Limitations in Employing Natural Language Supervision for Sensor-Based Human Activity Recognition – And Ways to Overcome Them

Harish Haresamudram[1], Apoorva Beedu[1], Mashfiqui Rabbi[2]
Sankalita Saha[2], Irfan Essa[1], Thomas Ploetz[1]
[1]Georgia Institute of Technology
[2]Optum AI
hharesamudram3@gatech.edu, abeedu3@gatech.edu, mashfiqui.rabbi@optum.com
sankalita.saha@optum.com, irfan@gatech.edu, thomas.ploetz@gatech.edu

Abstract

Cross-modal contrastive pre-training between natural language and other modalities, e.g., vision and audio, has demonstrated astonishing performance and effectiveness across a diverse variety of tasks and domains. In this paper, we investigate whether such natural language supervision can be used for wearable sensor based Human Activity Recognition (HAR), and discover that–surprisingly–it performs *substantially worse* than standard end-to-end training and self-supervision. We identify the primary causes for this as: *sensor heterogeneity* and the *lack of rich, diverse text descriptions of activities.* To mitigate their impact, we also develop strategies and assess their effectiveness through an extensive experimental evaluation. These strategies lead to significant increases in activity recognition, bringing performance closer to supervised and self-supervised training, while also enabling the recognition of unseen activities and cross modal retrieval of videos. Overall, our work paves the way for better sensor-language learning, ultimately leading to the development of foundational models for HAR using wearables.

1 Introduction

Learning joint embedding spaces by pairing modalities with natural language descriptions of their contents (e.g., image captions or sound descriptions for audio) has proven successful across modalities (Radford et al. 2021; Xu et al. 2021; Wu et al. 2022). Here, the task is to predict which description goes with which input, for large-scale datasets. The expressiveness of natural language enables it to oversee a wider array of concepts, culminating in highly effective representations (Radford et al. 2021).

Some advantages of such setups include: *(i) zero-shot prediction of classes* through text descriptions, potentially aided by useful auxiliary and context information (Shen et al. 2022); and *(ii) cross-modal retrieval*, where natural language can be utilized to retrieve relevant images, audio, or sensor data, akin to performing search. The promise of these models is that they are *plug-and-play*: they can be utilized without further training or adaptation in diverse application scenarios.

This paper focuses on Natural Language Supervision (NLS) for wearables based Human Activity Recognition (HAR), which involves automatically predicting which activity is being performed. It has numerous applications, including health and fitness monitoring (Koskimäki, Siirtola, and Röning 2017) and eating detection (Bin Morshed et al. 2022). From a wearables standpoint, advantages of NLS such as predicting unseen activities and performing recognition through text queries are desirable, as they would allow new capabilities to be added on the fly.

Despite the astonishing success of NLS across modalities, domains, and applications, we discover and demonstrate in this paper that it is *highly challenging* to apply it in a plug-and-play manner to wearables based HAR. Pre-training on a large-scale dataset (Capture-24) followed by zero shot prediction of activities is *drastically worse* by around *30-40%*, than simple end-to-end and self-supervised training, across six target datasets. We discover two reasons that explain this reduction in performance, underpinned by challenges unique to wearable sensors and the HAR task: *(i) Sensor Heterogeneity:* Diversity in sensors results in significant differences in data distributions due to hardware constraints and settings such as gain, data and signal processing, differences in sampling rates – even if the sensor locations and activities are the same (Stisen et al. 2015). This renders zero shot prediction very difficult, as pre-trained models cannot deal well with shifting distributions causing substantial performance degradation when there is no opportunity for adaptation to (vastly) different test conditions; and *(ii) Lack of Rich Descriptions of Activities:* Learning such joint embedding spaces is data intensive, relying on diverse and unique text descriptions to learn wide ranging concepts. However, many HAR datasets contain only a handful of activity labels and (in some cases) demographics information (Kwon et al. 2020; Plötz 2023) – a far cry from the 400M image-text pairs in the original CLIP paper (Radford et al. 2021). Therefore, in scenarios with diverging data distributions, or when there is paucity of diverse descriptions of data, NLS can be an inferior option.

We develop strategies to tackle these challenges, leading to improved HAR performance, and more broader applicability across scenarios. To deal with varying data distributions between pre-training and target datasets, we show how *updating/adapting some layers of the pre-trained network on target data* with as little as 4 mins of data/activity leads to substantially improved recognition. This demonstrates that

adaptation with minimal amounts of target data can be sufficient for improved HAR, and potentially, across other applications with distribution differences. To improve diversity in text descriptions of activities, we not only explore the *generation of additional prompts through LLMs*, but also study how *external knowledge* can aid in improved recognition.

The contributions of our work are as follows:

- We adopt and adapt natural language supervision (NLS) for performing wearables-based HAR.
- We identify challenges rendering adaptation difficult and less successful, such as sensor heterogeneity and a lack of rich text descriptions accompanying sensor data.
- We develop strategies to tackle these challenges, enabling more successful application of NLS to sensor-based HAR, and potentially opening up such cross-modal training to other applications facing similar challenges.

