The full version of this paper is available at: arxiv.org/abs/2411.15127

PRIMUS: Pretraining IMU Encoders with Multimodal and Self-Supervised Learning

Arnav M. Das[1]* Chi Ian Tang[2] Fahim Kawsar[2,3] Mohammad Malekzadeh[2]

[1]University of Washington, USA [2]Nokia Bell Labs, UK [3]University of Glasgow, UK

Abstract

Sensing human motions through Inertial Measurement Units (IMUs) embedded in personal devices has enabled significant applications in health and wellness. While labeled IMU data is scarce, we can collect unlabeled or weakly labeled IMU data to model human motions. For video or text modalities, the "pretrain and adapt" approach utilizes large volumes of unlabeled or weakly labeled data for pretraining, building a strong feature extractor, followed by adaptation to specific tasks using limited labeled data. However, for IMU data, pretraining methods are poorly understood, and pretraining pipelines are rarely evaluated on out-of-domain tasks. We propose PRIMUS: a method for PRetraining IMU encoderS that uses a novel pretraining objective that is empirically validated based on downstream performance on both in-domain and out-of-domain datasets. The PRIMUS objective effectively enhances downstream performance by combining self-supervision, multimodal, and nearest-neighbor supervision. With fewer than 500 labeled samples per class, PRIMUS can improve test accuracy by up to 15%, compared to state-of-the-art baselines. To benefit the broader community, we open-source our code at github.com/nokia-bell-labs/pretrained-imu-encoders.

1 Introduction

Wearable devices embed Inertial Measurement Unit (IMU) sensors, including accelerometers and gyroscopes, which track the movement, acceleration, and orientation of the human body. When modeled using machine learning (ML) methods, IMU data provides valuable insights into human physical and emotional behaviors, playing a crucial role in health monitoring and overall well-being [13, 6, 20, 31, 19]. For example, step-counting data from IMU sensors is one of the most effective indicators of cognitive impairment progression in elderly individuals [1]. Such potential has motivated the community to collect vast amounts of IMU data in time-series form. However, obtaining large amounts of *labeled* IMU data remains a major challenge, because IMU time series are inherently difficult to interpret and annotate, even by experts [30].

A promising solution for label scarcity is the "pretrain once, adapt many times" approach. This involves initially training an *encoder* on a large corpus of unlabeled or weakly labeled data. Afterward, a smaller ML model is trained on top of the (typically frozen) encoder for specific tasks, using relatively small amounts of labeled data. While this approach has shown significant success in image, video, audio, and natural language processing, its potential for IMU data remains underexplored, primarily because of challenges in curating large volumes of quality datasets.

Difficulties in collecting labeled data have motivated *representation learning* methods for IMU encoders by using supervisory signals from IMU data itself (*self-supervised learning*), or other concurrent modalities (*multimodal learning*). Self-supervised (SS) learning approaches based on

*Work has been done during the author's internship at Nokia Bell Labs.

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

multi-task learning [22, 25], contrastive learning [26, 29], and masked reconstruction [9], have yet to be evaluated on cross-domain use cases. Multimodal (MM) learning has become popular in the field of representation learning [21, 11, 28, 15, 5], and has recently been used for pretraining IMU encoders by utilizing supervisory signals from multiple devices [10, 23] or multimodal data [3]. IMU2CLIP [17] aligns the latent representations of IMU data with those coming from text annotations or those from egocentric videos, where they show enhanced capabilities in multimodal data retrieval.

While both SS and MM approaches for representation learning have shown promising results, neither one fully leverages diverse sources of information present in IMU data. Given the promising synergistic relationship between SS and MM learning shown in computer vision and natural language processing fields [18, 12, 27], and with the recent publicly available large multimodal datasets EgoExo4D [8], which includes synchronized video, text, and IMU segments, we explore the combination of SS and MM learning for pretraining IMU encoders.

