ZeroHAR: Sensor Context Augments Zero-Shot Wearable Action Recognition

Ranak Roy Chowdhury[1], Ritvik Kapila[1], Ameya Panse[1], Xiyuan Zhang[1], Diyan Teng[2], Rashmi Kulkarni[2], Dezhi Hong[3]*, Rajesh K. Gupta[1], Jingbo Shang[1]

[1] University of California San Diego
[2] Qualcomm
[3] Amazon
{rrchowdh, rkapila, apanse, xiz032, rgupta, jshang}@ucsd.edu, {diyateng, rashmik}@qti.qualcomm.com, hondezhi@amazon.com

Abstract

Wearable Human Action Recognition (wHAR) uses motion sensor data to identify human movements, which is essential for mobile and wearable devices. However, traditional wHAR systems are only trained on a limited set of activities. Hence, they fail to generalize to diverse human motions, prompting Zero-Shot Learning (ZSL). Existing ZSL methods for wHAR focus solely on augmenting labels, such as representing them as attribute matrices, images, videos, or text. We propose ZeroHAR that enhances ZSL by not just focusing on activity labels, but also augmenting motion data with sensor context features. Our approach incorporates information about the sensor type, the Cartesian axis of the data, and the sensor's body position, providing the model with crucial spatial and biomechanical insights. This helps the model to generalize better to new actions. First, we train the model by aligning the latent space of the motion time series with its corresponding sensor context, while distancing it from unrelated sensor contexts. Then, we train the model using the target activity descriptions. We tested our method against eight baselines on five benchmark HAR datasets with various sensors, placements, and activities. Our model shows exceptional generalizability across the 18 motion time series classification benchmark datasets, outperforming the best baselines by 262% in the zero-shot setting.

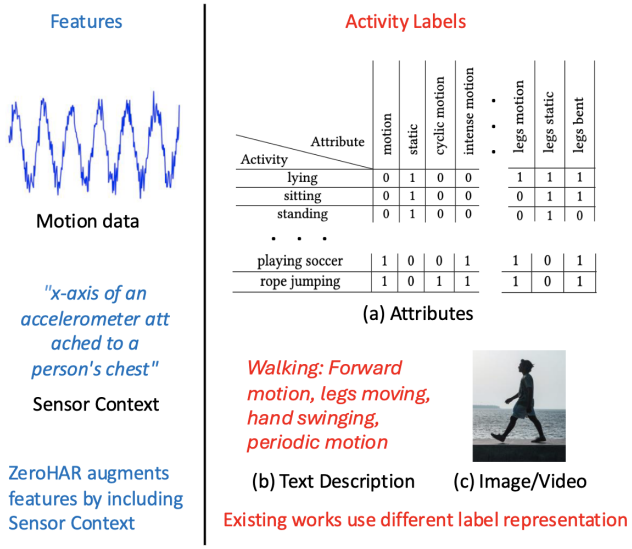


Figure 1: Different Approaches to Zero-shot Inertial HAR.

Introduction

Wearable HAR predicts human activities using data from Inertial Measurement Units (IMUs). Existing inertial HAR systems are often trained on data from a limited set of motions, collected and annotated within a controlled setting. These models do not recognize the richer and more diverse set of motions that humans exhibit in the real world. Annotating vast amounts of IMU data for all possible human movements to train a HAR model is not plausible. Hence, in HAR, although a model may be trained on a limited set of activities, we expect it to recognize unseen activities after deployment, a process known as Zero-Shot Learning (ZSL).

ZSL involves training a model on data from seen classes and evaluating it on test data from unseen classes. To achieve this objective, ZSL is trained to learn fine-grained attributes shared among classes that can be generalized to recognize unseen classes. For example, a model may be trained on IMU data for "walking" but is expected to recognize "running" during test time. Both these activities share several similarities in terms of body posture and limb movements. For example, both involve an upright posture with a straight spine and head facing forward, the arms swing alternately with the legs, and they both require a cyclic movement pattern of the legs. Such basic body movements constitute activities at large. Hence, a model that learns these low-level fine-grained knowledge from the activities that it were trained on, can recognize unseen activities at test time.

Figure 1 shows how existing work in ZSL extracts fine-grained details about human activity from different representations of these activities: **(a)** Activity-attribute matrices that encode basic limb movements and body posture information in a binary matrix (Wang, Miao, and Hao 2017; Cheng et al. 2013b,a); **(b)** Text describing the nature and type of limb and joint movements involved in various activity and feeding the text through a language model to extract embeddings (Matsuki, Lago, and Inoue 2019; Wu et al. 2020; Moon et al. 2022); **(c)** Images or videos of people

*Work unrelated to Amazon.

