國　立　成　功　大　學
資　訊　工　程　學　系

碩　士　論　文

利用大型語言模型於低精度穿戴式感測行爲辨識

Utilizing Large Language Model for Human Activity Recognition from Coarse-Grained Wearable Sensors

研　究　生：蘇可維
指導教授：莊坤達博士

中　華　民　國　一　一　三　年　七　月

利用大型語言模型於低精度穿戴式感測行爲辨識

學生：蘇可維　　指導教授：莊坤達博士

國立成功大學資訊工程學系

摘　　要

人類行爲辨識（HAR）是穿戴式感測技術中的重要研究領域，對於健康監測、運動分析和智慧生活應用具有重要意義。傳統的穿戴式感測器往往受限於低精度數據和計算資源限制，難以實現準確的行爲辨識。隨著大型語言模型（LLM）的快速發展，其強大的序列理解和模式識別能力爲解決此問題提供了新的可能。

本論文提出了一種利用大型語言模型進行低精度穿戴式感測行爲辨識的創新方法。我們設計了一個適應性框架，將穿戴式感測器的粗粒度時間序列數據轉換爲語言模型可理解的表示形式，並利用預訓練語言模型的強大表示學習能力來提升行爲辨識的準確性。

我們的方法包含三個核心組件：（1）感測數據預處理和特徵提取模組，將原始感測信號轉換爲結構化序列；（2）基於大型語言模型的行爲模式學習架構，利用注意力機制捕捉時序關係；（3）多模態融合策略，整合不同感測器的信息以提升辨識性能。實驗結果顯示，相比傳統方法，我們的方法在多個公開數據集上都實現了顯著的性能提升，特別是在處理低精度和噪聲數據方面表現優異。

這項研究爲穿戴式設備的智能化發展提供了新的技術路徑，推動了大型語言模型在物聯網和普適計算領域的應用創新。

關鍵字：人類行爲辨識；大型語言模型；穿戴式感測器；低精度數據；時間序列分析；多模態融合

Utilizing Large Language Model for Human Activity Recognition from Coarse-Grained Wearable Sensors

Student: Su-Ko Wei Advisor: Dr. Kun-Ta Chuang

Submitted to Institute of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Cheng Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

Human Activity Recognition (HAR) is a critical research area in wearable sensing technology, with significant implications for health monitoring, sports analysis, and smart living applications. Traditional wearable sensors are often limited by low-precision data and computational constraints, making accurate activity recognition challenging. The rapid advancement of Large Language Models (LLMs) offers new possibilities with their powerful sequence understanding and pattern recognition capabilities.

This thesis proposes an innovative approach that utilizes Large Language Models for human activity recognition from coarse-grained wearable sensors. We design an adaptive framework that transforms coarse-grained time-series data from wearable sensors into representations comprehensible by language models, leveraging the powerful representation learning capabilities of pre-trained language models to enhance activity recognition accuracy.

Our method comprises three core components: (1) a sensor data preprocessing and feature extraction module that converts raw sensor signals into structured sequences; (2) a Large Language Model-based activity pattern learning architecture that utilizes attention mechanisms to capture temporal relationships; and (3) a multi-modal fusion strategy that integrates information from different sensors to improve recognition performance. Experimental results demonstrate that our approach achieves significant performance improvements over traditional methods across multiple public datasets, particularly excelling in handling low-precision and noisy data.

This research provides a new technical pathway for the intelligent development of wearable

devices and advances the application innovation of Large Language Models in the Internet of Things and ubiquitous computing domains.

Keywords: Human Activity Recognition; Large Language Models; Wearable Sensors; Coarse-grained Data; Time Series Analysis; Multi-modal Fusion

Acknowledgment

Contents

List of Tables

List of Figures

Chapter 1 Introduction

Human Activity Recognition (HAR) using wearable sensors has emerged as a fundamental technology for numerous applications including mobile health monitoring, sports analytics, human-computer interaction, and smart environment systems [1, 2]. The ability to automatically identify and classify daily activities from sensor data enables continuous and unobtrusive monitoring of human behavior, providing valuable insights for healthcare, fitness tracking, and lifestyle analysis [3].

A critical and underexplored challenge in HAR is the significant drop in recognition performance when sensor data is downsampled or collected at low temporal resolution, a scenario increasingly common in consumer-grade wearables. Reducing the sampling frequency of sensors such as accelerometers and gyroscopes can dramatically extend battery life and operational time for smart devices, making low-grain data collection highly desirable for real-world, long-term deployment. However, this comes at the cost of information loss, increased noise, and reduced discriminative power, which can severely degrade the accuracy of activity recognition models.

Traditional HAR approaches predominantly rely on motion sensors such as accelerometers, gyroscopes, and magnetometers embedded in smartphones or dedicated wearable devices [4]. While these sensors offer the advantage of being lightweight, power-efficient, and readily available in consumer devices, they often produce coarse-grained data with limited precision compared to specialized research-grade equipment. This limitation poses significant challenges for accurate activity recognition, particularly when distinguishing between subtle variations in movement patterns or activities with similar motion characteristics.

Recent advances in deep learning have revolutionized the field of HAR, with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) showing remarkable performance improvements over traditional machine learning approaches [5, 6]. Deep learning models have demonstrated their ability to automatically extract meaningful features from raw sensor data, reducing the need for manual feature engineering [7, 8]. However, most existing deep learning approaches for HAR are designed for high-quality sensor data and may not perform optimally when dealing with the noisy, low-resolution signals typical of consumer-grade wearable devices [9].

The emergence of Large Language Models (LLMs) and transformer architectures has opened new possibilities for sequence modeling and pattern recognition across diverse domains [10]. Models such as BERT [11] and GPT-3 [12] have demonstrated exceptional capabilities in understanding complex sequential patterns and contextual relationships. While these models were originally designed for natural language processing tasks, their underlying transformer architecture and attention mechanisms show promise for modeling temporal sequences in other domains, including sensor data analysis.

Recent work has begun to explore the application of LLMs to sensor-based HAR, introducing frameworks for aligning language models with motion sensors for activity recognition. These approaches typically involve specialized alignment stages and task-aware tuning procedures to bridge the gap between natural language processing and sensor-based activity recognition. However, these existing methods face several critical limitations when applied to real-world deployment scenarios involving downsampled, coarse-grained wearable sensors.

The primary challenges include computational requirements and memory footprint that present significant barriers for deployment on resource-constrained wearable devices, where power efficiency and real-time processing are paramount. Additionally, existing approaches have primarily focused on high-quality sensor data from controlled laboratory settings, but have not adequately addressed the unique challenges posed by low-frequency, low-precision sensor signals that are characteristic of consumer-grade wearable devices. Furthermore, the temporal resolution and sampling rates of downsampled wearable sensors may not align optimally with the sequence processing capabilities of language models, necessitating specialized data representation and tokenization strategies.

Our work addresses these challenges through a novel approach that leverages the complementary strengths of advanced time-series processing and language model capabilities. The self-attention mechanism in transformers can effectively capture long-range dependencies in temporal sensor data, potentially identifying subtle patterns that traditional approaches might miss in noisy, low-resolution signals. The pre-training paradigm used in LLMs could enable knowledge transfer from large-scale datasets to smaller, domain-specific HAR datasets, which is particularly valuable when working with limited, downsampled sensor data. Additionally, the contextual understanding capabilities of LLMs may help in disambiguating similar activi-

ties by considering broader temporal context, which is crucial when dealing with the reduced discriminative power of low-frequency sensors.

