Inertia-1: An Open Exploration of Wearable Motion Foundation Models

Zongzhe Xu[*]
University of California, Los Angeles

Aakarsh Anand[*]
University of California, Los Angeles

Sarah Jiang[*]
Duke University

Chuntung Zhuang
Johns Hopkins University

Zitao Shuai
University of California, Los Angeles

Sriram Sankararaman
University of California, Los Angeles

Yuzhe Yang[†]
University of California, Los Angeles
yuzhey@ucla.edu

Abstract

Wearable motion sensing provides a continuous and scalable window into human behavior and health, making it a natural fit for foundation models, yet its pretraining and scaling principles remain poorly understood. Prior work studies isolated design choices, such as sensor placement or sampling frequency, often under fixed settings and narrow downstream tasks that fail to capture real-world sensing diversity. We introduce `Inertia-1`, a fully open exploration of wearable motion foundation models. Using massive corpora of accelerometer data from global sources spanning more than 18.2M hours, we build a controlled framework for studying the *full life-cycle* of wearable motion foundation models, covering *data choices* such as sensor modality, device placement, sampling rate, window length; *model choices* such as architectures and model size; and *training choices* such as pretraining objective and data scale. Extensive evaluations across 15 datasets spanning *human activity recognition*, *freezing-of-gait detection*, and *disease prediction* reveal intriguing findings for building motion foundation models that generalize across tasks and sensing conditions. Collectively, `Inertia-1` not only presents state-of-the-art recipes for diverse downstream tasks, but also serves as a comprehensive, practical, and open cookbook for wearable motion representation learning.

1 Introduction

Foundation models turn scale into reusable structure [45,50,58,63]. For wearable intelligence, *motion* is one of the most natural sources of such structure: continuous accelerometer and gyroscope streams record how people move, rest, and function in daily life, linking everyday behavior to mobility [10,42], physiology [32,61], and personal health [17,29]. Large population cohorts now collect longitudinal motion recordings over days or weeks [13,17,37], supporting applications from daily activity analysis and mobility assessment to disease risk stratification. Major studies such as UK Biobank [17] contain *millions of hours* of motion signals, yet most of these data remain unlabeled and underused. This creates a clear potential for wearable motion foundation models: learning general-purpose models from large unlabeled cohorts and transferring them across diverse behavioral and health tasks.

Realizing this potential, however, is hindered by a *fragmented* research landscape. Such fragmentation is layered: ***data pipelines*** differ in sampling rate, window length, sensor modality, body placement,

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

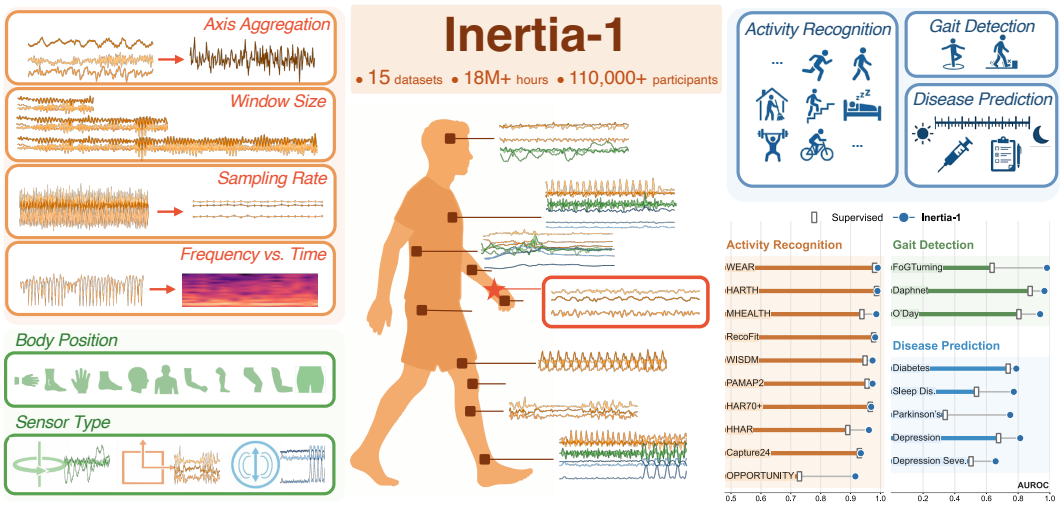


Figure 1: **Overview of `Inertia-1`.** We present a unified and fully open exploration of wearable motion FMs, spanning large-scale accelerometer pretraining, diverse self-supervised objectives, controlled sensing setups, and downstream evaluation across activity recognition, gait analysis, and disease prediction. `Inertia-1` covers 18.2M hours of motion data from over 115,000 individuals and 15 datasets, enabling systematic study of how data, model, and training choices shape transferable motion representations. More details are in Appendix A.

and axis representation; ***methodologies*** differ in pretraining objective, architecture, representation domain, and training recipe; ***datasets*** differ in scale, duration, label density, and clinical context; and ***downstream evaluations*** differ across human activity recognition (HAR), gait analysis, and disease prediction. As summarized in Table 1, existing motion studies typically cover only a subset of sensors, placements, tasks, and scales, leaving the field with strong local findings but limited guidance for building a general-purpose wearable motion foundation model.

Specifically, the nature of this fragmentation is twofold. **First**, *data configurations* and *modeling methods* are rarely examined under a shared protocol. Prior work has studied diverse but isolated design factors such as windowing [3], sensor displacement [5], and device heterogeneity [46] across disjoint datasets, while recent foundation-model efforts explore reconstructive [30], contrastive [60], self-distillation [49], and frequency-domain objectives [25], often under fixed input settings. Yet, when objectives, preprocessing choices, sensing setups, and representation domains all vary together, it becomes difficult to attribute progress to *any* single design choice. **Second**, *dataset scale* and *task definition* are split across regimes. *Large-scale* cohorts provide population-level longitudinal data and health outcomes, but often release compressed motion representations, such as downsampled traces or vector-magnitude summaries [34, 37]. In contrast, datasets for HAR and freezing-of-gait (FoG) offer fine-grained, high-frequency signals, but are inherently *small-scale*, with limited participants and short sessions [42]. This leaves open whether massive unlabeled cohorts can produce representations that bridge both small, label-scarce motion tasks and population-level disease modeling.

To fill the gaps, we present `Inertia-1`, an open exploration of wearable motion foundation models. Rather than simply proposing a new architecture, `Inertia-1` is designed to disentangle how pretraining objectives, sensing configurations, scale, and downstream tasks jointly shape transferable motion representations (see Fig. 1): ❶ **Pretraining.** We integrate 10 representative methods under the same training and evaluation settings, covering 5 classes of pretraining objectives, including general self-supervised as well as motion-specific approaches. ❷ **Data pipelines.** We evaluate models across a rigorous grid of experimental axes, including sampling rate, window length, sensor modality, sensor axis dimensionality, and body placement (Table 7). ❸ **Scale.** We aggregate, to our knowledge, the largest waveform-level wearable motion data collection to date, spanning 18.2M hours of motion data from more than 115,000 people and 15 datasets (Table 1). ❹ **Downstream tasks.** We evaluate across 10 human activity recognition datasets, 3 freezing-of-gait datasets, and 7 disease prediction tasks, spanning both fine-grained motion understanding and population-level health modeling.

`Inertia-1` is designed to be easily extensible, supporting new models, datasets, and tasks as the field evolves. With the unified framework and more than 1,000 trained models, we reveal intriguing lessons and future directions for building state-of-the-art wearable motion foundation models:

Table 1: **Comparisons of studies on wearable motion data modeling.** We compare a subset of major device placements; the full set of covered placements in `Inertia-1` is provided in Table 7.

Study†	# People	# Hours (000s)	Datasets	Sensors			Placements						Tasks		
				ACC	GYR	MAG	WR	AR	CH	HP	LG	AN	HAR	FoG	DP
Banos *et al.* [5]	17	≪1	1	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗
Stisen *et al.* [46]	9	≪1	1	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗
Haresamudram *et al.* [19]	320	4	10	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗
Yuan *et al.* [62]	100,000	15,700	9	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗
Logacjov *et al.* [25]	35,000	100	6	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✗
Hoddes *et al.* [20]	60	1.6	4	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗
Xu *et al.* [56]	87,380	720	7	✓	✗	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗
`Inertia-1` **(Ours)**	**115,450**	**18,224**	**15**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

ACC: Accelerometer. GYR: Gyroscope. MAG: Magnetometer. WR: Wrist. AR: Arm. CH: Chest. HP: Hip. LG: Leg. AN: Ankle.
HAR: Human Activity Recognition. FoG: Freezing of Gait Detection. DP: Disease Prediction. † Only raw-waveform studies considered.

- *Pretraining is beneficial, but not universal* – Self-supervised pretraining consistently improves over supervised counterparts, yet no single objective dominates across all task families and metrics.
- *Data representation is a first-order modeling choice* – Triaxial motion signals outperform magnitude summaries, while sampling rate and window length matter differently across task granularities.
- *Scale helps, but not all scale is equal* – More (diverse) pretraining data yields steadier gains than larger model sizes, and multi-sensor fusion lead to strong gains in sensor-expansion settings.
- *State-of-the-art motion foundation models require coordinated design* – Performance depends on the joint choice of data fidelity, sensing setup, objective, and scale, rather than any single recipe.
- *Inertia-1 as an open cookbook* – As to date the most diverse and unified testbed, `Inertia-1` offers a practical foundation for studying, extending, and deploying wearable motion foundation models.

2 Related Work

Applications of Wearable Motion Modeling. Wearable motion signals support tasks across very different temporal and clinical scales. HAR classifies short windows of inertial data into activities such as walking, sitting, running, or household actions [16, 23], while FoG detection identifies brief clinically meaningful gait events in Parkinson's disease [22, 40]. At a larger scale, population studies use multi-day accelerometry to quantify rest-activity rhythms, sleep, mobility, sedentary behavior, and disease risk [17, 28, 44]. These areas have usually evolved separately: HAR emphasizes fine-grained labels and high-resolution windows, FoG emphasizes symptom-specific detection in targeted cohorts, and population-health studies emphasize aggregate longitudinal behavior. In contrast, `Inertia-1` studies transfer across these regimes under *one unified* framework, with to date the broadest coverage, including ten HAR datasets, three FoG datasets, and six disease prediction tasks (Table 1).

Sensing Configurations and Deployment Settings. Wearable motion models are shaped not only by architecture, but also by how signals are collected and represented. Prior work has shown that window length affects recognition accuracy and latency [3], body placement and displacement can cause large performance shifts [5, 15], and device heterogeneity leads to variation in sensing behavior, hardware, and operating-system processing [46]. Sampling rate, axis representation, and sensor modality also determine how much temporal and spatial motion structure is available [21, 59]. These factors are usually studied isolated at a time, often in supervised HAR settings and outside modern large-scale pretraining. `Inertia-1` revisits these sensing and deployment choices under a unified pretraining setup, testing sampling frequency, window length, triaxial *vs.* reduced-axis input, time- *vs.* frequency-domain representation, sensor modality, placement, model size, and data scale (Table 1).

Self-Supervised Learning and Motion Foundation Models. Self-supervised learning (SSL) is a natural fit for wearable motion because unlabeled accelerometry is abundant, while dense labels are costly and task-specific. Early wearable SSL studies explored pretext and contrastive objectives for HAR [19, 47], while recent methods adapt masked reconstruction, contrastive, temporal prediction, self-distillation, and frequency-domain modeling to inertial signals [18, 25, 60, 62]. A parallel line scales pretraining to large cohorts such as UK Biobank [17], showing that large accelerometer corpora can improve activity recognition and health prediction [17, 33, 57, 62]. Related biosignal foundation models (FMs) further show the promise of large-scale physiological pretraining [1, 24, 51–53]. However, existing motion FM studies often focus on a single objective, input format, corpus, or task

setting. In contrast, Inertia-1 compares representative objectives under controlled conditions and studies how they transfer across activity recognition, gait analysis, and disease prediction, providing practical guidance for building wearable motion FMs that are robust, generalizable, and useful across datasets, tasks, input modalities, and deployment settings.

3 *Inertia*-1

The goal of Inertia-1 is to turn the fragmented landscape of wearable motion modeling into a unified space of controlled exploration. Rather than searching for one best architecture, we study motion FMs a joint problem over data, sensing configurations, pretraining algorithms, and downstream tasks. This is essential for motion sensing: the same behavior can appear differently across placements, sampling rates, sensor modalities, and temporal windows, while the same pretrained representation should transfer from short-term activity recognition to long-term disease prediction.

Inertia-1 is designed as an open, extensible framework for studying these interactions. It aggregates **15** datasets across **3** task families and **18** downstream tasks, pretrains a broad set of representative objectives, and evaluates models across a controlled grid of sensing and scaling choices. Together, these components allow us to ask not only which model performs best, but which design choices make wearable motion FMs robust, transferable, and useful across real-world conditions. Fig. 2 and Table 2 summarize the breadth of the explored space.

3.1 Data Regimes and Tasks

A central design goal of Inertia-1 is to bridge motion sensing regimes that have been studied separately: large cohorts provide scale, longitudinal coverage, and health outcomes, while targeted motion datasets provide high-frequency signals and fine-grained labels. We use NHANES [33] as the primary pretraining source because it connects these regimes, providing raw high-frequency accelerometer data from 14,000 participants over multiple days, together with rich health and survey variables. We further incorporate the UK Biobank [17] accelerometer data (more than 100,000 participants) in scaling studies to examine how massive lower-resolution cohort data complement high-resolution waveform pretraining. Our downstream evaluation spans three task families, as summarized in Table 2. More datasets details and experimental setups are provided in Appendix A.