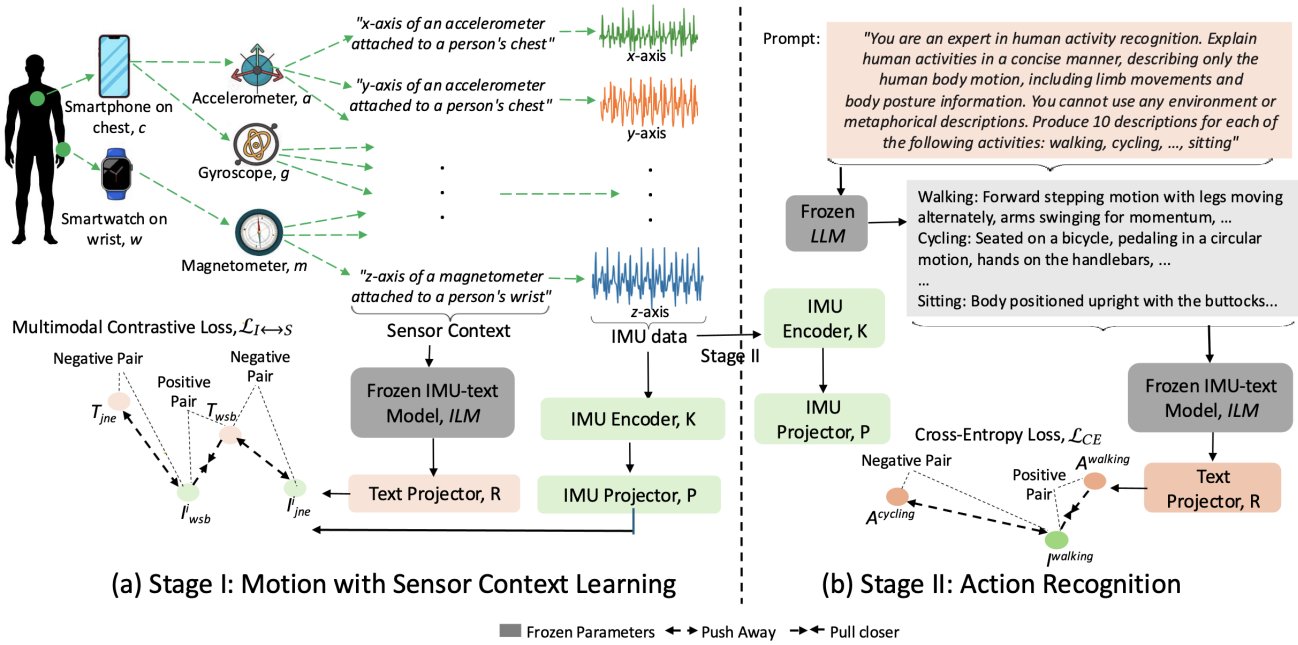


Figure 2: Overview of ZeroHAR: (a) Stage I (Motion with Sensor Context Learning): IMU and Sensor Context embeddings generated by the respective model branches are trained through Multimodal Contrastive Learning. (b) Stage II (Action Recognition): IMU and Activity Description embeddings are trained through Cross-Entropy Loss.

	Accelerometer	Gyroscope	Magnetometer
Metric	Linear Acc	Angular Vel	Magnetic Field Strength
X-axis	Left-Right	Forward-Backward	East-West
Y-axis	Up-Down	Left-Right	North-South
Z-axis	Forward-Backward	Clock-Counterclock	Upward-Downward

Table 1: Metric and directionality of IMU sensors.

conducting activities and feeding image frames to a vision model to extract embeddings.

The above methods use motion data as input features to the model. These motion data are generated by different IMU sensors located at different body positions. Table 1 shows the different types of IMU sensors, their metric and directionality. We use the term *sensor context* to refer to the information about sensor type, axis, and body position where the sensor is located.

We hypothesize that adding sensor context to IMU data provides valuable spatial and biomechanical information, helping the model learn to recognize actions more effectively. For instance, when identifying the action "picking up an object," IMU data alone might produce similar patterns to actions like "tying shoelaces" or "squatting," causing confusion. However, if the model knows that the upward movement is detected by an accelerometer on the wrist and the downward movement by an accelerometer on the thigh, it can interpret the overall motion as "upward wrist movement paired with thigh bending," allowing it to accurately distinguish "picking up" from similar actions.

To this end, we propose ZeroHAR, to leverage sensor context knowledge in ZSL for wearable HAR. Fig 2 shows ZeroHAR's two-stage training setup. In Stage I, we compliment each IMU measurement with its corresponding sensor context provided as text. ZeroHAR is trained on this IMU and sensor context through multimodal contrastive loss. During training, the model learns a joint IMU-text latent space by bringing the latent space of IMU representation closer to its corresponding sensor context information. In Stage II, we prompt an LLM to generate precise, fine-grained bio-mechanical information about human activity. The pre-initialized model from Stage I is then trained to recognize actions from an IMU with its corresponding activity description as the label.

We extensively evaluate ZeroHAR with 12 baselines on 18 benchmark HAR datasets, covering a wide variety of number and type of IMU sensors and range of human motions. ZeroHAR resulted in a 262% average improvement in Zero-Shot Accuracy over the 2^{nd} best results. We also conduct ablations and case study to show how the addition of sensor context in the IMU latent space helps ZeroHAR to generalize to unseen classes. Our main contributions are as follows:

- We enhance HAR by integrating motion data with sensor context for spatial and biomechanical insights.
- We propose a two-stage ZSL framework: combining IMU with sensor context, then performing activity recognition.
- We present new state-of-the-art performance for ZSL on 18 benchmark wearable HAR datasets.

Related Works

Zero-Shot Learning (ZSL) extends learned knowledge from known training classes to unknown classes in testing. ZSL techniques capture shared low-level semantics among classes (Socher et al. 2013; Gupta et al. 2023). One-hot encoding of classes is inadequate to represent such semantics. Some ZSL studies for Inertial HAR manually designs class-attribute matrix, including body posture and limb motion details. (Wang, Miao, and Hao 2017; Cheng et al. 2013b,a) (Fig 1(a)). While insightful, manual design is impractical for large-scale HAR due to diverse range of human motions.

Others represent class names as embeddings to learn semantic space (Matsuki, Lago, and Inoue 2019; Wu et al. 2020) (Fig 1(b)), mitigating domain expertise. Pre-trained models for motion time series are inspired by the recent success of large language or multimodal models. ImageBind (Girdhar et al. 2023) and IMU2CLIP (Moon et al. 2022) leverage recent large vision-language models (Radford et al. 2021) to learn a joint embedding across multiple modalities including motion time series and text. However, both ImageBind and IMU2CLIP are trained on motion time series collected from head-mounted devices (Grauman et al. 2022), limiting their generalizability across different device locations and orientations. Furthermore, several studies have explored directly applying LLMs for motion time series classification. For example, HARGPT (Ji, Zheng, and Wu 2024) processes raw motion time series through LLMs and incorporates role-play and chain-of-thought strategies for prompting. ContextGPT (Arrotta et al. 2024) designs prompt engineering approaches leveraging context information. However, since LLMs are not directly trained on raw motion time series, such methods require extensive context information that is not usually available, and struggle with accurately recognizing complex activities.

ZeroHAR incorporates context by complementing motion data with sensor metadata, providing crucial spatial and anatomical insights for understanding biomechanics. Its two-stage pipeline first integrates sensor context into the latent space, enhancing action recognition in the next stage.

Methodology

Fig 2 shows a schematic diagram of ZeroHAR. We first present the problem setting, followed by our two-stage training recipe, namely Stage I: IMU-Text Alignment (i.e., Fig 2(a)) and Stage II: Action Recognition (i.e, Fig 2(b)).

Problem Statement

Let O and U be two disjoint sets of observed and unobserved activities, respectively, i.e. $O \cap U = \emptyset$. Let $G = O \cup U$ be the set of all activities, and g denote a given activity, $g \in G$. All the labeled instances for training are from the observed activities in O. More formally, the training set is $\mathcal{D}_{tr} = \{(X^i, Y^i)|i = 1, 2, 3, ..., N\}$ where N is the number of training data points, X^i is a multi-variate time series, and $Y^i \in O$ is the activity label corresponding to X^i. In ZSL, the data and labels for the test set come only from the *unobserved* classes: $\mathcal{D}_{te} = \{(X^i, Y^i)|i = 1, 2, 3, ..., Z\}$, where Z is the number of test data points and $Y^i \in U$.

