation aligns context vectors from sensing modalities with semantic embeddings from pretrained language models—a relatively basic approach that serves as a proof of concept. Future research could explore deeper integration with existing foundation models, potentially contributing to mobile foundation model firmware development [63] and enabling more sophisticated multi-modal learning capabilities through integration with text and vision data [54].

7 Conclusions

In this paper, we explore the concept of training a foundation model for human activity recognition (HAR) using fragmented open-source datasets. We propose a two-stage training framework consisting of: (A) self-supervised learning with a Quantizer Module to effectively model patterns from raw, unlabeled sensory data, and (B) multi-task supervised learning designed to retain discretized representations, complete downstream tasks, and align the representations with corresponding semantic embeddings. The implementation and evaluation were conducted in a device placement-wise manner with appropriate customizations, demonstrating the effectiveness of the proposed HAR foundation model. The results offer valuable insights and contribute to advancing research in this emerging field.

Acknowledgments

This research is supported in part by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), China NSFC Grant (No.62472366), 111 Center (No.D25008), the Project of DEGP (No.2024GCZX003, 2023KCXTD042), Shenzhen Science and Technology Foundation (ZDSYS20190902092853047). Kaishun Wu is the corresponding author.

References

[1] Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. 2023. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409* (2023).

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*, Vol. 3. 3.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[4] Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: A Novel Framework for Agile Development of Mobile Health Applications. In *Ambient Assisted Living and Daily Activities*, Leandro Pecchia, Liming Luke Chen, Chris Nugent, and José Bravo (Eds.). Springer International Publishing, Cham, 91–98.

[5] Oresti Baños, Miguel Damas, Héctor Pomares, Ignacio Rojas, Máté Attila Tóth, and Oliver Amft. 2012. A benchmark dataset to evaluate sensor displacement in activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 1026–1035. doi:10.1145/2370216.2370437

[6] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948* (2023).

[7] Shing Chan, Yuan Hang, 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 (Oct. 2024), 1135. doi:10.1038/s41597-024-03960-3

[8] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469* (2023).

[9] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042. doi:10.1016/j.patrec.2012.12.014

[10] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 244–250.

[11] Hristo Dimitrov, Anthony M. J. Bull, and Dario Farina. 2023. High-density EMG, IMU, kinetic, and kinematic open-source data for comprehensive locomotion activities. *Scientific Data* 10, 1 (10 Nov 2023), 789. doi:10.1038/s41597-023-02679-x

[12] Aiden Doherty, Karl Smith-Byrne, Teresa Ferreira, Michael V. Holmes, Chris Holmes, Sara L. Pulit, and Cecilia M. Lindgren. 2018. GWAS identifies 14 loci for device-measured physical activity and sleep duration. *Nature Communications* 9, 1 (Dec. 2018), 5257. doi:10.1038/s41467-018-07743-4

[13] Azul Garza and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589* (2023).

[14] 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, Article 64 (May 2024), 26 pages. doi:10.1145/3659597

[15] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. arXiv:1611.01144 [stat.ML] https://arxiv.org/abs/1611.01144

[16] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 289–304.

[17] Wenchao Jiang and Zhaozheng Yin. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1307–1310.

[18] Yonghang Jiang, Zhenjiang Li, and Jianping Wang. 2018. Ptrack: Enhancing the applicability of pedestrian tracking with wearables. *IEEE Transactions on Mobile Computing* 18, 2 (2018), 431–443.

[19] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[20] Herve Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 117–128. doi:10.1109/TPAMI.2010.57

[21] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. 2016. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. 164–175.

[22] Mengxi Liu, Sizhen Bian, Bo Zhou, and Paul Lukowicz. 2024. iKAN: Global Incremental Learning with KAN for Human Activity Recognition Across Heterogeneous Datasets. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers* (Melbourne VIC, Australia) *(ISWC '24)*. Association for Computing Machinery, New York, NY, USA, 89–95. doi:10.1145/3675095.3676618

[23] Miaomiao Liu, Sikai Yang, Wyssanie Chomsin, and Wan Du. 2023. Real-Time Tracking of Smartwatch Orientation and Location by Multitask Learning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22)*. Association for Computing Machinery, New York, NY, USA, 120–133. doi:10.1145/3560905.3568548

[24] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.