Table 2: **Overview of Inertia-1 datasets.** Detailed statistics and setups are in Appendix A and Table 8.

Dataset	Task	Sensor	Placement
CAPTURE-24 [11]	HAR	ACC	WR
HAR70+ [27]	HAR	ACC	CH LG
HARTH [26]	HAR	ACC	CH LG
HHAR [6]	HAR	ACC GYR	WR
MHEALTH [4]	HAR	ACC GYR MAG	AR CH AN
OPPORTUNITY [41]	HAR	ACC GYR MAG	WR HN AR BK HP KN
PAMAP2 [38]	HAR	ACC GYR MAG	HN CH AN
RecoFit [31]	HAR	ACC GYR	AR
WISDM [54]	HAR	ACC GYR	WR
WEAR [7]	HAR	ACC	AR LG
Daphnet FoG [42]	FoG	ACC	HP LG AN
OdayFoG [36]	FoG	ACC GYR	WR CH HP LG AN
FoGTurning [39]	FoG	ACC GYR	SH
NHANES [33]	Disease	ACC	WR
UK Biobank [17]	Disease	ACC	WR

- **Human activity recognition (*short-term*):** We include **10** HAR datasets spanning daily routines, free-living behavior, laboratory activities, and fine-grained exercise movements across diverse environments and sensor setups. These tasks evaluate whether pretrained representations capture short-term motion semantics such as walking, sitting, running, and household activity.
- **Freezing-of-gait detection (*transient*):** We select **3** FoG datasets targeting brief, clinically meaningful disruptions in locomotion. These tasks test whether motion representations pretrained on broad wearable data can detect subtle, transient abnormalities beyond standard activity categories.
- **Disease prediction (*long-term*):** We include **7** disease-related tasks from NHANES and UK Biobank, covering population-level health modeling from longitudinal motion patterns. Unlike HAR and FoG, these tasks are not tied to isolated short-window actions, but to broader behavioral signatures.

3.2 Pretraining Algorithms

To study how different learning principles interact with wearable motion data, Inertia-1 compares representative pretraining algorithms under shared data, sensing, and evaluation settings. We include both (1) *general SSL* objectives, and (2) *wearable-specific SSL* methods, and compare them with *supervised baselines* trained from scratch on each downstream task. In total, we cover **10** (pre)training approaches that represent the state-of-the-art in wearable motion FM. This design lets us separate

whether pretraining helps from which form of pretraining is most useful for transferable motion representations. Additional details for each model are provided in Appendix B.

- **Autoregressive prediction.** GRU-based (AR_G [14]) and Transformer-based (AR_T [48]) models learn representations by predicting future motion from past context, testing whether sequential forecasting objectives capture transferable behavioral dynamics.
- **Contrastive learning.** SimCLR [12], SSL-Wearables [62], and RelCon [56] learn by aligning related motion views while separating unrelated ones. They cover both general contrastive learning and wearable-specific objectives designed around motion augmentations and temporal structure.
- **Self-distillation.** DINO [8] learns stable representations through teacher-student consistency training, testing whether non-contrastive representation learning transfers to wearable motion.
- **Masked reconstruction.** PatchTST [35], Self-PAB [25], and LSM [32] learn by recovering missing signal segments, covering both time-domain and time-frequency reconstruction-style objectives.

3.3 Axes of Exploration

Disentangling Sensing Configurations. To understand how motion FMs interact with with the physical constraints of wearable data, we evaluate selected pretraining objectives across four core input axes below:

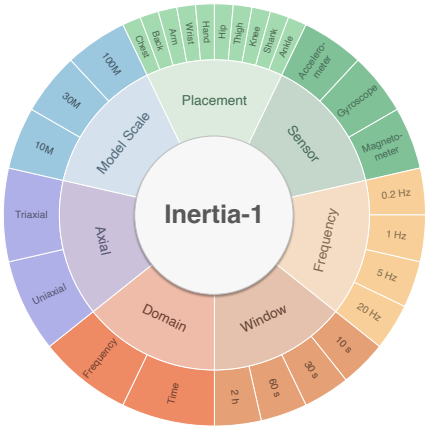


Figure 2: **Coverage breadth of `Inertia-1`.** Full settings are included in Table 7.

- *Uniaxial vs. Triaxial Input:* Large population cohorts often release vector-magnitude summaries of acceleration to reduce data footprint and privacy risk [17, 37], whereas targeted motion datasets usually provide full triaxial signals [10, 41]. We compare these settings to quantify information loss from axis aggregation.
- *Window Length:* Prior work often favors short windows to isolate atomic movements and balance accuracy with latency [3]. $\{10s, 30s, 60s, 2h\}$ windows are used to test how temporal context affects diverse tasks across HAR, FoG, and disease prediction.
- *Sampling Frequency:* Motion datasets vary widely in sampling rate, from high-frequency segments to low-resolution monitoring. We downsample raw accelerometer to $\{20Hz, 5Hz, 1Hz, 0.2Hz\}$ to test when temporal resolution is essential and whether pretraining can offset downsampling [59].
- *Representation Domain:* Time-domain models are widely used for modern wearable SSL [47, 62], while frequency-domain methods may improve invariance to noise and phase shifts [25, 60]. We compare them under matched settings to answer which domain transfers better across tasks.

Spatial and Modality Robustness. Large-scale pretraining datasets are often dominated by wrist-worn accelerometers, whereas real-world and clinical studies involve more diverse placements and sensor suites. We therefore evaluate whether representations learned from common consumer-style sensing transfer to heterogeneous body locations and additional inertial modalities.

- *Body Placements:* Prior work often treats placement shift as a domain adaptation problem [2, 9, 43]. In contrast, we study direct transfer to other placements after large-scale pretraining on wrist-worn accelerometers, and test whether multi-placement fusion improves over single-placement ones.
- *Sensor Modalities:* Most cohort-scale data are accelerometer-based, whereas many high-fidelity datasets also include gyroscopes and magnetometers. We test whether accelerometer-pretrained models transfer to these modalities and whether additional sensors provide complementary benefits.

Model and Data Scaling. Finally, we examine whether scale can help resolve fragmented design choices in wearable motion modeling. We vary model capacity from small (∼5M) to medium (∼30M) and large (∼100M), and scale pretraining data volume using `NHANES` [33] and `UK Biobank` [17].

Table 3: **Comparisons of SSL models and supervised baselines across all HAR datasets.** We report the linear probing AUROC of each model. AUPRC and full finetuning results are in Table 15 & 18.

Model	HHAR	MHEALTH	OPPORTUNITY	PAMAP2	RecoFit	CAPTURE-24	HAR70+	HARTH	WEAR	WISDM
Supervised Models										
ViT	83.9	93.8	72.9	92.0	97.7	92.9	96.6	98.7	97.7	94.1
CNN	89.1	90.7	69.5	95.5	97.0	90.9	94.5	98.7	98.0	94.8
General Pretraining Methods										
PatchTST [35]	93.5	**98.2**	89.8	**97.4**	98.5	91.6	95.8	98.9	99.2	95.5
AR_T [48]	88.6	96.0	**92.1**	95.6	**98.9**	**92.3**	95.5	**99.0**	**99.3**	94.8
AR_G [14]	89.2	96.0	89.4	95.0	98.7	91.3	95.3	98.5	99.2	94.7
DINO [8]	73.4	87.2	87.8	82.7	98.1	92.0	95.4	98.9	98.5	85.2
SimCLR [12]	68.6	90.0	90.0	86.8	97.8	92.2	95.9	98.6	98.0	86.4
Domain-specific Methods										
SSL-Wearables [62]	90.0	89.9	73.1	91.8	91.0	87.0	85.4	92.0	88.5	90.2
RelCon [56]	87.8	96.3	85.3	94.5	97.0	89.9	**95.9**	98.0	97.9	90.5
Self-PAB [25]	**96.8**	91.8	74.7	95.4	97.4	89.7	93.4	98.6	97.7	**96.7**

Table 4: **Comparison of SSL models and supervised baselines across all FoG tasks.** We report the linear probing AUROC of each model. AUPRC and full finetuning results are in Table 16 & 19.

Model	FoGTurning	OdayFoG	DaphnetFoG
Supervised Models			
ViT	63.6	68.0	83.5
CNN	51.2	80.5	87.6
General Pretraining Methods			
PatchTST [35]	85.7	65.4	**95.0**
AR_T [48]	80.2	58.6	94.3
AR_G [14]	83.0	71.6	93.4
DINO [8]	61.4	**83.2**	89.2
SimCLR [12]	91.4	66.7	94.8
Domain-specific Methods			
SSL-Wearables [62]	**91.6**	62.0	71.4
RelCon [56]	87.5	78.2	93.0
Self-PAB [25]	80.4	55.0	92.5

Table 5: **Comparison of SSL models and supervised baselines across disease prediction tasks.** We report the AUROC of MIL on frozen backbone results for each model. Full results are in Table 17.

Model	Depr. Sev.	Depr.	Diabetes	PD	Sleep
Supervised Models					
ViT	49.5	57.9	49.4	45.4	38.3
CNN	51.7	59.5	61.9	53.7	54.0
General Pretraining Methods					
PatchTST [35]	55.0	77.0	53.8	72.8	59.6
AR_T [48]	58.4	**81.3**	67.7	**75.0**	61.7
AR_G [14]	60.6	76.2	53.2	72.4	61.1
DINO [8]	62.9	75.8	74.4	54.5	76.5
SimCLR [12]	58.0	70.7	**78.8**	51.3	**77.2**
Domain-specific Methods					
SSL-Wearables [62]	**65.8**	60.5	78.2	74.4	65.3
RelCon [56]	57.8	72.6	70.5	51.0	56.8
Self-PAB [25]	52.0	65.3	59.6	60.3	58.5

4 Results and Analyses

4.1 Identifying Robust Pretraining Objectives

We first compare our suite of 10 model objectives against supervised baselines to identify which methods remain robust across the full task spectrum. Under the default setting of 20Hz sampling, 30s windows, and triaxial inputs, all models are pretrained on NHANES. We evaluate transfer through both full finetuning and linear probing across all HAR and FoG tasks. For disease prediction, we use frozen pretrained backbones with a specialized multi-instance learning (MIL) head to produce daily-level predictions [45]. Supervised baselines are trained from scratch on each downstream task. Additional training and evaluation details are provided in Appendix A & B. Task-family aggregate comparisons, additional LIMU-BERT and LSM-2 baselines, and uncertainty analyses are provided in Appendices D.1, D.2, and C.

Large-scale pretraining improves transfer across task families. As shown in Table 3, 4, and 5, self-supervised pretraining consistently outperforms supervised training from scratch across task categories. This suggests strong representations learned from unlabeled population cohorts *regardless of the specific downstream application*. While no single objective dominates all settings, specialized or predictive objectives are generally more robust for short-term recognition tasks than generic augmentation-based baselines, while these differences are less pronounced in disease prediction. Overall, objective design remains impor-

Table 6: **Task variation in model performance.** *SSL variance* measures the task-level average gap between the best and median SSL AUROC. *SSL gain* measures the average gap between the best SSL AUROC and the best supervised AUROC.

Task Type	SSL Variance	SSL Gains
HAR	**2.5**	4.0
FoG	11.0	**16.4**
Disease Prediction	10.0	**19.7**

tant, but its benefit depends on the temporal structure and semantic granularity of the downstream task rather than a universally optimal pretraining strategy.

Task variation widens SSL variances and gains. While differences between SSL objectives are modest for standard HAR, they grow significantly for other specialized tasks. Table 6 measures this using ***SSL variance***, defined as the gap between the best and median SSL AUROC. This value is only 2.5 for HAR, but increases to 11.0 for FoG and 10.0 for disease prediction. ***SSL gains*** over supervised training follow the same pattern, rising from 4.0 on HAR to 16.4 on FoG and 19.7 on disease prediction. These results suggest that basic HAR is closer to saturation, whereas clinical and disease-oriented tasks still benefit substantially from representations that capture subtle motion and physiological signatures.

4.2 When Does Scaling Help Motion Foundation Models?

Model capacity saturates under fixed data. We first ask whether increasing encoder size alone improves motion representations. Under the default setting, we scale AR_T from 10M to 30M and 100M parameters while keeping the pretraining corpus and input configuration fixed. As shown in Fig. 3, linear-probe performance largely plateaus across model sizes and task families, suggesting that additional capacity does not translate into better representations without additional data or stronger task alignment. Under full finetuning, larger models can even degrade performance, likely due to overfitting on smaller downstream datasets (Appendix E). These results suggest that motion FMs are less limited by parameter count than by the diversity and relevance of the pretraining signal.

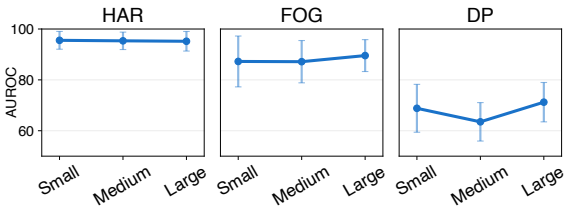


Figure 3: **Model scaling across tasks.** We vary the model size and report linear probe performance across all tasks.