Algorithm 1: Stage I: Motion with Sensor Context Learning

Input: $\mathcal{D}_{tr}$, B, M, W, ILM
Hyper-parameters: τ
Output: Trained (K and R)

1: K, P, and R initialized randomly
2: **for** X^i in $\mathcal{D}_{tr}$ **do**
3: **for** b in B **do**
4: **for** s in M_b **do**
5: **for** w in W **do**
6: $t_{wsb} \leftarrow$ "$w-$axis of s attached to b"
7: $T_{wsb} \leftarrow R(ILM(t_{wsb}))$, $T_{wsb} \in \mathbb{R}^h$
8: $I^i_{wsb} \leftarrow P(K(X^i_{wsb}))$, $I^i_{wsb} \in \mathbb{R}^h$
9: Compute $\mathcal{L}_{I^i_{wsb} \rightarrow T_{wsb}}$ & $\mathcal{L}_{T_{wsb} \rightarrow I^i_{wsb}}$ (Eq. (1) & (2), respectively)
10: **end for**
11: **end for**
12: **end for**
13: Compute $\mathcal{L}_{I^i \rightarrow T}$ & $\mathcal{L}_{T \rightarrow I^i}$ (Eq. (3) & (4), respectively)
14: Compute $\mathcal{L}_{I^i \leftrightarrow T}$ from $\mathcal{L}_{I^i \rightarrow T}$ and $\mathcal{L}_{T \rightarrow I^i}$ (Eq. (5))
15: Update K, P, and R based on $\mathcal{L}_{I^i \leftrightarrow T}$
16: **end for**
17: **return** Trained (K and R)

Let B be the set of body positions with wearable devices, and b denote a given body position, $b \in B$. Let M_b denote the set of IMU sensors at body position b. Then M denotes the set of IMU sensors at all body positions, hence $M = \{M_1, M_2, ..., M_b, \ldots, M_{|B|}\}$. Hence, the total number of IMU sensors attached to the body is $\sum_{b=1}^{|B|} |M_b|$. If s denotes a given IMU sensor, then M_{sb} denotes IMU sensor s at body position b. Let W denote the set of axes along which an IMU records measurements and w denote a given axis, $w \in W$. All IMUs record data along $x-$, $y-$, and $z-$ axes, so W is constant for every IMU sensor, $W = \{x, y, z\}$, $|W| = 3$. So the total number of channels in a given data point is $|W| \sum_{b=1}^{|B|} |M_b|$. If l is the number of timestamps in the data, then $X^i \in \mathbb{R}^{|W| \sum_{b=1}^{|B|} |M_b| \times l}$. And $X^i_{wsb} \in \mathbb{R}^l$ denote a uni-variate time series from axis, w, of IMU sensor, s, located at body position, b, for data X^i.

Stage I: Motion with Sensor Context Learning

Sensor Context Construction

Algorithm 1 outlines the procedure. For each IMU time series, we construct its sensor context, t_{wsb}, that corresponds to: "$w-$axis of s attached to b" (Line 10). Complimenting an IMU time series with corresponding sensor context provides rich spatial and anatomical information that helps model better understand biomechanics.

Training We embed t_{wsb} using the text encoder of ImageBind (Girdhar et al. 2023), a Pre-trained IMU-text Model (ILM). Unlike BERT (Kenton and Toutanova 2019), GPT (Radford et al. 2018), or other language models (LM), an ILM is pre-trained with both IMU and text. Hence, ILM's text embedding better aligns with IMU than that of LM's. To preserve the IMU-text alignment that ILM ex-

hibits, we freeze its parameters. The output of ILM is then passed through a text projection block, R, to extract sensor metadata embedding output, T_{wsb}, where $T_{wsb} \in \mathbb{R}^h$ and h is the hidden dimension of the model (Line 11).

ILM's IMU encoder is limited to specific sensors and body positions, lacking generalization to other IMU configurations. Hence, we use a separate IMU encoder, K, with a transformer backbone (Zerveas et al. 2021). The univariate IMU data, X^i_{wsb}, is passed through K, followed by an IMU projection block, P, to extract IMU embeddings I^i_{wsb}, $I^i_{wsb} \in \mathbb{R}^h$ (Line 12). Projection blocks R and P are learnable non-linear layers that project text and IMU, respectively, into a shared latent space.

Cross-Modal Contrastive Learning After obtaining T_{wsb} and I^i_{wsb}, we propose modality-mutual learning for IMU-text alignment. This involves a joint optimization process using a contrastive strategy to refine parameters in both language extraction and sensor encoders. The goal is to align the latent space of IMU embeddings with their corresponding textual sensor metadata embeddings while maintaining separation from unrelated sensor metadata. Contrastive learning (Chen et al. 2020; Jaiswal et al. 2020) pulls similar points closer (anchor and positive) while pushing dissimilar points away (anchor and negative), facilitating good representation learning. We utilize Cross-Modal Contrastive Multiview Coding (CMC) (Tian, Krishnan, and Isola 2020) to achieve similar representation learning capability across different modalities, maximizing the similarity between IMU embedding and its corresponding sensor metadata embedding while minimizing the similarity between all other pairs of embeddings via Information Noise Contrastive Estimation (InfoNCE) (Oord, Li, and Vinyals 2018).

For data X^i, the IMU embedding from w axis of s sensor at body position b, I^i_{wsb}, and its corresponding text embedding, T_{wsb}, are considered a positive pair. To compute the IMU-to-Text Loss, $\mathcal{L}_{I^i_{wsb} \to T_{wsb}}$, I^i_{wsb} and T_{wsb} are anchor and positive, respectively. (I^i_{wsb}, T_{wsb}) forms a positive pair. All other combinations of projections from different sensor channels (I^i_{wsb}, T_{jne}) are treated as negative pairs, where $j = \{1, 2, ..., |W|\}, n = \{1, 2, ..., |M_e|\}, e = \{1, 2, ..., |B|\}$ and $(j \neq w \wedge n \neq s \wedge e \neq b)$. For example, if I^i_{wsb} is the anchor, then T_{wsb} is its positive and T_{jne} is one of the $|W| \sum_{b=1}^{|B|} |M_b| - 1$ negatives. $\mathcal{L}_{I^i_{wsb} \to T_{wsb}}$ for X^i is,

$$\mathcal{L}_{I^i_{wsb} \to T_{wsb}} = -\log \frac{fn(I^i_{wsb}, T_{wsb})}{\sum_{e=1}^{|B|} \sum_{n=1}^{|M_e|} \sum_{j=1}^{|W|} fn(I^i_{wsb}, T_{jne})}, \tag{1}$$

where $fn(A, B) = \exp(sim(A, B))/\tau$ and sim is the cosine similarity function. The numerator computes the similarity score between IMU and and its corresponding sensor metadata embedding while the denominator considers similarities with all available sensor metadata. τ is a temperature parameter to scale the similarities.