[25] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM on Web Conference 2024*. 4095–4106.

[26] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 287–299.

[27] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*. 49–58.

[28] Mika A Merrill and Tim Althoff. 2023. Self-supervised pretraining and transfer learning enable\titlebreak flu and covid-19 predictions in small mobile sensing datasets. In *Conference on Health, Inference, and Learning*. PMLR, 191–206.

[29] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2017. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences* 7, 10 (2017), 1101.

[30] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. 2024. Scaling Wearable Foundation Models. *arXiv preprint arXiv:2410.13638* (2024).

[31] Stylianos Paraschiakos, Cláudio Rebelo de Sá, Jeremiah Okai, P. Eline Slagboom, Marian Beekman, and Arno Knobbe. 2022. A recurrent neural network architecture to model physical activity energy expenditure in older people. *Data Mining and Knowledge Discovery* 36, 1 (Jan. 2022), 477–512. doi:10.1007/s10618-021-00817-w

[32] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28, 6 (2010), 976–990.

[33] Xin Qin, Jindong Wang, Shuo Ma, Wang Lu, Yongchun Zhu, Xing Xie, and Yiqiang Chen. 2023. Generalizable Low-Resource Activity Recognition with Diverse and Discriminative Representation Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 1943–1953. doi:10.1145/3580305.3599360

[34] Zhen Qin, Lingzhou Hu, Ning Zhang, Dajiang Chen, Kuan Zhang, Zhiguang Qin, and Kim-Kwang Raymond Choo. 2019. Learning-aided user identification using smartphone sensors for smart homes. *IEEE Internet of Things Journal* 6, 5 (2019), 7760–7772.

[35] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

[36] Attila Reiss and Didier Stricker. 2012. Creating and benchmarking a new dataset for physical activity monitoring. In *Proceedings of the 5th international conference on pervasive technologies related to assistive environments*. 1–8.

[37] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.

[38] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.

[39] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 233–240.

[40] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[41] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.

[42] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *CoRR* abs/2004.09297 (2020). arXiv:2004.09297 https://arxiv.org/abs/2004.09297

[43] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[44] Scott Sun, Dennis Melamed, and Kris Kitani. 2021. IDOL: Inertial deep orientation-estimation and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6128–6137.

[45] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. [n. d.]. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. 5 2024. *arXiv preprint arXiv:2405.17766* ([n. d.]).

[46] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[48] Rosemary Walmsley, Shing Chan, Karl Smith-Byrne, Rema Ramakrishnan, Mark Woodward, Kazem Rahimi, Terence Dwyer, Derrick Bennett, and Aiden Doherty. 2020. Reallocating time from machine-learned sleep, sedentary behaviour or light physical activity to moderate-to-vigorous physical activity is associated with lower cardiovascular disease risk. *medRxiv* (2020). doi:10.1101/2020.11.10.20227769

[49] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (2024), 5362–5383. doi:10.1109/TPAMI.2024.3367329

[50] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. 65–76.

[51] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*. 1578–1585. doi:10.1109/IJCNN.2017.7966039

[52] Gary Weiss. 2019. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset.

[53] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. 2018. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific Reports* 8, 1 (May 2018), 7961. doi:10.1038/s41598-018-26174-1

[54] Kang Xia, Wenzhong Li, Shiwei Gan, and Sanglu Lu. 2024. TS2ACT: Few-Shot Human Activity Sensing with Cross-Modal Co-Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 188 (Jan. 2024), 22 pages. doi:10.1145/3631445

[55] Tianzhang Xing, Qing Wang, Chase Q Wu, Wei Xi, and Xiaojiang Chen. 2020. Dwatch: A reliable and low-power drowsiness detection system for drivers based on mobile devices. *ACM Transactions on Sensor Networks (TOSN)* 16, 4 (2020), 1–22.

