IMUZero: Zero-Shot Human Activity Recognition by Language-Based Cross Modality Fusion

JIE SU, Zhejiang University of Technology, China
FENGTONG GE, Zhejiang University of Technology, China
ZHENYU WEN*, Zhejiang University of Technology, China
TAOTAO LI, Zhejiang University of Technology, China
YANG BAI, Institute of High-Performance Computing (IHPC), ASTAR, Singapore
YEJIAN ZHOU, Zhejiang University of Technology, China
XIAOQIN ZHANG, Zhejiang University of Technology, China

Wearable-based human activity recognition (HAR) typically uses motion sensor data, such as inertial measurement unit (IMU) signals, to identify human movements. While effective in controlled scenarios, traditional HAR models are trained on a fixed set of activities and fail to generalize to new or unseen actions. This limitation motivates the use of zero-shot learning (ZSL), which aims to recognize unseen activities without direct training examples. Existing ZSL methods often rely on projecting seen and unseen classes into a shared latent space using external semantic information, such as visual or textual data. However, visual data are commonly unavailable in wearable settings, and text-based semantics from activity labels or coarse descriptions lack the detail needed for accurate recognition. Recent work explores large language models (LLMs) to provide prior knowledge through question-answering mechanisms. While promising, these approaches do not use raw sensor data directly and often miss important contextual signals. We propose *IMUZero* , a ZSL framework that fuses sensor signals with LLM-generated semantic attributes. Our method uses LLMs to produce fine-grained, decomposable activity attributes without additional LLM-based training, preserving sensor context. We also introduce a channel shuffle order constraint that models axial bias to improve generalization. Experiments on four public datasets show that our method outperforms existing ZSL approaches that rely on learned semantic embeddings. We release the code at https://github.com/Was-Lab/IMUZero.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

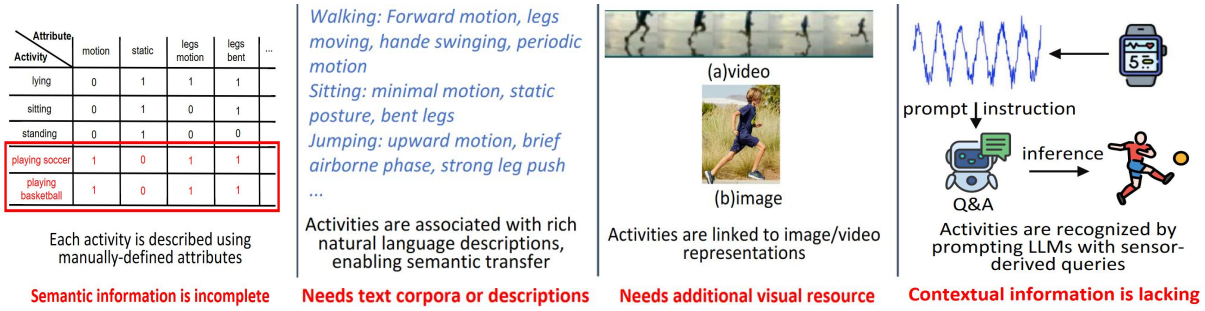
Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Jie Su, Fengtong Ge, Zhenyu Wen, Taotao Li, Yang Bai, Yejian Zhou, and Xiaoqin Zhang. 2018. *IMUZero*: Zero-Shot Human Activity Recognition by Language-Based Cross Modality Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1 (September 2018), 28 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding Author

Authors' Contact Information: Jie Su, jiesu@zjut.edu.cn, Zhejiang University of Technology, Hangzhou, China; Fengtong Ge, Zhejiang University of Technology, Hangzhou, China; Zhenyu Wen, Zhejiang University of Technology, Hangzhou, China; Taotao Li, Zhejiang University of Technology, Hangzhou, China; Yang Bai, Institute of High-Performance Computing (IHPC), ASTAR, Singapore; Yejian Zhou, Zhejiang University of Technology, Hangzhou, China; Xiaoqin Zhang, Zhejiang University of Technology, Hangzhou, China.

ACM 2474-9567/2018/9-ART
https://doi.org/XXXXXXX.XXXXXXX



(a) Action–Label Approach[7] (b) Text-Based Approach[24] (c) Visual-Based Approach[25] (d) LLM-Based Approach[17]

Fig. 1. Different Approaches to Zero-shot Inertial HAR.

1 Introduction

Wearable-based human activity recognition (HAR) [10, 19, 23, 29, 33, 46] uses inertial measurement unit (IMU) data to infer human behaviors. Most existing HAR systems are trained on a small set of predefined activities collected in controlled environments. However, human motion in daily life is highly diverse, unstructured, and context-dependent. Manually labeling IMU data for all possible activities is labor-intensive, time-consuming, and often infeasible at scale. This creates a fundamental gap between training data and real-world deployment: models must recognize activities that were never seen during training. Traditional supervised learning cannot address this gap, as it requires labeled examples for each activity class. Zero-shot learning (ZSL) offers a practical solution by enabling models to infer unseen activities based on semantic relationships with known classes. In this way, ZSL extends the applicability of HAR systems to more realistic and dynamic environments without requiring exhaustive annotation.

Fig 1 shows the existing solutions for ZSL for HAR which primarily explored four research directions: (a) *Action–Label Table* encode basic limb movements and body postures as binary indicators, but their semantic coverage remains limited [7, 38]; (b) *Textual descriptions* of joint and limb motions are processed by language models to produce semantic embeddings, which nonetheless depend on extensive annotated corpora [24, 41]; (c) *Visual inputs*, such as images or video frames of performed activities, yield rich feature embeddings via vision models, but require additional visual resources [25]; and (d)*LLM-based methods* leverage prompt engineering and QA mechanisms to infer unseen activities from expert knowledge, but they often lack sufficient contextual grounding and do not operate in an end-to-end fashion, complicating the recognition process. [17, 22].

Key Idea. We are inspired by the knowledge of kinematics [45]: *Human activities are defined by the spatial configuration and motion of specific body segments.* For example, “walking” can be defined by: "high transverse leg movement, high longitudinal leg movement, high coronal leg movement, medium transverse chest movement, medium longitudinal chest movement, high coronal chest movement, low coronal arm movement, low longitudinal arm movement, high coronal arm movement". This paper proposes to use a limited set of body segments—defined as *activity attributes*—that are observable from IMU signals to represent human activities, enabling ZSL for HAR. However, although the designed activity attributes establish a semantic basis, mapping raw signals to semantics remains challenging, as it requires accurate extraction of fine grained patterns and formal alignment. Specifically, building this mapping faces the following challenges:

1) Signal-Semantic Gap: Raw inertial sensor outputs generate continuous, high-dimensional time-series signals riddled with noise, inter-subject variability, and devoid of inherent semantic markers tied to human actions. Furthermore, the noisy, high-dimensional nature of these signals makes it difficult to extract

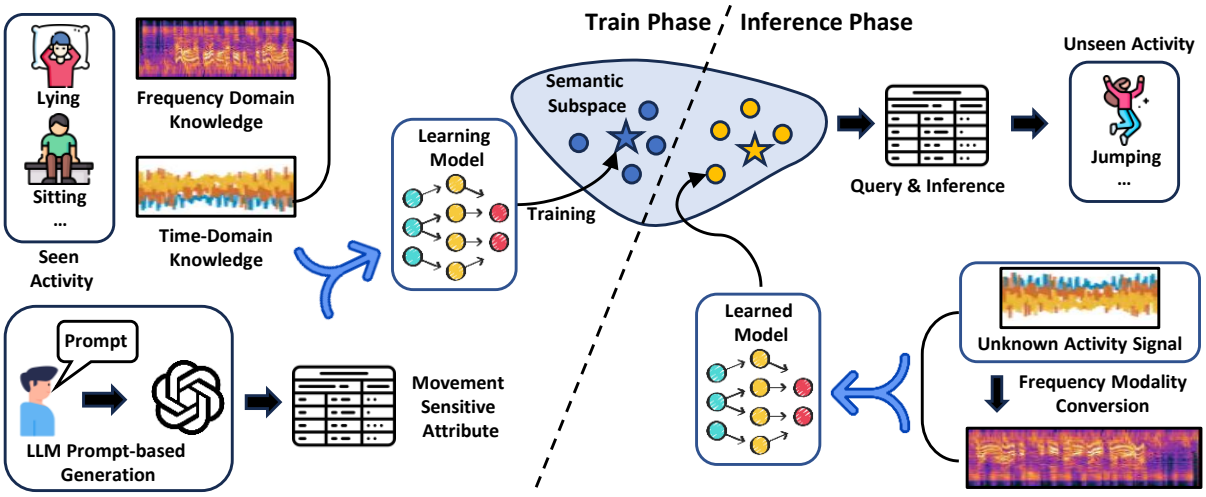


Fig. 2. The workflow of the proposed *IMUZero* framework. During the training phase, the learning model is optimized to construct a semantic bridge between movement-sensitive attributes generated by LLM and the input activity signals. In the inference stage, the unknown input activity signals are first converted into frequency modality and then simultaneously fed into the learned model for the nearest semantic search, resulting in the final inference outputs.

robust representations needed for meaningful alignment. Previous research using activity attribute matrices, textual or visual embeddings, and LLM driven QA has attempted to bridge the semantic gap but offers limited semantic coverage and demands extensive annotations and lack of contextual grounding. Further, these methods rely exclusively on IMU data, which captures time series kinematics but omits the spectral information essential for understanding movement semantics. Consequently, these limitations produce poor semantic mapping and hinder accurate interpretation of complex activities across varied contexts.

2) Axial Bias: The axial bias is introduced when user variations in wearing behavior or activity-induced movement alter sensor alignment, causing shifts in the captured signal. (as demonstrated in sec. 4.3) The discrepancy between subjects in performing activities was neglected in previous studies. [43, 47] These variations can distort intrinsic signal patterns across activities, thereby limiting model generalizability.

Thus, in this work, we present *IMUZero* , a fine-grained, end-to-end zero-shot HAR framework that enhances ZSL by constructing robust projections through a cross-modality fusion mechanism between activity signals and LLM-generated fine-grained, decomposable attributes. We leverage LLMs and their powerful, cognitive-science–inspired prior knowledge to generate reliable and informative semantic attributes, in contrast to conventional naive activity–attribute matrices or simple textual descriptions. Since our method does not rely on direct LLM training or QA inference, it avoids the loss of contextual information. Furthermore, as a first attempt, we encode the IMU data in the frequency domain to provide extensive spectral information that supports alignment with the generated attributes. A fused spatial and spectral representation establishes a robust semantic bridge for subsequent zero-shot tasks. The workflow of the proposed *IMUZero* are demonstrated in Fig. 2. Specifically, the *IMUZero* framework comprises three major components: 1) The **Movement-sensitive Attribute Generation** component takes predefined category information as input to the LLM, facilitating fine-grained and informative semantic attribute generation; 2) The **Multi-Scale Time-Frequency Fusion** module enhances the integration of semantically related information from the input activity signal and its corresponding frequency representation, thereby supporting subsequent fine-grained signal-to-semantic mapping; 3). The **Sig2Text Alignment** module

subsequently receives the fused cross-modality features along with the encoded attribute feature to perform local feature alignment. Additionally, during the feature alignment process, we address the axial bias problem and introduce a channel shuffle constraint to promote channel-invariant (activity-invariant) information extraction, thereby enhancing the model's generalization ability. The main contributions of this paper can be summarized as follows:

- We present *IMUZero* , an end-to-end zero-shot HAR framework that flexibly recognizes unseen and novel classes by bridging the activity signals with the fine-grained activity attributes generated by LLM.
- We present a novel cross-modality multi-stage fusion mechanism to integrate multi-level frequency information, enabling fine-grained semantic alignment between signal and text representations.
- We investigate the 'axis bias' generated by the experimental wearing discrepancy and design a novel channel shuffle order constraint to extract axis-invariant features to improve the recognition generalization ability.
- Extensive experiments were conducted, and we studied our *IMUZero* framework in detail. The promising results suggested its effectiveness.

The rest of this paper is organized as follows. Section 2 introduces the related background knowledge. Section 3&4 presents the problem definition and the details of the proposed *IMUZero* framework. Section 5 gives the experimental settings as well as evaluation results, and Section 6 concludes.

2 Related Work

Human Activity Recognition has a long-standing history in the broader fields of ubiquitous and wearable computing. In the following section, we will review the specific background for this paper, which spans two main subject areas: i) Deep learning for HAR in ubiquitous and wearable computing; ii) Zero-shot Learning; and iii) Zero-Shot Human Activity Recognition.

2.1 Human Activity Recognition

Traditional HAR methods mainly rely on conventional machine-learning approaches such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN), etc [1, 12, 20]. However, a significant limitation of these models is their dependence on hand-crafted features or heuristic information. With the recent surge of deep learning, which can automatically extract features from raw activity signals, the efforts required for feature engineering have been significantly reduced. One of the most popular deep learning models is the convolutional neural network (CNN), which can extract the HAR representation by stacking multiple convolutional layers [44]. Numerous works have investigated the CNN variants [11, 13, 26, 27, 34, 49] by integrating with specific novel network architecture (e.g., Long-Short Term Memory(LSTM) [14], Attention Network [26], etc) to enhance the activity pattern extraction ability. Shao et al. [31] present the ConvBoost framework, which aims to enhance per-epoch training examples through a novel sampling layer designed to improve the model's generalization ability. However, conventional HAR approaches still focus on predefined class assignments, which limits their application in the real world.