Similarly, to compute the Text-to-IMU Loss, $\mathcal{L}_{T_{wsb} \to I^i_{wsb}}$, for data point X^i, (I^i_{wsb}, T_{wsb}) forms a positive pair. All other combinations of projections from different input instances (I^i_{jne}, T_{wsb}) are treated as negative pairs, where $j = \{1, 2, ..., |W|\}, n = \{1, 2, ..., |M_e|\}, e = \{1, 2, ..., |B|\}$ and $(j \neq w \wedge n \neq s \wedge e \neq b)$. If T_{wsb} is the anchor, then I^i_{wsb} is its positive and I^i_{jne} is one of the $|W| \sum_{b=1}^{|B|} |M_b| - 1$ negatives. So $\mathcal{L}_{T_{wsb} \to I^i_{wsb}}$ is calculated as,

$$\mathcal{L}_{T_{wsb} \to I^i_{wsb}} = -\log \frac{fn(I^i_{wsb}, T_{wsb})}{\sum_{e=1}^{|B|} \sum_{n=1}^{|M_e|} \sum_{j=1}^{|W|} fn(I^i_{jne}, T_{wsb})} \tag{2}$$

$\mathcal{L}_{I^i_{wsb} \to T_{wsb}}$ and $\mathcal{L}_{T_{wsb} \to I^i_{wsb}}$ are computed in Line 13. Lines 6 - 13 is parallelized by matrix vectorization for efficient computation.

The total IMU-to-Text Contrastive Loss, $\mathcal{L}_{I^i \to T}$, and Text-to-IMU Contrastive Loss, $\mathcal{L}_{T \to I^i}$, over training instance i are computed as,

$$\mathcal{L}_{I^i \to T} = \frac{1}{|W| \sum_{b=1}^{|B|} |M_b|} \sum_{b=1}^{|B|} \sum_{s=1}^{|M_b|} \sum_{w=1}^{|W|} \mathcal{L}_{I^i_{wsb} \to T_{wsb}} \tag{3}$$

$$\mathcal{L}_{T \to I^i} = \frac{1}{|W| \sum_{b=1}^{|B|} |M_b|} \sum_{b=1}^{|B|} \sum_{s=1}^{|M_b|} \sum_{w=1}^{|W|} \mathcal{L}_{T_{wsb} \to I^i_{wsb}} \tag{4}$$

The total Cross-Modal Contrastive Loss for X^i, $\mathcal{L}_{I^i \leftrightarrow T}$, is computed as the average of the two as follows,

$$\mathcal{L}_{I^i \leftrightarrow T} = \frac{1}{2}(\mathcal{L}_{I^i \to T} + \mathcal{L}_{T^i \to I}) \tag{5}$$

$\mathcal{L}_{I^i \to T}$ and $\mathcal{L}_{T \to I^i}$ are computed in Lines 14 and 15, respectively. $\mathcal{L}_{I^i \leftrightarrow T}$ updates model components, K, R, and P, in Line 16. The loss minimization ensures high similarity scores for correct IMU and textual sensor metadata pairs. This bridges the gap between IMU and text, facilitating cross-modal understanding by joint training.

Algorithm 2: Stage II: Action Recognition

Input: $\mathcal{D}_{tr}$, G, LLM, ILM, Trained K, P, R from Stage I
Hyper-parameters: c
Output: Trained (K, P, and R)

1: ϕ constructed from G
2: $\beta \leftarrow LLM(\phi)$
3: **for** (X^i, Y^i) in $\mathcal{D}_{tr}$ **do**
4: $\quad \alpha \leftarrow R(ILM(\beta))$, $\alpha \in \mathbb{R}^{c|G| \times h}$
5: $\quad A \leftarrow$ Mean c embeddings per class from α, $A \in \mathbb{R}^{|G| \times h}$
6: $\quad A^a \leftarrow$ Embedding for activity, a, where $a = Y^i$, $A^a \in \mathbb{R}^h$
7: $\quad I^i \leftarrow P(K(X^i))$, $I^i \in \mathbb{R}^h$
8: $\quad$ Compute $\mathcal{L}^i_{CE}$ (Eq. (6))
9: $\quad$ Update K, P, and R, based on $\mathcal{L}^i_{CE}$
10: **end for**
11: **return** Trained (K, P, and R)

Stage II: Action Recognition

With K, P, and R pre-initialized via joint IMU and sensor context training, the model has learnt a joint latent space shared between IMU and text modalities. This will help ZeroHAR to learn to recognize activities from IMU.

Activity Description Generation Algorithm 2 outlines our action recognition pipeline. We prompt GPT-4 (OpenAI 2023), a Large Language Model (LLM), to generate fine-grained description for each activity in G which consists of all observed and unobserved activities. Prompt ϕ is detailed in Fig 2(b). Since an activity can be explained in numerous ways, we generate c descriptions per activity to obtain a more stable and less noisy estimate of activity representation. Given the potential for environmental details in LLM-generated descriptions, such as "walk in the park" or "sleeping soundly," we design ϕ to focus strictly on bio-mechanics, body posture, and limb motion, explicitly instructing it to avoid environmental or metaphorical language. The prompt is outlined in Fig 2(b). If $\beta_g = \{\beta_{g_1}, \beta_{g_2}, ..., \beta_{g_c}\}$ denotes a set of c generated descriptions for activity g, then $\beta = \{\beta_1, \beta_2, ..., \beta_g, ..., \beta_{|G|}\}$ (Line 7). We manually check all the generated descriptions to verify their accuracy.