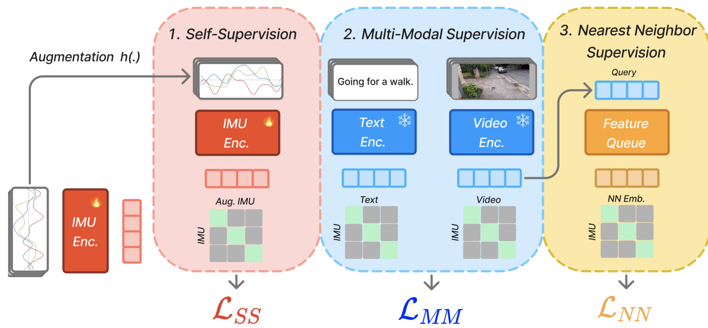


Figure 1: **PRIMUS** uses multi-objective pretraining including three terms, $\mathcal{L}_{SS}, \mathcal{L}_{MM}$, and $\mathcal{L}_{NN}$. Self-supervised losses encourage the IMU encoder to be augmentation invariant, while multimodal and nearest neighbor losses align the IMU data to co-occurring video and/or text data. We use open-source models, pretrained by others, for both text and video encoders.

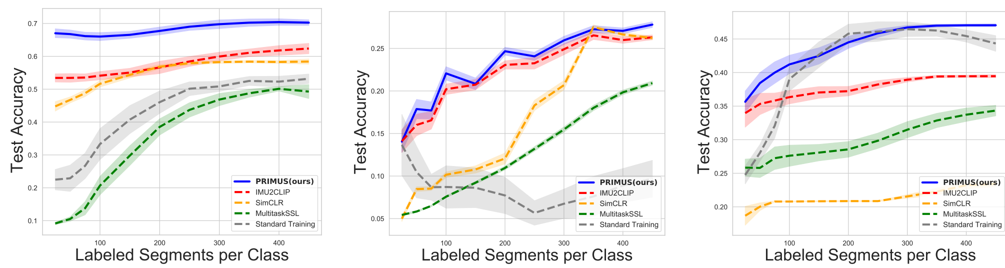
We propose **PRIMUS** (see Figure 1): a novel method for PRetraining IMU encoderS that produces transferable representations by building a multi-objective representation learning strategy that combines SS and MM losses to pretrain an IMU encoder. Pretrained on the recently released EgoExo4D dataset [8], we assess the effectiveness of our strategy by evaluating how well PRIMUS IMU encoder performs on both *in-domain* and *out-of-domain* classification tasks using only a small amount of labeled data (i.e., few-shot learning). A consistent performance improvement of up to 15% in test accuracy when compared to existing state-of-the-art multimodal and self-supervised training methods was observed throughout different levels of data availability. Our ablation study also showcased the superior performance of our proposed combined training objective. This demonstrates that PRIMUS enables encoders to learn highly transferable representations, allowing for various future adaptations.

2 Methodology

Let $\mathcal{I}$ denote an *encoder* that takes a segment of multivariate IMU time series as input and generates a latent representation as output. As shown in Fig. 1, we train $\mathcal{I}$ with three objectives: *self-supervision loss* ($\mathcal{L}_{SS}$), *multimodal loss* ($\mathcal{L}_{MM}$), and *nearest-neighbour loss* ($\mathcal{L}_{NN}$). (1) $\mathcal{L}_{SS}$ ensures that $\mathcal{I}$ remains invariant to noise, similar to those that are introduced by slight changes in sensor position or type. (2) $\mathcal{L}_{MM}$ pushes IMU representations towards aligned text and video representations, allowing $\mathcal{I}$ to learn the rich semantic information present in other modalities. (3) $\mathcal{L}_{NN}$ uses the closest examples in representation space as positive pairs, enabling the model to leverage natural data similarities for more adaptive contrastive learning.

In our implementation, $\mathcal{I}$ is a Stacked RNN consisting of convolutional, group normalization, and max-pooling layers, topped with a GRU layer, based on the architecture of the IMU2CLIP model [17], with a total of 1.4M parameters (see Appendix Fig. 4). For pretraining, we use the EgoExo4D dataset [8], a multimodal dataset containing IMU from head-placed sensors, egocentric videos, and free-form text annotations. After pre-processing, this dataset consists of around 250K segments, each of 5-second length, providing aligned IMU, video, and text triplets. We denote the pretraining dataset as $\mathcal{D} = \{(m_i, v_i, t_i)\}_{i=1}^{N}$ where m_i, v_i, t_i correspond to a single segment of time-aligned IMU, video, and text, respectively.