2.2 Zero-Shot Learning

Zero-shot Learning[21] is proposed to utilize a semantic modality to connect the visual and label spaces. Early ZSL methods [32, 42] focus on mapping visual and semantic domains to transfer knowledge from seen to unseen classes. These methods extract global visual features from either pre-trained or end-to-end trainable networks, with the latter generally performing better by fine-tuning features and reducing cross-dataset bias. However, they often struggle to capture subtle distinctions between class types. Recent attention-based ZSL methods like DAZLE [16] and Composer [15], use attribute descriptions to identify discriminative features but typically overlook the localization of crucial attributes, focusing instead on broader region embeddings. Chen et al. present

HSVA [5] that applies structure adaptation and distribution adaptation to solve distribution-aligned space shifting problems so to improve generalization ability. Transformer models [36] have shown remarkable performance across various tasks, benefiting from self-supervision and self-attention mechanisms. Chen et al. propose an attribute-guided Transformer, termed TransZero [4], which reduces the entangled relationships among regional features to improve their transferability. The HRT [6] improves the transformer structure by presenting a hybrid of top-down and bottom-up attention pathways to strengthen the modality bridge. Later, the SHIP [39] has been proposed to reconstruct the visual features by inputting the synthesized prompts and the corresponding class names to the textual encoder of LLM. While the aforementioned approaches achieve satisfactory performance in the vision community, their adaptation to HAR tasks remains limited due to the lack of attribute information, the uninterpretable nature of signal characteristics, and the significant gap between signal and semantic modalities.

2.3 Zero-Shot Human Activity Recognition

Zero-Shot Human Activity Recognition was initially proposed by Cheng et al. [7] to use a graphical model to map IMU sensor data to human-defined attribute sequences. However, due to the lack of attribute information in many popular HAR datasets, relatively few researchers have pursued this task. Some studies have investigated the potential of transferring knowledge from vision-based zero-shot activity recognition models [35, 40]. Recent advancements in multi-modal pre-training, such as IMU2CLIP [25], leverage contrastive learning to align IMU sensor data with visual and textual modalities in a shared representation space, enabling zero-shot transfer to unseen classes by bridging the semantic gap between signals and human-interpretable descriptions. However, corresponding vision modality data is often impractical for daily use. Another prominent approach relies on label-level word embeddings to predict unseen activities through a nearest word embedding search [37]. Yet, label-level information often lack sufficient semantic richness to bridge the semantic gap, resulting in coarse mappings. To address the scarcity of labeled IMU data, IMUGPT [22] employs LLM and motion synthesis to generate diverse virtual IMU data from textual descriptions, facilitating zero-shot HAR by augmenting training datasets with synthetic samples.With recent advances in LLM, a few studies have started exploring the possibility of directly inferring unseen activities via "Question & Answering" under prompt engineering [17]. Specifically, HARGPT [17] demonstrates the capability of LLM to perform zero-shot HAR by directly processing raw IMU data with role-playing and chain-of-thought prompting, achieving superior performance compared to traditional baselines without requiring fine-tuning. However, these approaches require highly accurate prompts, and the non-end-to-end process complicates the inference stage.

In contrast to prior studies, our work introduces the use of LLM-generated attribute descriptions as an intermediate, structured semantic representation. Unlike label-level embeddings, which lack extensive semantic information, and direct prompting methods, which depend heavily on prompt quality and complex role-playing, our approach leverages the LLM's capacity to generate rich, human-like attribute descriptions that are both interpretable and semantically aligned with activity concepts. These attributes serve as a bridge between raw sensor data and class labels, enabling end-to-end zero-shot learning with improved generalization and interpretability. This introduces a novel and practical alternative to existing methods that either lack semantic granularity or require complex prompting mechanisms.

3 Problem Definition

Assumption:A general assumption in ZSL is that both seen and unseen classes share a common semantic space, where samples and class prototypes are projected to facilitate the recognition task [18]. This common semantic space is typically constructed using external knowledge (a.k.a. side information), such as textual descriptions, visual annotations, or taxonomies, which enable the transfer of knowledge learned from seen classes to unseen

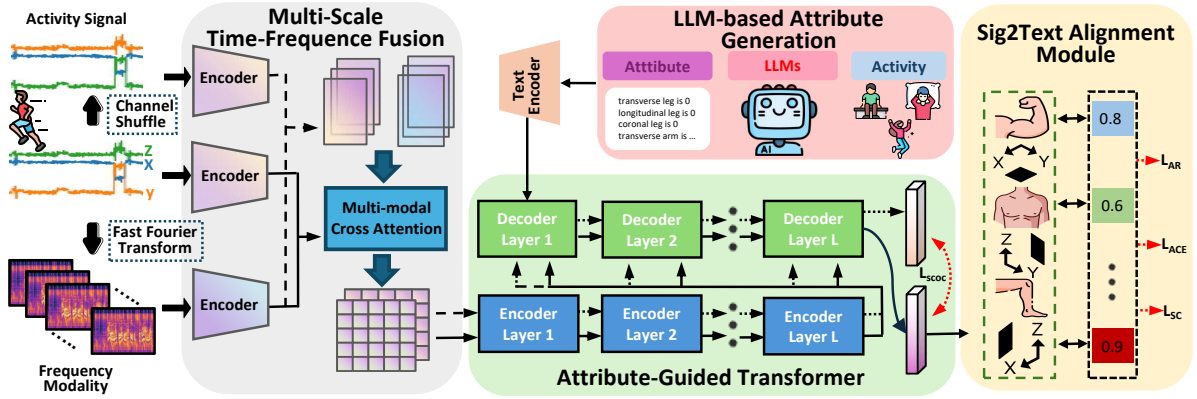


Fig. 3. Structure of our proposed *IMUZero* framework. A detailed description of *IMUZero* would be presented in Section 4.

ones [2]. In this work, we adopt the same assumption and leverage LLM-generated attributes as side information to enable knowledge transfer.

Definition: Zero-shot Human Activity Recognition seeks to recognize previously unseen activities by mapping sensory data into a semantic space defined by decomposable attributes rather than fixed categorical labels. Multi-modal signals—e.g., 3-axis accelerometer, gyroscope, and magnetometer readings—are collected from multiple body locations and segmented into fixed-length windows via a sliding-window approach (see experimental settings for details). Formally, let $\{x^s, y^s\}^N$ denote the training set of seen activities, where x^s is a multivariate sensor segment and $y^s \in C^s$ its corresponding label in the seen label set C^s. The goal of ZSL is to learn a mapping $f(x;\theta)$ that projects x^s into the semantic attribute space, enabling generalization to unseen activities $\{x^u, y^u\}^M$ where x^u represents sensor data from unseen classes and $y^u \in C^u$ their labels in the unseen label set C^u. It should be noted that there are no overlapping between seen and unseen classes $C^u \cap C^s = \emptyset$.

4 Methodology

Previous work has primarily relied on activity–attribute matrices, textual or visual embeddings, or LLM-based QA methods, but these approaches suffer from limited semantic coverage, high annotation or resource demands, and poor contextual grounding. To overcome these challenges, we propose the *IMUZero* framework, which leverages LLMs' cognitive-science–inspired knowledge to generate fine-grained semantic attributes capturing body movement frequency—thereby preserving contextual information by avoiding direct LLM training or QA inference. Additionally, to better align sensor data with these attributes, we encode IMU signal frequency patterns as supplementary projection cues. Fig. 3 illustrates the main pipeline of our framework. Our framework first encourage the LLM to produce attribute descriptions for each human activity (Sec. 4.1). Then, multi-scale time-frequency fusion is performed for IMU data to extract more comprehensive signal features (Sec. 4.2). Finally, the attribute descriptions are aligned with the signal time-frequency features to achieve Zero-Shot Human Activity Recognition (Sec. 4.3).

4.1 Movement-sensitive Attribute Generation:

Category attributes are a prerequisite for the ZSL task as they provide the essential semantic information required for generalizing to unseen classes. These attributes have to carry clear semantic meanings, often rooted in cognitive science-inspired patterns (e.g., the frequency movement pattern of body part), to enable the model to

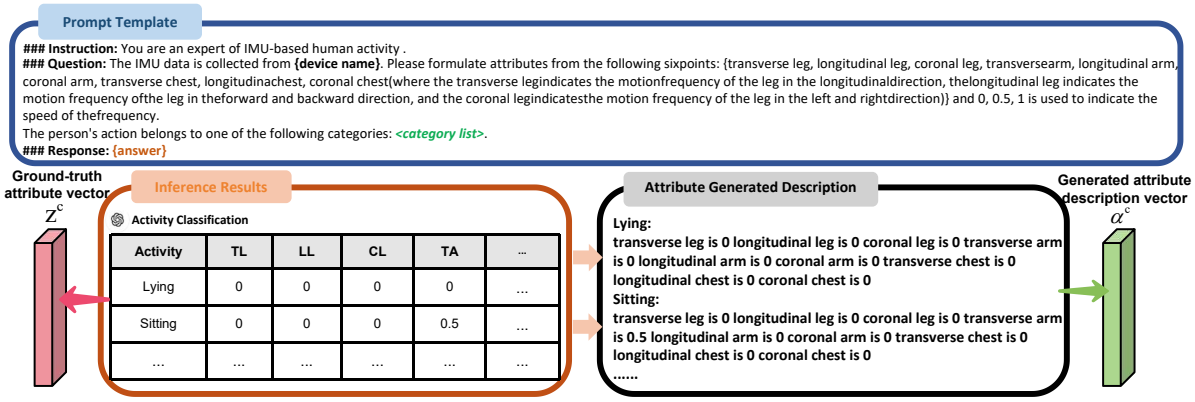


Fig. 4. Chain-of-thought prompt design for *IMUZero* .

effectively map input data to natural language descriptions. Previous approaches predominantly leveraged visual modalities—such as videos, 3D skeletons, or images—as an intermediary to bridge the gap between IMU sensor data and textual descriptions (IMU↔Image↔Text). However, those methods encounter significant challenges: aligned video sources are frequently unavailable, and the substantial variability in online activity videos can introduce domain shift issues, undermining model performance. Inspired by [8], we propose a novel movement-guided semantic attribute generation module to bridge the gap between sensor readings and text description directly. This module leaverage the prior cognitive activity knowledge embedded in existing LLM to generate meaningful and reliable semantic attributes tailored for activity recognition, facilitating robust generalization to unseen activities without dependence on visual intermediaries.

Generation Workflow: The generation workflow of the movement-sensitive attribute, as demonstrated in Figure 4, consists of three stage. Firstly, we prompt an LLM to generate precise, fine-grained bio-mechanical information about human activity. Specifically, we construct a prompt template with instructions and questions to guide the LLM in generating attributes with maximum discrimination. Existing datasets for activity recognition predominantly collect data from sensors placed on key body parts such as the arm, leg, and chest. These locations are chosen because they are central to human movement, providing rich data about motion and posture that are essential for identifying various activities. Thus, in the prompt, we encourage the LLM to generate attributes corresponding to arm, leg, chest and their movements along different axes (i.e., transverse, longitudinal, and coronal). Then, the LLM produces a detailed activity attribute table that quantifies the cognitive activity frequency for each body part. Finally, the generated values are integrated into a coherent paragraph of text, providing a semantic representation that supports the projection in the proposed framework. The process could be represented as:

$$\begin{aligned} z^c &= LLM(prompt[c \in C]) \\ a^c &= LLM(prompt[z^c]) \end{aligned} \tag{1}$$

where $LLM(\cdot)$ denotes the prompt-guided generation by the language model, prompt$[c \in C]$ represents the prompt constructed based on class c, and z^c and a^c denote the attribute vector and the attribute-based generated description vector, respectively.

Attribute Value Constraint: The expression of frequency or movement patterns through attribute values is critical, as these values encapsulate the discriminative information needed to differentiate activities. The reliability of these values significantly impacts both the projection quality and the construction of a discriminative latent

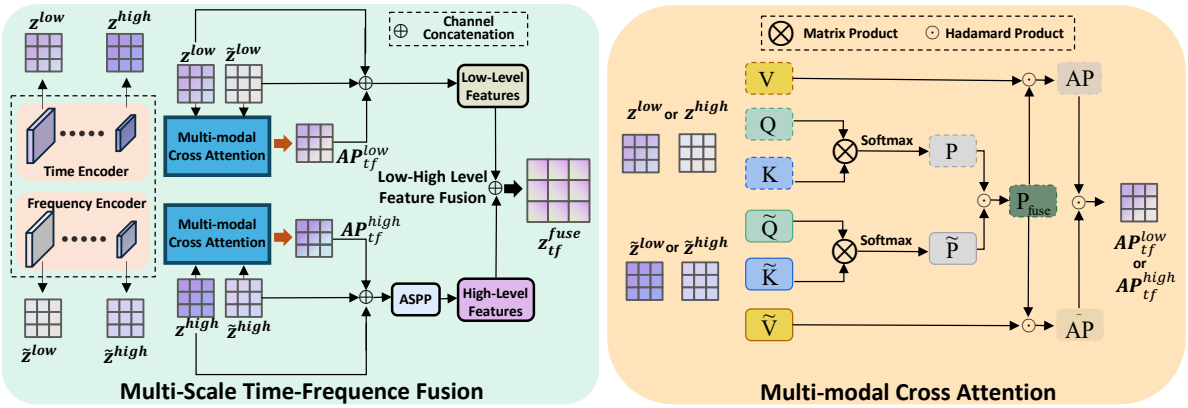


Fig. 5. Architecture of Multi-Modal Feature Fusion. (best viewed in color)

space. To address this, we constrain attribute values to three discrete levels: "0" (nearly no movement), "0.5" (mid-level movement), and "1" (high-level movement). This choice is supported by empirical evidence (i.e., Sec. 5.6) showing that continuous values—whether derived from statistical analyses or LLM—fail to establish clear decision boundaries. Continuous values often lead to coarse boundaries, complicating the model's ability to fit them via constrained regression. While statistical analyses of accelerometer data can provide a baseline by capturing movement frequencies along the x, y, and z axes, their continuous nature and potential variability for unseen activities limit their effectiveness. Thus, we select discrete attribute values that offer a more robust and generalizable solution for the ZSL task.