Action Recognition Training β is fed through the text encoder of the frozen ILM to extract IMU-aligned text embeddings, followed by the pre-initialized text projector, R, from Stage I. The extracted embeddings, α, represent c embeddings corresponding to c descriptions for each activity in G, $\alpha \in \mathbb{R}^{c|G| \times h}$, where h is the hidden dimension (Line 9). We average c embeddings per activity from α to extract A, a single embedding for each activity, $A \in \mathbb{R}^{|G| \times h}$ (Line 10). Representing an activity by the average of c embeddings instead of one, helps in reducing the variance. Similarly, the IMU measurements for data point X^i are also passed through the pre-initialized IMU encoder, K, from Stage I, followed by an IMU projector, P, which consists of some learnable non-linear layers, to extract I^i, $I^i \in \mathbb{R}^h$ (Line 12).

If a is the true activity corresponding to X^i, then (I^i, A^a) are the corresponding IMU and activity embedding. We compute cross entropy loss, $\mathcal{L}^i_{CE}$, as the cosine similarity of (I^i, A^g), with all the activities in G (Line 13). K, R, and P, are trained by backpropagating $\mathcal{L}^i_{CE}$ in Line 14.

$$\mathcal{L}^i_{CE} = -\log \frac{exp(sim(I^i, A^a))}{\sum_{g=1}^{|G|} exp(sim(I^i, A^g))} \tag{6}$$

We use different optimizers to train the IMU-related layers (K, P from Stage I and II) and text-related layers (R in Stage I and II) of ZeroHAR.

Zero-Shot Action Recognition During testing, we predict activity with the largest cosine similarity score,

$$\hat{Y}^i = \arg\max_{g \in U} F(X^i, A^g) \tag{7}$$

, where g is an activity from the set of unobserved activities U, $X^i \in \mathcal{D}_{te}$, A^g is the embedding of activity g, F is ZeroHAR, and $\hat{Y}^i$ is the prediction for X^i.

Experimental Results

Dataset, Baselines, and Experimental Settings

We evaluate on a comprehensive motion time series classification benchmark, comprising of 18 real-world datasets that cover diverse activities. These datasets are collected from various body locations such as head, chest, back, arm, wrist, waist, hip, leg, knee and ankle. We categorize these datasets into three difficulty levels: (1) easy level (with fewer than 10 activities): Opportunity (Roggen et al. 2010), UCI-HAR (Anguita et al. 2013), MotionSense (Malekzadeh et al. 2019), w-HAR (Showail 2022), Shoaib (Shoaib et al. 2014), HAR70+ (Ustad et al. 2023), RealWorld (Sztyler and Stuckenschmidt 2016), TNDA-HAR (Yang et al. 2024); (2) medium level (with 10 to 20 activities): PAMAP2 (Reiss and Stricker 2012), USC-HAD (Zhang and Sawchuk 2012), Mhealth (Oresti et al. 2014), Harth (Logacjov et al. 2021), UT-Complex (Shoaib et al. 2016), Wharf (Bruno et al. 2013), WISDM (Weiss 2019), DSADS (Altun, Barshan, and Tunçel 2010); (3) hard level (with more than 20 activities): UTD-MHAD (Chen, Jafari, and Kehtarnavaz 2015), MMAct (Kong et al. 2019). We provide the specific number of activities for each dataset in Table 2 and detail their collection settings in Appendix.

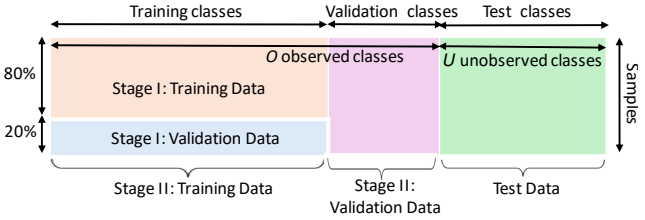


Figure 3: Training, validation, and test data splits. Each row represents #samples and column represent #classes. Note that #samples in each set may be different. (# - no. of)

We compare ZeroHAR against classification models with zero-shot capabilities: NuActiv (Cheng et al. 2013b), SemAtt (Cheng et al. 2013a), NCBM (Wang, Miao, and Hao 2017), LETS-GZSL (Bhaskarpandit, Gupta, and Gupta 2022), SemHAR (Matsuki, Lago, and Inoue 2019), SHARE (Zhang et al. 2023a), NonViz (Al Machot, R. Elkobaisi, and Kyamakya 2020), ImageBind (Girdhar et al. 2023), IMU2CLIP (Moon et al. 2022), IMUGPT (Leng, Kwon, and Plötz 2023) and HARGPT (Ji, Zheng, and Wu 2024). We also input the 2D visualizations of motion time series to pre-trained vision-language model LLaVA (Liu et al. 2024) for comparison. We detail the configurations of baselines in Appendix. As shown in Table 2, ZeroHAR significantly outperforms all baselines in the zero-shot setting. We also apply the Wilcoxon-signed rank test with Holm's α (5%) following previous works (Holm 1979; Zhang et al. 2023b). The Wilcoxon-signed rank test indicates that the improvement of ZeroHAR compared with all the baselines is statistically significant, with p-values significantly lower than 0.05 (e.g., p-value = 8×10^{-6} for ImageBind, which has the highest F1 score among the baselines).

We evaluate the ZSC performance of ZeroHAR and baselines using average per-class accuracy and macro-F1 score. Macro-F1 is defined as macro-F1 = $\frac{1}{|U|} \sum_{g=1}^{|U|} 2 \times \frac{Prec_g \times Rec_g}{Prec_g + Rec_g}$, where $Prec_g$ and Rec_g, represent the precision and recall for activity g, respectively, and $|U|$ is the total number of unseen activities in the test set, $\mathcal{D}_{te}$.