2 Related Work

Cross-modal contrastive training between natural language and other modalities has emerged as a highly effective training paradigm. Typically, the Internet is crawled for collating large-scale datasets (of hundreds of millions of samples) with corresponding text descriptions. Contrastive Language-Image Pre-training (CLIP (Radford et al. 2021)), in particular, delivered superb performance across target scenarios by contrastively training to match images with corresponding text captions, by using 400M image-text pairs. This training setup was subsequently adopted for video (Ma et al. 2022; Xu et al. 2021) and audio (Elizalde et al. 2023; Wu et al. 2022), as well. However, these methods require substantial training data containing rich and diverse text captions. When only keywords/class names are available, converting them into sentences for effective contrastive training has also been explored (Wu et al. 2023). These methods are largely 'plug-and-play', i.e., they can perform zero-shot recognition.

This setup has been extended to wearable sensors, through methods like IMU2CLIP (Moon et al. 2022), which uses the Ego4d dataset (Grauman et al. 2022), containing head-mounted IMU and fine-grained text descriptions of activities. As it was recorded at the head, it cannot be directly used for HAR, where common recording locations include the wrist or the waist. ImageBind (Girdhar et al. 2023) also utilizes the Ego4d dataset to learn a joint embedding space between six modalities (incl. text and IMU) through contrastive pairwise training with vision as bridge. Alternatively, LLMs have also been fine-tuned to utilize pre-trained sensor embeddings and text to recognize both seen and unseen activities, e.g.., SensorLLM (Li et al. 2024) and LLaSA (Imran et al. 2024). However, early approaches (Mishra et al. 2020; Matsuki, Lago, and Inoue 2019) used pre-trained word embeddings for such unseen activity recognition.

For wireless sensors, TENT (Zhou et al. 2023) connects large language models (LLMs) to IoT sensors such as video, Radar, and LiDAR with text through a public dataset containing all modalities, and uses the pre-trained CLIP text encoder for obtaining text embeddings. More recently, Ts2Act (Xia et al. 2024) curated an image dataset for activity classes, and performed cross-modal contrastive pre-training with IMU data, for few shot recognition. Other works perform zero-shot learning by learning to align IMU embeddings with synchronized video (Tong, Ge, and Lane 2021). ContextGPT (Arrotta et al. 2024) inputs context information (location, weather, etc.) into LLMs for predicting the most probable activities, for use in neuro-symbolic HAR.

Both Ts2Act and IMU2CLIP utilize different splits (i.e., train and test) from the same dataset, and therefore, do not perform HAR *across datasets*. Further, IMU2CLIP has access to rich text annotations whereas Ts2Act trains with images, which may be hard to obtain for rare activities. Wearables datasets however, typically lack access to text data, beyond activity names. Consequently, our work focuses on directly pre-training with text sentences derived from activity labels. This brings the capability to flexibly describe (in natural language) the movements present in activities, for HAR. We also develop strategies to address challenges inherent to this setup, leading to the missing capability to function in truly plug-and-play fashion.

3 Natural Language Supervision for HAR

Our NLS setup involves two stages: *i)* cross-modal contrastive pre-training; and *ii)* Human Activity Recognition.

3.1 Cross-Modal Contrastive Pre-Training

Sensor data windows are first embedded through an encoder and a projection head, resulting in N vectors $S = \{S_i\}_{i=1...N}$, where N is the batch size. Correspondingly, the textual descriptions of activities are also encoded through a text encoder and a separate projection head, to obtain text representations $T = \{T_i\}_{i=1...N}$. Since only activity labels are available, simple text templates are employed to obtain sentences (Sec. 3.2). Both S_i and T_i have dimension D.

In such a batch of N sensor-text pairs, the task is to identify which of the $N \times N$ pairs are actual matches. We compute the cosine similarity ($C \in \mathbb{R}^{N \times N}$) for each S_i and T_i, and train a joint embedding space to maximize the similarity of the N actual pairs in the batch (i.e., the diagonal of the matrix in Fig. 1), while minimizing the cosine similarity of the $N^2 - N$ incorrect pairings (i.e., the off diagonal elements in Fig. 1). In line with related work (Elizalde et al. 2023), the similarity is computed as follows, where τ is the temperature parameter used to scale the logits:

$$C = \tau * (S \cdot T^{\top}) \tag{1}$$

τ is initialized to $1/0.07$ and updated during training.

A symmetric cross entropy loss is optimized over these similarity scores, to update the network parameters (Elizalde et al. 2023; Radford et al. 2021):

$$\mathcal{L} = 0.5 * (\ell_{\text{sensor}}(C) + \ell_{\text{text}}(C)) \tag{2}$$

where, $\ell = \frac{1}{D} \sum_{i=0}^{D} \log(\text{diag}(\text{softmax}(C)))$, along the sensor and IMU axes respectively.