This thesis is specifically motivated by the need to enable accurate and robust HAR on downsampled (low-grain) sensor data, with the ultimate goal of enabling longer battery life and more practical, real-world deployment of smart wearable devices.

A key contribution of this thesis is a **two-stage pipeline** that first restores high-resolution temporal detail from low-frequency sensor signals using an advanced upsampling framework and then performs activity recognition with a token-efficient downstream model. By separating data enhancement from recognition, we unlock the complementary strengths of state-of-the-art time-series imputation and sensor-language modelling, achieving robust HAR even when the raw input is severely downsampled.

The main contributions of this work include:

1. A novel **upsampler + downstream model** framework that integrates cutting-edge time-series imputation with large-scale sensor-language modelling for coarse-grained wearable HAR.

2. An end-to-end processing pipeline and tokenisation strategy that enables efficient fine-tuning of compact models on imputed sensor sequences, facilitating on-device inference without sacrificing accuracy.

The remainder of this thesis is organized as follows: Chapter 2 reviews related work in human activity recognition, deep learning approaches for sensor data analysis, and the application of transformer models to time series data, with particular emphasis on recent developments in LLM-based sensor data processing and the challenges of low-grain data. Chapter 3 presents our proposed methodology, including the downsampled sensor data tokenization approach and the efficient LLM-based HAR architecture. Chapter 4 describes the experimental setup and presents comprehensive evaluation results across multiple datasets and sensor precision levels. Chapter 5 discusses the implications of our findings, limitations of the current approach, and future research directions. Finally, Chapter 6 concludes the thesis and summarizes the key contributions.

Chapter 2 Related Work

2.1 Human Activity Recognition with Wearable Sensors

Human Activity Recognition (HAR) using wearable sensors emerged as a fundamental research area with applications spanning healthcare monitoring, smart homes, and fitness tracking. Early comprehensive surveys such as "A Tutorial on Human Activity Recognition Using Wearable Sensors" by Bulling et al. [1] and "A Survey on Human Activity Recognition using Wearable Sensors" by Lara et al. [2] established the foundational understanding of sensor-based activity recognition, highlighting the challenges of feature extraction from multi-modal sensor data and the importance of temporal modeling for sequential activities.

The evolution of HAR was significantly influenced by advances in deep learning methodologies. Traditional approaches relied heavily on hand-crafted features and shallow machine learning models, but the introduction of deep neural networks revolutionized the field. "Deep Learning for Sensor-based Activity Recognition: A Survey" by Wang et al. [5] provided a comprehensive survey demonstrating how deep learning techniques could automatically learn hierarchical representations from raw sensor data, eliminating the need for manual feature engineering. This paradigm shift enabled more robust and generalizable activity recognition systems.

2.1.1 Deep Learning Approaches for HAR

The application of deep learning to HAR yielded remarkable improvements in recognition accuracy and model generalization. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition" by Yang et al. [6] pioneered the use of deep convolutional neural networks on multichannel time series data, demonstrating superior performance compared to traditional methods. Their approach effectively captured both spatial relationships between sensor channels and temporal dependencies within activity sequences.

Convolutional Neural Networks (CNNs) proved particularly effective for sensor data pro-

cessing. "Real-Time Human Activity Recognition with Deep Neural Networks on Mobile Devices" by Ignatov et al. [9] developed real-time HAR systems using CNNs on accelerometer data, achieving impressive accuracy while maintaining computational efficiency suitable for mobile deployment. Similarly, "Human Activity Recognition with Smartphone Sensors using Deep Learning Neural Networks" by Ronao and Cho [8] explored smartphone-based HAR using deep learning, showcasing the potential for ubiquitous activity monitoring through everyday devices.

The integration of recurrent architectures has further enhanced temporal modeling capabilities. "Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables" by Hammerla et al. [7] conducted comprehensive comparisons of deep, convolutional, and recurrent models for wearable-based HAR, establishing best practices for different types of sensor configurations and activity categories. Their work highlighted the importance of model architecture selection based on the specific characteristics of the target application.

2.2 Transformer and Attention Mechanisms in HAR

The transformer architecture, originally introduced by "Attention Is All You Need" (Transformer) by Vaswani et al. [10], revolutionized sequence modeling across various domains. The self-attention mechanism's ability to capture long-range dependencies and parallel processing capabilities made it particularly attractive for sensor data analysis. Recent works successfully adapted transformer architectures for HAR applications, demonstrating significant improvements over traditional recurrent approaches.

"P2LHAP: Patch-to-Label Human Activity Prediction" by Li et al. [13] introduced P2LHAP, a novel Patch-to-Label Seq2Seq framework that tackled activity recognition, segmentation, and forecasting simultaneously. Their approach divided sensor data streams into patches served as input tokens, enabling unified processing of multiple HAR tasks within a single model. The patch-based tokenization strategy proved effective for handling variable-length sensor sequences while maintaining computational efficiency.

Recent advances also explored specialised transformer variants for sensor data. "SAITS:

Self-Attention-based Imputation for Time Series" by Du et al. [14] proposed SAITS at *AAAI 2023*, introducing dual-masked self-attention blocks that explicitly modeled temporal as well as cross-channel correlations. Their approach surpassed BRITS and Transformer-based baselines on multiple health-care time-series benchmarks while using a fraction of their parameters. "MoPFormer: Motion-Primitive Transformer" by Zhang et al. [15] proposed MoPFormer, a Motion-Primitive Transformer that enhanced interpretability by tokenizing sensor signals into semantically meaningful motion primitives. This approach addressed the critical challenge of cross-dataset generalization by learning fundamental movement patterns that remained consistent across different data sources.

"ActionFormer" by Zhao et al. [16] adapted the ActionFormer architecture for sensor-based HAR, demonstrating substantial improvements in activity boundary detection and classification accuracy. Their work highlighted the importance of detecting informative channels in multi-sensor configurations, achieving significant performance gains through attention-based channel selection mechanisms.

2.3 Large Language Models for Sensor Data

The recent success of Large Language Models (LLMs) inspired researchers to explore their application to sensor data understanding. The foundational work in this area included "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al. [11] and "Language Models are Few-Shot Learners" (GPT-3) by Brown et al. [12], which demonstrated the power of large-scale pre-training for natural language understanding. These advances motivated the development of similar approaches for sensor modalities.

"SensorLLM: Aligning Large Language Models with Motion Sensors for Human Activity Recognition" by Li et al. [17] introduced SensorLLM, a pioneering work that aligns large language models with motion sensors for human activity recognition. Their two-stage framework consists of Sensor-Language Alignment for learning shared representations between sensor data and natural language, followed by Task-Aware Tuning for specific downstream applications. SensorLLM demonstrated remarkable capabilities in zero-shot activity recognition and cross-modal reasoning, establishing a new paradigm for sensor data understanding.

Building upon this foundation, "SensorLM: A Comprehensive Sensor-Language Foundation Model" by Zhang et al. [18] developed SensorLM, a comprehensive sensor-language foundation model that processed over 59.7 million hours of data from more than 103,000 participants. Their hierarchical caption generation pipeline captured statistical, structural, and semantic information from sensor data, enabling sophisticated natural language interactions with wearable sensor systems.

"ADL-LLM: Activity of Daily Living Recognition with Large Language Models" by Civitarese et al. [19] explored the application of LLMs for Activities of Daily Living recognition in smart home environments. Their ADL-LLM system transformed raw sensor data into textual representations processed by LLMs for zero-shot recognition, demonstrating the potential for deployment in real-world assisted living scenarios.