Self-Supervision. The self-supervised learning objective is an unimodal loss that encourages the representations of augmented versions of the same data to be similar (the first block, shown in red in Fig. 1). For data augmentation, we define a stochastic transformation module $h(.)$ consisting of two transformations: (1) scaling by a random factor and (2) reversing the direction of time. These transformations were chosen after evaluating all pairs of augmentations proposed in [26].



(a) EgoExo4D Results (b) Ego4D Results (c) REALWORLD Results

Figure 2: **Main Results.** The few-shot learning performance on various classification datasets. PRIMUS generally outperforms SS methods (SimCLR, MultitaskSSL), and prior MM methods (IMU2CLIP), as well as training a randomly initialized model (standard training). The standard error is computed over 5 trials.

Given a batch $B = \{(m_i, v_i, t_i)\}_{i=1}^n$, and considering τ as a learnable temperature parameter, the self-supervised objective, adapted from SimCLR [2, 26], can be formally expressed as

$$\mathcal{L}_{SS}(B) = \sum_{i=1}^{n} \frac{\exp\left(\mathcal{I}(m_i) \cdot \mathcal{I}(h(m_i))\right)^{1/\tau}}{\sum_{k=1}^{n} \exp\left(\mathcal{I}(m_i) \cdot \mathcal{I}(h(m_k))\right)^{1/\tau}},$$

Multimodal Supervision. We use multimodal learning (the second block, shown in blue in Fig. 1) in order to allow the IMU encoder to learn semantic features that are present in rich modalities such as text and video, but difficult to learn with self-supervision alone [17]. Many open-source video and text encoders have been pretrained on web-scale data and can be used to produce rich representations for the video/text in each frame. Throughout this paper, we use an open-source video encoder $\mathcal{V}$ and text encoder $\mathcal{T}$ produced by CLIP4Clip [15] to instantiate our multimodal learning objective, since this model is designed to handle short video clips and is readily available. Given a batch $B = \{(m_i, v_i, t_i)\}_{i=1}^n$, the multimodal loss has two components which can be expressed as

$$\mathcal{L}_{m2v}(B) = \sum_{i=1}^{n} \frac{\exp\left(\mathcal{I}(m_i) \cdot \mathcal{V}(v_i)\right)^{1/\tau}}{\sum_{j=1}^{n} \exp\left(\mathcal{I}(m_i) \cdot \mathcal{V}(v_j)\right)^{1/\tau}}, \; \mathcal{L}_{m2t}(B) = \sum_{i=1}^{n} \frac{\exp\left(\mathcal{I}(m_i) \cdot \mathcal{T}(t_i)\right)^{1/\tau}}{\sum_{j=1}^{n} \exp\left(\mathcal{I}(m_i) \cdot \mathcal{T}(t_j)\right)^{1/\tau}},$$

where again τ is a learnable temperature. Intuitively, $\mathcal{L}_{m2v}$ (or $\mathcal{L}_{m2t}$) encourage $\mathcal{I}$ to map IMU data to representations that are close to corresponding video (or text) representations in the latent space. We use $\mathcal{L}_{MM}(B) = \mathcal{L}_{m2v}(B) + \mathcal{L}_{m2t}(B)$ as our MM objective.