4.2 Multi-Scale Time-Frequency Fusion

Apart from the quality of the attributes, the mapping between sensor readings and these attributes is fundamental to achieving generalization in the ZSL task. Previous approaches have predominantly relied on either vision modalities or raw IMU data to establish this mapping. However, vision data is often unavailable or difficult to collect, limiting its practicality, while IMU data alone provides only temporal information, which makes it difficult to capture periodic and oscillation patterns, resulting in coarse representation extraction. To overcome these challenges, we propose a Multi-Scale Time-Frequency Fusion module(MSTFF) (as shown in Fig. 3), which aims to generate time-frequency representation by integrating the frequency modality of IMU data as additional information. Since the attributes are specifically designed to correspond to frequency patterns—such as the rate or rhythm of body movements—this frequency-based representation captures more extensive and detailed movement dynamics.

Previous work [48] demonstrates that low-level and high-level features are complementary by nature—low-level features capture fine-grained spatial details but lack semantic information, while high-level features provide semantic abstraction at the expense of subtle nuances. Thus, inspired by this work, the MSTFF module is designed to extract both low- and high-level features from the time and frequency domains, respectively, and integrate them via a multi-modal cross-attention mechani In this module, the Fast Fourier Transform (FFT) [9] is first applied to the IMU signal segments to obtain their frequency-domain representations. Subsequently, high-level and low-level features are extracted from both the time-domain signals and the frequency-domain representations using dedicated time-domain and frequency-domain encoders (i.e., $E_{time}(\cdot), E_{freq}(\cdot)$), respectively. Both two encoders are implemented as 1D-convolutional ResNet-18 backbones, each comprising an initial convolutional

layer, four residual blocks (layer1–layer4) and an adaptive average pooling layer; layer 1 captures low-level signal features while layer 4 extracts high-level abstractions. The high-level features are primarily derived from the last layer of these encoders, while the low-level features are extracted from the third layer. This process can be mathematically expressed as:

$$\begin{aligned} \tilde{x}^s &= FFT(x^s) \\ z^{low}, z^{high} &= E_{time}(x^s) \\ \tilde{z}^{low}, \tilde{z}^{high} &= E_{freq}(\tilde{x}^s) \end{aligned} \tag{2}$$

where $\tilde{x}^s$ denotes frequency format IMU segment, z^{low} and z^{high} represent low/high level time-domain features, $\tilde{z}^{low}$ and $\tilde{z}^{high}$ represent low/high level freqeuncy-domain features, $FFT(\cdot)$ represents the FFT transformation.

Multi-modal Feature Fusion Module. Inertial Measurement Unit (IMU) signals offer rich temporal dynamics that reflect motion states, while frequency-domain features capture the instantaneous frequency content of activities. To fuse these complementary modalities, we employ a multi-modal cross-attention mechanism. It takes the low- and high-level features extracted by a dual-stream network as input and produces high-dimensional joint representations (see Fig. 5). Specifically, we first apply 1×1 convolutions to the single-modality features $z^{low}, z^{high}, \tilde{z}^{low}, \tilde{z}^{high}$ to generate twelve feature maps:

$$\begin{aligned} Q^{low} &= z^{low}W_q, \quad K^{low} = z^{low}W_k, \quad V^{low} = z^{low}W_v \\ Q^{high} &= z^{high}W_q, \quad K^{high} = z^{high}W_k, \quad V^{high} = z^{high}W_v \\ \tilde{Q}^{low} &= \tilde{z}^{low}W_q, \quad \tilde{K}^{low} = \tilde{z}^{low}W_k, \quad \tilde{V}^{low} = \tilde{z}^{L}W_v \\ \tilde{Q}^{high} &= \tilde{z}^{high}W_q, \quad \tilde{K}^{high} = \tilde{z}^{high}W_k, \quad \tilde{V}^{high} = \tilde{z}^{high}W_v \end{aligned} \tag{3}$$

We then compute the self-attention map by multiplying the transpose of each query feature with its corresponding key feature, and normalizing via a softmax. The low-level/high-level self-attention maps for the time (i.e., P^{low},P^{high}) and frequency (i.e., $\tilde{P}^{low}$,$\tilde{P}^{high}$) modalities are given by:

$$\begin{aligned} P^{low} &= softmax(Q^{low^\intercal} \otimes K^{low}), \quad P^{high} = softmax(Q^{high^\intercal} \otimes K^{high}) \\ \tilde{P}^{low} &= softmax(\tilde{Q}^{low^\intercal} \otimes \tilde{K}^{low}), \quad \tilde{P}^{high} = softmax(\tilde{Q}^{high^\intercal} \otimes \tilde{K}^{high}) \end{aligned} \tag{4}$$

Then, the self-attention maps generated from both time and frequency modalities are input into the cross-attention fusion mechanism. This results in a joint weighted feature map, which can be represented as:

$$\begin{aligned} P^{low}_{fuse} &= P^{low} \odot \tilde{P}^{low} \\ AP^{low}, \tilde{AP}^{low} &= (P^{low}_{fuse} \odot V^{low}), (\tilde{P}^{low}_{fuse} \odot \tilde{V}^{low}) \\ AP^{low}_{tf} &= AP^{low} \odot \tilde{AP}^{low} \end{aligned} \tag{5}$$

Similarly, the high-level features can be formulated as:

$$\begin{aligned} P^{high}_{fuse} &= P^{high} \odot \tilde{P}^{high} \\ AP^{high}, \tilde{AP}^{high} &= (P^{high}_{fuse} \odot V^{high}), (\tilde{P}^{high}_{fuse} \odot \tilde{V}^{high}) \\ AP^{high}_{tf} &= AP^{high} \odot \tilde{AP}^{high} \end{aligned} \tag{6}$$

Fusion of Low- and High-Level Features. To effectively integrate the extracted joint feature maps, we introduce a multi-scale feature fusion module that preserves the complementarity between low- and high-level representations. First, we concatenate the cross-modality low- and high-level features with their corresponding

joint feature maps:

$$\begin{aligned} z_{tf}^{low} &= \text{concat}\big(z^{low},\, \tilde{z}^{low},\, AP_{tf}^{low}\big), \\ z_{tf}^{high} &= \text{concat}\big(z^{high},\, \tilde{z}^{high},\, AP_{tf}^{high}\big), \end{aligned} \tag{7}$$

where $\text{concat}(\cdot)$ denotes channel-wise concatenation. Next, to enrich contextual information at multiple scales, we apply an Atrous Spatial Pyramid Pooling (ASPP) module [3] to the high-level joint feature map z_{tf}^{high}. Specifically, we employ parallel 3×3 convolutions with dilation rates of 6, 12, and 18 to capture features at different receptive fields, and a 1×1 convolution following global average pooling to aggregate channel-wise context. The output of the ASPP is denoted

$$z_{tf}^{high_{ASPP}} = \text{ASPP}\big(z_{tf}^{high}\big).$$

Finally, we fuse the low- and high-level representations by concatenating z_{tf}^{low} with $z_{tf}^{high_{ASPP}}$, yielding the overall fused feature:

$$z_{tf}^{fuse} = \text{concat}\big(z_{tf}^{low},\, z_{tf}^{high_{ASPP}}\big). \tag{8}$$

4.3 Sig2Text Alignment

The ZSL task relies on aligning signal semantics with textual attributes during training. This alignment enables category matching using the textual attributes of unseen classes during testing. To enhance alignment precision, we introduce the Sig2Text Alignment module, which localizes the activity pattern most relevant to each attribute.

We hypothesize that adding sensor context to IMU data provides valuable spatial and biomechanical information, helping the model learn to recognize actions more effectively. Thus, we first incorporate the attribute-based generated description vector a^c as semantic guidance for the fused time–frequency feature via a cross-attention mechanism, which can be expressed as

$$z^{cr} = \text{softmax}\Big[(a^c W_{cr}^q)^{\mathsf{T}} \otimes (U(z_{tf}^{\text{fuse}}) W_{cr}^k)\Big] \;\odot\; (U(z_{tf}^{\text{fuse}}) W_{cr}^v), \tag{9}$$

where U is the transformer encoder, z^{cr} is the attribute-integrated feature, and W_{cr}^q, W_{cr}^k, W_{cr}^v are the query, key, and value projection weights of the cross-attention layer, respectively.

To align the attribute-integrated feature with the ground-truth class attribute, we employ an attribute-based cross-entropy loss $\mathcal{L}_{\text{ACE}}$:

$$\mathcal{L}_{\text{ACE}} = -\log \frac{\exp\big(z^{cr} \cdot z^{c_i}\big)}{\sum_{c_i \in C} \exp\big(z^{cr} \cdot z^{c_i}\big)}, \tag{10}$$

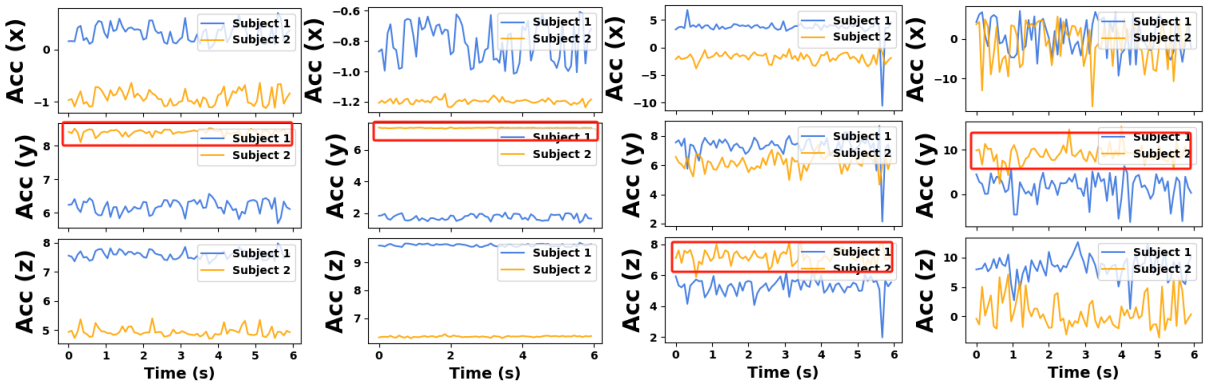
where z^{c_i} is the ground-truth attribute vector for class i and C is the set of all class attributes.

While $\mathcal{L}_{\text{ACE}}$ maximizes inter-class separation, it does not explicitly minimize intra-class variation. To address this, we introduce an attribute-regression loss $\mathcal{L}_{\text{AR}}$ that penalizes the distance between the integrated feature and its corresponding class attribute:

$$\mathcal{L}_{\text{AR}} = \| z^{cr} - z^{c_i} \|_2^2. \tag{11}$$

Since $\mathcal{L}_{\text{AR}}$ and $\mathcal{L}_{\text{ACE}}$ optimize the model's performance on known classes, the training process would inevitably overfit to these classes. To address this issue, we introduce self-calibration loss $\mathcal{L}_{\text{SC}}$, which explicitly shifts some prediction probability from known to unknown classes. $\mathcal{L}_{\text{SC}}$ is defined as:

$$\mathcal{L}_{\text{SC}} \;=\; -\log \frac{\exp\big(z^{cr} \cdot z^{c_i} \;+\; \mathbb{I}_{[c_i \in C^u]}\big)}{\sum_{j=0}^{|C|} \exp\big(z^{cr} \cdot z^{c_j} \;+\; \mathbb{I}_{[c_j \in C^u]}\big)}, \quad j = 0, 1, \ldots, |C|. \tag{12}$$



(a) Sitting (b) Lying (c) Cycling (d) Rowing

Fig. 6. Visualization of IMU data for two distinct individuals from the DSADS dataset. Ideally, accelerometer peak values for the same activity should exhibit consistent patterns (similar peak and trough values, or similar scale under ideal experimental data-collection settings). However, the recorded data reveal significant variability in these peaks—highlighted by the red boxes—indicating the presence of axis bias.

where C^u denotes the unknown (unseen) classes, $\mathbb{I}_{[c_j \in C^u]}$,$\mathbb{I}_{[c_i \in C^u]}$ is an indicator function, (i.e., it is 1 when c_i or c_j is unseen class, otherwise -1). Intuitively, $\mathcal{L}_{\text{ACE}}$ encourages assigning non-zero probabilities to unknown classes during training, enabling *IMUZero* to generate a significant non-zero probability for true unknown classes when encountering test samples from unknown categories.

Furthermore, to mitigate axis bias (as demonstrated in Fig. 6) introduced by variations in users' wearing behaviors, we introduce the Shuffle Channel Order Constraint (SCOC). This constraint encourages the model to extract consistent activity information by maximizing the mutual information between the attribute-integrated feature of original (i.e., x^s) and randomly permuted input signals (i.e., shuffle(x^s)). Specifically, we randomly shuffle the accelerometer channels of the input signal to produce channel-shuffled augmentations, and then optimize the mutual information between their attribute-integrated feature (i.e., z^{cr} and $\hat{z}^{cr}$):

$$\mathcal{L}_{\text{SCOC}} = \left\| z^{cr} - \hat{z}^{cr} \right\|_2^2 \tag{13}$$

where shuffle$(\cdot)$ denotes the channel shuffle operation.