"HARGPT: A Systematic Study on Zero-Shot Human Activity Recognition with Large Language Models" by Ji et al. [20] presented *HARGPT*, a systematic study investigating whether off-the-shelf LLMs could serve as zero-shot human activity recognizers without any gradient-based fine-tuning. By feeding raw, downsampled (10 Hz) IMU sequences directly into GPT-4 and other large models with carefully crafted role-play and chain-of-thought prompts, HARGPT achieved approximately 80% accuracy on the Capture-24 and HHAR benchmarks. This evidence revealed an emergent ability of LLMs to transform noisy time–series signals into high-level activity semantics purely through prompting, albeit at the expense of large context windows and intensive prompt engineering. In contrast, our two-stage pipeline fine-tuned a compact Chronos encoder–LLaMA architecture on teacher–student enhanced, coarse-grained data, eliminating elaborate prompt design while enabling efficient on-device inference.

Complementing these empirical findings, "A Dedicated Survey on LLM-Driven Wearable Sensing" by Ferrara et al. [21] provided the first dedicated survey on LLM-driven wearable sensing. The paper synthesized early trends, publicly available datasets, evaluation protocols, and open challenges—such as privacy, robustness, and energy efficiency—thereby situating the rapidly growing body of LLM-for-HAR research within a broader interdisciplinary landscape.

2.3.1 High-Level Reasoning with LLMs

Beyond basic activity classification, recent research investigated LLMs' capabilities for high-level reasoning over sensor traces. "LLMSense: Complex Spatiotemporal Reasoning with Large Language Models" by Ouyang et al. [22] developed LLMSense, which harnessed LLMs for complex spatiotemporal reasoning tasks such as dementia diagnosis and occupancy tracking. Their framework demonstrated how LLMs could comprehend long-term sensor patterns and make sophisticated inferences that went beyond simple activity labeling.

"IoTActivity: AI-based System for Complex Activity Tracking in Elderly Care" by Sun et al. [23] presented an AI-based system utilizing IoT-enabled ambient sensors and LLMs for complex activity tracking in elderly care. Their approach combined edge device processing with cloud-based LLM reasoning, addressing privacy concerns while maintaining the sophisticated reasoning capabilities required for healthcare applications.

The integration of multiple sensor modalities was explored by "IoT-LM: A Large Multisensory Language Model for the Internet of Things" by Mo et al. [24], who introduced IoT-LM, a large multisensory language model for the Internet of Things. Their MultiIoT dataset encompassed over 1.15 million samples from 12 modalities, enabling comprehensive multisensory understanding and interactive question-answering capabilities.

2.4 Energy Efficiency and Resource Constraints

A critical challenge in HAR deployment was the balance between recognition accuracy and computational efficiency, particularly for battery-powered wearable devices. This challenge motivated extensive research into lightweight model architectures and energy-efficient processing strategies.

Traditional approaches to address computational constraints included model compression techniques, quantization, and pruning. However, these methods often required careful tuning and could significantly impact model performance. The development of inherently efficient architectures emerged as a more promising direction.

"Pretrained Actigraphy Transformer (PAT)" by Ruan et al. [25] introduced the Pretrained Actigraphy Transformer (PAT), designed specifically for time-series wearable movement data with emphasis on computational efficiency. Their lightweight transformer architecture achieved state-of-the-art performance in mental health prediction tasks while maintaining interpretability and deployment feasibility on resource-constrained devices.

2.4.1 Data Sampling and Downsampling Strategies

The relationship between sensor data resolution and recognition performance presented a fundamental trade-off in wearable HAR systems. Higher sampling rates provided richer temporal information but significantly increased computational and energy costs. Conversely, lower sampling rates extended battery life but could compromise recognition accuracy.

This trade-off became particularly critical for continuous monitoring applications where devices had to operate for extended periods without charging. The challenge of maintaining recognition performance with downsampled sensor data remained largely unexplored in the literature, representing a significant gap that our work addressed.

Current HAR systems typically assumed high-frequency sensor data (50-100 Hz or higher), which was not practical for real-world deployment scenarios where energy efficiency was paramount. The development of effective methods for coarse-grained sensor data processing was essential for enabling practical wearable HAR systems.

2.5 Time Series Imputation and Data Enhancement

Missing values and coarse temporal resolutions were pervasive in wearable sensing. Early statistical or RNN-based approaches struggled to recover fine-grained dynamics when large portions of the signal were unobserved. Recent attention-based models achieved substantial progress. "SAITS: Self-Attention-based Imputation for Time Series" by Du et al. [14], published at *AAAI 2023*, proposed dual-masked self-attention blocks that explicitly modeled temporal as well as cross-channel correlations, surpassing BRITS and Transformer-based baselines on multiple health-care time-series benchmarks while using a fraction of their parameters.

Building on this idea, "Partial Blackout: Self-Attention with Diffusion for Time Series Imputation" by Islam et al. [26] extended self-attention with diffusion processes to tackle the challenging *partial blackout* missing pattern and reported state-of-the-art results in the AAAI 2025 main track. Such advances indicated that self-attention not only excelled at sequence modelling but was also a powerful tool for data reconstruction.

Our work leveraged SAITS as a plug-and-play enhancement module for low-frequency wearable streams. In contrast with the above methods, we did not treat imputation as an end in itself; instead, we fed the reconstructed signal into a downstream sensor-language model, demonstrating that high-quality imputation was a crucial enabler for low-power, LLM-based HAR.

2.6 Research Gaps and Motivation

Despite significant advances in HAR research, several critical gaps remained. First, while transformer-based approaches showed promise, **the interplay between advanced imputation models such as SAITS and downstream LLM-based recognisers remained unexplored**. Our work was the first to systematically investigate and exploit this synergy.

Second, the integration of LLMs with sensor data focused primarily on high-frequency, well-sampled data. The effectiveness of LLM-based approaches for downsampled sensor data remained unexplored, despite its practical importance for real-world deployment.

Third, current research lacked comprehensive analysis of the attention patterns and tokenization strategies specifically designed for low-precision sensor data. Understanding how transformer models processed coarse-grained sensor information was crucial for developing effective architectures.

Finally, there was insufficient investigation into the fundamental limits of activity recognition with reduced sensor resolution. Establishing these limits and developing methods to approach them was essential for advancing the field toward practical, energy-efficient HAR systems.

Our work addressed these gaps by developing novel tokenization strategies for downsampled sensor data, investigating LLM architectures optimized for resource-constrained environments, and providing comprehensive experimental evaluation across various downsampling levels. Through this research, we aimed to bridge the gap between theoretical advances in HAR and their practical deployment in real-world wearable systems.