Nearest Neighbor Supervision. The loss terms introduced so far, $\mathcal{L}_{SS}$ and $\mathcal{L}_{MM}$, both derive supervision from within the same triplet segment. To increase the diversity of supervision and go beyond a single instance, we leverage nearest-neighbor supervision [12, 4] (shown in the rightmost block in orange in Fig. 1 and in detail in Appendix Fig. 5). During training, we maintain a feature queue $\mathcal{Q} = \{(z_j^m, z_j^v, z_j^t)\}_{j=1}^K$, where z_j^m, z_j^v, and z_j^t are cached representations of IMU, video, and text produced from their respective encoders. For every given instance (m_i, v_i, t_i) in a batch B, we define $\eta(i) = \text{argmax}_{k \in [K]} \left(z_k^v \cdot \mathcal{V}(v_i)\right)$, which identifies the index k in $\mathcal{Q}$ corresponding to the video embedding that is the most similar to v_i. We leverage the video representations for identifying the closest pairs because the video encoder is pretrained on a large dataset, and therefore produces stable representations. Also, videos capture much finer details about human activities compared to text descriptions. We then push $\mathcal{I}(m_i)$ close to $z_{\eta(i)}^m, z_{\eta(i)}^v, z_{\eta(i)}^t$ by $\mathcal{L}_{NN}$, which consists of a unimodal and multimodal loss similar to $\mathcal{L}_{SS}$ and $\mathcal{L}_{MM}$ as

$$\mathcal{L}_{NN}(B) = \sum_{mod \in \{m,v,t\}} \sum_{i=1}^{n} \frac{\exp\left(\mathcal{I}(m_i) \cdot z_{\eta(i)}^{mod}\right)^{1/\tau}}{\sum_{j=1}^{n} \exp\left(\mathcal{I}(m_i) \cdot z_{\eta(j)}^{mod}\right)^{1/\tau}}. \tag{1}$$

The final **multi-objective loss** that we use in PRIMUS is $\mathcal{L}(B) = \alpha \mathcal{L}_{SS}(B) + \beta \mathcal{L}_{MM}(B) + \gamma \mathcal{L}_{NN}(B)$. We set $\alpha = \beta = \gamma = 1$, and leave the fine-tuning of hyperparameters to future studies. Note that, as these hyperparameters are not tuned, the results reported in the following sections represent lower bounds on the performance achievable with a more thorough hyperparameter search.

3 Experimental Evaluation

We evaluate on human activity recognition tasks **using only IMU data** (see the datasets' details in Appendix Table 1). We consider different levels of data scarcity by varying the number of labeled segments per class (i.e., few-shot learning). We compare with baselines by analyzing the performance of a linear classifier on the representations produced by the IMU encoder (i.e., linear probing [21]), a technique which requires few computational resources to train and retains the robustness of the pretrained encoder. We compare PRIMUS against other pretraining baselines (§3.1), evaluate data efficiency of PRIMUS (§3.2), and conduct ablations on each loss term of PRIMUS (Appendix C).

3.1 Main Results

We compare our **PRIMUS** against four baselines. (I) **SimCLR** [26] is a self-supervised training method based on data augmentations. (II) **IMU2CLIP** [17] is a multimodal training method for IMU data, corresponding to using only $\mathcal{L}_{SS}$ or $\mathcal{L}_{MM}$ as the pretraining objective on EgoExo4D. Moreover, our work leverages supervisory signals from different learning setups to train a better-performing feature extractor. Thus, we also compare PRIMUS against (III) **MultitaskSSL** [22], a well-established self-supervised approach for IMU signals. Finally, we compare PRIMUS against (IV) **Standard Training**, which starts from a randomly initialized model and updates *all the parameters* (as opposed to just the final layer) with standard supervised learning. This final baseline represents the standard procedure used to train a model in the absence of a pretrained IMU encoder.

Fig. 2 presents the comparison of PRIMUS with all four baselines. Across all experiments, we observe that our PRIMUS model, pretrained with the joint objective, significantly outperforms any pretraining strategy previously proposed. Our method consistently outperforms all other baselines by as much as 15% on the EgoExo4D dataset. On Ego4D, it performs on par with IMU2CLIP but still surpasses all other baselines. Additionally, on the REALWORLD dataset, our method generally outperforms all baselines, particularly in scenarios where labeled data is limited (fewer than 100 samples). Notably, standard training, which updates all the parameters, fails to generalize well in the low-data regime particularly for complex classification tasks (on EgoExo4D and Ego4D). We include additional results evaluating the importance of each objective in the PRIMUS loss in Appendix C.