Total Loss. Finally, we integrate Eq. 10, 11, 12 and 13 to obtain a final loss function as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SCOC}} + \mathcal{L}_{\text{ACE}} + \lambda_{\text{AR}}\mathcal{L}_{\text{AR}} + \lambda_{\text{SC}}\mathcal{L}_{\text{SC}}, \tag{14}$$

where λ_{AR} and λ_{SC} are the weighting factor.

5 Experiment

5.1 Dataset

To evaluate the effectiveness of our *IMUZero* framework, we perform it on four public datasets: PAMAP2 [47], DSADS [3], MHEALTH [1] and GOTOV [40].More details of the evaluated datasets can be found in Table 1.

- **PAMAP2 Dataset.** records 18 daily physical activities which include 12 protocol activities (walking, running, vacuum cleaning, rope jumping, etc.) and 6 optional activities (watching TV, computer work, folding laundry, etc.).The activity data were captured for 9 subjects from all the sensors (a Heart rate monitor and 3 inertial measurement units) The data recorded comprised the measurements of gyroscopes,

Dataset	#Subject	#Activity	Frequency	#Sample	#Dim	Seen Activity Pool	Unseen Activity Pool
PAMAP2	9	18	100Hz	2.84M	52	Rope Jumping, Lying, Sitting, Standing, Walking, Running, Ironing	Vacuum Cleaning, Cycling, Nordic Walking, Ascending Stairs, Descending Stairs
DSADS	8	19	25Hz	1.14M	45	Sitting, Standing, Lying on Back, Lying on Right Side, Ascending Stairs, Descending Stairs, Standing in an Elevator Still, Moving Around in an Elevator, Walking in a Parking Lot, Walking on a Treadmill (Flat), Walking on a Treadmill (15 deg), Running on a Treadmill (8 km/h), Exercising on a Stepper, Exercising on a Cross Trainer	Playing Basketball, Jumping, Rowing, Cycling (Horizontal), Cycling (Vertical)
MHEALTH	10	12	50Hz	0.34M	23	Standing Still, Sitting and Relaxing, Lying Down, Walking, Climbing Stairs, Waist Bends Forward, Frontal Elevation of Arms	Jumping, Running, Jogging, Cycling, Knees Bending
GOTOV	35	16	83Hz	5.9M	3	Jumping, Standing, Step, Lying, Sitting, Walking Stairs Up	Washing Dishes, Stacking Shelves, Vacuum Cleaning, Walking, Cycling

Table 1. Description of the four public HAR datasets used in our study

accelerometers, magnetometers, heart rate monitor, and temperature. The dataset has a total of 52 features and was captured at a sampling rate of 100 Hz.

- **DSADS Dataset.** contains 19 physical activities performed by 8 subjects (4 female, 4 male, aged 20-30) for 5 minutes each. The activities, which include various daily movements and exercises, were recorded using five sensor units (torso, arms, and legs) equipped with accelerometers, gyroscopes, and magnetometers. The data was captured at a sampling rate of 25 Hz and is organized into segments, resulting in a total of 45 features per segment.
- **MHEALTH Dataset.** contains recordings of 12 physical activities performed by 10 volunteers using wearable sensors placed on the chest, right wrist, and left ankle. The data includes measurements of acceleration, rate of turn, magnetic field orientation, and 2-lead ECG from the chest sensor. Captured at a sampling rate of 50 Hz, the dataset provides insights into body motion and vital signs.
- **GOTOV Dataset.** contains 16 daily activities collected from thirty-five elder-age participants. The subjects were instructed to wear accelerometer sensors at three locations: ankle, chest, and wrist with a 9-dimensional recording, at the sampling rate of 83Hz.

5.2 Experiment Setting.

Data Segmentation. In our study, we follow previous work [27] by employing a sliding window approach to segment the raw sensory data streams from each dataset into smaller segments. Specifically, for the PAMAP2 dataset, the sliding window is configured with a window length of 5.12 seconds (170 timesteps) and a slide length of 1 second (33 timesteps). The DSADS dataset uses a window length of 5 seconds (125 timesteps) and a slide length of 1 second (25 timesteps). For the MHEALTH dataset, the sliding window is configured with a window length of 1.2 seconds (timesteps not specified) and a slide length of 0.6 seconds (25 timesteps). Lastly, for the GOTOV dataset, the sliding window is set with a window length of 5 seconds (25 timesteps) and a slide length of 2.5 seconds (25 timesteps). These segments can be fed directly into the network without the need for hand-crafted feature engineering or transformation.

Zero-Shot Class Partition. For zero-shot class separation, we create four partition choices corresponding to 2, 3, 4, and 5 unseen classes from the aforementioned datasets. The increasing number of unseen classes makes model prediction more challenging, thereby reflecting the model's generalization ability. The seen and unseen activity pool is detailed in Table 1, and the first two, three, and four classes are designated for the 2, 3, and 4 settings, respectively. For the rest classes that are not used in each setting will be added to the seen class set.

Dataset	Setting	Class Split		Unseen Only		Both Seen and Unseen		
				#Sample				
		Seen	Unseen	Train	Test	Train	Test(Seen)	Test(Unseen)
PAMAP2	5-class	7	5	11415	8452	6845	4570	8452
	4-class	8	4	13342	6525	8001	5341	6525
	3-class	9	3	14545	5322	8722	5823	5322
	2-class	10	2	15624	4243	9369	6255	4243
DSADS	5-class	14	5	6720	2400	4032	2688	2400
	4-class	15	4	7200	1920	4320	2880	1920
	3-class	16	3	7680	1440	4608	3072	1440
	2-class	17	2	8160	960	4896	3264	960
MHEALTH	5-class	7	5	7044	4394	4223	2821	4394
	4-class	8	4	8022	3416	4809	3213	3416
	3-class	9	3	9046	2392	5423	3623	2392
	2-class	10	2	10070	1368	6037	4033	1368
GOTOV	5-class	6	5	12610	15371	7562	5048	15371
	4-class	7	4	14697	13284	8814	5883	13284
	3-class	8	3	16570	11411	9937	6633	11411
	2-class	9	2	18387	9594	11027	7360	9594

Table 2. Detailed evaluation setting for "Unseen Only" and "Both Seen and Unseen" tasks.

Evaluation setting. To make the evaluation more realistic, we conduct two settings: "Unseen Only" (often referred to as the zero-shot learning setting) and "Both Seen and Unseen":(often referred to as generalized zero-shot learning [30]). The "Unseen Only" setting measures performance exclusively on classes never seen during training. This simulates cases like a detector deployed to identify entirely new product types or newly observed species. The "Both Seen and Unseen" setting measures performance when the test set contains a mix of seen and unseen classes (a more realistic setting), so the model must recognize familiar categories while also handling novel ones. This simulates cases like a production-line anomaly detector, where most defects are familiar but novel failure modes occasionally appear; GZSL can identify or prompt unseen failure types by utilizing existing defect attributes and sensor metadata. For the “Unseen Only” setting, we train on all samples of seen classes and test on all samples of unseen classes. For the “Both Seen and Unseen” setting, we reserve 40% of each seen class’s samples as the seen test set, while using the same unseen test samples as in the “Unseen Only” setting. Details of the train/test splits are summarized in Table 2.

Evaluation metrics. We evaluate all models using average per-class accuracy and F1 score for multi-class activity classification in the ZSL framework. Accuracy represents the average proportion of correctly classified samples across classes, while the F1 score, capturing class imbalances, is the harmonic mean of precision and recall, computed as

$$\mathrm{F1} = \frac{2 \cdot \mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{15}$$

where TP, FP, and FN denote true positives, false positives, and false negatives for class i, respectively. For the setting of "Both Seen and Unseen", we apply the harmonic mean (H), which is a way to balance the performance of a model on seen and unseen classes, providing a more holistic measure than just looking at individual accuracies.

which could be computed as

$$H = \frac{2 \times acc_{seen} \times acc_{unseen}}{acc_{seen} + acc_{unseen}} \tag{16}$$

where acc_{seen} and acc_{unseen} represent the accuracy on seen and unseen classes respectively. Models are trained on the source dataset (seen classes) and tested on the target dataset (unseen classes), with source dataset results reported. For robustness and better reproducibility, experiments are conducted ten times with different random seeds. Random seeds are used to initialize the random number generators (RNGs) that influence many stochastic operations during training. Setting a random seed ensures that these operations behave deterministically (i.e., produce the same output each time the code is run under the same conditions). Furthermore, by repeating experiments with different seeds, we can evaluate whether the proposed model consistently performs well. The results are averaged across all runs for both metrics.

Implementation details. Our model is designed in an end-to-end manner, utilizing the Adam optimizer with specific hyperparameters (weight decay of 0.0001) for optimization. We have also set the same parameters for all comparison methods except IMUGPT. For IMUGPT, we used GPT4o to generate motion descriptions. We provided the activities from all four datasets, along with the descriptions of the samples, as text to the LLM. Then we asked the LLM to describe the person performing the activity. The generation of one thousand descriptions for each activity was completed in 50 batches. For all datasets, we empirically set λ_{SC} to 0.3 and λ_{AR} to 0.05. Both the encoder and decoder are configured with a single layer and one attention head.

5.3 Comparison Methods

To validate the effectiveness of the proposed approach, we compared our proposed *IMUZero* framework with the closely related baselines. DeepConvLSTM [27] is the state-of-the-art feature learning approach for human activity recognition; Composer [15] is the conventional ZSL approach that aims to compose features of unseen classes from only relevant training samples so to increase the unseen prediction performance. DAZLE [16] is an attention-based approach that introduces a dense attribute-based attention mechanism, allowing each attribute to focus on the most relevant regions of the image. HSVA [5] is an adaptation-based approach that applies distribution adaptation to solve distribution-aligned space-shifting problems so to improve generalization ability. TransZero [4], HRT [6], and SHIP [39] are three newly and state-of-the-art transformer-based ZSL approaches. IMU2CLIP [25] leverages contrastive learning to align IMU sensor data with visual and textual modalities in a shared representation space, enabling zero-shot human activity recognition by transferring knowledge across modalities. IMUGPT [22] generates diverse virtual IMU data using LLM and motion synthesis, providing augmented training samples to enhance zero-shot HAR performance. HARGPT [17] employs LLM with role-playing and chain-of-thought prompting to directly process raw IMU data for zero-shot activity recognition, offering a non-fine-tuned alternative to traditional methods. For all baseline methods, we used the released code if available, and reproduced the unavailable methods using Pytorch [28].

5.4 Empirical Result

In this section, we present a series of empirical results and test the proposed *IMUZero* framework with the aforementioned representative methods. Through experiments, we have the following observations:

- The proposed *IMUZero* framework demonstrates significant improvements over baseline approaches across four datasets. Specifically, *IMUZero* achieves a 3~5% enhancement over SOTA ZSL methods in complex settings (i.e., 4-5 unseen classes).
- For a more realistic setting, where the test set contains both seen and unseen classes, the proposed *IMUZero* framework still achieves competitive performance compared to other methods, even under complex settings.

Dataset	Class	DeepConvLSTM [27]	Composer [15]	DAZLE [16]	HSVA [5]	HRT [6]	SHIP [39]	IMU2CLIP [25]	IMUGPT [22]	HARGPT [17]	IMUZero
Pamap2	2-class	71.97%	81.30%	55.66%	80.39%	81.86%	91.83%	90.13%	90.12%	85.60%	**92.51%±0.3%**
	3-class	60.42%	68.59%	31.13%	62.55%	55.24%	80.51%	80.31%	78.35%	72.80%	**82.03%±0.3%**
	4-class	49.36%	53.56%	23.25%	54.10%	48.91%	56.45%	57.25%	55.83%	53.40%	**57.87%±0.4%**
	5-class	42.92%	45.05%	30.15%	45.50%	28.04%	51.92%	51.10%	49.70%	47.25%	**52.55%±0.3%**
DSADS	2-class	91.11%	93.16%	94.15%	94.22%	93.72%	94.07%	92.88%	93.20%	93.50%	**95.10%±0.5%**
	3-class	85.80%	85.94%	83.70%	86.19%	85.39%	89.21%	89.30%	88.45%	87.20%	**90.33%±0.3%**
	4-class	54.43%	55.00%	44.27%	55.53%	51.34%	66.36%	66.12%	65.20%	60.85%	**68.50%±0.6%**
	5-class	39.12%	41.20%	42.06%	42.05%	33.19%	50.55%	52.30%	52.00%	48.30%	**54.33%±0.8%**
Mhealth	2-class	80.58%	80.34%	83.98%	82.26%	66.59%	86.52%	85.37%	84.58%	82.40%	**86.63%±0.2%**
	3-class	65.30%	68.06%	54.74%	68.45%	51.00%	**78.78%**	72.52%	70.15%	71.60%	73.81%±0.3%
	4-class	55.63%	54.47%	40.73%	57.62%	46.38%	61.92%	60.01%	62.10%	57.90%	**62.57%±0.3%**
	5-class	41.41%	40.17%	16.76%	41.60%	31.32%	**51.90%**	48.33%	46.20%	41.80%	49.16%±0.3%
GOTOV	2-class	82.20%	85.48%	77.93%	81.15%	80.66%	90.15%	89.45%	90.20%	86.90%	**91.26%±0.3%**
	3-class	71.91%	73.45%	56.57%	72.79%	63.93%	86.54%	85.03%	86.00%	84.30%	**88.12%±0.3%**
	4-class	51.78%	53.97%	36.06%	55.48%	50.42%	58.61%	60.70%	60.50%	60.50%	**62.79%±0.2%**
	5-class	43.36%	43.37%	29.68%	40.60%	30.77%	51.81%	51.92%	52.50%	52.90%	**53.25%±0.3%**

Table 3. Performance comparison (accuracy %) on different datasets and classification settings. Best results are highlighted in **bold**, and second best results are underlined.