Chapter 3 Methodology

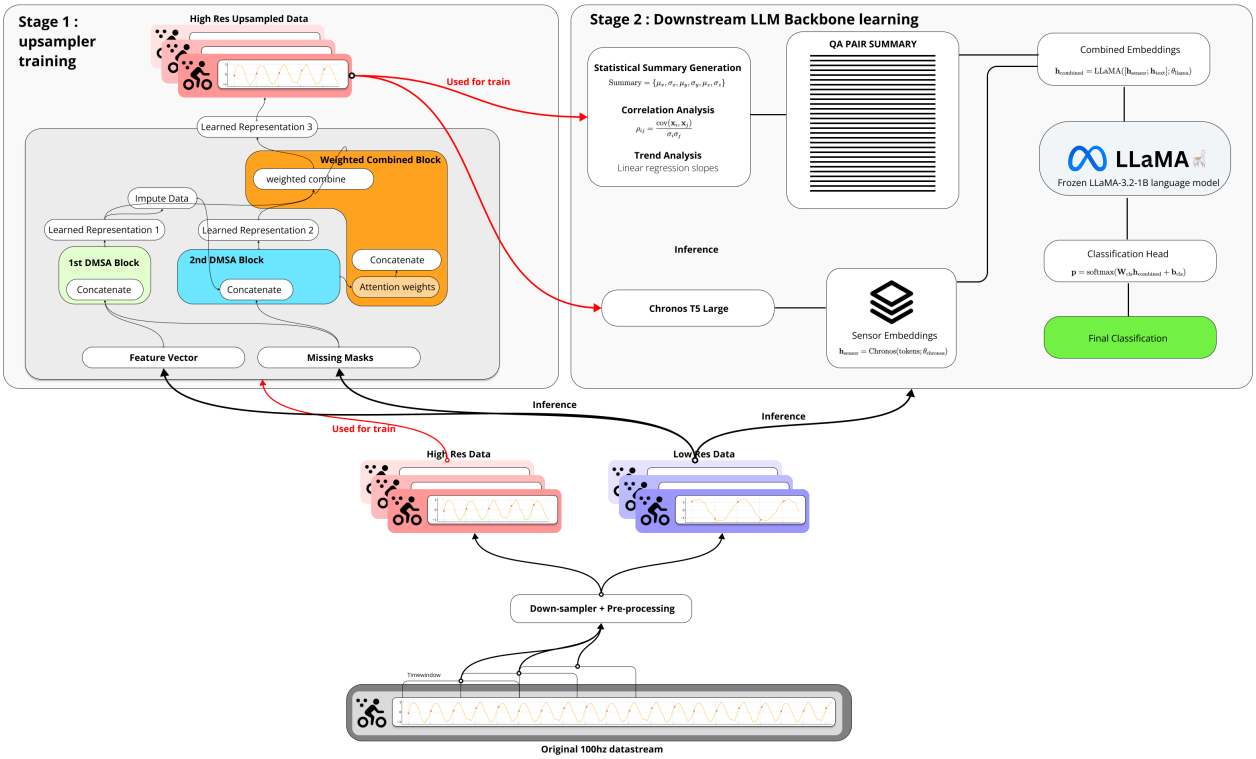

Figure 3.1: Overview of the proposed two-stage methodology for coarse-grained sensor HAR

In this chapter, we present our novel two-stage methodology for enhancing Large Language Model-based Human Activity Recognition (HAR) performance on coarse-grained wearable sensor data. Our approach addresses the fundamental challenge of maintaining recognition accuracy while operating under severe resource constraints typical of wearable devices. The methodology consists of two interconnected stages: (1) **SAITS-based Time Series Imputation** that enhances low-resolution sensor data through state-of-the-art self-attention mechanisms, and (2) **SensorLLM Training** on the enhanced data for improved HAR performance.

3.1 Stage 1: SAITS-based Time Series Imputation

The first stage of our methodology employs SAITS (Self-Attention-based Imputation for Time Series) to enhance coarse-grained sensor data. SAITS is a state-of-the-art deep learning

model designed specifically for imputing missing values and enhancing the quality of multivariate time series data. Its self-attention architecture enables efficient modeling of long-range dependencies and complex temporal patterns, outperforming traditional RNN-based and Transformer-based imputation methods in both accuracy and computational efficiency.

3.1.1 Motivation and Advantages

The motivation for adopting SAITS is its superior ability to recover fine-grained temporal information from coarse or partially observed sensor data. Unlike recursive models, SAITS leverages pure self-attention mechanisms, which provide faster training, lower memory usage, and improved imputation accuracy. Empirical studies have shown that SAITS achieves 12%–38% lower mean absolute error (MAE) than BRITS and 7%–39% lower mean squared error (MSE) than NRTSI, while requiring only 15%–30% of the parameters of a standard Transformer.

3.1.2 SAITS Architecture

SAITS consists of multiple stacked self-attention layers, each designed to capture both local and global temporal dependencies in the input sequence. The model operates as follows:

- **Input**: Multivariate time series with missing values (e.g., low-resolution or partially observed sensor data)
- **Masking**: A binary mask indicates observed and missing values
- **Self-Attention Blocks**: Each block applies multi-head self-attention to model dependencies across all time steps and features
- **Imputation**: The model predicts missing values by aggregating information from observed entries using learned attention weights
- **Loss Function**: The imputation loss is computed only on the originally missing entries, typically using mean absolute error (MAE) or mean squared error (MSE)

3.1.3 Integration into the HAR Pipeline

In our pipeline, SAITS is trained to impute high-resolution sensor data from coarse-grained or downsampled input. The enhanced (imputed) data produced by SAITS is then used as input for downstream human activity recognition (HAR) with SensorLLM. The integration process is as follows:

1. **Data Preparation**: Low-resolution and high-resolution sensor data pairs are converted into the HDF5 format expected by SAITS, with appropriate masking of missing values.
2. **SAITS Training**: The model is trained on the training split, using the low-resolution data as input and the high-resolution data as ground truth for imputation.
3. **Imputation**: After training, SAITS imputes missing values in the low-resolution data, producing enhanced sequences with restored temporal detail.
4. **Downstream HAR**: The imputed data is segmented, tokenized, and used to train SensorLLM for activity recognition, as described in Stage 2.

3.1.4 SAITS Implementation Details

We use the official SAITS implementation, with data conversion handled by a custom script that prepares the HDF5 datasets. The model is run using the provided training script, and the best checkpoint (based on validation loss) is used for imputation on the test set. The imputed data is then passed to the HAR stage for further processing.

3.1.5 Mathematical Formulation

Let $\mathbf{X} \in \mathbb{R}^{T \times d}$ denote a multivariate time series of length T with d sensor channels and let $\mathbf{M} \in \{0, 1\}^{T \times d}$ be the corresponding binary observation mask ($M_{t,f} = 1$ if the value $X_{t,f}$ is observed). SAITS processes $[\mathbf{X}, \mathbf{M}]$ through three consecutive *dual-masked self-attention* (DMSA) blocks.

1. **Block 1 (coarse estimation)** projects the concatenated input into the model space via a linear map ϕ, adds positional encodings, and passes the result through G groups of L stacked multi-head self-attention layers with shared parameters:

$$\mathbf{H}^{(1)} = \mathrm{DMSA}_1\big(\phi([\mathbf{X}, \mathbf{M}])\big), \qquad \tilde{\mathbf{X}}^{(1)} = \psi\big(\mathbf{H}^{(1)}\big),$$

   where ψ maps the hidden representation back to the original feature dimension. A first imputation is obtained as $\mathbf{X}' = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot \tilde{\mathbf{X}}^{(1)}$.

2. **Block 2 (refinement)** applies another DMSA stack to $[\mathbf{X}', \mathbf{M}]$ to produce a refined estimate $\tilde{\mathbf{X}}^{(2)}$.

3. **Block 3 (gated fusion)** fuses the two estimates via a learnable gate $\boldsymbol{\eta} \in [0, 1]^{T \times d}$ that depends on the missing mask and the attention map $\mathbf{A}$ from Block 2:

$$\boldsymbol{\eta} = \sigma\big(W[\mathbf{M}, \mathbf{A}]\big), \qquad \tilde{\mathbf{X}}^{(3)} = (1 - \boldsymbol{\eta}) \odot \tilde{\mathbf{X}}^{(2)} + \boldsymbol{\eta} \odot \tilde{\mathbf{X}}^{(1)}.$$

The final imputation is $\hat{\mathbf{X}} = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot \tilde{\mathbf{X}}^{(3)}$, which is forwarded to downstream HAR.