[56] 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* (Madrid, Spain) *(ACM MobiCom '23)*. Association for Computing Machinery, New York, NY, USA, Article 85, 15 pages. doi:10.1145/3570361.3613299

[57] 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 *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[58] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[59] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek 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*. 351–360.

[60] Chin-Chia Michael Yeh, Xin Dai, Huiyuan Chen, Yan Zheng, Yujie Fan, Audrey Der, Vivian Lai, Zhongfang Zhuang, Junpeng Wang, Liang Wang, et al. 2023. Toward a foundation model for time series data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4400–4404.

[61] Yinggang Yu, Dong Wang, Run Zhao, and Qian Zhang. 2019. RFID based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 298–310.

[62] Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. 2024. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine* 7, 1 (2024), 91.

[63] Jinliang Yuan, Chen Yang, Dongqi Cai, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, Shangguang Wang, and Mengwei Xu. 2024. Mobile Foundation Model as Firmware. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 279–295. doi:10.1145/3636534.3649361

[64] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 1036–1043.

[65] Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2024. UniMTS: Unified Pre-training for Motion Time Series. arXiv:2410.19818 [eess.SP] https://arxiv.org/abs/2410.19818

[66] Yi Zhang, Zheng Yang, Guidong Zhang, Chenshu Wu, and Li Zhang. 2021. XGest: Enabling cross-label gesture recognition with RF signals. *ACM Transactions on Sensor Networks (TOSN)* 17, 4 (2021), 1–23.

[67] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 313–325.

[68] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[69] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* 36 (2023), 43322–43355.

[70] Yexu Zhou, Tobias King, Haibin Zhao, Yiran Huang, Till Riedel, and Michael Beigl. 2024. MLP-HAR: Boosting Performance and Efficiency of HAR Models on Edge Devices with Purely Fully Connected Layers. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers* (Melbourne VIC, Australia) *(ISWC '24)*. Association for Computing Machinery, New York, NY, USA, 133–139. doi:10.1145/3675095.3676624

[71] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. 2022. TinyHAR: A Lightweight Deep Learning Model Designed for Human Activity Recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers* (Cambridge, United Kingdom) *(ISWC '22)*. Association for Computing Machinery, New York, NY, USA, 89–93. doi:10.1145/3544794.3558467

A Complementary Dataset Descriptions

We summarize the introduction to the datasets in Table 11.

A.1 Unique *[ACTIVITY]* Label.

As mentioned in Table 1 and detailed in Table 11, some datasets include multiple task definitions that annotate various relationships between activities. To avoid ambiguity when generating descriptions for each activity, we extract the fine-grained labels or create labels (by combining class labels) that uniquely align with all labeled signal samples.

A.2 Cross-Validation Subject Partitions for Downstream Tasks

Table 12 details the user-level data splits adopted for 5-fold cross-validation across all benchmark datasets, ensuring reproducible evaluation of downstream task performance.

B Incremental or Continuous Learning?

In our experimental setup, HAR-FM serves as part of the downstream classification model, trained on specific modality tasks. This raises an important question: can HAR-FM maintain its pretrained capabilities while adapting to new tasks? To investigate this, we evaluate HAR-FM's performance on the original pretraining corpus after downstream adaptation.

Fig. 15 compares performance before downstream adaptation (PRIOR group and the WORST case of random guessing) with performance after adaptation to the mHealth dataset (No. 6) under different few-shot conditions (1, 2, 3, 5, and FULL-shot) using different sensor modality (marked as Seen A, G, and M). The results demonstrate the inevitable degradation after getting retrained with downstream data. Generally, the influence gets pronounced as the amount of unseen data increases. Notably, the gyroscope data leads to the most significant degradation even for the downstream tasks implemented with gyroscope modality itself, indicating completely different patterns encountered in the downstream datasets.