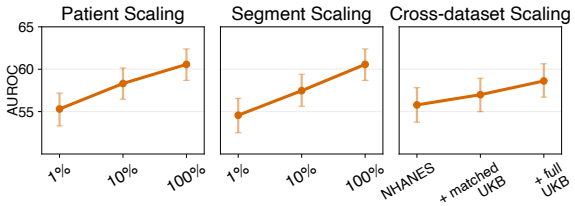


Figure 4: **Data scaling across settings.** We compare `UK Biobank` disease-prediction performance when scaling **(left)** the number of individuals, **(middle)** the number of segments per individual, and **(right)** cross-dataset data mixtures.

Data scale improves transfer through both population diversity and behavioral coverage. We next test whether the scaling bottleneck can be reduced by expanding the pretraining data, using `UK Biobank` as a large-cohort testbed and evaluating frozen representations on patient-disjoint disease prediction tasks. As demonstrated in Fig. 4 & 11, increasing both (1) the *number of individuals*, and (2) the *number of segments* per individual improves AUROC, suggesting that wearable disease representations benefit from population diversity as well as richer behavioral coverage within each person. We also observe gains when mixing `UK Biobank` with `NHANES` during pretraining, indicating that data diversity can improve transfer even when the original corpus is already large.

Scaling is data-first, but not model-free. Together, these experiments suggest that data volume and diversity provide a more reliable path to improved transfer than increasing model size alone. This does not mean model capacity is unimportant; rather, larger encoders may require richer pretraining data, better objective design, or more careful tuning before their capacity yields consistent downstream gains. In this sense, scaling wearable motion FMs is not only a question of bigger networks, but of matching model capacity to signal diversity, temporal resolution, and downstream task structure. We provide more results and experimental details in Appendix E.

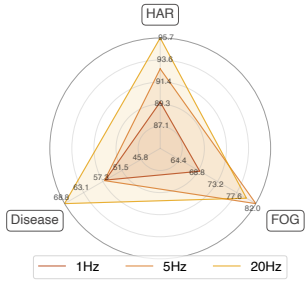
4.3 A Unified Lens over Sensing Design Space

To isolate the effect of core sensing configurations, we conduct controlled experiments using AR_T and PatchTST, two robust methods from the objective comparison. We use 30s windows, 20Hz sampling, and triaxial accelerometer inputs as the default setting, then vary one axis at a time, including sampling rate, window length, axis dimensionality, and representation domain.

For each setting, we pretrain a new model on the same corpus and fine-tune it across downstream tasks following the protocol in Sec. 4.1. We report AR_T results in Fig. 5 and Fig. 6; PatchTST results and additional details are provided in Appendix F.

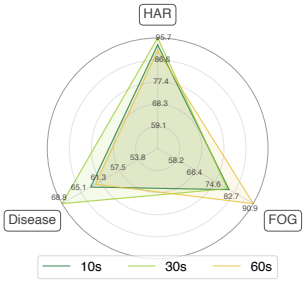
Pretraining improves robustness to low-frequency sampling. As shown in Fig. 5, performance generally improves with sampling frequency, but the degradation at lower rates is smaller than expected for HAR. Unlike prior studies that report sharp drops below 10Hz [59], pretrained representations remain competitive even at 1Hz, suggesting that large-scale pretraining captures contextual motion patterns that partly compensate for lost high-frequency detail. This robustness is weaker for disease prediction, where higher sampling rates provide clearer gains, indicating that clinical outcomes may depend on subtle biomechanical signatures that are attenuated by aggressive downsampling. Activity-stratified analyses also show greater downsampling sensitivity for dynamic activities (Appendices F.3 and C).

Window length is task-dependent. Window size has a more mixed effect. Shorter windows preserve discrete movement semantics and reduce label mixing, while longer windows provide broader temporal context. As shown in Fig. 5, 30s and 60s windows perform strongly overall, but no single duration is optimal across HAR, FoG, and Disease Prediction. This suggests that temporal context should be matched to task granularity: short-window activity labels favor localized motion structure, whereas gait and health tasks may benefit from longer behavioral context.



Freq.	HAR	FOG	DP	Overall
1Hz	89.5	69.1	56.5	71.7
5Hz	92.7	**82.0**	54.4	76.4
20Hz	**95.7**	79.8	**68.8**	**81.5**

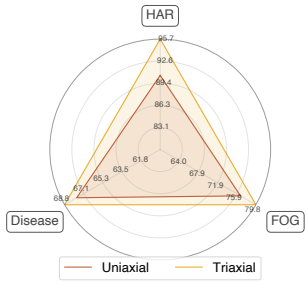
Sampling frequency



Window	HAR	FoG	DP	Overall
10s	92.9	80.6	63.2	78.9
30s	**95.7**	79.8	**68.8**	**81.5**
60s	90.5	**90.9**	62.2	81.2

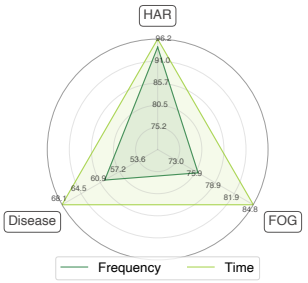
Window size

Figure 5: **Control studies: sampling frequencies and window sizes. (Left)** Full-finetuning AUROC across sampling rates. **(Right)** Full-finetuning AUROC across window sizes. Lower sampling rates remain competitive for HAR, while disease prediction benefits from higher temporal resolution; window length effects are task-dependent.



Axes	HAR	FoG	DP	Overall
Uniaxial	90.6	76.7	67.7	78.3
Triaxial	**95.7**	**79.8**	**68.8**	**81.5**

Axes dimensionalities



Input	HAR	FoG	DP	Overall
Freq.	94.3	76.3	60.0	76.9
Time	**96.2**	**84.8**	**68.1**	**83.0**

Representation domain

Figure 6: **Control studies: axis dimensionality and representation domain. (Left)** Full-finetuning AUROC for uniaxial versus triaxial inputs. **(Right)** Full-finetuning AUROC for frequency-domain versus time-domain representations. Triaxial inputs preserve stronger transferable motion structure, while time-domain modeling performs better for FoG and disease prediction.

Triaxial inputs preserve transferable motion structure. Following common cohort preprocessing practice, we compare full triaxial acceleration with vector magnitude, which collapses three axes into one channel. As shown in Fig. 6, triaxial inputs consistently outperform uniaxial inputs across task families. This indicates that directional and orientation-dependent motion patterns carry transferable information that cannot be fully recovered after magnitude aggregation. Interestingly, AR_T is relatively more resilient to this loss than PatchTST (Fig. 12c), suggesting that architecture can influence robustness to compressed input representations, but preserving raw axes remains the stronger default choice.

Time-domain modeling better preserves gait and disease signals. We compare frequency-domain reconstruction with direct temporal modeling by evaluating Self-PAB and PatchTST under matched settings. As shown in Fig. 6, the representation domain has limited impact on HAR, but time-domain modeling performs better for FoG and disease prediction. This suggests that gait events and disease-relevant movement signatures depend on precise temporal structure that may be weakened

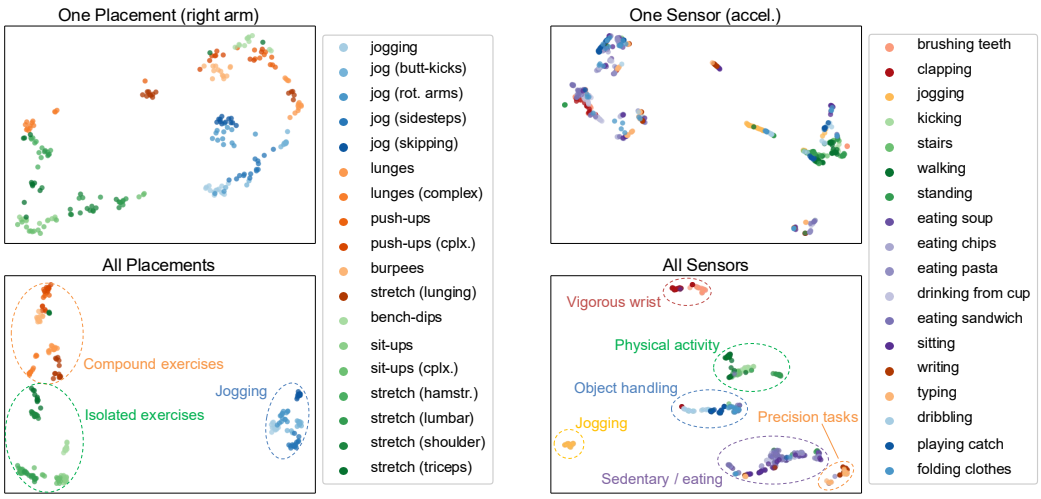


Figure 7: **Multi-stream fusion improves representation geometry. Left:** `WEAR` embeddings from a single right-arm placement *vs.* fused multi-placement. **Right:** `WISDM` embeddings from a single accelerometer stream *vs.* fused multi-sensor. In both datasets, aggregating multiple streams produces more separated activity clusters.

by spectrogram-style transformations. Episode-duration analyses further show a larger performance deficit for Self-PAB on brief FoG events, consistent with sensitivity to temporal aggregation (Appendix F.4). Overall, these results show that sensing choices are not secondary implementation details; they directly shape what information motion foundation models can transfer.

4.4 Multi-Stream Sensing Reveals Complementary Motion Structure

Multiple synchronized streams add complementary motion views. Wearable datasets often contain richer signals than a single wrist-worn accelerometer stream, including multiple sensor modalities and body placements. We therefore ask whether representations improve when additional streams are fused, and whether a model pretrained only on wrist accelerometry can transfer to these different physical views of motion.

To separate the effects of pretraining and stream fusion, we instantiate four settings: ❶ supervised single-stream, ❷ pretrained single-stream, ❸ supervised multi-stream fusion, and ❹ pretrained multi-stream fusion. For the single-stream setting, we use the wrist or accelerometer stream when available, falling back to the most common placement or modality otherwise. For multi-stream fusion, we temporally align windows across all available streams, process each stream with a separate encoder, and concatenate the resulting embeddings for downstream classification. Supervised models are trained from scratch, while pretrained models are initialized from the same pretrained wrist-accelerometer checkpoint.

As shown in Fig. 8, adding streams consistently improves performance for both sensor-modality fusion and body-placement fusion. This shows that different sensors and placements provide complementary task-relevant information rather than redundant copies of the same signal. More importantly, the wrist-accelerometer pretrained model improves performance across both single-stream and multi-stream settings, even when evaluated on unseen placements or sensor modalities. This suggests that large-scale wrist pretraining learns a general representation of human motion, rather than a representation tied only to one device location or hardware channel.

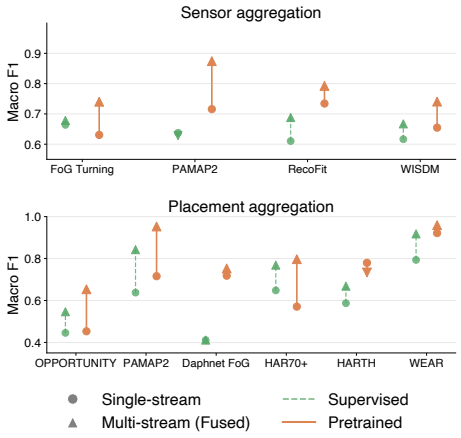


Figure 8: **Wrist-accelerometer pretraining transfers across sensors and placements.** We report macro-F1 for default single-stream inputs and fused multi-stream inputs.

Fusion improves representation geometry. We further examine whether multi-stream fusion changes the structure of the learned representation space. As shown in Fig. 7, embeddings from fused streams form better-separated activity clusters than embeddings from the default single stream. This pattern is visible for both *placement fusion* in `WEAR` and *sensor fusion* in `WISDM`, where multi-stream embeddings separate compound exercises, isolated exercises, jogging, object handling, physical activity, and sedentary behaviors more clearly. Quantitative cluster statistics in Table 24 further support this improved representation geometry. Together, these results suggest that multi-stream sensing provides complementary views of movement, and that pretraining helps align these views into a more separable and transferable motion representation.

5 Discussion

Limitations. While `Inertia-1` provides a controlled exploration of wearable motion FMs, several limitations remain. First, not all pretraining objectives may respond similarly to each design axis. Our detailed ablations focus on top-performing representative methods, but contrastive, self-distillation, reconstructive, and predictive objectives may each have different sensitivities to sampling rate, window length, placement, modality, and scale. A finer objective-specific sweep may reveal additional design patterns. Second, our downstream evaluations are classification-based, whereas real-world wearable applications also include continuous health prediction, biomarker regression, longitudinal trajectory modeling, and event forecasting. Extending `Inertia-1` to these settings is an important next step for understanding how motion FMs support broader health and behavior modeling.

Conclusion. We present `Inertia-1`, a fully open exploration of wearable motion FMs that addresses fragmentation across the full lifecycle of motion representation learning. Across more than 1,000 models and evaluations on 18.2M hours of accelerometry data from 115,000+ individuals, we find that self-supervised pretraining consistently improves transfer over supervised training from scratch, but no single objective dominates across all task families. Our analyses further show that data representation, temporal resolution, placement, and pretraining scale can be as important as model architecture itself. Collectively, `Inertia-1` provides state-of-the-art recipes for diverse downstream tasks and serves as a comprehensive, practical, and open cookbook for motion representation learning.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge the support by the Amazon Science Hub, the NVIDIA Academic Grant Program, and UCLA DataX. Sarah Jiang is support by the National Science Foundation Graduate Research Fellowship Program under Grant No.XXX. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

[1] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals, 2024.