3.2 PRIMUS Pretraining Data Efficiency

Obtaining large-scale IMU datasets that are temporally aligned with videos and text could be challenging. Therefore, we explore the possibility of training an effective IMU encoder using only a small portion of the data that includes aligned video and text. Specifically, we remove the aligned video and text data for different fractions of the pretraining data and evaluate the efficacy of the resulting IMU encoder in few-shot learning. Fig. 3 shows that there is no statistically significant difference between an encoder pretrained with only 10% of the data aligned with video/text with the PRIMUS and an encoder pretrained in the style of IMU2CLIP on EgoExo4D in terms of few-shot learning performance.

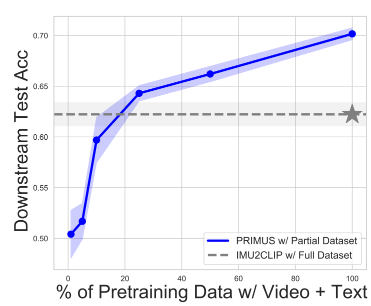


Figure 3: **Data Efficiency.** We report few-shot performance on the EgoExo4D classification task at 500 segments per class for PRIMUS models trained with various amounts of multimodal data. Models pretrained with the PRIMUS objective require far fewer IMU segments with aligned video/text than IMU2CLIP.

4 Conclusion

We study pretraining objectives for IMU time series that can be adapted to unseen tasks with limited labeled data. We empirically demonstrate the superiority of our pretraining method against existing approaches on in-domain and out-of-domain tasks, and identify some of the components that were critical to its success. See our discussion on future work in Appendix D.

References

[1] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2145–2155, 2019.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/chen20j.html`.

[3] Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora D Salim, and Akhil Mathur. Crossl: Cross-modal self-supervised learning for time-series through latent masking. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 152–160, 2024.

[4] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597, October 2021.

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022. URL `https://arxiv.org/abs/2206.04769`.

[6] Hemant Ghayvat, Jie Liu, Subhas Chandra Mukhopadhyay, and Xiang Gui. Wellness sensor networks: A proposal and implementation for smart home for assisted living. *IEEE Sensors Journal*, 15(12):7341–7348, 2015.

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL `https://arxiv.org/abs/2110.07058`.

[8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. URL `https://arxiv.org/abs/2311.18259`.

[9] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 45–49, 2020.

[10] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–28, 2022.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL `https://arxiv.org/abs/2201.12086`.

[12] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=zq1iJkNk3uN`.

[13] Yunji Liang, Xingshe Zhou, Zhiwen Yu, and Bin Guo. Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare. *Mobile Networks and Applications*, 19: 303–317, 2014.

[14] Wang Lu, Jindong Wang, and Yiqiang Chen. Local and global alignments for generalizable sensor-based human activity recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3833–3837. IEEE, 2022.

[15] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[16] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3): 90–126, 2006.

[17] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text, 2022. URL `https://arxiv.org/abs/2210.14395`.

[18] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[19] Andre GC Pacheco, Frank AC Cabello, Adriana MO Fonoff, Paula G Rodrigues, Otávio AB Penatti, and Paula R Pinto. Towards low-power heart rate estimation based on user's demographics and activity level for wearables. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[20] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290, 2010.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

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

[23] Yifan Song, Songpengcheng Xia, Jiarui Yang, and Ling Pei. A learning-based multi-node fusion positioning method using wearable inertial sensors. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1976–1980. IEEE, 2024.

[24] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–9. IEEE, 2016.

[25] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1), March 2021. doi: 10.1145/3448112. URL `https://doi.org/10.1145/3448112`.

[26] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare, 2021. URL `https://arxiv.org/abs/2011.11542`.

[27] Sahil Verma, Gantavya Bhatt, Avi Schwarzschild, Soumye Singhal, Arnav Mohanty Das, Chirag Shah, John P Dickerson, and Jeff Bilmes. Effective backdoor mitigation depends on the pre-training objective, 2023. URL `https://arxiv.org/abs/2311.14948`.