This setup contains three major components: *(i) IMU Encoder*, which provides embeddings of sensor data windows for contrastive pre-training. We utilize a convolutional encoder developed and utilized extensively in prior work (Saeed, Ozcelebi, and Lukkien 2019; Haresamudram, Essa,

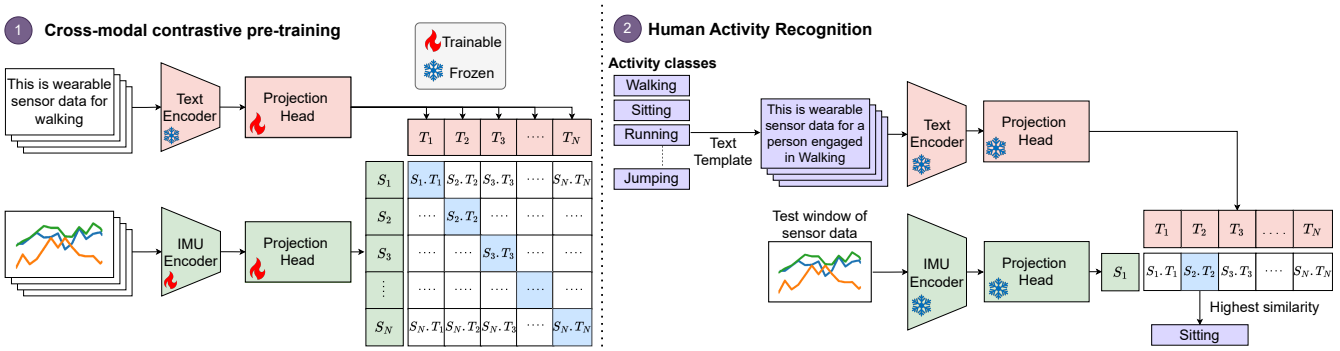

Figure 1: **Natural language supervision for sensor-based HAR:** the network is pre-trained by learning to accurately match windows of sensor data to the corresponding ground truth activities in form of textual descriptions. HAR is then performed by computing cosine similarity scores between windows of test sensor data and all activity sentences. The sentence with highest similarity score determines the final activity output (lower right part in phase 2). This figure is inspired by (Radford et al. 2021).

and Plötz 2022); *(ii) Text Encoder*, which is used to encode sentences of activities – we employ the DistilBERT model (Sanh et al. 2019); and *(iii) Projection layers*, which are used to project embeddings from the modalities to a common representation space. In line with recent work, we use an MLP-based projection head. Detailed descriptions of the architecture are given in the Appendix (Haresamudram et al. 2024).

3.2 HAR Through Text-Based Classification

HAR is performed by using sentences derived from activity labels (e.g., sitting, walking) and text templates (Fig. 1). Unless specified differently we use the following, hand crafted template to obtain activity sentences: `This is wearable sensor data for a person engaged in {activity_name}` where `activity_name` is replaced with activities, e.g., walking, running, etc.

For a target dataset containing C classes we generate the sentences using the text template, and compute embeddings from the pre-trained text encoder and the learned projection head. Similarly, we compute embeddings for N windows of target data using the learned encoder and projection heads. As both embeddings are in a common space, we compute cosine similarity between embeddings from each window of the target data (i.e., N embeddings), and the embeddings for all target classes (i.e., C embeddings). The predicted label for each window is the class embedding with the highest cosine similarity, i.e., the label is assigned based on the nearest match to the window. This is shown in Part 2 of Fig. 1.

4 Experimental Settings

Sensor data are segmented into overlapping windows using a sliding window approach, and used for training and evaluation. We now provide details of the datasets and segmentation setup used for our experimental study, with further details added to the Appendix (Haresamudram et al. 2024).

Datasets For pre-training, we use the large-scale Capture-24 dataset (Willetts et al. 2018), which has 177 fine-grained labels after some data cleaning (correcting minor typos and removing semi-colons, etc.), as they allow us to learn more concepts present in data. Following (Haresamudram, Essa, and Plötz 2022), we evaluate on a diverse set of six target datasets, covering different recording locations (wrist, waist, ankle/leg), conditions, number of users, and types of activities. Primarily, they contain locomotion-style activities (HHAR (Stisen et al. 2015), Mobiact (Chatzaki et al. 2016), Motionsense (Malekzadeh et al. 2018), PAMAP2 (Reiss and Stricker 2012)) and gym exercises (Myogym (Koskimäki, Siirtola, and Röning 2017), MHEALTH (Banos et al. 2014)), in addition to daily living activities (PAMAP2). Details of the datasets are given in the Appendix.

Sampling Rate and Segmentation We utilize raw accelerometer data, and downsample them (if necessary) to 50Hz to match the lowest frequency across datasets. Following (Haresamudram, Essa, and Plötz 2022), the sliding window size is set to 2 seconds, with 50% overlap.

5 Plug-and-Play NLS for HAR

We evaluate the effectiveness of NLS for HAR using the standard setup (refer to the Appendix for details (Haresamudram et al. 2024)), and compare against supervised training and self-supervision with a large dataset in Tab. 1.