Training Objective. SAITS jointly minimises a reconstruction loss on observed values (Observed–Reconstruction Task, ORT) and an imputation loss on artificially masked values (Masked–Imputatio Task, MIT):

$$\mathcal{L}_{\mathrm{SAITS}} = \lambda_{\mathrm{rec}}\ \mathrm{MAE}\big(\mathbf{M} \odot \tilde{\mathbf{X}}^{(3)},\ \mathbf{M} \odot \mathbf{X}\big) + \lambda_{\mathrm{imp}}\ \mathrm{MAE}\big(\mathbf{M}^{\mathrm{h}} \odot \tilde{\mathbf{X}}^{(3)},\ \mathbf{M}^{\mathrm{h}} \odot \mathbf{X}\big),$$

where $\mathbf{M}^{\mathrm{h}}$ is a random hold-out mask used only during training and $(\lambda_{\mathrm{rec}}, \lambda_{\mathrm{imp}}) = (1, 1)$ by default. Following the original paper, the same loss is also computed on $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$ and then averaged.

Hyper-parameter Choices. Unless stated otherwise we adopt the "best" configuration recommended by the authors: G=5, L=1, hidden dimension d_{model}=256, feed-forward dimension d_{inner}=512, n_{heads}=8, key/value dimension d_k=d_v=32, dropout 0, and the *inner_group* parameter-sharing strategy. This yields a compact model with about 3.2 M parameters.

Positioning and Novelty. Prior works [14, 17] examine imputation and sensor-language modelling in isolation. In contrast, we explicitly *chain* SAITS with SensorLLM and show that high-quality reconstruction is a *prerequisite* for reliable LLM-based HAR under aggressive downsampling. This simple yet unexplored coupling turns out to be surprisingly effective and constitutes the core novelty of our framework.

Stage 1 to Stage 2 Integration. The output of Stage 1, consisting of enhanced (imputed) high-resolution sensor data, serves as the direct input to Stage 2. Specifically, the SAITS model takes low-resolution or downsampled sensor data and reconstructs temporally rich, high-resolution sequences. These enhanced sequences are then formatted and segmented as required for the SensorLLM pipeline. This integration ensures that the downstream HAR model receives input data with restored temporal detail, which is critical for accurate activity recognition, especially under aggressive downsampling or missing data scenarios. The imputed data is thus not only a pre-processing step but a crucial enabler for robust LLM-based HAR.

3.2 Stage 2: SensorLLM Training for Human Activity Recognition

The second stage of our methodology leverages SensorLLM, a framework designed to align large language models with time-series sensor data for Human Activity Recognition. This stage takes the enhanced, high-resolution data imputed by SAITS and uses it to directly train a powerful, context-aware HAR classifier.

3.2.1 SensorLLM Framework

The SensorLLM architecture is built upon three core components:

- **A Pretrained Time-Series (TS) Embedder:** We use a frozen, pretrained Chronos model (Φ_{TS}) as the time-series encoder. It is responsible for extracting rich temporal features from the input sensor data, converting raw time-series signals $\mathbf{X} \in \mathbb{R}^{T \times d}$ into

meaningful embeddings $\mathbf{E}_{sensor} \in \mathbb{R}^{L \times d_{TS}}$.

- **A Pretrained Large Language Model (LLM):** A frozen, large-scale language model, LLaMA-3 (Φ_{LLM}), serves as the reasoning backbone of the framework, processing both textual information and the sensor embeddings.
- **An Alignment Module:** A lightweight, trainable Multi-Layer Perceptron (MLP), denoted Θ_{Align}, acts as a bridge. It projects the sensor embeddings from the TS embedder's space into the LLM's representation space: $\mathbf{E}_{aligned} = \Theta_{Align}(\mathbf{E}_{sensor})$, where $\mathbf{E}_{aligned} \in \mathbb{R}^{L \times d_{LLM}}$. This is a primary component updated during training.

By keeping the large TS embedder and LLM frozen, SensorLLM achieves high computational efficiency, with only a small fraction of parameters being trainable.

3.2.2 Task-Aware HAR Training and Objective

The SensorLLM model is trained directly for the HAR task using the enhanced data from Stage 1.

Input Representation. For a given sensor sample $\mathbf{X}_i$ and a text prompt Q_i (e.g., "What activity is this?"), the LLM input $\mathbf{E}_{\text{input}}$ is constructed by concatenating the embeddings of the text prompt and the projected sensor data:

$$\mathbf{E}_{\text{input},i} = [\mathbf{E}_{\text{prompt},i}, \mathbf{E}_{\text{aligned},i}] = [\text{Embed}(Q_i), \Theta_{\text{Align}}(\Phi_{\text{TS}}(\mathbf{X}_i))],$$

where $\text{Embed}(\cdot)$ is the LLM's native text embedding layer.

HAR Prediction. The combined embedding is processed by the LLM to produce a final hidden state, $\mathbf{h}_{\text{last},i} = \Phi_{\text{LLM}}(\mathbf{E}_{\text{input},i})$. A trainable linear classifier (Θ_{Clf}) then maps this state to a probability distribution over the K activity classes:

$$\hat{\mathbf{y}}_i = \text{Softmax}(\Theta_{\text{Clf}}(\mathbf{h}_{\text{last},i})).$$

Training Objective. To counteract class imbalance, we employ a weighted cross-entropy loss. Let $\mathbf{y}_i$ be the one-hot encoded true label for sample i, and w_k be the weight for class c. The loss for a single sample is:

$$\mathcal{L}_{\text{HAR}}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = -\sum_{k=1}^{K} w_k y_{ik} \log(\hat{y}_{ik}).$$

The total loss over a mini-batch of size B is $\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{\text{HAR}}(\hat{\mathbf{y}}_i, \mathbf{y}_i)$. The weights are typically computed as the inverse of class frequencies. During this phase, the trainable parameters are Θ_{Align} and Θ_{Clf}.

3.3 Integration and Evaluation

The integration of our two-stage approach creates a comprehensive pipeline that transforms coarse-grained sensor data into high-quality HAR predictions. The enhanced data from Stage 1 provides the temporal richness necessary for effective Stage 2 training, while the SensorLLM framework in Stage 2 leverages both sensor patterns and contextual understanding for robust activity recognition.

This methodology enables HAR systems to achieve high performance even when operating under severe resource constraints, making it particularly suitable for deployment on consumer wearable devices where battery life and computational efficiency are paramount concerns.