Dataset	Class	DeepConvLSTM [27]	Composer [15]	DAZLE [16]	HSVA [5]	HRT [6]	SHIP [39]	IMU2CLIP [25]	IMUGPT [22]	HARGPT [17]	IMUZero
Pamap2	2-class	68.1%	76.5%	48.7%	74.8%	75.3%	84.9%	83.7%	82.4%	78.6%	**85.2%±0.4%**
	3-class	54.7%	61.2%	25.3%	55.6%	47.5%	72.1%	71.6%	69.1%	64.3%	**74.6%±0.5%**
	4-class	41.3%	44.9%	18.2%	45.2%	40.1%	48.3%	49.1%	47.5%	44.2%	**50.3%±0.5%**
	5-class	33.8%	36.4%	22.9%	37.1%	21.8%	43.2%	42.5%	40.3%	38.5%	**45.1%±0.6%**
DSADS	2-class	87.9%	88.7%	87.6%	88.3%	87.4%	87.5%	86.1%	86.3%	85.9%	**88.9%±0.6%**
	3-class	79.4%	79.8%	75.8%	79.1%	78.2%	81.7%	81.2%	80.5%	79.1%	**83.4%±0.4%**
	4-class	46.2%	47.6%	36.4%	47.9%	43.1%	56.4%	56.9%	56.1%	51.7%	**59.8%±0.6%**
	5-class	30.5%	33.1%	33.5%	34.0%	26.3%	42.1%	43.6%	42.9%	39.4%	**46.5%±0.9%**
Mhealth	2-class	75.2%	74.3%	76.1%	75.8%	58.7%	79.6%	78.5%	77.2%	75.1%	**80.7%±0.3%**
	3-class	57.6%	60.8%	45.9%	60.3%	43.2%	64.4%	64.3%	61.8%	63.1%	**67.3%±0.4%**
	4-class	47.5%	45.3%	32.8%	48.7%	38.4%	52.8%	51.2%	53.1%	48.9%	**54.6%±0.5%**
	5-class	32.9%	31.6%	11.5%	33.5%	23.9%	40.1%	39.7%	38.2%	40.8%	**43.9%±0.5%**
GOTOV	2-class	76.8%	79.9%	70.2%	74.1%	73.8%	83.2%	81.9%	83.1%	79.3%	**84.3%±0.3%**
	3-class	64.1%	65.4%	48.3%	64.5%	55.6%	77.8%	76.3%	77.1%	74.9%	**80.5%±0.4%**
	4-class	43.1%	45.6%	29.1%	46.4%	41.5%	49.6%	51.4%	51.3%	50.6%	**55.1%±0.3%**
	5-class	34.7%	34.8%	22.4%	32.9%	24.1%	42.7%	43.0%	43.5%	43.7%	**45.8%±0.4%**

Table 4. Performance comparison (F1 Score %) on different datasets and classification settings. Best results are highlighted in **bold**, and second best results are underlined.

- The attributes generated by the LLM enable the transfer of the vision-based ZSL framework to the HAR task. Moreover, the quality of semantic embeddings is crucial for zero-shot recognition tasks, as label-level word embeddings often fail to capture fine-grained semantic attributes, thereby diminishing the ability to predict unseen classes.
- Through ablation studies, we demonstrate that each component in *IMUZero* contributes to recognition performance. Specifically, the Shuffle Channel Order Constraint (SCOC) contributes comparably to the attribute regression constraint, underscoring that biases introduced by wearing patterns can significantly impact data quality and, consequently, performance.
- The "Test-Time" Zero-Shot HAR setting (i.e., the semantic information, like labels and attributes, of unseen classes is not given) remains a significant challenge for SOTA ZSL approaches.

5.4.1 ***"Unseen Only" Setting Result***. We evaluated the proposed *IMUZero* framework on four public HAR datasets. Table 3 reports the average accuracy of both baseline models and our *IMUZero* framework under four zero-shot settings on the PAMAP2, MHEALTH, DSADS, and GOTOV datasets. The results show significant

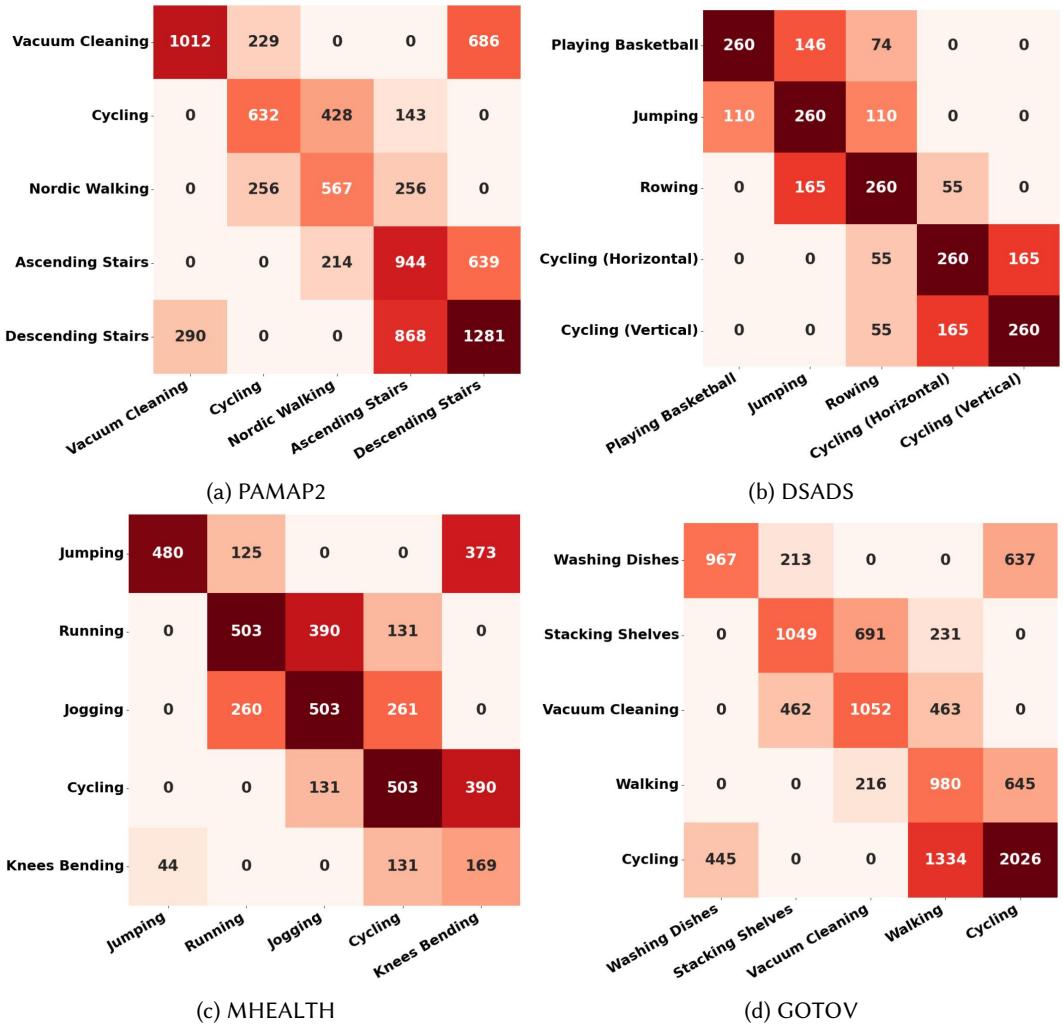


Fig. 7. Confusion matrix of "Seen Only" task for the proposed *IMUZero* framework in four datasets under 5-class setting. (best viewed in color)

improvements—an average gain of 2% in accuracy—across all settings achieved by *IMUZero* . Furthermore, we evaluate the F1 score (as presented in Table 4), which penalizes class imbalance more strongly, on all four datasets. Compared to accuracy, *IMUZero* demonstrates an even larger average performance gain of 3%, achieving the best F1 scores in every setting. Such observations demonstrate the effectiveness of integrating semantic descriptions for generalizable latent space construction together with frequency-domain information support.

It is important to note that as the number of unseen classes increases, the inference for these unseen classes becomes more complex and challenging, leading to a significant degradation in performance for all methods. As the number of target classes grows, a zero-shot model must carve its feature space into ever more finely divided regions—each corresponding to one unseen activity prototype—using only indirect knowledge from seen classes.

		Pamap2		DSADS		Mhealth		GOTOV	
		4-class	5-class	4-class	5-class	4-class	5-class	4-class	5-class
DeepConvLSTM [27]	U	33.0%	30.1%	25.0%	20.0%	39.3%	33.7%	40.0%	37.8%
	S	39.6%	36.1%	44.7%	38.8%	60.9%	52.2%	60.0%	56.7%
	H	36.0%	33.0%	32.1%	26.4%	47.6%	40.9%	48.0%	45.5%
Composer [15]	U	35.0%	31.9%	39.0%	25.0%	40.3%	34.5%	43.7%	40.9%
	S	42.0%	38.3%	47.0%	39.8%	62.5%	53.5%	65.8%	61.4%
	H	38.3%	34.8%	42.6%	30.7%	48.8%	41.9%	52.4%	49.0%
DAZLE [16]	U	20.8%	19.1%	39.5%	25.0%	16.5%	13.5%	30.2%	27.8%
	S	25.0%	22.2%	47.5%	40.2%	25.6%	20.9%	45.3%	41.7%
	H	22.7%	20.8%	43.1%	30.8%	20.0%	16.4%	36.2%	33.3%
HSVA [5]	U	34.1%	31.2%	39.5%	37.8%	40.5%	34.6%	39.4%	36.7%
	S	40.9%	37.4%	47.5%	40.0%	62.8%	53.6%	59.1%	55.1%
	H	37.2%	34.2%	43.1%	38.9%	49.0%	42.0%	47.3%	44.0%
HRT [6]	U	20.1%	18.4%	25.0%	29.9%	24.7%	21.2%	39.1%	35.8%
	S	24.1%	22.1%	37.9%	32.0%	29.2%	32.9%	35.2%	53.7%
	H	22.0%	20.1%	30.1%	30.9%	26.8%	25.7%	37.0%	42.8%
SHIP [39]	U	49.8%	**48.2%**	49.0%	**50.5%**	52.5%	**48.8%**	50.9%	**47.3%**
	S	59.8%	56.7%	51.0%	48.0%	81.4%	45.3%	76.4%	44.8%
	H	54.3%	**50.6%**	50.0%	49.2%	63.3%	46.5%	61.1%	46.0%
IMU2CLIP [25]	U	48.9%	47.4%	50.0%	49.7%	51.7%	48.0%	42.9%	39.1%
	S	58.7%	54.4%	49.0%	47.0%	80.1%	45.5%	64.4%	62.6%
	H	53.3%	45.5%	49.5%	48.3%	62.3%	**46.7%**	51.5%	48.2%
IMUGPT [22]	U	47.5%	46.2%	46.0%	49.4%	50.1%	46.5%	46.8%	43.6%
	S	57.0%	53.6%	49.0%	47.0%	77.7%	44.0%	70.2%	60.8%
	H	51.9%	44.3%	47.5%	48.2%	60.5%	45.2%	56.2%	50.8%
HARGPT [17]	U	44.8%	43.9%	47.0%	45.9%	47.3%	44.0%	41.8%	38.7%
	S	53.8%	50.2%	47.5%	44.0%	73.3%	41.5%	62.7%	61.9%
	H	48.8%	42.0%	47.2%	44.9%	57.6%	42.7%	50.2%	47.6%
IMUZero	U	**51.1%±0.7%**	46.4%±0.5%	**50.1%±0.8%**	49.0%±1.2%	**53.5%±0.7%**	35.3%±0.8%	**53.5%±0.7%**	46.5%±0.8%
	S	**61.3%±0.5%**	**58.2%±0.5%**	**55.0%±0.5%**	**51.6%±0.5%**	**82.9%±0.5%**	**54.7%±0.6%**	**80.3%±0.4%**	**64.0%±0.5%**
	H	**55.6%±0.4%**	47.4%±0.5%	**52.4%±0.5%**	**50.3%±0.7%**	**64.3%±0.6%**	43.0%±0.7%	**64.2%±0.6%**	**53.3%±0.5%**

Table 5. Performance comparison (U: Unseen accuracy %, S: Seen accuracy %, H: Harmonic mean %) on different datasets for 4-class and 5-class settings in "Both Seen and Unseen" setting. Best results are highlighted in **bold**, and second-best results are underlined. More detailed performance can be seen in Appendix A (Table 8,9)

This "crowding" makes it harder to keep inter-class distances large enough for reliable discrimination: prototypes for different activities end up closer together, decision boundaries become more fragile, and small shifts in the embedding can cause a sample to be assigned to the wrong class. Moreover, with more classes there's inevitably greater semantic diversity (and often imbalance) among activities, so the transfer of alignment learned on seen classes becomes less precise for each new unseen class. In short, finer-grained multi-class splits demand higher resolution in the learned representation—something inherently limited when you have zero direct examples of the target categories—so zero-shot performance naturally degrades as class count increases.