5.1 Human Activity Recognition Experiments

Here, we summarize the baselines of our evaluation (details in the Appendix), followed by a discussion of the results.

Baselines *(i) Supervised learning:* the Conv. classifier contains 1D convolutional layers whereas DeepConvLSTM (Ordóñez and Roggen 2016) has 2D convolutional layers followed by an LSTM network. These methods are trained end-to-end on the target datasets; and *(ii) Self-supervised learning:* the Autoencoder (Haresamudram, Essa, and Plötz 2022) is trained to reconstruct the input window after being passed through encoder and decoder layers. SimCLR (Tang et al. 2020; Chen et al. 2020) contrasts randomly augmented versions of the same input window whereas Enhanced CPC (Haresamudram, Essa, and Plötz 2023) performs contrastive training on future timesteps of sensor data. They use 1D convolutional encoders along with an MLP classifier for HAR.

Method	Wrist		Waist		Leg	
	HHAR	Myogym	Mobiact	Motionsense	MHEALTH	PAMAP2
Baselines						
Conv. classifier *	55.63 ± 2.05	38.21 ± 0.62	78.99 ± 0.38	89.01 ± 0.89	48.71 ± 2.11	59.43 ± 1.56
DeepConvLSTM *	52.37 ± 2.69	39.36 ± 1.56	82.36 ± 0.42	84.44 ± 0.44	44.43 ± 0.95	48.53 ± 0.98
Autoencoder + MLP classifier *	53.64 ± 1.04	46.91 ± 1.07	72.19 ± 0.35	83.10 ± 0.60	40.33 ± 0.37	59.69 ± 0.72
SimCLR + MLP classifier *	56.34 ± 1.28	47.82 ± 1.03	75.78 ± 0.37	87.93 ± 0.61	42.11 ± 0.28	58.38 ± 0.44
Enhanced CPC + MLP classifier *	59.25 ± 1.31	40.87 ± 0.50	78.07 ± 0.27	89.35 ± 0.32	53.79 ± 0.83	58.19 ± 1.22
Natural language supervision + zero shot prediction						
NLS w/ pre-training on Capture-24	31.05	1.47	16.93	38.97	11.15	10.88
NLS w/ pre-train. on train split of target data	29.05	33.30	59.09	73.36	41.72	48.36

Table 1: **HAR performance of natural language supervision (NLS) [mean F1]**: zero shot prediction is substantially worse than supervised and self-supervised baselines. *: from (Haresamudram, Essa, and Plötz 2023) (five random classifier runs).

Dataset	#classes	Vocab. size	Vocab. + template size
Capture-24	177	282	292
HHAR	6	9	19
Myogym	31	55	65
Mobiact	11	23	33
Motionsense	6	8	18
PAMAP2	12	16	26
MHEALTH	13	27	37
ImageNet-21k*	$\sim$19.2k	13.5k	–
YFCC 14M*	$\sim$14.2M	2.41M	–

Table 2: Vocab. sizes of datasets. *: from (Shen et al. 2022)

Pre-training is performed on Capture-24 whereas the classifier layers are updated during classification on target data.

Results Both end-to-end training and self-supervision substantially outperform NLS – *across datasets*. In particular, NLS-based pre-training on Capture-24 and zero shot prediction on target datasets leads to very poor performance. This is because zero shot prediction involves *no further training on target data.* Baselines perform substantially better than pre-training on Capture-24, as they have the chance to adapt to target conditions, by training at least some parts of the network with target data. Similarly, pre-training on train splits of target datasets is also much better, as the model does not have to contend with (vastly) differing data distributions between training and testing. This leads to substantial increases of 30-50%, yet under performing baselines. Overall, we find that zero shot prediction of activities for wearables based HAR is difficult and less effective than existing baselines. Consequently, we conduct an analysis into the challenges affecting the performance of NLS.

5.2 Challenges

Differences in distributions between pre-training and target data have substantial impact on HAR performance. Such differences are unique to wearables, where there is high diversity in sensors deployed, along with associated hardware constraints and settings (Fig. 6 in the Appendix shows these differences for three datasets). As a result, pre-training on Capture-24 and performing zero shot HAR on target datasets results in substantial performance drop, because there is no opportunity for the model to adapt to the new target data.

The resulting diversity in recorded data, even for similar activities and movements, is referred to as *'Sensor Heterogeneity'* (Koskimäki, Siirtola, and Röning 2017). This is a *challenge* and it prevents the use of large-scale pre-training, followed by zero shot prediction in diverse target conditions without adaptation – in contrast to other domains, e.g., computer vision, where this is standard practice.