These continuous learning challenges currently suggest that incremental learning represents the more viable approach for HAR-FM development at this stage. By training with aggregated data corpora, the model can better preserve performance on previously encountered tasks and modalities. Looking ahead, future research could

Table 11. Detailed Introduction to the Datasets

ID	**Descriptions**
1	OPPTY /OPPORTUNITY[9]
	The authors collects complex naturalistic activities with a particularly large number of atomic activities in a sensor rich environments. The datasets involve recordings from 4 subjects. The number of available modalities could be 1 or 3 for different placements. There are two types of devices with shared modality involved, resulting in 2 domain conditions at most. ***Tasks:*** We choose two task definitions, i.e., Locomotion (4, including Stand, Sit, Walk and Lie) and Gestures (17 activities, e.g., Open Fridge, Close Fridge and so on).
2	realdisp[5]
	It is an open benchmark dataset to investigate inertial sensor displacement effects in HAR. The recordings covers ideal-, self- and mutual-displacement settings. While we do not explicitly discriminate the placement condition for all 17 subjects, the domain conditions mainly derive from the simultaneous collection with senors mounted on both left and right wrists. ***Tasks:*** We adopt several task definitions based on the body parts, which are whole body (10), trunk (10), and upper extremes (7 activities) oriented activity recognition.
3	WISDM[52]
	The authors collect the accelerometer and gyroscope sensory data with 51 subjects performing activities of daily living while carrying smartphone and wearing smartwatch. ***Tasks:*** Following the paper[52], we partition the activities into three groupings, which are Non-hand-oriented (5), Hand-oriented in general meaning (7), and Hand-oriented during eating (5 activities).
4	capture24[7]
	The researchers release a large HAR dataset collected in the wild, involving wrist-worn accelerometers, wearable cameras and sleep diaries. The study involves 151 participants, amounting 3883 hours of accelerometer data, of which 2562 hours are annotated. ***Tasks:*** We construct tasks with the detailed annotations (206 different descriptions) and adopt six other task definitions (the involved activities number ranges from 4 to 11) previously proposed in different studies[12, 48, 53].
5	PAMAP2[37]†
	The dataset involves IMUs data recorded from 18 human activities performed by 9 subjects with 3 wireless sensors.
6	mHealth[4]†
	The dataset comprises body motion recordings by wearable sensors from 10 volunteers while performing 12 physical activities in an out-of-lab environment with limited constraints.
7*	shoaib[41]†
	The authors construct the dataset that contains 7 activities in daily living with sensors embedded in the commercial-off-the-shelf smartphone and smartwatch.
8	HHAR [43]†
	The authors construct a dataset of HAR to discuss the variance between different devices. The dataset consists of recordings from IMUs of at most 4 devices (two devices mounted on each side of wrist) while performing 6 physical activities in total.
9	GOTOV[31]†
	The dataset consists of accelerometer sensory data collected from 35 healthy elderly (older than 60 years old). Each individual performed a set of 16 everyday life activities with sensors mounted on his/her body.

† *We construct HAR task with all activities by default.*

explore several promising directions to improve continuous learning capabilities. Advanced fine-tuning strategies incorporating regularization or rehearsal mechanisms may offer one path forward. Architectural innovations also show potential, particularly approaches inspired by recent work applying Kolmogorov-Arnold Networks to HAR continuous learning [22]. These findings establish important baselines for HAR-FM's ongoing development while highlighting concrete avenues for enhancing its learning capabilities.

Table 12. Subject-wise data partitioning for 5-fold cross-validation across benchmark datasets. The table shows the distribution of participants into evaluation folds (1-5) for each dataset.