Fig 3 illustrates the train, validation, and test sets for

Dataset	Metrics	*Opportunity*	*UCI-HAR*	*MotionSense*	*w-HAR*	*Shoaib*	*HAR70+*	*RealWorld*	*TNDA-HAR*	*PAMAP2*	*USC-HAD*	*Mhealth*	*Harth*	*UT-Complex*	*Wharf*	*WISDM*	*DSADS*	*UTD-MHAD*	*MMAct*	*Average*
Num Classes		4	6	6	7	7	7	8	8	12	12	12	12	13	14	18	19	27	35	
Level		Easy								Medium								Hard		Avg
NuActiv	Acc	10.4	10.1	13.4	9.5	12.3	0.0	7.9	11.6	8.5	10.1	4.8	3.6	5.8	2.1	3.6	1.4	2.3	1.7	6.6
	F1	11.5	2.7	8.5	4.8	10.6	0.0	7.8	9.4	6.7	5.2	0.8	3.2	3.9	1.1	2.4	0.9	1.2	0.8	4.5
SemAtt	Acc	25.4	7.6	14.6	9.8	17.3	0.0	13.4	11.3	9.5	9.4	6.8	4.7	5.8	2.7	4.6	1.8	2.7	2.2	8.3
	F1	18.6	3.4	11.1	6.8	10.1	0.0	9.2	9.7	7.5	6.8	1.4	5.3	2.7	0.6	3.7	1.1	0.4	0.8	5.5
NCBM	Acc	18.4	8.6	10.4	7.8	10.5	0.0	11.2	13.4	9.4	3.7	4.1	5.6	7.1	1.7	4.2	1.5	2.2	1.8	6.8
	F1	11.4	3.7	5.6	4.9	7.8	0.0	6.4	7.1	7.5	2.1	1.1	2.9	3.5	0.7	2.6	0.3	0.8	1.1	3.9
LETS-GZSL	Acc	24.8	10.6	15.2	19.3	11.8	0.0	14.5	10.1	11.3	8.6	5.9	7.5	7.8	2.4	4.5	1.8	2.1	1.5	8.9
	F1	12.5	4.2	10.2	6.3	9.4	0.0	8.7	7.8	5.8	3.6	1.2	5.3	2.8	1.4	2.6	1.1	0.8	1.2	4.7
SemHAR	Acc	20.7	8.6	13.4	10.5	13.8	7.5	13.2	8.7	12.4	6.9	5.4	6.2	7.1	2.5	4.6	1.3	2.4	1.6	8.2
	F1	11.4	3.2	10.2	5.9	9.6	3.6	7.2	6.8	5.9	4.8	1.2	5.1	4.5	1.4	4.3	0.8	0.3	0.4	4.8
SHARE	Acc	28.5	10.2	15.1	12.7	16.8	8.4	14.1	15.6	12.7	9.2	6.5	7.2	8.4	2.6	5.8	1.4	2.5	1.3	9.9
	F1	14.7	4.6	10.9	6.7	12.4	6.4	8.6	9.8	6.5	5.7	1.2	4.8	5.4	1.3	2.1	0.6	1.1	1.2	5.8
NonViz	Acc	24.6	9.4	12.6	9.4	12.8	7.3	10.5	12.6	7.7	6.8	3.8	4.6	3.8	1.3	4.2	1.5	1.2	0.8	7.5
	F1	12.4	3.2	8.9	5.7	8.6	4.8	7.3	9.2	4.8	3.6	0.8	3.5	2.8	0.5	2.1	0.4	0.6	0.7	4.4
ImageBind	Acc	28.8	14.0	16.7	14.2	18.7	0.0	17.8	18.7	15.1	10.9	7.6	8.7	9.4	3.5	6.8	1.8	2.5	3.1	11.0
	F1	27.6	7	13.8	9.1	14.8	0.0	9.9	13.1	6.7	6.8	1.1	5.9	5.8	2.0	4.1	0.9	1.4	1.7	7.3
IMU2CLIP	Acc	24.8	15.9	16.1	6.4	14.7	6.2	5.9	8.3	2.0	11.9	7.5	2.4	6.8	1.7	4.2	4.1	3.5	5.9	8.2
	F1	9.7	9.4	8.9	3.1	8.9	2.0	4.5	4.0	1.5	9.2	1.1	0.9	2.5	0.8	2.8	0.9	1.8	1.7	4.1
IMUGPT	Acc	9.6	1.3	10.9	**64.2**	11.4	0.0	17.3	13.8	9.2	6.1	9.3	4.5	11.7	2.4	8.5	7.2	3.9	2.1	10.7
	F1	9.8	0.7	3.8	<u>36.9</u>	9.2	0.0	4.3	5.8	1.2	7.2	2.3	1.5	7.8	1.6	6.2	2.1	0.2	0.9	5.6
HARGPT	Acc	29.7	14.8	10.7	3.9	21.5	32.8	11.9	12.6	10.7	9.2	10.8	29.3	6.8	5.3	5.7	5.9	3.1	2.2	12.6
	F1	17.1	12.3	5.8	2.8	11.6	9.9	5.6	5.1	2.4	3.3	6.8	7.3	4.1	1.7	3.2	3.8	1.3	1.5	5.9
LLaVA	Acc	39.7	17.4	23.5	0.0	14.8	12.5	15.9	11.8	9.6	10.7	19.4	16.8	1.9	3.4	6.1	5.2	3.8	3.7	12.0
	F1	13.7	6.1	6.7	0.0	4.3	3.4	3.8	2.7	1.1	2.8	7.3	5.2	0.9	0.1	0.4	0.8	0.2	0.3	3.3
ZeroHAR	Acc	**72.6**	**28.9**	**38.5**	<u>54.0</u>	**57.2**	**65.1**	**42.9**	**53.6**	**70.0**	**61.8**	**68.2**	**69.0**	**32.5**	**23.8**	**26.0**	**28.9**	**18.7**	**8.6**	**45.6**
	F1	**59.0**	**20.7**	**30.8**	**38.6**	**54.9**	**32.6**	**35.3**	**50.1**	**59.4**	**50.2**	**57.5**	**41.1**	**33.1**	**12.4**	**23.7**	**22.4**	**17.0**	**7.7**	**35.9**
Only Stage-II	Acc	28.4	11.4	14.7	10.6	8.7	13.2	14.8	18.2	7.8	8.9	6.3	4.9	7.8	3.2	4.1	5.4	2.7	3.1	9.7
	F1	11.2	7.3	8.2	4.6	6.1	10.9	7.4	6.8	3.5	5.2	7.6	3.2	2.4	0.4	1.7	2.6	1.7	0.4	5.1
Multitask	Acc	<u>64.2</u>	<u>22.6</u>	<u>29.4</u>	38.9	<u>46.2</u>	<u>54.2</u>	<u>32.1</u>	<u>44.8</u>	<u>28.9</u>	<u>38.8</u>	<u>54.7</u>	<u>57.8</u>	<u>26.8</u>	<u>12.1</u>	<u>19.7</u>	<u>14.2</u>	<u>14.5</u>	<u>6.2</u>	<u>33.7</u>
	F1	<u>37.5</u>	<u>13.2</u>	<u>21.4</u>	21.7	<u>45.3</u>	<u>21.4</u>	<u>26.8</u>	<u>31.2</u>	<u>19.0</u>	<u>32.5</u>	<u>26.4</u>	<u>34.7</u>	<u>24.1</u>	<u>9.5</u>	<u>18.8</u>	<u>13.7</u>	<u>14.6</u>	<u>5.9</u>	<u>23.2</u>

Table 2: Zero-Shot performance. We bold the **best** and underline the <u>second best</u>. ZeroHAR performs the best compared with both baselines and our model ablations. The last column shows the average performance across 18 datasets.