NLS benefits strongly from pre-training on large datasets containing *rich and descriptive text* about diverse concepts, leading to generalization and effective zero shot prediction. In contrast, most wearable datasets only contain activity names, thereby limiting what the model can learn. Consequently, effectively learning sensor-language joint embedding spaces is difficult, as seen in Tab. 1. We summarize the number of classes and vocab. size used for pre-training in Tab. 2, and observe that the vocab. size is 3-4 orders of magnitude smaller than for vision datasets like ImageNet-21K (Deng et al. 2009) and YFCC-14M (Thomee et al. 2016) (as tabulated by (Shen et al. 2022)). As such, the *lack of rich text descriptions* is another challenge limiting performance.

6 Tackling NLS for HAR Challenges

Here, we detail strategies to tackle challenges in employing NLS for sensor-based HAR.

6.1 Tackling Sensor Heterogeneity: Adapting Projection Layers on Target Data

Self-supervised methods shown in Tab. 1 were also pre-trained on Capture-24, yet, training classifier layers with target data enables them to perform effective HAR. Clearly, learning / adapting at least some layers of the network (e.g., the classifier) on target data is necessary for usable HAR.

We propose to perform additional cross-modal contrastive training on the train split of the labeled target data, updating the weights of *only the text and sensor projection heads*, while keeping the IMU and text encoders frozen. This is

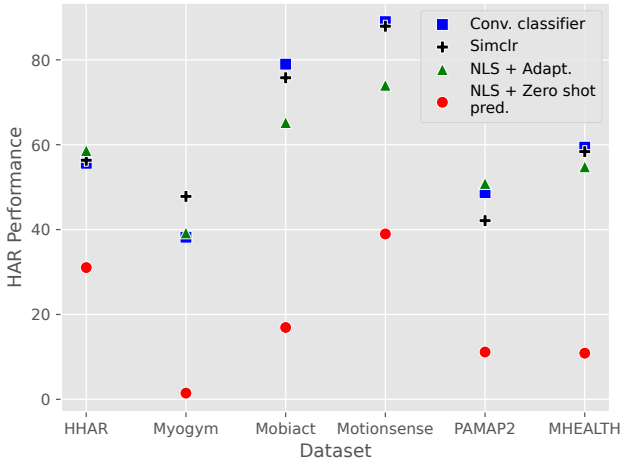

Figure 2: **Adapting projection layers** increases HAR performance of sensor-based NLS by 20-40%.

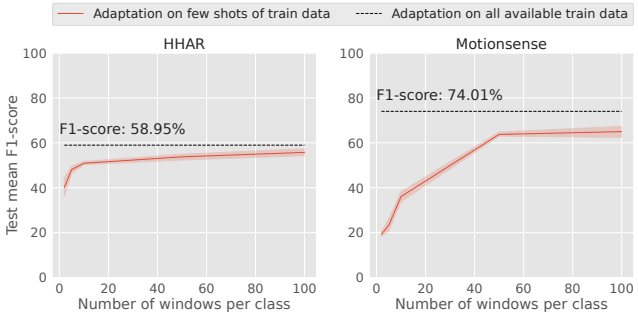

Figure 3: **Adaptation on target data:** access to even small quantities of target data (<2 min) substantially improves performance. Full figure in the Appendix (Fig. 10).

an established practice in multi-modal setups, as it enables the alignment of modalities, especially when the encoders for different modalities are already pre-trained on different datasets (Moon et al. 2023; Verma et al. 2024). We note that such an adaptation step, strictly speaking, violates the zero shot setup which does not allow further training. Rather, it resembles few-shot learning, where small quantities of annotated target data can vastly improve HAR.

For Fig. 2, we utilized the entire train split of target datasets for adaptation. Consistently, we observe *substantial performance improvements resulting from adaptation*. HHAR, Mobiact, MHEALTH, and PAMAP2 see increases of around 30-50%, indicating the necessity of access to target annotations, bringing NLS closer to / beyond baselines.

For scenarios where it can be impractical to collect and annotate multiple hours of data, we also study if adaptation with small quantities of data can be useful. To do so, we randomly sample $\{2, 5, 10, 25, 50, 100\}$ windows/class for adaptation and report the performance on the test split, for five randomized runs in Fig. 3. Clearly, *adapting projecting layers is highly advantageous*, with performance increases of 20-40% (across datasets) with just 100 labeled windows, i.e., *less than 4 minutes per activity*.

6.2 Tackling Lack of Rich Activity Descriptions: Increasing Text Diversity with LLMs

We propose two measures to increase text diversity, and tabulate their impact in Tab. 3. For this setting, pre-training is performed on the train split of the target datasets, whereas the performance is reported on the test split:

(i) Additional text templates: We hand-crafted eight text templates, and further employed ChatGPT to generate 25 additional (similar) templates, leading to a total of 33 templates for both pre-training and recognition. The underlying content of the sentences remains largely similar, while the activity information is presented in more diverse variations. As a result, any resulting performance improvements indicate the importance of text diversity. We list the text templates in Tab. 15 in the Appendix (Haresamudram et al. 2024).

Similar to (Fan et al. 2023), we also explore further diversification of the activity sentences (i.e., the sentences obtained after applying the template to activities) via ChatGPT, by rewriting into 15 different variations. Consequently, each of the 33 templates results in 15 variations per activity.