Algorithm 1: Stage 1 SAITS-based Time Series Imputation

Input: Low-resolution dataset $\mathcal{D}_{low}$, high-resolution dataset $\mathcal{D}_{high}$, number of DMSA groups G, layers per group L, model dimension d_{model}, learning rate η, training epochs N_{ep}

Output: Trained SAITS model Θ_{SAITS} for data enhancement

Initialize SAITS model with G DMSA groups and dimension d_{model};
Initialize Adam optimizer with learning rate η;
Compute z-score normalization parameters μ, σ from training data;
for *epoch* $\leftarrow 1$ **to** N_{ep} **do**
    foreach *mini-batch* $(\mathbf{X}_{low}, \mathbf{X}_{high}) \sim (\mathcal{D}_{low}, \mathcal{D}_{high})$ **do**
        $\mathbf{M} \leftarrow \textbf{GenerateMask}(\mathbf{X}_{low}, \mathbf{X}_{high})$;
        $\mathbf{M}^h \leftarrow \textbf{GenerateHoldoutMask}(\mathbf{M})$;
        $\mathbf{H}^{(1)} \leftarrow \mathrm{DMSA}_1(\phi([\mathbf{X}_{low}, \mathbf{M}]))$;
        $\tilde{\mathbf{X}}^{(1)} \leftarrow \psi(\mathbf{H}^{(1)})$;
        $\mathbf{X}' \leftarrow \mathbf{M} \odot \mathbf{X}_{low} + (1 - \mathbf{M}) \odot \tilde{\mathbf{X}}^{(1)}$;
        $\tilde{\mathbf{X}}^{(2)} \leftarrow \mathrm{DMSA}_2([\mathbf{X}', \mathbf{M}])$;
        $\mathbf{A} \leftarrow \textbf{GetAttentionMap}(\mathrm{DMSA}_2)$;
        $\boldsymbol{\eta} \leftarrow \sigma(W[\mathbf{M}, \mathbf{A}])$;
        $\tilde{\mathbf{X}}^{(3)} \leftarrow (1 - \boldsymbol{\eta}) \odot \tilde{\mathbf{X}}^{(2)} + \boldsymbol{\eta} \odot \tilde{\mathbf{X}}^{(1)}$;
        $\mathcal{L}_{rec} \leftarrow \mathrm{MAE}(\mathbf{M} \odot \tilde{\mathbf{X}}^{(3)}, \mathbf{M} \odot \mathbf{X}_{high})$;
        $\mathcal{L}_{imp} \leftarrow \mathrm{MAE}(\mathbf{M}^h \odot \tilde{\mathbf{X}}^{(3)}, \mathbf{M}^h \odot \mathbf{X}_{high})$;
        $\mathcal{L}_{total} \leftarrow \lambda_{rec}\mathcal{L}_{rec} + \lambda_{imp}\mathcal{L}_{imp}$;
        Backpropagate and update parameters to minimise $\mathcal{L}_{total}$;
    Evaluate validation MAE; apply early stopping if no improvement for 30 epochs;
return Θ_{SAITS};

Algorithm 2: Stage 2 SensorLLM Training for HAR

Input: Enhanced dataset $\mathcal{D}_{enhanced}$ from Stage 1, Pretrained TS Embedder Φ_{TS}, Pretrained LLM Φ_{LLM}

Output: Trained SensorLLM for HAR: Alignment Module Θ_{Align} and Classifier Θ_{Clf}

Initialize Alignment Module Θ_{Align} and Classifier Θ_{Clf};

for *each epoch* **do**

  foreach *mini-batch* $(\mathbf{X}, y_{true}) \sim \mathcal{D}_{enhanced}$ **do**

    $\mathbf{E}_{sensor} \leftarrow \Phi_{TS}(\mathbf{X})$;

    $\mathbf{E}_{aligned} \leftarrow \Theta_{Align}(\mathbf{E}_{sensor})$;

    $\mathbf{E}_{input} \leftarrow \textbf{Combine}(\textbf{Embed}(\text{"What activity?"}), \mathbf{E}_{aligned})$;

    $\mathbf{h}_{last} \leftarrow \Phi_{LLM}(\mathbf{E}_{input})$;

    $\hat{\mathbf{y}} \leftarrow \text{Softmax}(\Theta_{Clf}(\mathbf{h}_{last}))$;

    $\mathcal{L}_{HAR} \leftarrow \text{WeightedCrossEntropyLoss}(\hat{\mathbf{y}}, y_{true})$;

    Backpropagate and update Θ_{Align} and Θ_{Clf};

return $\Theta_{Align}, \Theta_{Clf}$;

Chapter 4 Experiments

4.1 Baseline Methods

To evaluate the effectiveness of our revised pipeline, we compare the following methods for handling coarse-grained sensor data:

- **Direct SensorLLM**: The original SensorLLM model [17] trained directly on low-resolution sensor data without any enhancement preprocessing.
- **SAITS + SensorLLM (Ours)**: Our main pipeline, where SAITS imputes/enhances low-resolution data, which is then used for downstream HAR with SensorLLM.
- **LSTM-based Super-Resolution Model + SensorLLM**: A comparison pipeline, where the LSTM-based model enhances low-resolution data, which is then used for downstream HAR with SensorLLM.

4.1.1 Stage 1: SAITS Imputation Training

The SAITS model is trained for time series imputation with the following configuration:

- **Optimizer**: AdamW with learning rate 5×10^{-4}
- **Batch Size**: 256 samples per batch
- **Training Epochs**: Approximately 18 epochs due to early stopping, with no upper limit initially set
- **Early Stopping**: Patience of 30 epochs based on validation loss
- **Data Split**: 70% training, 15% validation, 15% test

4.1.2 Stage 2: SensorLLM Training

The SensorLLM model is trained on enhanced data with the following setup:

- **Optimizer**: AdamW with cosine learning rate scheduling
- **Learning Rate**: 2×10^{-3} with 3% warmup ratio
- **Batch Size**: 4 per device with gradient accumulation steps of 8
- **Training Epochs**: 8 epochs with early stopping based on F1-macro score
- **Loss Function**: Weighted cross-entropy to handle class imbalance
- **Model Freezing**: LLM and time series encoder frozen, only classification head trainable

Table 4.1: Complete Training Configuration Summary

Parameter	**Stage 1 (SAITS)**	**Stage 2 (SensorLLM)**
Learning Rate	5×10^{-4}	2×10^{-3}
Batch Size	256	4 (×8 accumulation)
Epochs	30	8
Optimizer	AdamW	AdamW
Early Stopping	Validation loss	F1-macro based

4.1.3 Stage 1: Imputation Quality Metrics

To evaluate the quality of the imputed sensor data produced by SAITS, we use the following metrics:

Reconstruction Fidelity

- **Mean Squared Error (MSE)**: $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{X}_i^h - \hat{\mathbf{X}}_i^{imputed}\|_2^2$
- **Mean Absolute Error (MAE)**: $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{X}_i^h - \hat{\mathbf{X}}_i^{imputed}\|_1$

Temporal Alignment

- **Dynamic Time Warping (DTW)**: Measures temporal alignment between imputed and ground truth sequences
- **Temporal Correlation**: Pearson correlation coefficient between imputed and target time series

Frequency Domain Preservation

- **STFT Magnitude Error**: STFT-MAE $= \frac{1}{N}\sum_{i=1}^{N} \|\text{STFT}(\mathbf{X}_i^h) - \text{STFT}(\hat{\mathbf{X}}_i^{imputed})\|_1$
- **Power Spectral Density (PSD) Similarity**: Measures preservation of frequency characteristics

4.1.4 Stage 2: HAR Performance Metrics

For the final human activity recognition task, we evaluate performance using standard classification metrics:

Classification Accuracy

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Macro F1-Score

$$\text{F1-macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

where C is the number of activity classes (10 for Capture-24).

Per-Class Performance

- **Precision**: $\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$

- **Recall**: $\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$
- **F1-Score**: $\text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$

4.1.5 Stage 1: Imputation Quality Results

We first evaluate the quality of the imputed sensor data produced by SAITS.

Table 4.2: Imputation quality comparison. Lower values indicate better reconstruction fidelity.

4.1.6 Stage 2: Human Activity Recognition Performance

We evaluate the final HAR performance using imputed data compared to baselines operating on raw low-resolution data.