[2] Sungtae An, Alessio Medda, Michael N Sawka, Clayton J Hutto, Mindy L Millard-Stafford, Scott Appling, Kristine LS Richardson, and Omer T Inan. Adaptnet: Human activity recognition via bilateral domain adaptation using semi-supervised deep translation networks. *IEEE Sensors Journal*, 21(18):20398–20411, 2021.

[3] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014.

[4] Oresti Banos, Rafael Garcia, and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TW22.

[5] Oresti Banos, M. Akos Toth, Miguel Damas, Hector Pomares, and Ignacio Rojas. Dealing with the effects of sensor displacement in wearable activity recognition. *Sensors*, 14(6):9995–10023, 2014.

[6] Henrik Blunck, Sourav Bhattacharya, Thor Prentow, Mikkel Kjrgaard, and Anind Dey. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

[7] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition, 2024.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[9] Avijoy Chakma, Abu Zaher Md Faridee, Indrajeet Ghosh, and Nirmalya Roy. Domain adaptation for inertial measurement unit-based human activity recognition: A survey. *arXiv preprint arXiv:2304.06489*, 2023.

[10] Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data*, 11(1):1135, 2024.

[11] Chan Chang, S Walmsley, R Gershuny, J Harms, T Thomas, E Milton, K Kelly, P Foster, C Wong, A Gray, N Haque, S Hollowell, and S Doherty. Capture-24: Activity tracker dataset for human activity recognition, 2021.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[13] Zhengming Chen, Junshi Chen, Rory Collins, Yu Guo, Richard Peto, Fan Wu, and Liming Li. China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International journal of epidemiology*, 40(6):1652–1666, 2011.

[14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[15] Anis Davoudi, Mamoun T Mardini, David Nelson, Fahd Albinali, Sanjay Ranka, Parisa Rashidi, and Todd M Manini. The effect of sensor placement and number on physical activity recognition and energy expenditure estimation in older adults: Validation study. *JMIR Mhealth Uhealth*, 9(5):e23681, May 2021.

[16] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access*, 8:210816–210836, 2020.

[17] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, 12(2):e0169649, 2017.

[18] 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*, ISWC '20, page 45–49, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Harish Haresamudram, Irfan Essa, and Thomas Plotz. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3), 2022.

[20] Tom Hoddes, Alex Bijamov, Saket Joshi, Daniel Roggen, Ali Etemad, Robert Harle, and David Racz. Scaling laws in wearable human activity recognition. *arXiv preprint arXiv:2502.03364*, 2025.

[21] Emily J. Huang, Kebin Yan, and Jukka-Pekka Onnela. Smartphone-based activity recognition using multistream movelets combining accelerometer and gyroscope data. *Sensors*, 22(7):2618, 2022.

[22] Tinghuai Huang, Meng Li, and Jianwei Huang. Recent trends in wearable device used to detect freezing of gait and falls in people with parkinson's disease: A systematic review. *Frontiers in Aging Neuroscience*, Volume 15 - 2023, 2023.

[23] Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[24] Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

[25] Aleksej Logacjov, Sverre Herland, Astrid Ustad, and Kerstin Bach. Selfpab: Large-scale pre-training on accelerometer data for human activity recognition. *Applied Intelligence*, 54(6):4545–4563, 2024.

[26] Aleksej Logacjov, Atle Kongsvold, Kerstin Bach, Hilde Bremseth Bårdstu, and Paul Jarle Mork. HARTH. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC90.

[27] Aleksej Logacjov and Astrid Ustad. HAR70+. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5CW3D.

[28] Hiral Master, Jeffrey Annis, Shi Huang, Joshua A. Beckman, Francis Ratsimbazafy, Kayla Marginean, Robert Carroll, Karthik Natarajan, Frank E. Harrell, Dan M. Roden, Paul Harris, and Evan L. Brittain. Association of step counts over time with the risk of chronic disease in the all of us research program. *Nature Medicine*, 28(11):2301–2308, November 2022.

[29] Ahmed A Metwally, A Ali Heydari, Daniel McDuff, Alexandru Solot, Zeinab Esmaeilpour, Anthony Z Faranesh, Menglian Zhou, Girish Narayanswamy, Maxwell A Xu, Xin Liu, et al. Insulin resistance prediction from wearables and routine blood biomarkers. *Nature*, pages 1–11, 2026.

[30] Shenghuan Miao, Ling Chen, and Rong Hu. Spatial-temporal masked autoencoder for multi-device wearable human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–25, 2024.

[31] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. Recofit. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 3225–3234, Apr 2014.

[32] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[33] National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES) 2011-2014 Physical Activity Monitor Data. Centers for Disease Control and Prevention, 2014.

[34] Camille Nebeker, Shahin Samiei, and Santosh Kumar. Sharing digital health data responsibly: Balancing open science with participant privacy. *PLOS Digital Health*, 5(4):e0001377, 2026.

[35] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[36] Johanna O'Day, Marissa Lee, Kirsten Seagers, Shannon Hoffman, Ava Jih-Schiff, Łukasz Kidziński, Scott Delp, and Helen Bronte-Stewart. Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of neuroengineering and rehabilitation*, 19(1):20, 2022.

[37] Theresa Patten, Edward A Preble, Hiral Master, Jennifer Adjemian, Andrea Ramirez, James McClain, and Amy Rose Price. The all of us research program's wearables dataset. *Nature Medicine*, pages 1–9, 2026.

[38] Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

[39] Caroline Ribeiro De Souza, Runfeng Miao, Júlia Ávila De Oliveira, Andrea Cristina De Lima-Pardini, Débora Fragoso De Campos, Carla Silva-Batista, Luis Teixeira, Solaiman Shokur, Bouri Mohamed, and Daniel Boari Coelho. A public data set of videos, inertial measurement unit, and clinical scales of freezing of gait in individuals with parkinson's disease during a turning-in-place task. *Frontiers in Neuroscience*, Volume 16 - 2022, 2022.

[40] Daniel Rodriguez-Martin, Albert Sama, Celia Perez-Lopez, Andreu Catala, Juan M. Moreno Arostegui, Joan Cabestany, and Alejandro Rodriguez-Molinero. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PLOS ONE*, 12(2):e0171764, 2017.

[41] Daniel Roggen, Alberto Calatroni, Long-Van Nguyen-Dinh, Ricardo Chavarriaga, and Hesam Sagha. OPPORTUNITY Activity Recognition. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5M027.

[42] Daniel Roggen, Meir Plotnik, and Jeff Hausdorff. Daphnet Freezing of Gait. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C56K78.

[43] Andrea Rosales Sanabria, Franco Zambonelli, and Juan Ye. Unsupervised domain adaptation in activity recognition: A gan-based approach. *IEEE Access*, 9:19421–19438, 2021.

[44] Jinjoo Shim, Elgar Fleisch, and Filipe Barata. Wearable-based accelerometer activity profile as digital biomarker of inflammation, biological age, and mortality using hierarchical clustering analysis in nhanes 2011–2014. *Scientific Reports*, 13(1):9326, Jun 2023.

[45] Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On pre-training and scaling of sleep foundation models. *arXiv preprint arXiv:2603.00190*, 2026.

[46] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjaergaard, Anind K. Dey, Tobias Sonne, and Mads Moller Jensen. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 127–140, 2015.

[47] 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. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–30, 2021.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[49] Hoang-Dieu Vu, Duc-Nghia Tran, Quang-Tu Pham, Hieu H Pham, Nicolas Vuillerme, and Duc-Tan Tran. Smooth-distill: A self-distillation framework for multitask learning with wearable sensor data. *arXiv preprint arXiv:2507.00061*, 2025.

[50] Xiaoda Wang, Ching Chang, Defu Cao, Kaiqiao Han, Fang Sun, Yue Huang, Minxiao Wang, Chang Xu, Xiao Luo, Runze Yan, Xiangliang Zhang, Xiao Hu, Yan Liu, Yizhou Sun, Wei Wang, and Carl Yang. Position: Beyond prediction: Toward verifiable physiological waveform reasoning with foundation models and agentic llms. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[51] Xiaoda Wang, Kaiqiao Han, Yuhao Xu, Xiao Luo, Yizhou Sun, Wei Wang, and Carl Yang. Se-diff: Simulator and experience enhanced diffusion model for comprehensive ecg generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[52] Xiaoda Wang, Minxiao Wang, Kaiqiao Han, Defu Cao, Ching Chang, Yidan Shi, Runze Yan, Xiao Luo, Yan Liu, Xiao Hu, et al. Pg-lrf: Physiology-guided latent rectified flow for electro-hemodynamic ppg-to-ecg generation. *arXiv preprint arXiv:2605.12541*, 2026.

[53] Xiaoda Wang, Minxiao Wang, Maxwell A Xu, Patrick Langer, Kaiqiao Han, Defu Cao, Xiao Luo, Yuzhe Yang, Yan Liu, Xiao Hu, et al. Ppg-lm: A photoplethysmography-language model with multi-level clinical alignment. *arXiv preprint arXiv:2609.33516*, 2026.

[54] Gary Weiss. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset . UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5HK59.

[55] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233, 2021.

[56] Maxwell A Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J Raghuram, James M Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. *arXiv preprint arXiv:2411.18822*, 2025.

[57] Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. Lsm-2: Learning from incomplete wearable sensor data, 2025.

[58] Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S Aysola, Rajesh Kumar, and Yuzhe Yang. SleepLM: Natural-language intelligence for human sleep. *arXiv preprint arXiv:2602.23605*, 2026.

[59] Takahiro Yamane, Moeka Kimura, and Mizuki Morita. Effects of sampling frequency on human activity recognition with machine learning aiming at clinical applications. *Sensors*, 25(12):3780, 2025.

[60] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations (ICLR)*, 2023.

[61] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G. Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R. Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature Medicine*, 28(10):2207–2215, 2022.

[62] 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.

[63] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *Neural Information Processing Systems (NeurIPS)*, 2025.

A Experimental Scope, Data Regimes, and Evaluation Setup

A.1 Experimental axes

To make the design space explicit, `Inertia-1` evaluates wearable motion foundation models across the major choices that can be made in wearables pipelines: self-supervised objective, model capacity, sampling frequency, window length, sensor axis dimensionality, sensor modality, body placement, and input representation domain. Table 7 summarizes the resulting grid of experimental axes used throughout the paper. The list of Device Placements in order of badges are: wrist, hand, arm, back, chest, hip, knee, leg, shank, ankle.

Table 7: **Complete summary of `Inertia-1` breadth and coverage.**

Component	Coverage
Sensors	ACC; GYR; MAG
Device Placements	WR; HN; AR; BK; CH; HP; KN; LG; SH; AN
Sampling Frequency	20 Hz; 5 Hz; 1 Hz; 0.2 Hz
Window Length	10 s; 30 s; 60 s; 2 h
Representation Domains	Time; Time–Frequency
Model Families	GRU; Transformer; ViT (time); ViT (spectrogram); wave2vec CNN; wave2vec Transformer
Pretraining Objectives	***Autoregressive***: AR_G, AR_T ***Contrastive***: RelCon, SSL-Wearables, SimCLR ***Self-Distillation***: DINO ***Masked Reconstruction***: PatchTST, Self-PAB, LSM
Model Scale	Small ($\sim$5M); Medium ($\sim$30M); Large ($\sim$100M)
Downstream Tasks	HAR; FoG; Disease Prediction
Evaluation Protocol	Linear Probing; Full Fine-tuning

A.2 Pretraining data

Our pretraining corpus is mainly derived from the National Health and Nutrition Examination Survey (`NHANES`) accelerometry cohort from year 2011-2014. Participants wore a physical activity monitor continuously for seven days. Raw triaxial accelerometer signals were collected at 80 Hz, low-pass filtered using a 4th-order Butterworth filter with a 10 Hz cutoff, and downsampled to 20 Hz. The processed data are stored in ten-minute parquet segments, comprising 14,688 subjects and approximately 2.47 million subject-hours of recordings. This dataset provides large-scale, high-resolution unlabeled motion data suitable for representation learning.

In addition, we utilize the `UK Biobank` accelerometer dataset for data scaling experiments, including cross-dataset mixing with `NHANES`. Approximately 100,000 participants wore a physical activity monitor continuously for seven days. Raw 100 Hz triaxial accelerometer signals were auto-calibrated to local gravity, gravity-corrected, and low-pass filtered using a 4th-order Butterworth filter with a 20 Hz cutoff. The processed Euclidean norm data available to us are stored as pre-computed five-second epoch time-series files, with non-wear periods imputed using time-of-day averages. Filtered for subjects meeting a minimum 72-hour wear-time threshold, this cohort comprises approximately 15.7 million subject-hours of recordings, further expanding the scale and diversity of our unlabeled pretraining corpus.

A.3 Downstream datasets and task families

We evaluate learned representations on thirteen public HAR and FoG benchmarks, spanning diverse sensor placements, sampling frequencies, and activity label sets. In addition, we evaluate on five patient-level Disease Prediction tasks from `NHANES` and two patient-level Disease Prediction tasks derived from `UK Biobank` ICD10 codes, using multiple-instance pooling over windows from each

subject. Train/val/test splits are selected on the subject level for all datasets. A comprehensive summary of downstream datasets is provided in Table 8.