→*Insight: Diversity in the text descriptions is important*. From Tab. 3, we observe that increasing diversity through multiple text templates generally results in improved performance. We observe only minor changes in recognition performance when using the base template vs. randomly sampling from the larger set. However, utilizing ChatGPT to diversify activity sentences results in substantial boosts in performance across target datasets. For example, for Motionsense and HHAR, we see increases between 5-7% over utilizing the base template (row 1 vs row 3 in Tab. 3).

(ii) Leveraging external knowledge about activities: Additional context about classes, obtained through external knowledge sources, potentially enables improved generalization to *new concepts* as they can be described using *known concepts* (Shen et al. 2022). In our experiments, we utilize ChatGPT as the source of external knowledge, and query it to obtain the following information about activities: *(i)* body parts; and *(ii)* description of movements required, for performing the activities. For reference, Tab. 12 and Tab. 13 in the Appendix show external knowledge for Mobiact. We observe that the information about body parts focuses on the muscles such as quadriceps, etc., but places less emphasis on how the movements are performed. In contrast, description of the movements details the sequence of actions more clearly, which are made using the body parts.

→*Insight: External knowledge about activities improves performance*. Information about body parts and movements used for activities generally results in increased performance (rows 4-6 in Tab. 3). Even while using the base handcrafted template, HHAR, Mobiact, and Motionsense, observe improvements of around 6-10% over the base setup. There are further increases when randomly sampling the set of templates during pre-training (i.e., from the 33 templates), and tuning for the best text template during activity recognition. We also observe that information about body parts leads to

#	Pretrain Setup	HAR Setup	Wrist		Waist		Leg	
			HHAR	Myogym	Mobiact	Mot.sense	MHEALTH	PAMAP2
1	Base handcrafted template	Base handcrafted template	29.05	33.30	59.09	73.36	41.72	48.36
2	Randomly sample text template	Best template	33.88	33.40	59.55	72.19	39.69	48.42
3	Randomly sample ChatGPT diversified sentences	Best template	36.01	**33.89**	61.93	78.25	**43.84**	47.9
4	Base handcrafted template + body parts info	Base handcrafted template + body parts info	34.73	31.54	62.77	81.00	39.61	48.78
5	Randomly sample text template + body parts info	Best template + body parts info	**36.43**	32.80	**65.68**	**82.93**	40.36	**51.90**
6	Randomly sample text template + movement descriptions	Best template + movement descriptions	27.01	32.11	65.29	78.13	39.41	51.62

Table 3: **Increasing text diversity in activity descriptions:** Additional information about activities leads to better outcomes.

Method	Modifications	HHAR	Mobiact	PAMAP2
NLS w/ pre-training on train split of target data	–	29.05	59.09	48.36
NLS w/ pre-training on train split of target data	Improved activity sentences + SLIP objective + CLIP text encoder	50.2	63.69	51.95
NLS w/ pre-training on Capture-24	–	31.05	16.93	10.88
NLS w/ pre-training on Capture-24 + adapt. on target data	–	58.63	65.22	54.80
NLS w/ pre-training on Capture-24 + adapt. on target data	Improved activity sentences	63.43	65.28	54.99

Table 4: **Improving sensor-language systems:** Putting together findings from our exploration results in improved HAR.

more accurate HAR than the movements involved, as the underlying description of the body parts is more distinct across activities. Further, the addition of external knowledge is most advantageous for waist-based datasets, which contain locomotion-style activities along with transitions (in the case of Mobiact). From Tables 1 and 3, we see that overcoming the impact of Sensor Heterogeneity is more difficult than dealing with the Lack of rich descriptions of activities, even though Capture-24 is significantly larger than target datasets. Overall, using train splits of target datasets for pre-training is a better option, as LLMs can be utilized to improve text diversity as well as a source of useful external knowledge.

7 Going Beyond HAR through NLS

In Sec. 6, we presented our experimental evaluation of NLS for standard HAR. Here, we go beyond, by exploring alternatives to network components, examining how our adapted NLS method recognizes unseen activities, and performing cross-modal search (full details available in Sec. A.4 and Sec. A.8 of the Appendix (Haresamudram et al. 2024)).

7.1 Improving the NLS Setup

We explore alternatives to components of the NLS setup, in order to further improve performance: *(i) IMU Encoder:* Surprisingly, using simple convolutional encoders results in better performance than more complex ResNets; *(ii) Text Encoder:* BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), and the CLIP Text encoder (Radford et al. 2021) perform better than DistilBERT, though they can have more parameters; and *(iii) Training Objective:* Advancements to the CLIP objective, e.g., allowing multiple matches between sensor windows and sentences (UniCL (Yang et al. 2022)), and using an additional SimCLR loss (SLIP (Mu et al. 2022)) are advantageous. SLIP is the overall best option.