[28] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567, 2022. doi: 10.1109/ICASSP43922.2022.9747669.

[29] Cong Xu, Yuhang Li, Dae Lee, Dae Hoon Park, Hongda Mao, Huyen Do, Jonathan Chung, and Dinesh Nair. Augmentation robust self-supervised learning for human activity recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

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

[31] Shibo Zhang, Ebrahim Nemati, Minh Dinh, Nathan Folkman, Tousif Ahmed, Mahbubur Rahman, Jilong Kuang, Nabil Alshurafa, and Alex Gao. Coughtrigger: Earbuds imu based cough detection activator using an energy-efficient sensitivity-prioritized time series classifier. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2022.

APPENDIX

A PRIMUS Methodology

A.1 IMU Encoder Architecture

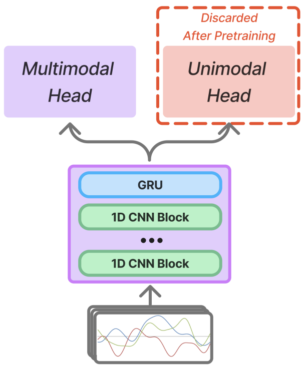


Figure 4: **The architecture of IMU Encoder $\mathcal{I}$.** The backbone consists of both 1D-CNN and GRU layers. During pretraining, the IMU encoder has two MLP heads: one for multimodal loss and the other for unimodal loss. After pre-training, only the output of the multimodal head is kept for training downstream tasks, as it offers a more generalized latent representation. The architecture is adopted from [17].

Our motivation for this architecture (Figure 4) is its efficiency in deployment on mobile and wearable devices, which are the target platforms for collecting IMU data [14]. Moreover, it has shown effective generalization performance in processing ML tasks on IMU data. During pretraining, $\mathcal{I}$ has two MLP heads: the first head is used to compute the *unimodal* self-supervision loss and the second head is used to compute the *multimodal* loss. For downstream tasks, only the latter is retained as it provides a richer latent representation.

A.2 Nearest Neighbor Supervsion

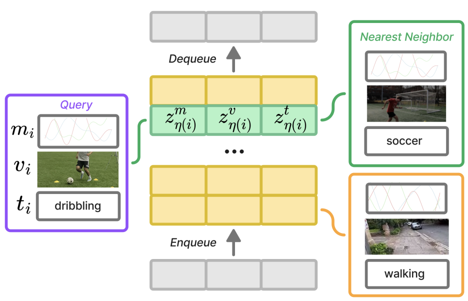


Figure 5: **Nearest neighbor supervision.** Given a query segment, we retrieve the most similar segment in the queue, based on video-to-video similarity, and use all modalities to derive supervisory signals for the IMU segment. Features are retrieved from a fixed-size queue.

We illustrate the queuing mechanism for nearest neighbor retrieval in Fig. 5.

B Experimental Setup

B.1 Datasets and Setup

Table 1: **Downstream Tasks.** Classification datasets for our evaluation. Unlike previous work, we consider tasks that have IMU data collected from unseen devices and have novel output domains.

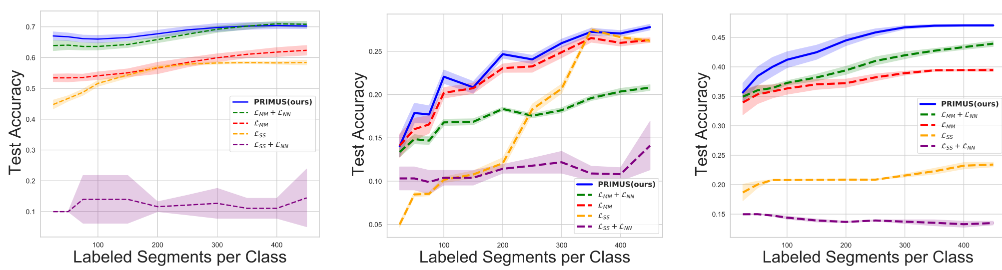
Test Set	Activities	Input Domain	Output Domain	Sample Size
EgoExo4D [8]	8: {play music, cook, medical test, perform CPR, repair bike, climb rock, soccer, dance}	Same	Same	Train: 195K–Test: 53K
Ego4D [7]	10: {play music, cook, eat, clean, carpenter, craft, farmer, household, walk, construction}	Same	*Different*	Train: 555K–Test: 57K
REALWORLD [24]	8: {climbing up, climbing down, jumping, lying down, run, walk, sit, down}	*Different*	*Different*	Train: 8.3K–Test: 2.6K