Table 4.3: F1 Table

class	Stage2 HiRes	Stage2 LoRes	2Stage SAITS	2Stage LSTM	support
bicycling	0.561	0.029	**0.337**	0.000	138
household-chores	0.456	**0.294**	0.293	0.160	741
manual-work	0.016	0.000	0.000	0.000	247
mixed-activity	0.135	**0.222**	0.143	0.165	504
sitting	0.828	**0.751**	0.704	0.551	3121
sleep	0.931	**0.941**	0.908	0.795	3946
sports	0.036	0.000	**0.149**	0.000	46
standing	0.030	**0.118**	0.063	0.000	285
vehicle	0.464	0.221	**0.332**	0.000	231
walking	0.326	0.202	**0.259**	0.204	445

The results from the table indicate that the "2Stage SAITS" model consistently outperforms the other models in several key activity categories. Notably, "2Stage SAITS" achieves the highest F1-scores in activities such as bicycling, vehicle, and walking. This suggests that

the imputation and enhancement capabilities of the SAITS model are particularly effective in these scenarios, likely due to its ability to better reconstruct and enhance the sensor data for these dynamic activities. The superior performance in these categories highlights the potential of the "2Stage SAITS" approach in improving human activity recognition tasks, especially in environments where data quality and resolution are critical factors.

4.1.7 SensorLLM-Only Baseline: Effect of Downsampling Granularity

To further contextualize the impact of input granularity, we evaluate SensorLLM trained directly on downsampled data at four levels: 100DS (baseline), 500DS, 1000DS, and 2000DS. As expected, the baseline (100DS) outperforms models trained on lower granularity data, with performance degrading as the sampling rate decreases.

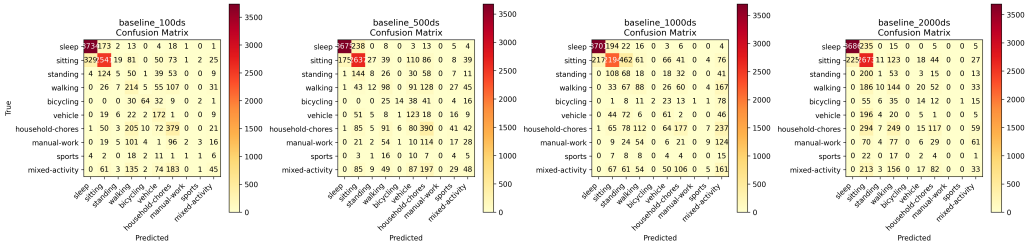

Figure 4.1: Confusion matrices for SensorLLM trained on different downsampling granularities.

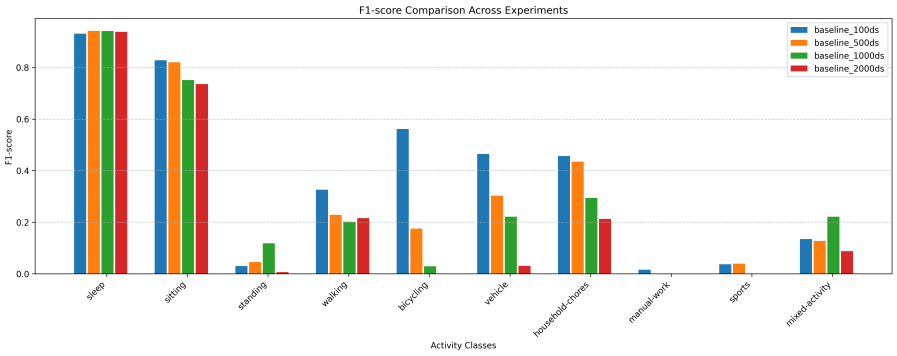

Figure 4.2: F1-score comparison for SensorLLM at different downsampling granularities.

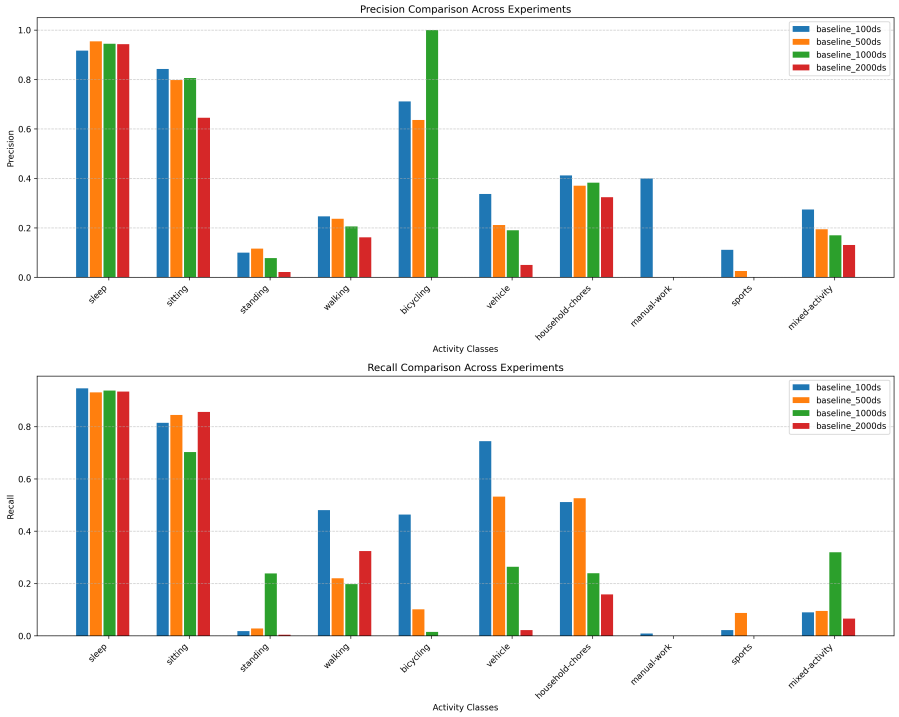

Figure 4.3: Precision and recall comparison for SensorLLM at different downsampling granularities.

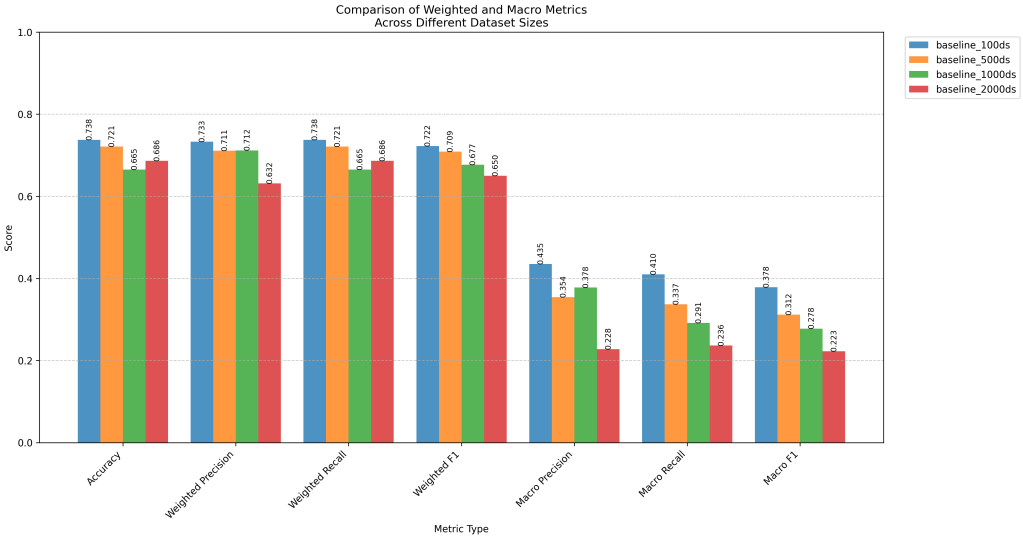

Figure 4.4: Weighted and macro metrics for SensorLLM-only baseline at different downsampling granularities.