Table 8: **Summary of the datasets used in `Inertia-1`.** Train and test samples are reported in subjects for `NHANES` and `UK Biobank`, 30 second windows for HAR and FoG downstream datasets, and subject-days for the Disease Prediction tasks.

Dataset	Task	# Class	# Train	# Val	# Test	Rate	Sensors	Placement
Pretraining:								
NHANES [33]	Pretraining	–	11.7K	1.5K	1.5K	1-20 Hz	ACC	WR
UK Biobank [17]	Pretraining	–	80K	–	20K	0.2 Hz	ACC	WR
Downstream:								
CAPTURE-24 [11]	HAR	10	214.8K	46.9K	45.8K	30 Hz	ACC	WR
HAR70+ [27]	HAR	5	27.3K	5.6K	6.0K	50 Hz	ACC	CH LG
HARTH [26]	HAR	9	86.9K	18.2K	18.0K	50 Hz	ACC	CH LG
HHAR [6]	HAR	6	10.4K	2.2K	2.2K	30 Hz	ACC GYR	WR
MHEALTH [4]	HAR	12	2.8K	2.0K	2.0K	50 Hz	ACC GYR MAG	AR CH AN
OPPORTUNITY [41]	HAR	4	25.6K	12.2K	9.8K	30 Hz	ACC GYR MAG	WR HN AR BK HP KN
PAMAP2 [38]	HAR	12	16.0K	9.4K	13.5K	100 Hz	ACC GYR MAG	HN CH AN
RecoFit [31]	HAR	22	63.1K	13.6K	13.4K	50 Hz	ACC GYR	AR
WISDM [54]	HAR	18	111.8K	24.6K	25.1K	20 Hz	ACC GYR	WR
WEAR [7]	HAR	18	29.0K	6.7K	6.5K	50 Hz	ACC	AR LG
Daphnet FoG [42]	FoG	2	12.5K	2.6K	2.8K	64 Hz	ACC	HP LG AN
OdayFoG [36]	FoG	2	1.7K	.8K	.8K	128 Hz	ACC GYR	WR CH HP LG AN
FoGTurning [39]	FoG	2	4.5K	1.0K	1.0K	128 Hz	ACC GYR	SH
NHANES [33] (Sleep)	Disease	2	1525	207	207	20 Hz	ACC	WR
NHANES [33] (Depression)	Disease	2	1764	168	221	20 Hz	ACC	WR
NHANES [33] (Dep. Severity)	Disease	5	2548	315	289	20 Hz	ACC	WR
NHANES [33] (Parkinson's)	Disease	2	1267	162	153	20 Hz	ACC	WR
NHANES [33] (Diabetes)	Disease	2	1317	180	151	20 Hz	ACC	WR
UK Biobank [17] (Osteoarthritis)	Disease	2	89.6K	19.2K	19.2K	0.2 Hz	ACC	WR
UK Biobank [17] (Osteoporosis)	Disease	2	27.8K	6K	6K	0.2 Hz	ACC	WR

A.4 Training and evaluation protocol

Unless otherwise specified, models are pretrained for $170k$ optimization steps using AdamW on 4 H200 GPUs. Default controlled experiments use $30s$ windows of 20 Hz triaxial accelerometer data from wrist-worn devices, or the closest available placement when a dataset does not include a wrist-worn IMU. The default encoder capacity is the Small setting, corresponding to approximately 5–8M parameters, with Medium ($\sim$30M) and Large ($\sim$100M) variants used for scaling analyses. Downstream evaluation uses linear probing, where the pretrained encoder is frozen and only a lightweight classifier is trained, and full fine-tuning, where all model parameters are updated end-to-end.

For patient-level Disease Prediction in `NHANES`, we use multiple-instance learning by sampling 1024 windows from each subject at predetermined times of the day and pooling their embeddings into a subject-level prediction. Labels are derived from `NHANES` survey reports and medications data. Parkinson's disease, depression (binary), and diabetes are inferred through use of anti-Parkinson's agents, antidepressants, and insulin, respectively. Additionally, we report the average metric across 5 run seeds for Disease Prediction to reduce performance variance due to low positive sample size. See Appendix E.3 for details on `UK Biobank` Disease Prediction.

B Pretraining Algorithms and Model Implementations

Unless otherwise specified, method implementations use the shared default preprocessing and training configuration: 30-second windows sampled at 20 Hz, triaxial accelerometer input, a batch size of 1024, AdamW optimization with learning rate 2×10^{-4} and weight decay 0.04, and 5 pretraining epochs. All pretraining runs utilized a learning rate scheduler with linear warmup and cosine decay. Downstream evaluation uses the linear-probing and full-finetuning protocols described in Appendix A.4. The probing and finetuning runs use learning rate 1×10^{-3} and weight decay 0.01.

Size-controlled experiments use the corresponding small, medium, and large configuration presets when applicable.

B.1 General self-supervised methods

PatchTST. We implement PatchTST as a masked patch-reconstruction model for accelerometer windows [35]. The input window is divided into temporal patches and a random subset of patches is masked before being passed to the PatchTST encoder. We use patch length 10, stride 10, and mask ratio 0.4. The model predicts the original unmasked patch values, and the pretraining loss is mean-squared error evaluated only on masked patches. For downstream evaluation, we use the mean-pooled PatchTST backbone representation. In model-size experiments, the preset-controlled PatchTST sizes are: small ($d = 512, L = 4, H = 4, d_{ff} = 1024$), medium ($d = 768, L = 6, H = 8, d_{ff} = 2048$), and large ($d = 1024, L = 8, H = 16, d_{ff} = 4096$).

$\mathbf{AR_T}$. The $\mathrm{AR_T}$ is trained with a causal next-patch prediction objective. The input is divided into non-overlapping temporal patches of size 10 by default. The pretraining loss is mean-squared error between the predicted and target next patches. The model uses a linear patch projection, a causal Transformer encoder, and a linear prediction head. For downstream evaluation, the default representation is generated by mean pooling over all final layer positions. In model-size experiments, the preset-controlled $\mathrm{AR_T}$ sizes are: small ($d = 512, L = 4, H = 4, d_{ff} = 1024$), medium ($d = 768, L = 6, H = 8, d_{ff} = 2048$), and large ($d = 1024, L = 8, H = 16, d_{ff} = 4096$).

$\mathbf{AR_G}$. The $\mathrm{AR_G}$ uses the same next-patch prediction objective as $\mathrm{AR_T}$, but replaces the Transformer encoder with a recurrent GRU. The default patch length is 10 samples. Each patch is projected into the GRU hidden space, the GRU processes the patch sequence causally, and a linear head predicts the next patch. The pretraining loss is mean-squared error on next-patch prediction. For downstream evaluation, the default representation is the mean of the GRU hidden states over time.

DINO. We adapt DINO to wearable motion using a student–teacher self-distillation setup [8]. Both student and teacher use the same ViT-style motion encoder, and the teacher parameters are initialized from the student and then updated as an exponential moving average of the student parameters. Under the default configuration, both encoders use a patch size of 10 on the temporal axis for input embedding. The base augmentations are adapted from genertic augmentation strategies such as adding noises and masking inputs. Specifically, we use jitter with standard deviation 0.02, scaling with standard deviation 0.1, time masking with ratio sampled from $[0.3, 0.6]$, and channel dropout over 0 or 1 channels. The DINO loss matches all student views to the teacher outputs from the global views. For downstream evaluation, the implementation uses the raw backbone representation from the student encoder by default and discards the DINO head.

SimCLR. We adapt SimCLR to wearable motion by generating two augmented views of each input window and training a shared encoder with a contrastive objective [12]. The default encoder is the same ViT-style motion encoder used in DINO, with patch size of 10. Training uses the NT-Xent loss with temperature 0.1. We use the same genertic augmentation strategies as DINO by default. For downstream evaluation, the implementation returns the raw encoder representation and discards the SimCLR projection head.

B.2 Domain-specific self-supervised methods

SSL-Wearables. Following the original implementation by [62], we frame the SSL-Wearables pre-training objective as a three-task temporal transformation prediction problem. For each input window, the model constructs views corresponding to arrow-of-time, permutation, and time-warping tasks. We also follow the official codebase to dictate the transformation parameters and loss calculations. For downstream evaluation, we follow standard procedure by extracting features directly from the raw ResNet encoder and discarding the self-supervised prediction heads.

RelCon. Following the original implementation by [56], we employ a two-stage self-supervised approach that replaces fixed distance metrics with a learned reconstructability criterion. In the first stage, a distance model is trained to identify motion-specific motifs across accelerometer windows. In the second stage, this model is frozen and used to rank candidate windows for a relative contrastive

Table 9: **Model sizes in millions of trainable parameters.** Medium and large configurations are reported for best performing SSL methods.

Model	Small (M)	Medium (M)	Large (M)
AR_T	8.4	33.1	100.8
PatchTST	8.5	33.2	100.9
AR_G	3.9	–	–
DINO	8.1	–	–
SimCLR	5.8	–	–
RelCon	3.6	–	–
SSL-Wearables	10.5	–	
Self-PAB	3.3	–	–
Supervised CNN	5.0	–	–
Supervised ViT	5.4	–	–

loss. We strictly adhere to the official codebase to implement the loss formulation and default hyperparameters over a ResNet1D encoder.

Self-PAB. We implement Self-PAB as masked reconstruction in a spectrogram representation [25]. The module computes the short-time Fourier transform internally. The default STFT uses $n_{\text{fft}} = 128$, hop length 64, no centering, and magnitude-only input without phase to adjust for our longer window length. The masking procedure alters both time and frequency regions, and the reconstruction objective is masked L1 loss by default. For downstream evaluation, the model converts the raw window to its STFT representation without masking, applies through the model encoder, and returns the mean-pooled frame representation.

LIMU-BERT. We incorporate LIMU-BERT as an additional self-supervised method developed for inertial sensing. The method is pretrained using its original learning objective under our shared data and training settings, and the resulting encoder is evaluated with the downstream protocol used for the other pretrained models. This comparison tests how its representations transfer beyond a single activity-recognition setting to FoG and disease prediction.

LSM-2. We incorporate LSM-2 as an additional wearable representation-learning method, following its original pretraining formulation and released implementation. As with the other methods, pretraining uses the shared corpus and downstream evaluation uses the same task-specific protocol. This allows us to compare transfer across task families under common sensing and training conditions.

B.3 Supervised baselines

CNN. The CNN baseline implementation uses a ResNet-style 1D convolutional encoder with input channels equal to the selected sensor axes. The method is trained by the downstream evaluation code with the task-specific supervised objective.

ViT. The ViT baseline divides each motion window into temporal patches and processes them with a 1D ViT encoder. The patch size is default to 10. As with the CNN baseline, the method is trained by the downstream evaluation code with the task-specific supervised objective, rather than by a self-supervised pretraining loss.

B.4 Model capacity definitions

We report model capacity using the Small, Medium, and Large settings used in the scaling experiments. Table 9 lists the parameter counts for each backbone.

Table 10: **Uncertainty of central HAR comparisons.** We report AUROC (%) with 95% bootstrap confidence intervals and paired-comparison p-values for AR_T (ART), except where a supervised CNN is specified.

Comparison	Setting A	AUROC (95% CI)	Setting B	AUROC (95% CI)	p
Sampling rate	20 Hz	95.3 (95.2–95.4)	1 Hz	91.3 (91.2–91.4)	< 0.01
	20 Hz	95.3 (95.2–95.4)	5 Hz	93.6 (93.4–93.7)	< 0.01
	5 Hz	93.6 (93.4–93.7)	1 Hz	91.3 (91.2–91.4)	< 0.01
Window length	30 s	95.3 (95.2–95.4)	10 s	92.7 (92.6–92.8)	< 0.01
	30 s	95.3 (95.2–95.4)	60 s	90.3 (90.2–90.4)	< 0.01
Axis representation	Triaxial	95.3 (95.2–95.4)	Magnitude	91.4 (91.3–91.5)	< 0.01
Pretraining	ART	95.2 (95.1–95.3)	CNN	92.0 (91.8–92.1)	< 0.01

Table 11: **Uncertainty of central FoG comparisons.** We report AUROC (%) with 95% bootstrap confidence intervals and paired-comparison p-values. Configuration ablations and the SSL-versus-supervised comparison are reported separately.

Comparison	Setting A	AUROC (95% CI)	Setting B	AUROC (95% CI)	p
Sampling rate	20 Hz	80.0 (78.4–81.6)	1 Hz	68.8 (66.9–70.7)	< 0.01
	20 Hz	80.0 (78.4–81.6)	5 Hz	82.0 (80.5–83.4)	0.0850
	5 Hz	82.0 (80.5–83.4)	1 Hz	68.8 (66.9–70.7)	< 0.01
Window length	30 s	80.0 (78.4–81.6)	10 s	80.5 (78.9–82.0)	0.6800
	30 s	80.0 (78.4–81.6)	60 s	88.1 (86.2–89.9)	< 0.01
Axis representation	Triaxial	80.0 (78.4–81.6)	Magnitude	78.1 (76.4–79.8)	0.0430
Pretraining	ART	77.7 (76.0–79.4)	CNN	73.1 (71.4–74.7)	< 0.01

C Statistical uncertainty of central comparisons

To assess whether the central comparisons are supported beyond their point estimates, we compute bootstrap confidence intervals for AUROC. For the comparisons in Tables 10–12, we use 2,000 bootstrap replicates. Each replicate resamples test rows with replacement and recomputes the evaluation metric. The 95% confidence interval is defined by the 2.5th and 97.5th percentiles of the bootstrap estimates. For paired model comparisons, both models are evaluated using the same resampling indices, and we compute the difference in AUROC within each replicate.