ZeroHAR. The test set contains novel classes, U, unseen during training. To enable early stopping, we reserve data from novel classes, O, to form a validation set, O_{va}, at the start of training. During Stage II, ZeroHAR trains only on O_{tr}, where $O_{tr} = O - O_{va}$. Additionally, part of O_{tr} is reserved for Stage I validation.

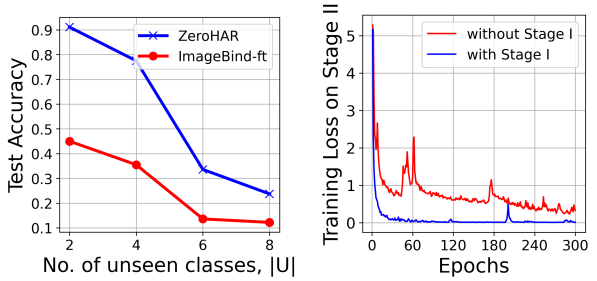
We normalize our datasets and train all baselines with sufficient hyper-parameter tuning. Since our datasets are widely heterogeneous in terms of number of data points, sensors, body positions, and sampling frequency, we obtain better performance via cursory tuning of dataset-specific hyper-parameters. We set the temperature parameter τ, in Algorithm 1, to 0.05 and the number of descriptions per activity, c, in Algorithm 2 to 10. We use Adam optimizers to update the IMU modality (IMU Encoder and IMU projectors P and Q for Stage I and II, respectively) and the text modality (text projector R). We use a batch size of 128, learning rate of 0.001, 8 self-attention layers with 8 heads for the IMU Encoder, a dropout of 0.01 and a hidden dimension, h, of 128, for both Stage I and II. We save the model with the lowest validation loss and evaluate it on the test set.

Results

Table 2 summarizes the results. Activity-attribute methods underperform, while pre-trained models fare better. ImageBind and IMU2CLIP, trained on head-mounted data, lack generalization to other sensor locations. IMUGPT struggles with cross-dataset generalization and requires separate training per dataset. HARGPT and LLaVA focus on simple activities but are limited by their training on non-motion data and reliance on careful prompt design. All these models also fail to handle varying device orientations. In contrast, ZeroHAR

	Average Accuracy	Average Macro F1
2^{nd} best results	12.6	7.3
ZeroHAR's results	45.6	35.9
% relative improvement of ZeroHAR	+262%	+392%

Table 3: Relative performance comparison of ZeroHAR with 2^{nd} best results in Table 2).



(a) Test Accuracy vs #unseen classes, $|U|$ (b) Stage II's Loss Convergence

Figure 4: (a) compares how the zero-shot accuracy changes with the number of unseen classes in test set, $|U|$. (b) shows the effect of Stage I training on loss convergence of Stage II.

achieves state-of-the-art performance, demonstrating robust generalization across device locations, orientations, and activities.

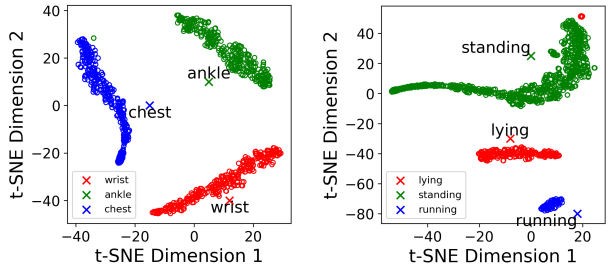
Table 3 highlights ZeroHAR's performance improvements over the second-best results. Relative improvements are measured from Table 2. Among baselines, HARGPT (Ji, Zheng, and Wu 2024) leads in Accuracy, and ImageBind (Girdhar et al. 2023) excels in Macro F1. Comparing ZeroHAR's average performance across 18 datasets, it achieves a remarkable relative improvement of 262% in Accuracy and 392% in Macro F1 over these baselines.

Test Accuracy vs no. of unseen classes, $|U|$ To compare how test accuracy on unseen classes vary with the number of unseen classes, $|U|$, we compare ZeroHAR with ImageBind on Opportunity dataset. Result in Fig 4(a) shows that accuracy goes down with increase in $|U|$ for both models, but ZeroHAR surpasses ImageBind for all $|U|$.

Loss Convergence To assess the effect of Stage I (Motion with Sensor Context Training) on ZeroHAR, we compare its loss convergence when trained on both Stage I and II versus Stage II alone. Fig 4(b) shows results on Opportunity dataset. ZeroHAR trained with Stage I converges faster and achieves a lower loss during Stage II compared to the model trained solely on Stage II. Stage I used contrastive training of IMU with textual sensor context to initialize the shared space, simplifying Stage II learning.

Ablations

We analyze ZeroHAR's performance through various ablations to justify our training decisions. The last two rows of Table 2 what happens if we conduct 1) only Stage II training and 2) train Stage I and Stage II parallelly in a multitask



(a) IMU with Body Position - Stage I (b) IMU with Class Description - Stage II

Figure 5: t-SNE vizualization of ZeroHAR on PAMAP2 showing 'o' - IMU embeddings with (a) 'x' - embedding of body positions for Stage I and (b) 'x' - embedding of *unseen* test classes' description in fold 3 for Stage II.

fashion. Conducting only Stage II training gives poor results as model is not trained on the sensor context. Training for both Stage I and Stage II parallelly significantly improves performance because of the addition of sensor context. But the performance still falls short of ZeroHAR which trains Stage I and stage II sequentially. This is because by first training for Stage I, it provides a joint IMU-text latent space that helps to recognize actions in Stage II.

Case Study on IMU-Text Alignment

Fig 5 depicts the IMU-text latent space learned by ZeroHAR on PAMAP2. Fig 5(a) demonstrates ZeroHAR's ability to align IMU data from different body parts with their corresponding word embeddings, highlighting that joint training of IMU with its corresponding sensor context can bring the latent space of IMU closer to text. Fig 5(b) shows that despite not being trained on any data from these unseen classes, ZeroHAR can align their IMU data with their respective textual activity description embeddings.