All encoders are pretrained on EgoExo4D, which contains IMU data (triaxial accelerometer and triaxial gyroscope) collected from head-placed sensors. Thus, we focus on downstream tasks that use IMU data of head-placed sensors. A summary of datasets is given in Table 1.

EgoExo4D [8] and **Ego4D [7].** From each dataset, we choose a held-out test set for human activity recognition, where IMU data is labeled according to the activities indicated in the filenames. Note that Ego4D is captured using the same device, Project-Aria smartglass[2], the same as EgoExo4D (pretraining dataset), but Ego4D includes some activities that are not present in EgoExo4D.

REALWORLD [24]. The REALWORLD dataset is a human activity recognition dataset with 8 predefined classes, that contain data captured by various Samsung Galaxy-S4 and LG G-Watch-R placed at different positions on the body. For our analysis, we use the data from the head-placed sensor. We adopted the well-established user-based dataset-splitting strategy for our evaluations, in which data from a held-out set of users are reserved for testing, measuring the performance of the model on unseen users. This dataset also evaluates the out-of-domain performance of our pretrained models since both the set of activities and device type are different from EgoExo4D.

C Ablations



(a) EgoExo4D Results (b) Ego4D Results (c) REALWORLD Results

Figure 6: **Ablations.** The importance of each individual term in the PRIMUS objective. We pretrain encoders with different losses and evaluate them based on few-shot learning performance; each experiment over 5 trials.

Fig. 6 presents an ablation study on the pretraining objectives to understand which components of the loss are most critical. We find that $\mathcal{L}_{MM}$ is a key component, indicating that future studies for developing IMU foundation models should incorporate aligned video, text, audio, and potentially other under-explored wearable sensors. We also find that $\mathcal{L}_{NN}$ is generally helpful, but only when we have a reliable estimate of similarity. With $\mathcal{L}_{SS}+\mathcal{L}_{NN}$, we observe some form of collapse (with accuracy around 10-15%) since this setting does not exploit any multimodal signals, and using the IMU representations itself to find similar segments from the queue can make training unstable. Finally, while $\mathcal{L}_{SS}$ is not particularly helpful in EgoExo4D (evident from the fact that $\mathcal{L}_{MM}+\mathcal{L}_{NN}$ nearly matches the performance of PRIMUS), self-supervision seems to make a significant difference on *out-of-domain* tasks, offering up to 5% of accuracy improvement in REALWORLD. We hypothesize that this is due to the fact that this loss term explicitly encourages the IMU encoder to be invariant to some of the types of noise that may be observed due to changing devices or positions on the body.

D Future Work

Our work has its limitations and there are several promising directions for future work. First, while we evaluate on out-of-domain downstream tasks, all of our evaluation schemes assume that the sensor position on the human body is similar to that of the pretraining set. Training a model that is capable of generalizing across human body positions is an important future direction, but pretraining datasets to enable this are not yet available. Second, our evaluation focuses on activities of medium granularity (corresponding to '*actions*' according to the hierarchy of activities proposed in [16]). To recognize more abstract or primitive activities, a different processing pipeline would be needed to accommodate the different time scales in which these activities occur. Further studies might be needed to adapt our proposed method for these different scenarios. Despite open challenges, our work contributes to developing generalizable IMU models by introducing a highly adaptable pretraining strategy. By open-sourcing our framework, we aim to encourage the community to further build upon our efforts.

[2] `https://www.projectaria.com`