Incorporating Better Alternatives: We obtain substantial improvements in performance with the use of the aforementioned alternatives, as shown in Tab. 4. Using the CLIP text encoder and adding the SimCLR loss is a consistently superior option when pre-training on the train split of target datasets. When studying adaptation on target data, using improved activity sentences (Sec. 6.2) is also advantageous.

7.2 Recognizing Unseen Activities

A core advantage of NLS based classification is the ability to predict classes not seen during (pre-)training. To evaluate this capability, we utilize the protocol detailed in Sec. A.4, where we partition datasets into 3-4 groups where *the test set contains unseen activities*, which are non-overlapping across groups. We tabulate results in Tab. 5, and present the average of mean F1-scores obtained across groups. We also modify the activity sentences based on the findings from Sec. 6.2.

Solely using the base template performs well across the datasets, and is generally a good option. Incorporating external knowledge from ChatGPT about the body parts involved is overall the best option, giving substantial increases

Setup	Wrist		Waist		Leg	
	HHAR	Myogym	Mobiact	Mot.sense	MHEALTH	PAMAP2
Pre-training and HAR using the base text template	55.01	36.96	56.71	40.68	39.54	55.33
Pre-training and HAR using base text template + body parts utilized	33.16	19.96	64.01	51.01	30.33	53.73
Randomly sample text template during pre-training, and HAR with base template + body parts utilized	44.82	28.85	54.46	63.86	41.13	54.02
Randomly sample text template during pre-training, and mean of all test sentence embeddings for HAR	34.45	35.69	52.51	41.48	35.85	53.14

Table 5: **Recognizing unseen activities:** We obtain improved recognition on some datasets with addition of external knowledge.

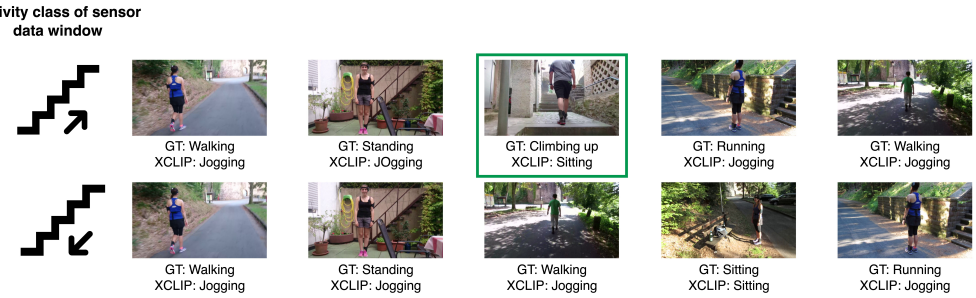

Figure 4: **Evaluating cross modal retrieval capabilities:** For four of the six activities from Motionsense, correct videos are retrieved among top-5 matches. 'GT' is the ground truth label from RealWorld, whereas 'XCLIP' comprises predictions from the pre-trained X-CLIP model. Full figure in the Appendix (Haresamudram et al. 2024).

throughout. Similarly, adding descriptions of movements involved in activities is useful for some datasets. This is in line with works from other domains such as computer vision (Shen et al. 2022), where auxiliary information about the classes (e.g., specific birds having red feathers) typically results in performance increases under zero shot conditions. While natural language supervision for wearables-based HAR has challenges, the ability to perform zero shot prediction is a big plus for practical wearable systems.

7.3 Cross Modal Retrieval of Videos

We investigate whether we can use the text encoder from a pre-trained video-language model (X-CLIP (Ma et al. 2022)) for contrastive training with sensor data, and subsequently *retrieve* similar videos from a *different dataset*, i.e., *perform search.* We evaluate on the *RealWorld* dataset (Sztyler and Stuckenschmidt 2016), which contains videos for locomotion-style activities, e.g., sitting, jogging. For sensor-language pre-training, we use the X-CLIP text encoder and the *Motionsense* dataset. We compute cosine similarity between sensor and video embeddings, and show the 5 closest video matches in Fig. 4 for a random window per activity in Motionsense (for details, see Sec. A.4).

For four classes–Walking up the stairs, Walking, Running, and Sitting–the correct video is retrieved among the top five matches. On the other hand, the Standing and Sitting have many incorrect matches, usually with other static activities, where even supervised methods are routinely confused. Predictions by X-CLIP are often incorrect (despite being trained on large scale and diverse video-text datasets), and, consequently, the closest matches obtained after sensor-language training can also have reduced accuracy. In summary, this experiment shows that NLS aids in cross modal retrieval of videos, even though *both sensor as well as the video encoders were not trained on same datasets.*

8 Summary and Conclusion

We explored whether natural language supervision (NLS) can be employed for zero shot prediction of activities from sensor data in the typically promised plug-and-play manner. We found that this is a *very challenging* endeavor and identified its two primary causes: *Sensor heterogeneity*, and the *Lack of rich, diverse text descriptions of activities.* Sensor heterogeneity causes HAR in diverging target conditions to be poor. To tackle this, we proposed to use small amounts of labeled target data for adaptation. To increase diversity in activity descriptions, we explored augmentation and incorporating external knowledge from pre-trained LLMs. Both strategies resulted in substantial improvements. While sensor-language modeling does not outperform state-of-the-art supervised and self-supervised training for some datasets, its additional capabilities, like recognizing unseen activities and performing cross-modal search, are clearly advantageous for real world scenarios. Our solutions result in improved sensor-language learning, paving the way for foundational models of human movements.