Conclusion

We present ZeroHAR, a two-stage framework to tackle the Zero-Shot Learning problem in Inertial HAR. In Stage I: Motion with Sensor Context Training, we compliment IMU with sensor context information to learn spatial and biomechanical information about motion. It brings the latent spaces of IMU and text closer to each other, which facilitates mapping IMUs to textual activity representations in the subsequent stage. We compared ZeroHAR with 12 baselines on 18 benchmark HAR datasets to evaluate its efficacy on Zero-Shot HAR. Our ablations and case study highlight the superior alignment of IMU with text-based sensor context and activity representations. Using sensor context as additional features to aid action recognition provides a new avenue to explore for similar IoT-based applications. This will enable us to engage in more advanced natural language queries, reasoning, and responses related to sensory data.

Acknowledgments

Our work is supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Our work is also supported by Qualcomm Innovation Fellowship and is sponsored in part by NSF CAREER Award 2239440, NSF Proto-OKN Award 2333790, NIH Bridge2AI Center Program under award 1U54HG012510-01, Cisco-UCSD Sponsored Research Project, as well as generous gifts from Google, Adobe, and Teradata. Ranak is partially funded by a Graduate Prize Fellowship from Halıcıoğlu Data Science Institute. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes not withstanding any copyright annotation hereon.

References

Al Machot, F.; R. Elkobaisi, M.; and Kyamakya, K. 2020. Zero-Shot Human Activity Recognition Using Non-Visual Sensors. *Sensors*, 20(3).

Altun, K.; Barshan, B.; and Tunçel, O. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10): 3605–3620.

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J. L.; et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, 3.

Arrotta, L.; Bettini, C.; Civitarese, G.; and Fiori, M. 2024. ContextGPT: Infusing LLMs Knowledge into Neuro-Symbolic Activity Recognition Models. *arXiv preprint arXiv:2403.06586*.

Bhaskarpandit, S.; Gupta, P.; and Gupta, M. 2022. Lets-gzsl: A latent embedding model for time series generalized zero shot learning. *arXiv preprint arXiv:2207.12007*.

Bruno, B.; Mastrogiovanni, F.; Sgorbissa, A.; Vernazza, T.; and Zaccaria, R. 2013. Analysis of human behavior recognition algorithms based on acceleration data. In *2013 IEEE International Conference on Robotics and Automation*, 1602–1607. IEEE.

Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, 168–172. IEEE.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cheng, H.-T.; Griss, M.; Davis, P.; Li, J.; and You, D. 2013a. 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.

Cheng, H.-T.; Sun, F.-T.; Griss, M.; Davis, P.; Li, J.; and You, D. 2013b. Nuactiv: Recognizing unseen new activities using semantic attribute-based learning. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 361–374.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Gupta, G.; Kapila, R.; Gupta, K.; and Raskar, R. 2023. Domain Generalization In Robust Invariant Representation. *arXiv preprint arXiv:2304.03431*.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2.

Ji, S.; Zheng, X.; and Wu, C. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? *arXiv preprint arXiv:2403.02727*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kong, Q.; Wu, Z.; Deng, Z.; Klinkigt, M.; Tong, B.; and Murakami, T. 2019. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8658–8667.

Leng, Z.; Kwon, H.; and Plötz, T. 2023. Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*, 39–43.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Logacjov, A.; Bach, K.; Kongsvold, A.; Bårdstu, H. B.; and Mork, P. J. 2021. HARTH: A Human Activity Recognition Dataset for Machine Learning. *Sensors*, 21(23): 7853.

Malekzadeh, M.; Clegg, R. G.; Cavallaro, A.; and Haddadi, H. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, 49–58.

Matsuki, M.; Lago, P.; and Inoue, S. 2019. Characterizing word embeddings for zero-shot sensor-based human activity recognition. *Sensors*, 19(22): 5043.

Moon, S.; Madotto, A.; Lin, Z.; Dirafzoon, A.; Saraf, A.; Bearman, A.; and Damavandi, B. 2022. IMU2CLIP:

Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text. *arXiv preprint arXiv:2210.14395*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Oresti, B.; Rafael, G.; Juan, A.; Miguel, D.; Hector, P.; Ignacio, R.; Alejandro, S.; and Claudia, V. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. *Ambient Assisted Living and Daily Activities*, 91–98.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Reiss, A.; and Stricker, D. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, 108–109. IEEE.

Roggen, D.; Calatroni, A.; Rossi, M.; Holleczek, T.; Förster, K.; Tröster, G.; Lukowicz, P.; Bannach, D.; Pirkl, G.; Ferscha, A.; et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, 233–240. IEEE.

Shoaib, M.; Bosch, S.; Incel, O. D.; Scholten, H.; and Havinga, P. J. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6): 10146–10176.

Shoaib, M.; Bosch, S.; Incel, O. D.; Scholten, H.; and Havinga, P. J. 2016. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, 16(4): 426.

Showail, A. J. 2022. Solving hajj and umrah challenges using information and communication technology: a survey. *IEEE Access*, 10: 75404–75427.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26.

Sztyler, T.; and Stuckenschmidt, H. 2016. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE international conference on pervasive computing and communications (PerCom)*, 1–9. IEEE.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Ustad, A.; Logacjov, A.; Trollebø, S. Ø.; Thingstad, P.; Vereijken, B.; Bach, K.; and Maroni, N. S. 2023. Validation of an activity type recognition model classifying daily physical behavior in older adults: the HAR70+ model. *Sensors*, 23(5): 2368.

Wang, W.; Miao, C.; and Hao, S. 2017. Zero-shot human activity recognition via nonlinear compatibility based method. In *Proceedings of the International Conference on Web Intelligence*, 322–330.

Weiss, G. M. 2019. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7: 133190–133202.

Wu, T.; Chen, Y.; Gu, Y.; Wang, J.; Zhang, S.; and Zhechen, Z. 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*, 210–221. Springer.

Yang, J.; Liao, T.; Zhao, J.; Yan, Y.; Huang, Y.; Zhao, Z.; Xiong, J.; and Liu, C. 2024. Domain Adaptation for Sensor-Based Human Activity Recognition with a Graph Convolutional Network. *Mathematics*, 12(4): 556.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2114–2124.

Zhang, M.; and Sawchuk, A. A. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 1036–1043.

Zhang, X.; Chowdhury, R. R.; Hong, D.; Gupta, R. K.; and Shang, J. 2023a. Modeling Label Semantics Improves Activity Recognition. *arXiv preprint arXiv:2301.03462*.

Zhang, X.; Chowdhury, R. R.; Zhang, J.; Hong, D.; Gupta, R. K.; and Shang, J. 2023b. Unleashing the Power of Shared Label Structures for Human Activity Recognition. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 3340–3350. ISBN 9798400701245.