Under the reported analysis, HAR performance improves with sampling frequency, and 30-second windows perform better than the tested 10-second and 60-second alternatives. Window-length effects differ across task families: FoG improves from 30 to 60 seconds, while the disease-prediction comparisons do not provide evidence of a window-length advantage. For FoG, 5 Hz performs numerically above 20 Hz, but the difference is not statistically significant ($p = 0.085$). Triaxial inputs improve HAR and show a nominal advantage for FoG. Together, these results support task-dependent sensing choices rather than one configuration that is optimal across all task families.

D Supplementary Metrics for Main Results

The main paper reports linear probing AUROC as the primary metric because it allows comparison across all task families. Since several downstream datasets are class-imbalanced, we also report both linear probing and full finetuning AUPRC and task-specific breakdowns here (Table 15, Table 16, Table 17). In additon, we also include full finetuning AUROC across HAR (Table 18) and FoG (Table 19) tasks. Finally, we include the raw AUROC and AUPRC across all axes ablations discussed in the main paper (Table 20, Table 21).

D.1 Task-family aggregate comparisons

The main paper reports per-dataset results to preserve task-specific variation. To provide a complementary high-level comparison, Table 13 summarizes mean AUROC within each task family and the average rank of the ten approaches in the core comparison. The disease-prediction summary covers the five `NHANES` targets in the main objective comparison; the additional `UK Biobank` tasks

Table 12: **Uncertainty of central disease-prediction comparisons.** We report AUROC (%) with 95% bootstrap confidence intervals and paired-comparison p-values.

Comparison	Setting A	AUROC (95% CI)	Setting B	AUROC (95% CI)	p
Sampling rate	20 Hz	68.8 (61.6–76.0)	1 Hz	56.5 (51.0–62.0)	0.0117
	20 Hz	68.8 (61.6–76.0)	5 Hz	54.4 (50.3–58.5)	< 0.01
Window length	30 s	68.8 (61.6–76.0)	10 s	63.2 (58.8–67.5)	0.2131
	30 s	68.8 (61.6–76.0)	60 s	62.3 (55.2–69.3)	0.2224
Axis representation	Triaxial	68.8 (61.6–76.0)	Magnitude	67.7 (64.1–71.3)	0.7564
Pretraining	ART	68.8 (61.6–76.0)	CNN	53.8 (50.7–56.8)	< 0.01

Table 13: **Aggregate comparison across task families.** We report mean AUROC (%) within each task family and average rank for the ten approaches in the core comparison. Lower average rank is better. Disease prediction includes the five `NHANES` targets.

Model	HAR	FoG	Disease Pred.	Avg. Rank
ViT	92.03	71.70	48.10	7.64
CNN	91.87	73.10	56.16	6.96
PatchTST	95.84	82.03	63.64	3.93
AR_{T}	95.21	77.70	68.82	3.98
AR_{G}	94.73	82.67	64.70	4.61
DINO	89.92	77.93	68.82	5.17
SimCLR	90.43	84.30	67.20	4.53
SSL-Wearables	87.89	75.00	68.84	6.04
RelCon	93.31	86.23	61.74	5.46
Self-PAB	93.22	75.97	59.14	6.68

are evaluated separately in the scaling studies. We retain the per-dataset results because an aggregate statistic can obscure differences between datasets and downstream settings.

As shown in Table 13, PatchTST achieves the highest mean AUROC on HAR, while RelCon performs best on FoG. For disease prediction, SSL-Wearables, AR_{T}, and DINO achieve similar aggregate performance. These results further support our observation that no single pretraining objective dominates across all task families. In particular, a strong average rank does not imply that the same model is the strongest choice for each downstream task.

D.2 Additional wearable-specific pretraining methods

To further examine whether our conclusions hold across a broader set of wearable-specific methods, we additionally evaluate LIMU-BERT [55] and LSM-2 [57]. Both methods are pretrained and evaluated under the same data, compute, and downstream evaluation protocol used in the core comparison. We follow each method's original formulation and released implementation as closely as possible, adapting the input and output interfaces to the shared framework.

As shown in Table 14, LSM-2 achieves competitive HAR and disease-prediction performance, while both additional methods perform less strongly on FoG than the best-performing methods in the core comparison. Their relative performance continues to vary across task families, reinforcing the need to match pretraining choices to downstream requirements.

E Supplementary Scaling Analyses

E.1 Parameter scaling under fixed pretraining data

We investigate scaling behavior by varying model capacity (Small $<$ 10M, Medium $\sim$30M, Large $\sim$100M parameters) and pretraining corpus size. Fig. 9 compares full fine-tuning against linear probing for AR_{T} under the same fixed pretraining data and input configuration. The full-fine-tuning plots show a slight degradation in performance, suggesting that larger encoders alone are not a reliable recipe without corresponding gains in data scale, diversity, or task-specific optimization. All error bars

Table 14: **Additional wearable-specific baselines.** We report mean AUROC (%) across the same task families used in the core objective comparison.

Model	HAR	FoG	Disease Pred.
LIMU-BERT	92.3	73.9	63.2
LSM-2	94.4	75.2	67.8

Table 15: **Full results for AUPRC on HAR tasks with both linear probing and full finetuning results.**

Model	HHAR	MHEALTH	OPPORTUNITY	PAMAP2	RecoFit	CAPTURE-24	HAR70+	HARTH	WEAR	WISDM
Supervised Models										
ViT	60.2	70.7	38.1	59.5	80.1	55.3	58.7	82.2	86.2	66.5
CNN	76.1	69.9	34.9	76.2	76.2	50.7	57.2	87.2	88.1	74.7
Linear Probing										
General Pretraining Methods										
PatchTST	77.3	**86.8**	70.8	**82.4**	85.0	53.9	61.3	85.9	94.0	**77.9**
AR_T	67.9	74.2	**79.5**	73.1	**87.8**	**55.2**	59.3	**86.6**	**94.8**	72.8
AR_G	68.0	82.1	71.6	71.1	85.7	53.2	55.9	82.4	94.0	72.1
DINO	43.5	54.2	70.1	39.4	80.6	49.9	63.1	84.6	92.5	37.4
SimCLR	38.2	60.2	76.9	48.1	77.2	50.7	61.6	80.9	89.1	38.8
Domain-specific Methods										
SSL-Wearables	64.3	57.5	42.6	65.2	44.3	39.9	61.4	56.7	42.0	51.1
RelCon	67.1	77.0	68.3	70.0	77.5	48.0	**65.2**	77.2	87.6	57.0
Self-PAB	**87.8**	55.4	40.3	74.0	75.7	48.7	64.1	84.7	87.1	77.0
Full Fine-tuning										
General Pretraining Methods										
PatchTST	**86.2**	89.0	62.9	**86.8**	83.6	**58.1**	67.4	85.0	92.4	**85.7**
AR_T	73.4	85.1	72.7	77.7	84.5	56.9	57.1	**88.5**	**94.9**	82.0
AR_G	75.0	83.6	63.9	83.0	83.4	54.4	57.9	84.4	93.1	79.7
DINO	43.5	67.6	64.5	44.8	83.8	57.0	65.4	78.7	91.0	57.1
SimCLR	40.0	67.4	68.8	47.4	80.4	56.2	62.3	79.0	92.2	64.1
Domain-specific Methods										
SSL-Wearables	78.9	81.1	**79.1**	62.4	78.9	56.4	66.5	82.6	89.2	51.8
RelCon	77.2	**90.2**	60.9	82.8	**84.5**	51.9	**67.9**	78.4	92.5	79.9
Self-PAB	79.3	68.7	40.7	80.6	72.0	54.2	59.8	82.1	88.5	81.2

Table 16: **Full results for AUPRC on FoG tasks.**

FoGTurning	OdayFoG	DaphnetFoG
Supervised Models		
39.4	72.4	59.2
22.6	72.4	76.8
Linear Probing		
General Pretraining Methods		
61.4	66.0	87.5
61.3	56.5	86.2
71.4	69.5	82.9
37.7	**77.8**	77.6
87.8	66.4	**87.7**
Domain-specific Methods		
84.6	52.3	39.1
76.2	77.1	79.3
63.3	61.0	77.3
Full Fine-tuning		
General Pretraining Methods		
78.8	70.8	89.0
54.5	78.9	85.0
40.7	65.5	83.4
69.9	**85.3**	72.5
67.9	49.8	86.1
Domain-specific Methods		
96.1	57.4	85.2
83.4	75.6	**90.9**
68.4	80.4	58.3

in scaling plots reflect the standard deviation across specific disease tasks and trials, calculated via function call. Standard deviations were calculated using the Python function `statistics.stdev()`.

E.2 PatchTST capacity under linear probing and full fine-tuning

Fig. 10 provides the capacity ablation for PatchTST. This analysis is included as a robustness check because the main text focuses primarily on AR_T for controlled ablations. Across both pretrained models evaluated, we find that scaling model parameters alone does not yield meaningful improvements in representation. On the contrary, in the case of finetuning the entire model backbone, both AR_T and PatchTST performances begin to degrade as the models overfit limited downstream training data. We find, however, that AR_T, performance remains comparable as model capacity increases, regardless of finetuning method, suggesting that AR_T scales more gracefully than does PatchTST.

E.3 Disease prediction scaling protocol

The `UK Biobank` disease scaling experiments use a scaling-specific pretraining and evaluation protocol. In this setting, we separate two sources of additional unlabeled data: the number of pretraining individuals and the number of motion segments per individual. For these experiments, we train an AR_T using 2-hour windows of accelerometer data downsampled to 0.2 Hz, corresponding to 1440 time steps per window. Each setting is pretrained for two epochs, and all model settings are checked to ensure normal training behavior. The pretraining split contains approximately 80% of eligible `UK Biobank` patients, or roughly 80k patients, while the remaining eligible patients are held out from pretraining and used for downstream disease evaluation.

Table 17: **Disease prediction tasks.** We report **AUPRC** per task. Best results are **bolded** and second best are underlined.

Model	Depr. Severity	Depression	Diabetes	Parkinson's	Sleep
Supervised Models					
ViT	28.3	62.6	35.2	27.0	44.5
CNN	29.8	55.4	40.5	38.4	48.6
General Pretraining Methods					
PatchTST	26.6	70.7	42.9	49.8	59.1
AR_T	35.0	**80.1**	54.6	52.3	56.5
AR_G	33.7	69.9	45.2	50.1	61.7
DINO	34.5	76.2	67.9	45.1	**78.1**
SimCLR	32.4	71.9	**70.9**	44.9	77.3
Domain-specific Methods					
SSL-Wearables	**37.9**	60.4	59.9	**60.4**	62.7
RelCon	32.6	73.7	64.0	53.9	54.7
Self-PAB	30.2	68.8	47.1	34.5	58.9

Table 18: **Full reesults for AUROC on HAR tasks with both linear probing and full finetuning results.**

Model	HHAR	MHEALTH	OPPORTUNITY	PAMAP2	RecoFit	CAPTURE-24	HAR70+	HARTH	WEAR	WISDM
Supervised Models										
ViT	83.9	93.8	72.9	92.0	97.7	92.9	96.6	98.7	97.7	94.1
CNN	89.1	90.7	69.5	95.5	97.0	90.9	94.5	98.7	98.0	94.8
Linear Probing										
General Pretraining Methods										
PatchTST [35]	93.5	**98.2**	89.8	**97.4**	98.5	91.6	95.8	98.9	99.2	95.5
AR_T [48]	88.6	96.0	**92.1**	95.6	**98.9**	**92.3**	95.5	**99.0**	**99.3**	94.8
AR_G [14]	89.2	96.0	89.4	95.0	98.7	91.3	95.3	98.5	99.2	94.7
DINO [8]	73.4	87.2	87.8	82.7	98.1	92.0	95.4	98.9	98.5	85.2
SimCLR [12]	68.6	90.0	90.0	86.8	97.8	92.2	95.9	98.6	98.0	86.4
Domain-specific Methods										
SSL-Wearables [62]	90.0	89.9	73.1	91.8	91.0	87.0	85.4	92.0	88.5	90.2
RelCon [56]	87.8	96.3	85.3	94.5	97.0	89.9	**95.9**	98.0	97.9	90.5
Self-PAB [25]	**96.8**	91.8	74.7	95.4	97.4	89.7	93.4	98.6	97.7	**96.7**
Full Fine-tuning										
General Pretraining Methods										
PatchTST [35]	**96.1**	**98.6**	87.6	97.1	98.1	93.0	**96.9**	98.3	98.8	97.4
AR_T [48]	90.9	97.8	90.2	97.2	98.0	93.1	94.8	**99.0**	**99.1**	97.3
AR_G [14]	92.4	97.7	88.2	**97.4**	**98.2**	92.5	95.6	98.8	98.9	96.6
DINO [8]	71.1	93.1	84.7	84.4	96.9	93.1	90.9	94.0	97.6	92.1
SimCLR [12]	62.4	89.4	84.2	79.0	95.8	93.2	91.2	94.6	97.0	91.3
Domain-specific Methods										
SSL-Wearables [62]	94.8	96.1	**91.6**	90.3	98.1	**93.4**	92.7	98.6	97.9	90.8
RelCon [56]	85.8	98.4	86.4	95.3	96.5	89.8	76.5	95.5	97.4	94.0
Self-PAB [25]	92.9	94.4	67.3	97.0	97.4	91.5	93.4	98.5	97.8	**97.4**

Table 19: **Full results for AUROC on FOG tasks.**

FoGTurning	OdayFoG	DaphnetFoG
Supervised Models		
63.6	68.0	83.5
51.2	80.5	87.6
Linear Probing		
General Pretraining Methods		
85.7	65.4	**95.0**
80.2	58.6	94.3
83.0	71.6	93.4
61.4	**83.2**	89.2
91.4	66.7	94.8
Domain-specific Methods		
91.6	62.0	71.4
87.5	78.2	93.0
80.4	55.0	92.5
Full Fine-tuning		
General Pretraining Methods		
83.9	75.0	95.6
78.3	68.0	93.2
77.7	65.5	93.2
85.5	**86.0**	86.1
89.3	51.5	93.8
Domain-specific Methods		
98.2	73.3	93.6
89.5	81.2	**96.6**
83.9	85.5	80.7

For each disease trait, we construct a case-control cohort from the held-out patients, using as close to a 1:5 case-control ratio as possible. This cohort is then split at the patient level into train, validation, and test folds for downstream multiple-instance learning.