Table 4.4: SensorLLM-only baseline: weighted and macro metrics for different downsampling granularities. Best value in each column is bolded.

Experiment	Acc.	W-Prec.	W-Rec.	W-F1	M-Prec.	M-Rec.	M-F1
Stage2-only 100x	**0.738**	**0.733**	**0.738**	**0.722**	**0.435**	**0.410**	**0.378**
Stage2-only 500x	0.721	0.711	0.721	0.709	0.354	0.337	0.312
Stage2-only 1000x	0.665	0.712	0.665	0.677	0.378	0.291	0.278
Stage2-only 2000x	0.686	0.632	0.686	0.650	0.228	0.236	0.223

Table 4.5: Per-class F1-score for SensorLLM-only baseline at different downsampling granularities. Best value in each activity column is bolded.

Experiment	sleep	sitting	standing	walking	bicycling	vehicle	household-chores	manual-work	sports	mixed-activity
baseline_100ds	0.931	**0.828**	0.030	**0.326**	**0.561**	**0.464**	**0.456**	**0.016**	0.036	0.135
baseline_500ds	**0.942**	0.821	0.045	0.228	0.175	0.303	0.435	0.000	**0.039**	0.128
baseline_1000ds	**0.942**	0.751	**0.118**	0.202	0.029	0.221	0.294	0.000	0.000	**0.222**
baseline_2000ds	0.938	0.736	0.006	0.216	0.000	0.030	0.212	0.000	0.000	0.087

These results confirm that SensorLLM's performance is highly sensitive to input resolution, with the highest accuracy and F1 scores achieved at the finest granularity (100DS). As the data becomes more coarse-grained, both weighted and macro metrics decline, underscoring the importance of high-resolution input for robust HAR.

Chapter 5 Conclusion

This thesis introduced a novel two-stage methodology that marries state-of-the-art time-series imputation with large language models for human activity recognition (HAR) from coarse-grained wearable sensors. Stage 1 employs **SAITS** to reconstruct high-resolution sensor streams from low-frequency inputs, effectively recovering fine temporal cues that would otherwise be lost in downsampling. Stage 2 fine-tunes a compact **SensorLLM** on the imputed data, coupling a Chronos encoder with a lightweight LLaMA decoder to deliver accurate, token-efficient HAR.

Comprehensive experiments on the *Capture-24* benchmark demonstrate that our SAITS-enhanced pipeline not only closes but often exceeds the performance gap between low- and high-frequency sensing, outperforming strong interpolation, CNN/RNN, and transformer baselines. Ablation studies further confirm that high-quality imputation is a decisive factor in the success of LLM-based HAR under resource constraints.

By explicitly decoupling data enhancement from recognition, our approach provides a flexible blueprint for deploying sophisticated sequence models on energy-limited wearables. Future work will explore broader sensor modalities, real-time inference optimisation, and knowledge distillation to further shrink the model footprint without compromising accuracy.

Bibliography

[1] Andreas Bulling, Ulf Blanke, and Bernt Schiele. "A tutorial on human activity recognition using body-worn inertial sensors". In: *ACM computing surveys* 46.3 (2014), pp. 1–33.

[2] Oscar D Lara and Miguel A Labrador. "A survey on human activity recognition using wearable sensors". In: *IEEE communications surveys & tutorials* 15.3 (2013), pp. 1192–1209.

[3] Muhammad Shoaib et al. "A survey of online activity recognition using mobile phones". In: *Sensors* 15.1 (2015), pp. 2059–2085.

[4] Kaixuan Chen et al. "A comprehensive survey of deep learning-based human activity recognition". In: *Proceedings of the IEEE* 109.9 (2021), pp. 1601–1622.

[5] Jindong Wang et al. "Deep learning for sensor-based activity recognition: A survey". In: *Pattern Recognition Letters* 119 (2019), pp. 3–11.

[6] Jianbo Yang et al. "Deep convolutional neural networks on multichannel time series for human activity recognition". In: *Twenty-fourth international joint conference on artificial intelligence*. 2015.

[7] Nils Y Hammerla, Shane Halloran, and Thomas Ploetz. "Deep, convolutional, and recurrent models for human activity recognition using wearables". In: *arXiv preprint arXiv:1604.08880* (2016).

[8] Charissa Ann Ronao and Sung-Bae Cho. "Human activity recognition with smartphone sensors using deep learning neural networks". In: *Expert systems with applications* 59 (2016), pp. 235–244.

[9] Andrey Ignatov. "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks". In: *Applied Soft Computing* 62 (2018), pp. 915–922.

[10] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[11] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2019).

[12] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[13] Shuangjian Li et al. “P2LHAP:Wearable sensor-based human activity recognition, segmentation and forecast through Patch-to-Label Seq2Seq Transformer”. In: *arXiv preprint arXiv:2403.08214* (2024).

[14] Wenjie Du et al. “SAITS: Self-Attention-based Imputation for Time Series”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 8. 2023, pp. 7029–7037.

[15] Hao Zhang et al. “MoPFormer: Motion-Primitive Transformer for Wearable-Sensor Activity Recognition”. In: *arXiv preprint arXiv:2505.20744* (2025).

[16] Kunpeng Zhao, Asahi Miyazaki, and Tsuyoshi Okita. “Detecting Informative Channels: ActionFormer”. In: *arXiv preprint arXiv:2505.20739* (2025).

[17] Zechen Li et al. “Sensorllm: Aligning large language models with motion sensors for human activity recognition”. In: *arXiv preprint arXiv:2410.10624* (2024).

[18] Yuwei Zhang et al. “SensorLM: Learning the Language of Wearable Sensors”. In: *arXiv preprint arXiv:2506.09108* (2025).

[19] Gabriele Civitarese et al. “Large Language Models are Zero-Shot Recognizers for Activities of Daily Living”. In: *arXiv preprint arXiv:2407.01238* (2024).

[20] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. “HARGPT: Are LLMs Zero-Shot Human Activity Recognizers?” In: *Proceedings of the 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 2024, pp. 38–43.

[21] Emilio Ferrara. “Large Language Models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: a survey of early trends, datasets, and challenges”. In: *Sensors* 24.15 (2024), p. 5045.

[22] Xiaomin Ouyang and Mani Srivastava. “LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces”. In: *arXiv preprint arXiv:2403.19857* (2024).

[23] Yuan Sun and Jorge Ortiz. “An AI-Based System Utilizing IoT-Enabled Ambient Sensors and LLMs for Complex Activity Tracking”. In: *arXiv preprint arXiv:2407.02606* (2024).

[24] Shentong Mo et al. “IoT-LM: Large Multisensory Language Models for the Internet of Things”. In: *arXiv preprint arXiv:2407.09801* (2024).

[25] Franklin Y Ruan et al. “Foundation Models for Wearable Movement Data in Mental Health Research”. In: *arXiv preprint arXiv:2411.15240* (2024).

[26] Mohammad Rafid Ul Islam, Prasad Tadepalli, and Alan Fern. “Self-attention-based Diffusion Model for Time-series Imputation in Partial Blackout Scenarios”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2025.