Acknowledgements

This work was partially supported by NSF Grant IIS-2112633, and grants from Optum and Google.

References

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*.

Banos, O.; Garcia, R.; Holgado-Terriza, J. A.; Damas, M.; Pomares, H.; Rojas, I.; Saez, A.; and Villalonga, C. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, 91–98. Springer.

Bin Morshed, M.; Haresamudram, H. K.; Bandaru, D.; Abowd, G. D.; and Plötz, T. 2022. A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting. In *Proceedings of the 2022 ACM international symposium on wearable computers*, 11–16.

Chatzaki, C.; Pediaditis, M.; Vavoulas, G.; and Tsiknakis, M. 2016. Human daily activity and fall recognition using a smartphone's acceleration sensor. In *International Conference on Information and Communication Technologies for Ageing Well and e-Health*, 100–118. Springer.

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.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023. Improving CLIP Training with Language Rewrites. *arXiv preprint arXiv:2305.20088*.

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.

Haresamudram, H.; Beedu, A.; Rabbi, M.; Saha, S.; Essa, I.; and Ploetz, T. 2024. Limitations in Employing Natural Language Supervision for Sensor-Based Human Activity Recognition–And Ways to Overcome Them. *arXiv preprint arXiv:2408.12023*.

Haresamudram, H.; Essa, I.; and Plötz, T. 2022. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3): 1–47.

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

Imran, S. A.; Khan, M. N. H.; Biswas, S.; and Islam, B. 2024. LLaSA: Large Multimodal Agent for Human Activity Analysis Through Wearable Sensors. *arXiv preprint arXiv:2406.14498*.

Koskimäki, H.; Siirtola, P.; and Röning, J. 2017. Myogym: introducing an open gym data set for activity recognition collected using myo armband. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 537–546.

Kwon, H.; Tong, C.; Haresamudram, H.; Gao, Y.; Abowd, G. D.; Lane, N. D.; and Ploetz, T. 2020. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3): 1–29.

Li, Z.; Deldari, S.; Chen, L.; Xue, H.; and Salim, F. D. 2024. SensorLLM: Aligning Large Language Models with Motion Sensors for Human Activity Recognition. *arXiv preprint arXiv:2410.10624*.

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

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Malekzadeh, M.; Clegg, R. G.; Cavallaro, A.; and Haddadi, H. 2018. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, 1–6.

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

Mishra, R.; Gupta, A.; Gupta, H. P.; and Dutta, T. 2020. A sensors based deep learning model for unseen locomotion mode identification using multiple semantic matrices. *IEEE Transactions on Mobile Computing*, 21(3): 799–810.

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*.

Moon, S.; Madotto, A.; Lin, Z.; Nagarajan, T.; Smith, M.; Jain, S.; Yeh, C.-F.; Murugesan, P.; Heidari, P.; Liu, Y.; et al.

2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, 529–544. Springer.

Ordóñez, F. J.; and Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1): 115.

Plötz, T. 2023. If only we had more data!: Sensor-Based Human Activity Recognition in Challenging Scenarios. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 565–570. IEEE.

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.

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.

Saeed, A.; Ozcelebi, T.; and Lukkien, J. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2): 1–30.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35: 15558–15573.

Stisen, A.; Blunck, H.; Bhattacharya, S.; Prentow, T. S.; Kjærgaard, M. B.; Dey, A.; Sonne, T.; and Jensen, M. M. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, 127–140.

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.

Tang, C. I.; Perez-Pozuelo, I.; Spathis, D.; and Mascolo, C. 2020. Exploring Contrastive Learning in Human Activity Recognition for Healthcare. *arXiv preprint arXiv:2011.11542*.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Tong, C.; Ge, J.; and Lane, N. D. 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): 1–23.

Verma, G.; Choi, M.; Sharma, K.; Watson-Daniels, J.; Oh, S.; and Kumar, S. 2024. Mysterious Projections: Multimodal LLMs Gain Domain-Specific Visual Capabilities Without Richer Cross-Modal Projections. *arXiv preprint arXiv:2402.16832*.

Willetts, M.; Hollowell, S.; Aslett, L.; Holmes, C.; and Doherty, A. 2018. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific reports*, 8(1): 1–10.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4563–4567. IEEE.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Xia, K.; Li, W.; Gan, S.; and Lu, S. 2024. TS2ACT: Few-Shot Human Activity Sensing with Cross-Modal Co-Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4): 1–22.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19163–19173.

Zhou, Y.; Yang, J.; Zou, H.; and Xie, L. 2023. Tent: Connect language models with iot sensors for zero-shot activity recognition. *arXiv preprint arXiv:2311.08245*.