Dataset	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
2012-PAMAP2	subject102	subject101, subject105	subject104, subject109	subject103, subject108	subject106, subject107
2014-mHealth	subject8, subject9	subject5, subject6	subject1, subject4	subject10, subject3	subject2, subject7
2014-shoaib	Participant-3, Participant-4	Participant-8, Participant-9	Participant-5, Participant-6	Participant-1, Participant-10	Participant-2, Participant-7
2015-HHAR	h	b, g	e, i	d, f	a, c
2020-GOTOV	GOTOV05, GOTOV18, GOTOV20, GOTOV23, GOTOV24, GOTOV25, GOTOV26	GOTOV03, GOTOV08, GOTOV12, GOTOV17, GOTOV19, GOTOV29, GOTOV33	GOTOV02, GOTOV11, GOTOV14, GOTOV15, GOTOV28, GOTOV30, GOTOV34	GOTOV04, GOTOV09, GOTOV21, GOTOV27, GOTOV32, GOTOV35, OTOV36	GOTOV06, GOTOV07, GOTOV10, GOTOV13, GOTOV16, GOTOV22, GOTOV31

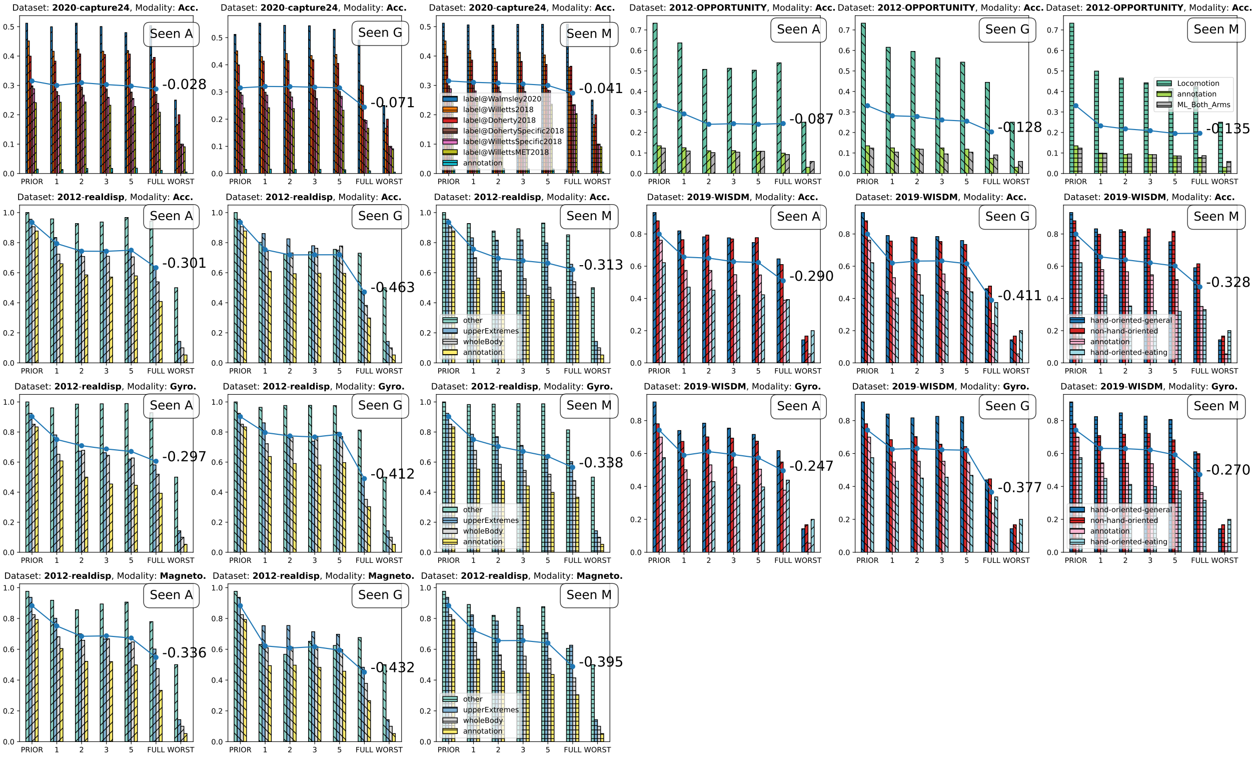


Fig. 15. Performance impact of downstream adaptation on pretrained tasks. While HAR-FM's current adaptation pipeline maintains basic functionality (evident when compared to the WORST-case baseline), the observed degradation on pretrained tasks highlights the need for enhanced continuous learning capabilities in future work.