For downstream evaluation, each patient-day is treated as an independent bag. A patient-day consists of 12 consecutive 2-hour windows, giving up to seven bags per patient. All bags from the same patient share the same disease label and are assigned to the same train, validation, or test fold to prevent leakage across days. The frozen encoder embeds each 2-hour window independently, and the MIL classifier is trained on bags of 12 window embeddings. Performance is measured using AUROC on the held-out test set, with 95% bootstrap confidence intervals computed using 1000 resamples.

The three scaling sweeps vary the number of pretraining patients, the number of windows available per patient, and the composition of the pretraining corpus. The patient-scaling sweep uses 1%, 10%, and 100% of the pretraining patients. The segment-scaling sweep caps the number of windows per patient at 1, 9, or 84. The mixing sweep compares `NHANES`-only pretraining, matched `NHANES`–`UK Biobank` mixtures, and full `UK Biobank` inclusion.

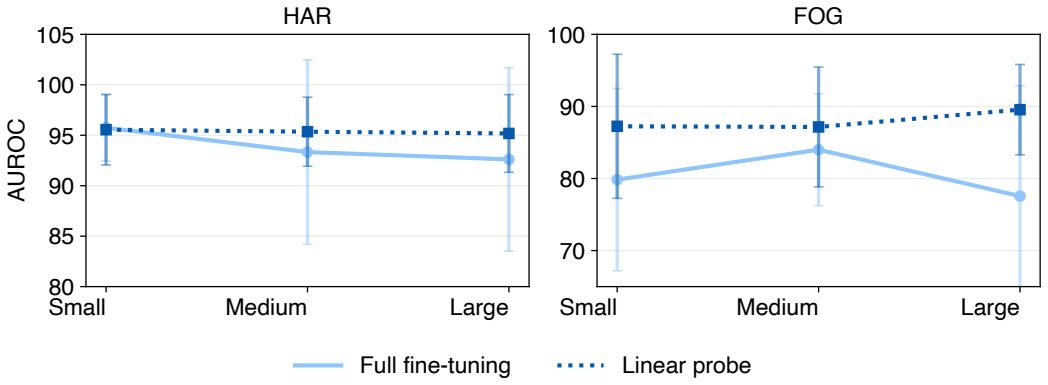


Figure 9: **Parameter scaling under fixed pretraining data.** Controlled ablations using AR_T. We vary encoder size and compare full fine-tuning against linear probing while holding the pretraining corpus and input configuration fixed.

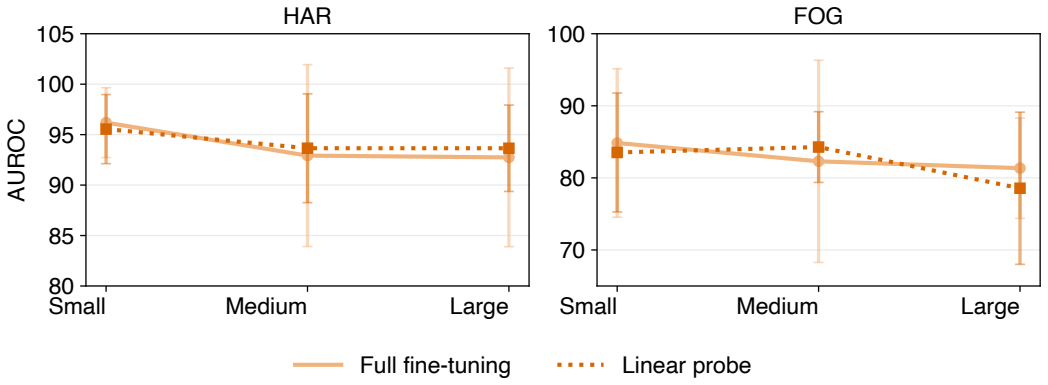


Figure 10: **Additional Model Capacity Ablations.** Controlled ablations using PatchTST; we vary the capacity of the encoder and style of finetuning while holding the rest of the protocol fixed.

E.4 Disease prediction scaling with additional UK Biobank targets

Fig. 11 additionally visualizes performance on osteoarthritis prediction for the `UK Biobank` scaling experiments. We observe trends similar to osteoporosis, indicating that increasing the number of pretraining individuals, the number of windows per individual, and the diversity of the pretraining corpus can improve disease-relevant wearable representations across multiple targets. This further suggests that wearable disease representations benefit from both population diversity and repeated behavioral sampling within individuals.

F Supplementary Controlled Configuration Studies

F.1 Sampling frequency and window length with PatchTST

The main text uses AR_T as the primary backbone for controlled configuration studies due to its strong and balanced performance across tasks in both linear probing and full finetuning. To test whether the same qualitative trends hold for another architecture, Fig. 12 repeats the sampling-frequency and window-length ablations with PatchTST. Similar to the results with AR_T in the main text, the highest temporal resolution of 20 Hz provides the greatest benefit to downstream performance, while window size effects are weaker and more task-dependent. Strikingly, model performance degrades significantly between 5Hz and 1Hz input streams. Even so, accounting for this degradation, the models pretrained at 1Hz are still able to perform sufficiently in HAR classification. This finding is critical for wearable foundation models as many large cohort studies of data report only aggregate

Table 20: **Ablation study across task families.** We report **AUROC** per dataset/task across axis settings.

Human Activity Recognition

Axis	Model	Axis Value	HHAR	MHEALTH	OPPORTUNITY	PAMAP2	RecoFit	CAPTURE-24	HAR70+	HARTH	WEAR	WISDM
Window Size												
	AR_T	10s	89.7	97.1	87.2	94.7	98.6	70.7	96.0	98.9	99.4	96.6
	AR_T	30s	90.9	97.8	90.2	97.2	98.0	93.1	94.8	99.0	99.1	97.3
	AR_T	60s	85.5	96.8	87.0	97.0	97.8	53.8	97.2	98.9	99.2	92.3
	PatchTST	10s	93.8	96.4	88.0	95.0	98.3	72.5	95.8	98.4	99.0	96.6
	PatchTST	30s	96.1	98.6	87.6	97.1	98.1	93.0	96.9	98.3	98.8	97.4
	PatchTST	60s	86.2	97.5	85.8	96.0	97.6	45.5	97.8	98.1	97.8	90.4
Sampling Frequency												
	AR_T	1Hz	76.3	96.1	88.6	94.5	97.6	51.4	99.2	97.5	98.5	95.4
	AR_T	5Hz	84.8	97.2	89.4	95.9	98.0	72.1	96.6	97.9	98.6	97.0
	AR_T	20Hz	90.9	97.8	90.2	97.2	98.0	93.1	94.8	99.0	99.1	97.3
	PatchTST	1Hz	51.3	83.0	67.3	74.9	81.1	50.2	88.4	89.5	70.6	76.0
	PatchTST	5Hz	90.6	94.3	92.6	94.0	97.7	53.6	96.7	97.7	98.1	96.3
	PatchTST	20Hz	96.1	98.6	87.6	97.1	98.1	93.0	96.9	98.3	98.8	97.4
Sensor Axes												
	AR_T	Uniaxial	93.1	95.5	84.7	95.8	96.4	62.9	91.4	94.2	96.4	95.3
	AR_T	Triaxial	90.9	97.8	90.2	97.2	98.0	93.1	94.8	99.0	99.1	97.3
	PatchTST	Uniaxial	71.8	76.5	54.9	72.6	86.6	49.0	54.8	85.5	70.8	84.2
	PatchTST	Triaxial	96.1	98.6	87.6	97.1	98.1	93.0	96.9	98.3	98.8	97.4

Freezing of Gait

Model	Axis Value	FoGTurning	OdayFoG	DaphnetFoG
Window Size				
AR_T	10s	79.9	68.5	93.2
AR_T	30s	78.3	68.0	93.2
AR_T	60s	94.4	82.6	95.8
PatchTST	10s	81.6	57.5	88.9
PatchTST	30s	83.9	75.0	95.6
PatchTST	60s	97.6	98.3	92.4
Sampling Frequency				
AR_T	1Hz	62.2	67.4	77.8
AR_T	5Hz	91.4	68.4	86.2
AR_T	20Hz	78.3	68.0	93.2
PatchTST	1Hz	56.0	68.5	59.6
PatchTST	5Hz	71.2	65.6	84.7
PatchTST	20Hz	83.9	75.0	95.6
Sensor Axes				
AR_T	Uniaxial	64.0	78.0	87.9
AR_T	Triaxial	78.3	68.0	93.2
PatchTST	Uniaxial	60.7	53.5	50.6
PatchTST	Triaxial	83.9	75.0	95.6

Disease Prediction

Model	Axis Value	Depr. Severity	Depression	Diabetes	Parkinson's	Sleep
Window Size						
AR_T	10s	58.9	76.2	52.7	71.5	56.6
AR_T	30s	58.4	81.3	67.7	75.0	61.7
AR_T	60s	53.8	74.2	62.3	69.6	51.3
PatchTST	10s	53.2	55.2	54.8	52.3	29.1
PatchTST	30s	55.0	77.0	53.8	72.8	44.7
PatchTST	60s	55.0	58.3	56.7	64.2	49.4
Sampling Frequency						
AR_T	1Hz	56.6	58.2	56.1	65.0	46.7
AR_T	5Hz	56.1	72.9	54.2	68.5	32.4
AR_T	20Hz	58.4	81.3	67.7	75.0	61.7
PatchTST	1Hz	50.7	50.7	50.4	50.4	50.0
PatchTST	5Hz	54.9	62.3	49.6	60.0	27.5
PatchTST	20Hz	55.0	77.0	53.8	72.8	59.6
Sensor Axes						
AR_T	Uniaxial	57.8	77.6	64.6	65.8	72.7
AR_T	Triaxial	58.4	81.3	67.7	75.0	61.7
PatchTST	Uniaxial	51.9	55.7	63.8	51.5	54.2
PatchTST	Triaxial	55.0	77.0	53.8	72.8	59.6

metrics or low frequency data, potentially erasing important movement information from the data. Across window size, we find less stark of a contrast; increasing context does not yield significant performance gains. We note that this saturation of performance should not be attributed to the cleanliness of the data, i.e. recordings collected of a pure action, as our window cleaniness check reports an average >80% label purity for all of our HAR and FoG splits. Detailed results are reported in Table 20 and Table 21.

F.2 Sensor-axis dimensionality across additional backbones

Similar to the experiment in Fig. 6, we test for qualitative trends in sensor-axis dimensionality across additional model architectures in Fig. 12 (PatchTST). Triaxial inputs generally preserve useful orientation and movement structure that can be lost when reducing the signal to one channel. Both models substantially improve with triaxial input data, with PatchTST appearing to collapse when given only a single axis representation of the data.

F.3 Activity-dependent sensitivity to downsampling

To better understand why HAR remains competitive at low sampling rates, we examine whether the effect of downsampling differs across activity groups. Using WISDM, we compare AR_T models pretrained at 20 Hz and 1 Hz and group activities into near-static, hand/eating, high-cadence, and

Table 21: **Ablation study across task families.** We report **AUPRC** per dataset/task across axis settings.