Although increasing the number of unseen classes poses significant challenges for zero-shot recognition, the proposed *IMUZero* framework still delivers promising results. In Table 3, we note a performance drop on the MHEALTH dataset under the 3- and 5-class ZSL settings. This may be attributed to the limited size of MHEALTH, which constrains model training and degrades generalization. Furthermore, Fig. 7 presents confusion matrices of unseen activity recognition under the "Unseen Only" setting for the proposed framework (5-class setting).While highly discriminative classes are classified accurately, activities with similar motion patterns exhibit

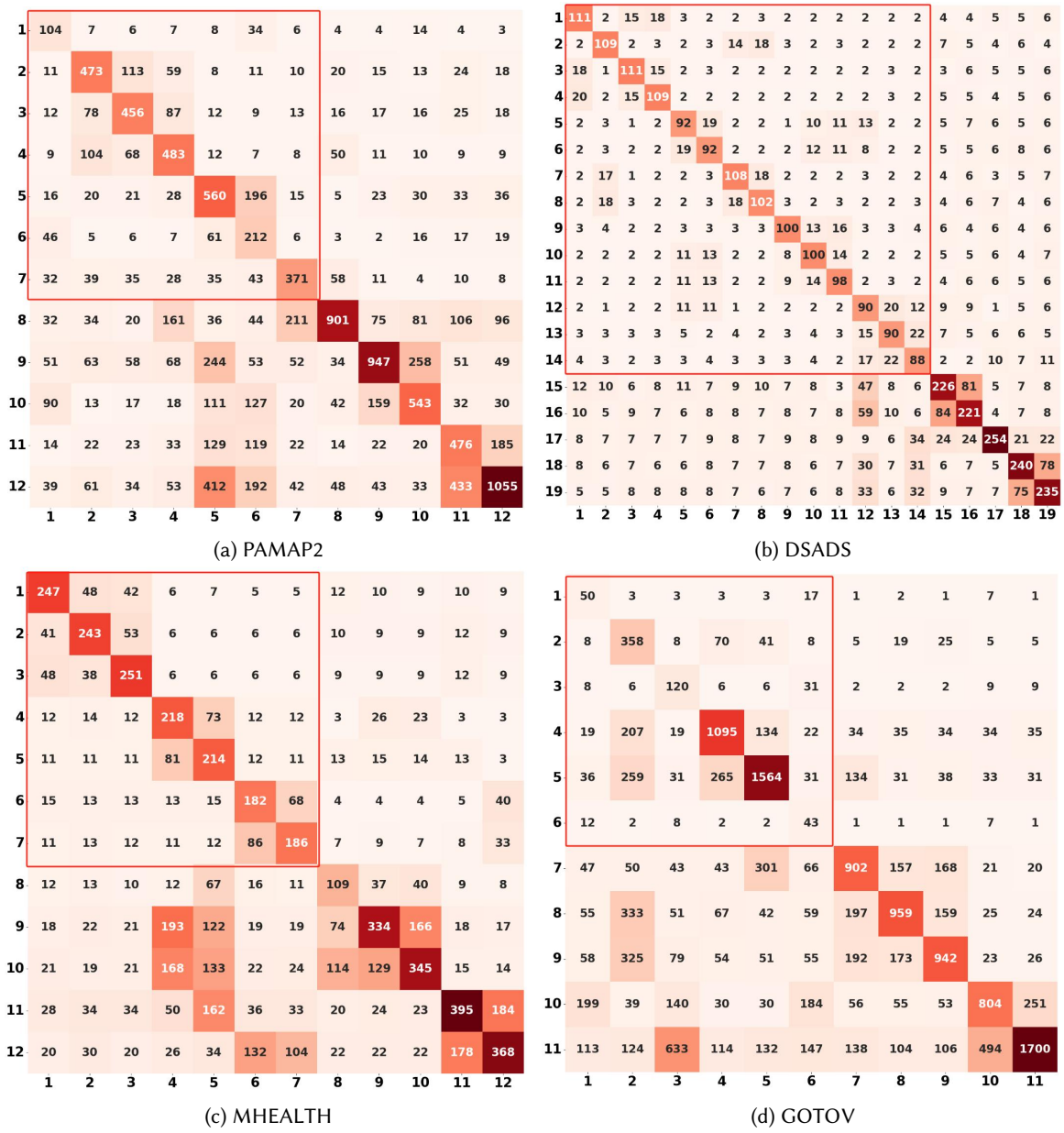


Fig. 8. Confusion matrix of "Both Seen and Unseen" task for the proposed IMUZero framework in four datasets under 5-class setting. (label number can refer to Table 6, best viewed in color)

higher misclassification rates. For example, in the PAMAP2 and DSADS datasets, ascending versus descending stairs and horizontal versus vertical cycling show increased confusion. Such errors likely stem from the limited

PAMAP2		DSADS		MHEALTH		GOTOV	
No.	**Activity**	**No.**	**Activity**	**No.**	**Activity**	**No.**	**Activity**
1	Rope Jumping (S)	1	Sitting (S)	1	Standing Still (S)	1	Jumping (S)
2	Lying (S)	2	Standing (S)	2	Sitting and Relaxing (S)	2	Standing (S)
3	Sitting (S)	3	Lying on Back (S)	3	Lying Down (S)	3	Step (S)
4	Standing (S)	4	Lying on Right Side (S)	4	Walking (S)	4	Lying (S)
5	Walking (S)	5	Ascending Stairs (S)	5	Climbing Stairs (S)	5	Sitting (S)
6	Running (S)	6	Descending Stairs (S)	6	Waist Bends Forward (S)	6	Walking Stairs Up (S)
7	Ironing (S)	7	Standing in Elevator (S)	7	Frontal Elevation of Arms (S)	7	Washing Dishes (U)
8	Vacuum Cleaning (U)	8	Moving in Elevator (S)	8	Jumping (U)	8	Stacking Shelves (U)
9	Cycling (U)	9	Walking in Parking Lot (S)	9	Running (U)	9	Vacuum Cleaning (U)
10	Nordic Walking (U)	10	Walking on Treadmill (Flat) (S)	10	Jogging (U)	10	Walking (U)
11	Ascending Stairs (U)	11	Walking on Treadmill (15 deg) (S)	11	Cycling (U)	11	Cycling (U)
12	Descending Stairs (U)	12	Running on Treadmill (S)	12	Knees Bending (U)	–	–
–	–	13	Exercising on Stepper (S)	–	–	–	–
–	–	14	Exercising on Cross Trainer (S)	–	–	–	–
–	–	15	Playing Basketball (U)	–	–	–	–
–	–	16	Jumping (U)	–	–	–	–
–	–	17	Rowing (U)	–	–	–	–
–	–	18	Cycling (Horizontal) (U)	–	–	–	–
–	–	19	Cycling (Vertical) (U)	–	–	–	–

Table 6. Activity Reference for "Both Seen and Unseen" Confusion Matrix (i.e., Fig. 8), Seen: S, Unseen: U

discriminative information in frequency-domain descriptions, which differ only by axis for these pairs. These findings underscore the urgent need for more fine-grained attribute descriptions to further improve zero-shot activity recognition.

5.4.2 ***"Both Seen and Unseen" Setting Result.*** To evaluate the proposed method in a more realistic scenario, we conduct experiments on a test set containing both seen and unseen classes. This setting is more challenging than the “Unseen Only” setting, as similarities between seen and unseen classes can significantly affect decision boundary formation, resulting in an accuracy drop compared to "Unseen Only". Table 5 (full result is in the appendix) presents the average accuracies for seen and unseen classes, along with the harmonic mean for the proposed *IMUZero* framework. Specifically, in the simpler settings (2–3 unseen classes), unseen class accuracy drops by nearly 30%, whereas in the more complex settings (4–5 unseen classes), the drop is only around 8%. One possible explanation is that when there are fewer unseen classes, the prior probability mass of seen classes overwhelms the unseen ones—causing even well-separated unseen examples to be misclassified. The results also show that, even under this more complex inference setting, our approach achieves a 3–5% improvement over the SOTA. These findings demonstrate that the generated semantic information substantially enhances the generalization ability of our framework in real-world scenarios.

It should be noted that the model can overfit to seen classes during semantic mapping, which may degrade inference performance. However, the proposed framework applies the self-calibration loss L_{SC} to explicitly shift some predictive probability from known to unknown classes. Consequently, in most cases, *IMUZero* achieves balanced accuracy on both seen and unseen classes. Furthermore, Figure 8 presents confusion matrices of activity recognition under the “Both Seen and Unseen” setting (5 classes) for our proposed framework. We observe that inference on unseen classes is indeed affected by interference from seen classes. For example, in the MHEALTH dataset, the model frequently confuses Running, Jogging, and Walking, as well as Climbing, since these activities share similar leg movement patterns. Similar patterns of misclassification occur in the other datasets. In contrast, the DSADS dataset shows better overall classification performance and fewer misclassifications between seen

and unseen classes. One possible explanation is that DSADS has more balanced samples and more discriminative raw-signal patterns.

5.5 Ablation Study on Key Components

To provide further insight into *IMUZero* , we conduct ablation studies to evaluate the effects of the: 1) Frequency modality fusion; 2) Shuffle Channel Order Constraint $\mathcal{L}_{SCOC}$; 3) Self Calibration Loss $\mathcal{L}_{SC}$; and 3) Attribute Regression Loss $\mathcal{L}_{AR}$. Due to the computational limitation, we used the PAMAP2 dataset with different ZS settings (from 2 to 5 classes). From Fig. 9a, we can observe that each key components earn its own credit for the final performance. Specifically, the attribute regression loss $\mathcal{L}_{AR}$ gains more credits compared to other components, as it aims to construct better signal-to-semantic mapping by narrowing the distance between the textual attributes of the signal feature domains. The Shuffle Channel Order Constraint $\mathcal{L}_{SCOC}$ and Self Calibration Loss $\mathcal{L}_{SC}$ contribute similarly to the overall framework, suggesting that extracting channel-invariant information and implementing constraints to prevent overfitting on seen classes can significantly enhance the model's generalization ability. The fusion of frequency modalities also demonstrates significant improvements in more complex zero-shot (ZS) settings (i.e., 3-4 classes). The incorporation of frequency modalities substantially enhances the semantic connection between signals and frequency-related movement attributes.

5.6 Ablation Study on Semantic Embedding Assessment

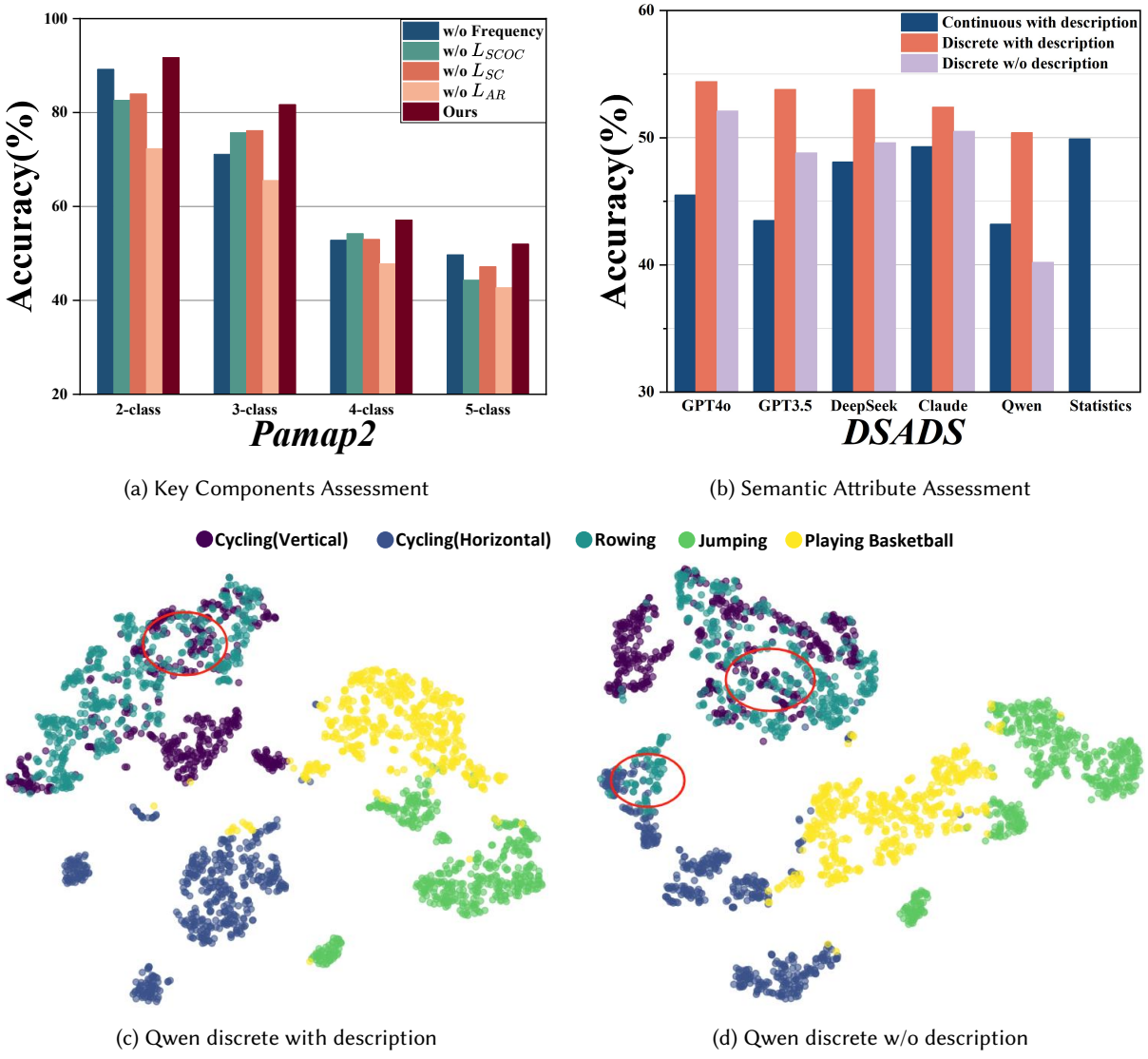
The quality of semantic attributes critically shapes the semantic embedding space and thus determines how well unseen activities can be generalized. To evaluate the fidelity of LLM-generated attributes, we perform a comprehensive quantitative analysis across several attribute formats—including: numerical attribute (i.e., z^c), continuous versus discrete outputs from different LLM, and statistical feature–based descriptors. The results are summarized in Figure 9b and discussed below:

Numerical Attribute vs. Attribute Description. We observe that direct use of the numerical attribute (i.e., z_c) leads to substantial performance degradation. For example, for Qwen-generated attributes, the performance gap between using purely numerical values and language descriptions can be as high as 10%. This suggests that language descriptions supply additional semantic information, enabling the generation of more discriminative decision boundaries for classification. Furthermore, we provide a t-SNE visualization to investigate interpretability. We observe that, with language descriptions, the latent space is more separable than with purely numerical attributes, further validating our assumption.

Continuous vs. Discrete Semantics. We observe that most LLM-generated attribute descriptions tend to favor discrete semantics. For instance, GPT4o and GPT3.5 exhibit discrete proportions of 54.4% and 53.8%, respectively, both higher than their continuous counterparts. This suggests that LLM inherently generate attribute semantics that emphasize categorical separability, which may enhance the discriminability of unseen classes. However, models such as Claude demonstrate a more balanced distribution (49.3% continuous vs. 52.4% discrete), potentially offering a compromise between generalization and specificity.

Statistical vs. LLM-Generated Attributes. To demonstrate the reliability of LLM-generated attributes, we compare them with statistical features extracted from accelerometer readings across different activities (see Fig. 10). For consistency, the mean value is used as the representative statistical attribute. These statistical descriptors exhibit a high continuous ratio (49.9%), indicating smooth semantic transitions that can support more nuanced mappings in complex activity distributions. In contrast, LLM-generated attributes lean toward discrete categorization, providing clearer semantic anchors that enhance zero-shot recognition performance, albeit with a potential trade-off in generalization across fine-grained action variations.

LLM Variant Differences. All evaluated LLM generate semantically coherent attribute descriptions that yield strong zero-shot recognition performance, confirming the reliability of LLM-driven semantics. Although there



(a) Key Components Assessment

(b) Semantic Attribute Assessment

(c) Qwen discrete with description

(d) Qwen discrete w/o description

Fig. 9. Ablation Study on *IMUZero* Key Components and Semantic Quality. (best viewed in color)

are minor variations in how each model balances continuous versus discrete attributes—Claude remains nearly balanced while qwen skews slightly toward discrete—GPT4o achieves the best overall trade-off, combining clear class separability with sufficient continuous nuance. Therefore, GPT4o is selected to produce the final semantic attributes in our framework.

5.7 Latent Space Analysis

To further verify the effectiveness of our *IMUZero* framework–(whether it presents better unseen semantic latent space), we applied t-SNE to generate visualizations of latent features for the PAMAP2 dataset. These

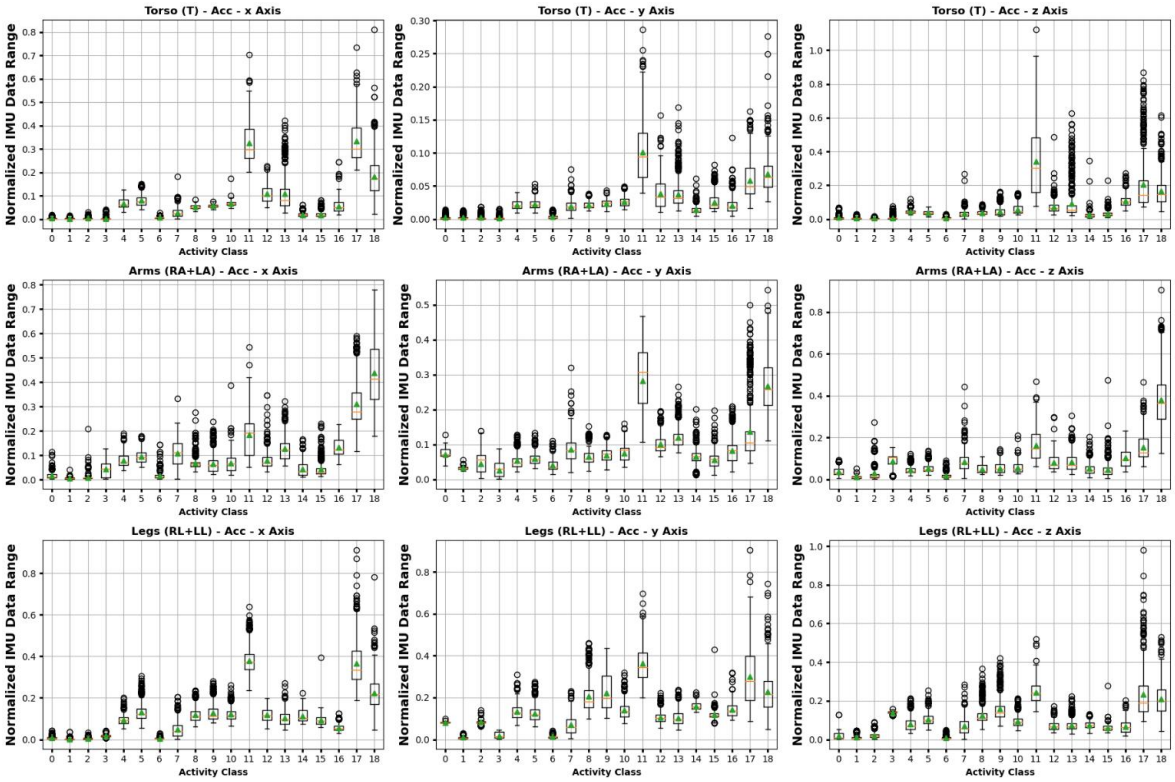


Fig. 10. Statistical Representation of Attributes

visualizations are based on the penultimate layer (i.e., the last layer before the final classifier) for our method and all baselines Fig. 11 illustrates the t-SNE plot for unseen activity distribution for the raw unseen activity data, HRT encoded features, SHIP encoded features, and *IMUZero* encoded features. We observe that the clusters of activity embeddings from *IMUZero* (see Fig. 11g) are more distinct and organized compared to those derived from HRT and SHIP features. In the case of SHIP-encoded features (see Fig. 11c), the decision boundaries between different classes are unclear, leading to potential confusion between activities such as "Vacuum Cleaning" and "Ironing," as well as "Ascending Stair" and "Descending Stair." The decision boundaries for HRT-encoded features are even less defined, resulting in smaller inter-class distances, which can lead to failures in inferring unseen activities. Additionally, the absence of extra frequency-sensitive information makes it challenging to extract discriminative and generalizable features from the raw activity signals.

5.8 Future Work Discussion: Test-time Zero-Shot Recognition

Although current zero-shot action recognition can address the issue of analyzing unknown motion categories to some extent, it relies too heavily on pre-defined labels or attribute descriptions, limiting its flexibility in responding to unknown motion categories in real life. Meanwhile, such methods are overly dependent on the quality of the attribute descriptions, and any bias in the descriptions can lead to significant differences in results. An ideal solution is one in which unseen activity recognition does not rely on predefined labels or attribute spaces (i.e., the labels or descriptions of unseen classes are not available during training). Instead, the model can

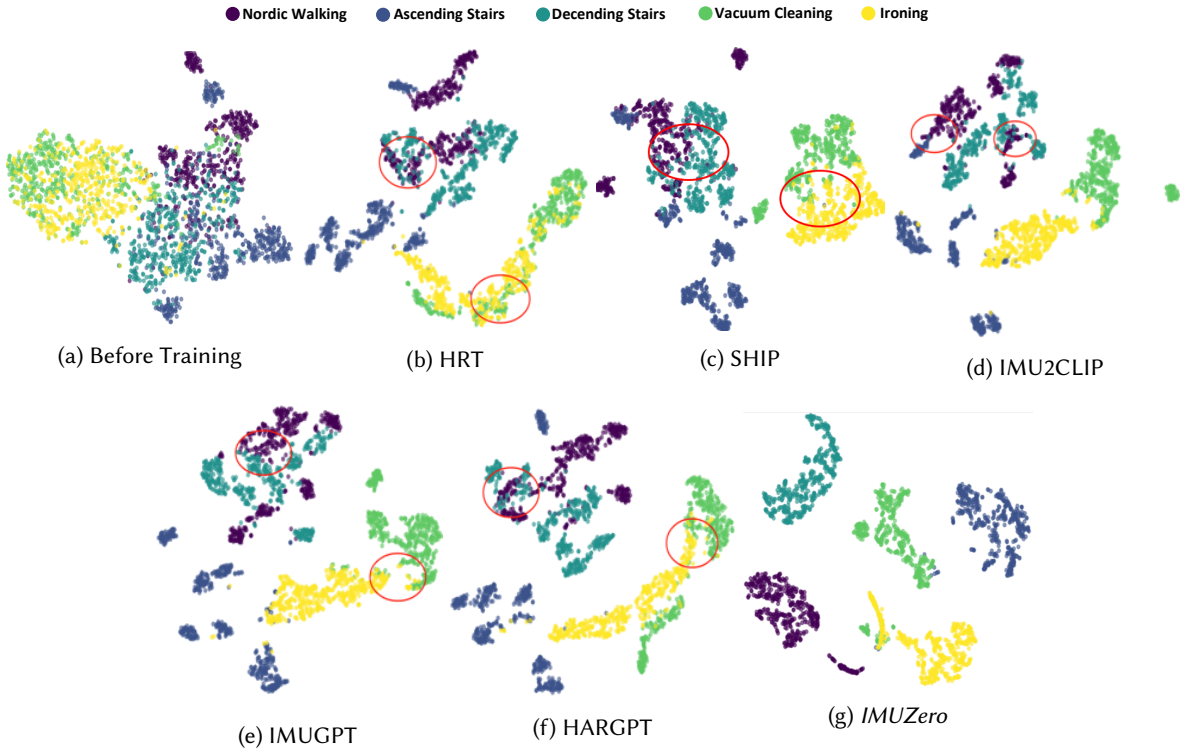


Fig. 11. Feature visualization: t-SNE plot of raw unseen activity feature, HRT encoded features, SHIP encoded features, IMU2CLIP encoded features, IMUGPT encoded features, HARGPT encoded features and *IMUZero* encoded features on PAMAP2 Dataset. We use different colors to denote different categories. (Best viewed in color.)

generate these descriptions during inference time for label/description/attribute nearest search. We define this setting as *"Test-time" Zero-Shot Recognition*. To investigate the performance of current state-of-the-art (SOTA) ZSL approaches, as well as the proposed *IMUZero* , we conduct experiments under the Test-time Zero-Shot Recognition setting using the PAMAP2 dataset. Table 7 demonstrates the performance of test-time zero-shot recognition, and we observe that all methods fail to generalize to unseen classes. This phenomenon indicates that, without prior knowledge, the ZSL system is prone to significant failures. Such challenges present a novel and urgent opportunity for the HAR community and could encourage the development of a universal HAR system.

6 Conclusion

In summary, this work introduces *IMUZero*, a fine-grained end-to-end zero-shot human activity recognition framework designed to enhance the recognition of unseen and novel activity classes. The proposed framework integrates three major components: the LLM-based Attribute Generation component that leverages predefined category information to produce fine-grained attributes through prompt engineering; the Multi-Scale Time-Frequency Fusion module that effectively consolidates semantically related information from both the input activity signals and their frequency representations, facilitating precise signal-to-semantic mapping; and the Sig2Text Alignment module, which aligns fused cross-modality features with encoded attributes while addressing

Method	*Pamap2*			
	2-class	3-class	4-class	5-class
DeepConvLSTM [27]	50.47%	30.71%	25.00%	20.00%
Composer [15]	48.21%	36.57%	28.96%	21.20%
DAZLE [16]	47.60%	33.33%	24.89%	14.76%
HSVA [5]	49.83%	37.57%	29.95%	19.09%
HRT [6]	50.35%	**44.75%**	33.23%	18.56%
SHIP [39]	46.57%	37.65%	30.33%	**24.52%**
IMU2CLIP [25]	49.80%	38.20%	31.50%	20.30%
IMUGPT [22]	**52.10%**	41.90%	33.80%	22.50%
HARGPT [17]	50.50%	42.80%	31.90%	20.10%
Ours	51.22%	43.66%	**35.63%**	19.86%

Table 7. Experimental Test on Test-time Zero-Shot Recognition Setting

the axial bias problem through a channel shuffle constraint to promote channel-invariant information extraction. Through our extensive experimental evaluation, we have demonstrated the effectiveness of the *IMUZero* framework, particularly in its ability to bridge the gap between activity signals and fine-grained attributes generated by LLM. The introduction of a novel cross-modality multi-stage fusion mechanism enables fine-grained semantic alignment, while our investigation into 'axis bias' and the implementation of a channel shuffle order constraint effectively enhance the model's generalization capabilities. The promising results of our experiments underscore *IMUZero* as a versatile and robust solution for the HAR research community, paving the way for broader applications in real-world scenarios.

References

[1] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[2] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. 2021. Knowledge-aware Zero-Shot Learning: Survey and Perspective. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4366–4373. https://doi.org/10.24963/ijcai.2021/597 Survey Track.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[4] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. 2022. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 330–338.

[5] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. 2021. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems* 34 (2021), 16622–16634.

[6] De Cheng, Gerong Wang, Bo Wang, Qiang Zhang, Jungong Han, and Dingwen Zhang. 2023. Hybrid routing transformer for zero-shot learning. *Pattern Recognition* 137 (2023), 109270.