Human Activity Recognition

Axis	Model	Axis Value	HHAR	MHEALTH	OPPORTUNITY	PAMAP2	RecoFit	CAPTURE-24	HAR70+	HARTH	WEAR	WISDM
Window Size												
	AR_T	10s	68.3	81.1	66.4	71.1	85.2	34.2	67.6	89.0	92.3	77.6
	AR_T	30s	73.4	85.1	72.7	77.7	84.5	56.9	57.1	88.5	94.9	82.0
	AR_T	60s	57.1	77.2	69.0	76.1	81.6	22.6	77.8	83.0	92.6	68.6
	PatchTST	10s	81.2	73.8	63.0	78.7	85.2	39.6	64.8	82.4	90.9	80.0
	PatchTST	30s	86.2	89.0	62.9	86.8	83.6	58.1	67.4	85.0	92.4	85.7
	PatchTST	60s	59.2	84.0	65.7	78.4	83.9	25.5	78.3	80.2	86.7	69.1
Sampling Frequency												
	AR_T	1Hz	39.7	76.1	70.6	63.7	71.3	15.8	95.5	74.6	83.7	73.3
	AR_T	5Hz	53.3	78.1	72.3	68.4	80.9	36.7	74.6	75.6	92.0	81.7
	AR_T	20Hz	73.4	85.1	72.7	77.7	84.5	56.9	57.1	88.5	94.9	82.0
	PatchTST	1Hz	20.5	29.6	32.7	29.8	17.7	15.3	60.7	41.7	12.6	18.3
	PatchTST	5Hz	68.3	64.7	74.6	70.4	79.8	24.4	76.2	76.3	90.9	76.5
	PatchTST	20Hz	86.2	89.0	62.9	86.8	83.6	58.1	67.4	85.0	92.4	85.7
Sensor Axes												
	AR_T	Uniaxial	68.6	69.1	59.9	70.4	65.1	29.3	58.0	68.5	72.2	66.5
	AR_T	Triaxial	73.4	85.1	72.7	77.7	84.5	56.9	57.1	88.5	94.9	82.0
	PatchTST	Uniaxial	33.1	28.5	27.4	23.8	29.2	21.5	27.9	35.2	12.5	28.1
	PatchTST	Triaxial	86.2	89.0	62.9	86.8	83.6	58.1	67.4	85.0	92.4	85.7

Freezing of Gait

Model	Axis Value	FoGTurning	OdayFoG	DaphnetFoG
Window Size				
AR_T	10s	70.2	70.8	65.3
AR_T	30s	54.5	78.9	85.0
AR_T	60s	90.3	45.7	91.3
PatchTST	10s	68.8	65.0	59.2
PatchTST	30s	78.8	70.8	89.0
PatchTST	60s	95.0	22.1	84.6
Sampling Frequency				
AR_T	1Hz	33.0	64.9	53.1
AR_T	5Hz	79.2	30.5	65.0
AR_T	20Hz	54.5	78.9	85.0
PatchTST	1Hz	28.5	55.7	34.5
PatchTST	5Hz	57.6	54.7	64.3
PatchTST	20Hz	78.8	70.8	89.0
Sensor Axes				
AR_T	Uniaxial	41.0	78.9	75.9
AR_T	Triaxial	54.5	78.9	85.0
PatchTST	Uniaxial	29.1	38.1	23.8
patchtst	Triaxial	78.8	70.8	89.0

Disease Prediction

Axis	Model	Axis Value	Depr. Severity	Depression	Diabetes	Parkinson's	Sleep
Window Size							
	AR_T	10s	35.0	72.5	48.0	40.0	56.3
	AR_T	30s	35.0	80.1	54.6	52.3	56.5
	AR_T	60s	31.4	66.6	46.8	44.9	53.4
	PatchTST	10s	27.0	59.4	35.8	28.2	36.6
	PatchTST	30s	26.6	70.7	42.9	49.8	46.6
	PatchTST	60s	27.6	61.4	43.4	37.7	49.9
Sampling Frequency							
	AR_T	1Hz	32.7	58.2	41.6	43.5	48.4
	AR_T	5Hz	31.2	67.2	37.3	43.7	41.3
	AR_T	20Hz	35.0	80.1	54.6	52.3	56.5
	PatchTST	1Hz	21.8	48.2	33.2	24.5	39.1
	PatchTST	5Hz	29.8	61.2	37.0	36.8	36.7
	PatchTST	20Hz	26.6	70.7	42.9	49.8	59.1
Sensor Axes							
	AR_T	Uniaxial	33.2	78.0	51.8	55.5	70.0
	AR_T	Triaxial	35.0	80.1	54.6	52.3	56.5
	PatchTST	Uniaxial	23.1	49.4	45.2	24.1	42.0
	PatchTST	Triaxial	26.6	70.7	42.9	49.8	59.1

Table 22: **Activity-dependent effects of downsampling on WISDM.** We report AUROC (%) at 20 Hz and 1 Hz. Δ is the AUROC change in percentage points (1 Hz minus 20 Hz), with a subject-bootstrap 95% confidence interval. Intensity denotes the reported range of gravity-removed acceleration power within each activity group.

Activity group	Intensity	20 Hz	1 Hz	Δ (95% CI)
Static	3.3–4.4	92.6	93.2	$+0.7\ [-0.1, +1.5]$
Hand/eating	4.6–74	89.8	87.3	$-2.5\ [-4.0, -1.3]$
High-cadence	25–478	95.6	92.8	$-2.8\ [-4.0, -1.8]$
Locomotion	116–160	97.4	94.0	$-3.4\ [-4.9, -2.1]$

locomotion categories. Activity intensity is quantified using gravity-removed acceleration power. For each group, we report AUROC and the change between the two sampling rates. Confidence intervals for the change are computed by bootstrapping the 51 subjects.

As shown in Table 22, near-static activities are relatively insensitive to aggressive downsampling, whereas dynamic activities show clearer reductions in AUROC. The small numerical improvement for static activities has a confidence interval that overlaps zero and should not be interpreted as evidence that 1 Hz is intrinsically superior. In contrast, hand/eating, high-cadence, and locomotion activities all show negative changes with confidence intervals below zero. These results suggest that aggregate HAR performance can remain competitive because different activities rely on different temporal

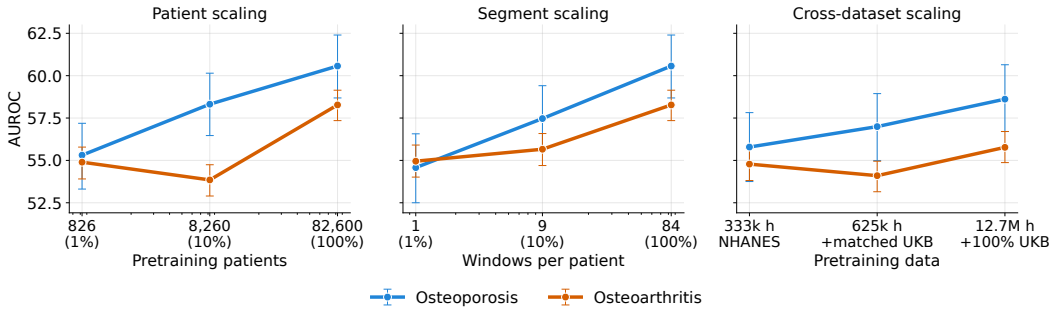
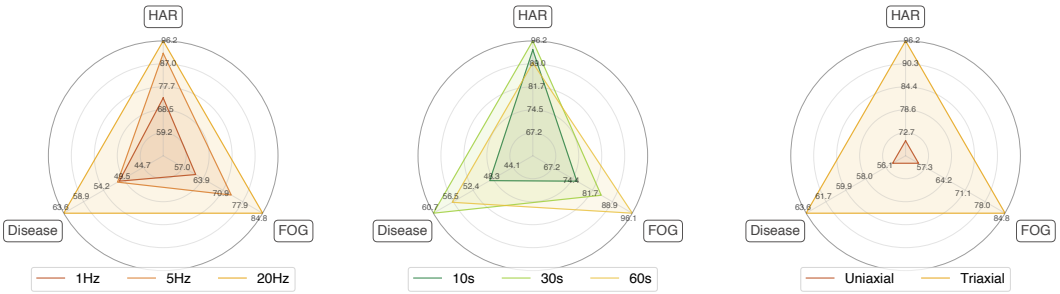


Figure 11: `UK Biobank` **disease scaling at 0.2 Hz.** We evaluate how downstream AUROC changes as pretraining varies by number of individuals, number of segments per individual, and mixture of `NHANES` and `UK Biobank` pretraining data. Individual and segment scaling both improve osteoporosis and osteoarthritis prediction, while adding in-domain `UK Biobank` data to `NHANES` pretraining provides a modest additional gain.



Freq.	HAR	FoG	DP	Overall
1Hz	73.2	61.4	50.4	61.7
5Hz	91.2	73.8	50.9	72.0
20Hz	**96.2**	**84.8**	**63.6**	**81.6**

(a) Sampling frequency

Window	HAR	FoG	DP	Overall
10s	93.4	76.0	48.9	72.8
30s	**96.2**	84.8	**60.7**	80.6
60s	89.3	**96.1**	56.7	**80.7**

(b) Window length

Axes	HAR	FoG	DP	Overall
Uniaxial	70.7	54.9	55.4	60.3
Triaxial	**96.2**	**84.8**	**63.6**	**81.6**

(c) Axis dimensionality

Figure 12: **Sensitivity to input and representation choices.** Controlled ablations using PatchTST. (a)(b) We vary sampling frequency and window length while holding the rest of the protocol fixed. (c) We compare performance of PatchTST trained on uniaxial inputs vs. triaxial inputs. AUROC is reported across task families.

information: posture-driven classes retain useful low-frequency structure, while more dynamic classes benefit from higher temporal resolution. Changes in input length, representation, and optimization may also contribute, so the activity grouping alone does not isolate the cause of this sensitivity.

F.4 FoG performance across episode durations

To examine whether representation differences are more pronounced for brief gait events, we stratify FoG performance by episode duration on the FoGTurning dataset. The duration-stratified comparison uses 30-second, 20 Hz triaxial inputs, frozen encoders, linear probes, and subject-grouped folds. We group positive windows by episode duration and report AUROC for each group, together with the number of positive windows.

As shown in Table 23, Self-PAB performance increases with episode duration, from 35.6 AUROC for episodes shorter than 3.2 seconds to 78.1 for episodes longer than 15 seconds. The time-domain encoders maintain higher performance across all duration groups, with a particularly large difference for the shortest episodes. These results are consistent with the possibility that temporal aggregation weakens brief FoG patterns by mixing them with surrounding gait. However, Self-PAB remains weaker for longer episodes as well, indicating that temporal resolution alone does not explain the full gap.

Table 23: **FoG detection across episode durations.** We report AUROC (%) on FoGTurning for a time-frequency method and time-domain encoders. Episode-duration bins are in seconds. Counts denote positive windows, not independent subjects or episodes.

Encoder	Domain	Overall	< 3.2	3.2–6.4	6.4–15	> 15
Self-PAB	Time-frequency	70.5	35.6	53.7	73.3	78.1
AR_T	Time	91.8	85.4	83.4	83.1	98.1
PatchTST	Time	92.3	82.0	81.4	90.0	97.7
AR_G	Time	90.8	82.1	84.1	80.3	97.5
Positive windows		1687	76	328	331	952

Table 24: **Overall cluster quality metrics on UMAP embeddings**, averaged across datasets. Incorporating additional placements yields the best-separated representations.

Condition	**Silhouette ↑**	**kNN Purity ↑**	**Calinski-Harabasz ↑**	**Davies-Bouldin ↓**
All Placements	**0.3686**	**0.8086**	**478.0**	**1.2991**
All Sensors	0.0881	0.6635	135.8	3.5703
Default	0.0634	0.6671	171.0	5.1265
Noise Control	-0.0534	0.5160	107.6	22.2653

G Representation Geometry and Cluster Analysis

To study placement robustness—and specifically whether wrist-based pretraining extrapolates to other sensors or placements—we finetune separate models using data from specific sensors (gyroscope, magnetometer) or body locations (waist, chest, leg, ankle) after pretraining on wrist data and train a linear classifier on concatenated embeddings for each downstream dataset with the additional available data streams. The main text interprets improvements from synchronized multi-stream fusion using representation geometry and cluster structure. Table 24 reports the corresponding cluster-quality metrics averaged across all datasets from Fig. 8, providing quantitative support for the qualitative embedding visualizations and cluster interpretations discussed in the main paper. The noise control row displays cluster metrics when concatenating random normal vectors of the same dimensionality as the extra placement or sensor embeddings that were concatenated in the other comparisons.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims we make in the abstract and introduction highlighting the contributions and scope of `Inertia-1` accurate reflect the panel of modern SSL objectives and the grid of experimental axes and downstream evaluation performed on every model utilized. Conclusions made are based on empirical results of such experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While we are excited by the potential of `Inertia-1` as a cookbook for future motion and wearable foundation models, there remain areas to study the extension of ablations across alternative SSL objectives and noisier, real world data. Scaling performance is discussed as a major arm of exploration, and all data leveraged by this work is publicly available and de-identified.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No theoretical results are presented by this work.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed descriptions of model architectures are provided in the appendix, along with the full list of datasets used for training and evaluation. Additionally, the source code, checkpoints, and data will be made available publicly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data is publicly accessible. Anonymized versions of the code are available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Specifics of every dataset's evaluation setup are listed in the appendix in detail, along with the specifics of the pretraining pipeline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars in plots are specified as the sample standard deviation across trials and tasks within a family. No statistical significance claims are made.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All necessary details of hardware and compute requirements of training and evaluation are listed in detail in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All data used by this work is publicly available and de-identified, and all code is reproducible and open. Our work serves as a foundation for future motion generalist models to improve quality of life and society.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work has direct impacts on the future of machine learning with wearable devices, in particular in the activity recognition, gait analysis, and health prediction spaces, all of which stand to benefit society.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not pose any risks regarding misuse of the released data or models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All data and models are credited to the original research teams who collect or design them and listed in the references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Specific documentation is provided in the anonymized codebase for leveraging the unified dataset for explorations, and more overarching details are listed in the appendix here.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This data did not involve crowdsourcing experiments or direct research with human subjects – all data was collected by external organizations and released publicly prior to this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methods development presented by this work did not involve LLMs as any important components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.