[7] Heng-Tze Cheng, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. Towards zero-shot learning for human activity recognition using semantic attribute sequence model. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 355–358.

[8] Heng-Tze Cheng, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. Towards zero-shot learning for human activity recognition using semantic attribute sequence model. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 355–358. https://doi.org/10.1145/2493432.2493511

[9] Pierre Duhamel and Martin Vetterli. 1990. Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing* 19, 4 (1990), 259–299.
[10] Yan Gao, Yang Long, Yu Guan, Anna Basu, Jessica Baggaley, and Thomas Ploetz. 2019. Towards Reliable, Automated General Movement Assessment for Perinatal Stroke Screening in Infants Using Wearable Accelerometers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 12 (March 2019), 22 pages. https://doi.org/10.1145/3314399
[11] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.
[12] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1112–1123.
[13] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016).
[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[15] Dat Huynh and Ehsan Elhamifar. 2020. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems* 33 (2020), 19849–19860.
[16] Dat Huynh and Ehsan Elhamifar. 2020. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4483–4493.
[17] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? *arXiv preprint arXiv:2403.02727* (2024).
[18] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. 2017. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision*. 4223–4232.
[19] Aftab Khan, James Nicholson, and Thomas Plötz. 2017. Activity Recognition for Quality Assessment of Batting Shots in Cricket Using a Hierarchical Representation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 62 (Sept. 2017), 31 pages. https://doi.org/10.1145/3130927
[20] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.
[21] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.
[22] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. 2024. Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–32.
[23] Taotao Li, Zhenyu Wen, Yang Long, Zhen Hong, Shilian Zheng, Li Yu, Bo Chen, Xiaoniu Yang, and Ling Shao. 2023. The importance of expert knowledge for automatic modulation open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13730–13748.
[24] Moe Matsuki, Paula Lago, and Sozo Inoue. 2019. Characterizing word embeddings for zero-shot sensor-based human activity recognition. *Sensors* 19, 22 (2019), 5043.
[25] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395* (2022).
[26] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 100–103.
[27] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.
[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
[29] Thomas Plötz, Nils Y. Hammerla, Agata Rozga, Andrea Reavis, Nathan Call, and Gregory D. Abowd. 2012. Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 391–400. https://doi.org/10.1145/2370216.2370276
[30] WJ Scheirer, A Rocha, A Sapkota, and TE Boult. 2013. Towards Open Set Recognition, TPAMI. *Cited on* (2013), 54.
[31] Shuai Shao, Yu Guan, Bing Zhai, Paolo Missier, and Thomas Plötz. 2023. ConvBoost: Boosting ConvNets for sensor-based activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–21.
[32] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. 2018. Transductive unbiased embedding for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1024–1033.

[33] Jie Su, Peng Sun, Yuting Jiang, Zhenyu Wen, Fangda Guo, Yiming Wu, Zhen Hong, Haoran Duan, Yawen Huang, Rajiv Ranjan, et al. 2024. A semantic-consistent few-shot modulation recognition framework for IoT applications. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[34] Jie Su, Zhenyu Wen, Tao Lin, and Yu Guan. 2022. Learning disentangled behaviour patterns for wearable-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–19.

[35] Catherine Tong, Jinchen Ge, and Nicholas D Lane. 2021. Zero-shot learning for imu-based activity recognition using video embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–23.

[36] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[37] Wei Wang and Qingzhong Li. 2023. Generalized zero-shot activity recognition with embedding-based method. *ACM Transactions on Sensor Networks* 19, 3 (2023), 1–25.

[38] Wei Wang, Chunyan Miao, and Shuji Hao. 2017. Zero-shot human activity recognition via nonlinear compatibility based method. In *Proceedings of the International Conference on Web Intelligence*. 322–330.

[39] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. 2023. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3032–3042.

[40] Nan Wu, Hiroshi Kera, and Kazuhiko Kawamoto. 2023. Improving zero-shot action recognition using human instruction with text description. *Applied Intelligence* 53, 20 (2023), 24142–24156.

[41] Tong Wu, Yiqiang Chen, Yang Gu, Jiwei Wang, Siyu Zhang, and Zhanghu Zhechen. 2020. Multi-layer cross loss model for zero-shot human activity recognition. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*. Springer, 210–221.

[42] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10275–10284.

[43] 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*. 1–15.

[44] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition.. In *Ijcai*, Vol. 15. Buenos Aires, Argentina, 3995–4001.

[45] Vladimir M Zatsiorsky. 2002. *Kinetics of human motion*. Human kinetics.

[46] Bing Zhai, Ignacio Perez-Pozuelo, Emma A. D. Clifton, Joao Palotti, and Yu Guan. 2020. Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 67 (June 2020), 33 pages. https://doi.org/10.1145/3397325

[47] Junru Zhang, Lang Feng, Zhidan Liu, Yuhan Wu, Yang He, Yabo Dong, and Duanqing Xu. 2024. Diverse Intra-and Inter-Domain Activity Style Fusion for Cross-Person Generalization in Activity Recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4213–4222.

[48] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. 2018. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 269–284.

[49] Yexu Zhou, Haibin Zhao, Yiran Huang, Tobias Röddiger, Murat Kurnaz, Till Riedel, and Michael Beigl. 2024. AutoAugHAR: Automated Data Augmentation for Sensor-based Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–27.

A GZSL result

		Pamap2				DSADS			
		2-class	3-class	4-class	5-class	2-class	3-class	4-class	5-class
DeepConvLSTM [27]	U	36.2%	34.5%	33.0%	30.1%	44.8%	48.0%	25.0%	20.0%
	S	43.4%	41.4%	39.6%	36.1%	79.5%	60.0%	44.7%	38.8%
	H	39.5%	37.6%	36.0%	33.0%	57.4%	53.3%	32.1%	26.4%
Composer [15]	U	37.2%	35.4%	35.0%	31.9%	46.7%	41.8%	39.0%	25.0%
	S	44.6%	42.5%	42.0%	38.3%	81.8%	61.9%	47.0%	39.8%
	H	40.0%	38.0%	38.3%	34.8%	59.5%	49.9%	42.6%	30.7%
DAZLE [16]	U	22.9%	21.8%	20.8%	19.1%	46.3%	41.7%	39.5%	25.0%
	S	27.5%	26.2%	25.0%	22.2%	81.7%	61.3%	47.5%	40.2%
	H	25.0%	23.9%	22.7%	20.8%	59.1%	49.4%	43.1%	30.8%
HSVA [5]	U	37.4%	35.7%	34.1%	31.2%	47.2%	42.3%	39.5%	37.8%
	S	44.9%	42.8%	40.9%	37.4%	82.3%	62.7%	47.5%	40.0%
	H	40.3%	38.3%	37.2%	34.2%	59.8%	50.7%	43.1%	38.9%
HRT [6]	U	22.1%	21.1%	20.1%	18.4%	48.1%	43.2%	25.0%	29.9%
	S	26.5%	25.3%	24.1%	22.1%	82.9%	62.8%	37.9%	32.0%
	H	24.2%	23.1%	22.0%	20.1%	60.7%	51.0%	30.1%	30.9%
SHIP [39]	U	54.8%	52.2%	49.8%	**48.2%**	49.3%	48.2%	49.0%	**50.5%**
	S	65.8%	62.6%	59.8%	56.7%	84.2%	64.3%	51.0%	48.0%
	H	59.8%	56.9%	54.3%	**50.6%**	62.0%	54.8%	50.0%	49.2%
IMU2CLIP [25]	U	53.8%	51.3%	48.9%	47.4%	48.7%	47.1%	50.0%	49.7%
	S	64.6%	61.6%	58.7%	54.4%	83.4%	63.7%	49.0%	47.0%
	H	58.8%	55.9%	53.3%	45.5%	61.2%	54.0%	49.5%	48.3%
IMUGPT [22]	U	52.2%	49.8%	47.5%	46.2%	49.6%	50.2%	46.0%	49.4%
	S	62.6%	59.8%	57.0%	53.6%	84.7%	65.1%	49.0%	47.0%
	H	57.2%	54.4%	51.9%	44.3%	62.3%	56.5%	47.5%	48.2%
HARGPT [17]	U	49.2%	46.9%	44.8%	43.9%	49.0%	48.4%	47.0%	45.9%
	S	59.0%	56.3%	53.8%	50.2%	84.0%	64.4%	47.5%	44.0%
	H	53.9%	51.2%	48.8%	42.0%	61.8%	55.2%	47.2%	44.9%
IMUZero	U	**56.2%±0.5%**	**53.5%±0.6%**	**51.1%±0.7%**	46.4%±0.5%	**50.8%±0.8%**	**62.2%±0.9%**	**50.1%±0.8%**	49.0%±1.2%
	S	**67.4%±0.3%**	**64.2%±0.4%**	**61.3%±0.5%**	**58.2%±0.5%**	**86.1%±0.5%**	**66.3%±0.6%**	**55.0%±0.5%**	**51.6%±0.5%**
	H	**61.3%±0.5%**	**58.3%±0.5%**	**55.6%±0.4%**	47.4%±0.5%	**63.9%±0.7%**	**64.2%±0.8%**	**52.4%±0.5%**	**50.3%±0.7%**

Table 8. Performance comparison (U: Unseen accuracy %, S: Seen accuracy %, H: Harmonic mean %) on Pamap2 and DSADS datasets for 2,3,4,5-class settings in "Both Seen and Unseen" setting. Best results are highlighted in **bold**, and second-best results are underlined.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

		MHEALTH				GOTOV			
		2-class	3-class	4-class	5-class	2-class	3-class	4-class	5-class
DeepConvLSTM [27]	U	43.9%	47.8%	39.3%	33.7%	44.7%	47.2%	40.0%	37.8%
	S	68.0%	43.0%	60.9%	52.2%	67.1%	42.5%	60.0%	56.7%
	H	53.2%	45.3%	47.6%	40.9%	53.6%	44.7%	48.0%	45.5%
Composer [15]	U	44.9%	42.6%	40.3%	34.5%	48.7%	46.2%	43.7%	40.9%
	S	69.6%	66.0%	62.5%	53.5%	73.1%	69.3%	65.8%	61.4%
	H	54.4%	51.6%	48.8%	41.9%	58.4%	55.3%	52.4%	49.0%
DAZLE [16]	U	18.3%	17.4%	16.5%	13.5%	33.7%	31.9%	30.2%	27.8%
	S	28.4%	27.0%	25.6%	20.9%	50.6%	47.9%	45.3%	41.7%
	H	22.2%	21.1%	20.0%	16.4%	40.4%	38.2%	36.2%	33.3%
HSVA [5]	U	45.1%	42.8%	40.5%	34.6%	44.0%	41.7%	39.4%	36.7%
	S	69.9%	66.3%	62.8%	53.6%	66.0%	62.6%	59.1%	55.1%
	H	54.6%	51.8%	49.0%	42.0%	52.8%	49.8%	47.3%	44.0%
HRT [6]	U	27.5%	26.1%	24.7%	21.2%	43.6%	41.3%	39.1%	35.8%
	S	42.6%	40.5%	29.2%	32.9%	65.4%	37.2%	35.2%	53.7%
	H	33.3%	31.6%	26.8%	25.7%	52.3%	39.1%	37.0%	42.8%
SHIP [39]	U	58.5%	55.5%	52.5%	**48.8%**	56.7%	**53.8%**	50.9%	**47.3%**
	S	90.7%	86.0%	81.4%	45.3%	85.1%	**80.7%**	76.4%	44.8%
	H	70.8%	67.5%	63.3%	46.4%	67.8%	**64.5%**	61.1%	46.0%
IMU2CLIP [25]	U	57.5%	54.6%	51.7%	48.0%	47.8%	45.3%	42.9%	39.1%
	S	89.1%	84.6%	80.1%	45.5%	71.7%	68.0%	64.4%	62.6%
	H	69.8%	66.5%	62.3%	**46.7%**	57.3%	54.3%	51.5%	48.2%
IMUGPT [22]	U	55.8%	52.9%	50.1%	46.5%	52.3%	49.5%	46.8%	43.6%
	S	86.5%	82.0%	77.7%	44.0%	78.5%	74.3%	70.2%	60.8%
	H	67.8%	64.5%	60.5%	45.2%	62.6%	59.6%	56.2%	50.8%
HARGPT [17]	U	52.7%	50.0%	47.3%	44.0%	46.7%	44.2%	41.8%	38.7%
	S	81.7%	77.5%	73.3%	41.5%	70.1%	66.3%	62.7%	61.9%
	H	64.0%	60.8%	57.6%	42.7%	56.0%	53.0%	50.2%	47.6%
IMUZero	U	**59.6%±0.5%**	**56.5%±0.5%**	**53.5%±0.7%**	35.3%±0.8%	**59.6%±0.5%**	52.6%±0.6%	**53.5%±0.7%**	46.5%±0.8%
	S	**92.4%±0.3%**	**87.6%±0.4%**	**82.9%±0.5%**	**54.7%±0.6%**	**89.4%±0.3%**	78.9%±0.4%	**80.3%±0.4%**	**64.0%±0.5%**
	H	**71.8%±0.4%**	**68.6%±0.4%**	**64.3%±0.6%**	43.0%±0.7%	**71.5%±0.4%**	63.3%±0.5%	**64.2%±0.6%**	**53.3%±0.5%**

Table 9. Performance comparison (U: Unseen accuracy %, S: Seen accuracy %, H: Harmonic mean %) on MHEALTH and GOTOV datasets for 2,3,4,5-class settings in "Both Seen and Unseen" setting. Best results are highlighted in **bold**, and second-best results are underlined.