LSM-2: Learning from Incomplete Wearable Sensor Data

Maxwell A. Xu[1,3*,†], Girish Narayanswamy[1,4*,†], Kumar Ayush[1], Dimitris Spathis[1], Shun Liao[1], Shyam A. Tailor[1], Ahmed Metwally[1], A. Ali Heydari[1], Yuwei Zhang[1], Jake Garrison[1], Samy Abdel-Ghaffar[1], Xuhai Xu[1], Ken Gu[1], Jacob Sunshine[1], Ming-Zher Poh[1], Yun Liu[1], Tim Althoff[1], Shrikanth Narayanan[2], Pushmeet Kohli[2], Mark Malhotra[1], Shwetak Patel[1], Yuzhe Yang[1], James M. Rehg[3], Xin Liu[1,◦], Daniel McDuff[1,◦]

[1]Google Research, [2]Google DeepMind, [3]University of Illinois Urbana-Champaign, [4]University of Washington
[†]Co-first authors, [◦]Co-last authors, [*]Work done during an internship at Google

Foundation models, a cornerstone of recent advancements in machine learning, have predominantly thrived on complete and well-structured data. Wearable sensor data frequently suffers from significant missingness, posing a substantial challenge for self-supervised learning (SSL) models that typically assume complete data inputs. This paper introduces the second generation of Large Sensor Model (LSM-2) with Adaptive and Inherited Masking (`AIM`), a novel SSL approach that learns robust representations directly from incomplete data without requiring explicit imputation. `AIM`'s core novelty lies in its use of learnable mask tokens to model both existing ("inherited") and artificially introduced missingness, enabling it to robustly handle fragmented real-world data during inference. Pre-trained on an extensive dataset of 40M hours of day-long multimodal sensor data, our LSM-2 with `AIM` achieves the best performance across a diverse range of tasks, including classification, regression and generative modeling. Furthermore, LSM-2 with `AIM` exhibits superior scaling performance, and critically, maintains high performance even under targeted missingness scenarios, reflecting clinically coherent patterns, such as the diagnostics value of nighttime biosignals for hypertension prediction. This makes `AIM` more reliable choice for real-world wearable data applications.

1. Introduction

In the real world, missing or incomplete data is a pervasive challenge across a variety of domains. In clinical settings for example, electronic health records frequently exhibit missingness due to factors such as loss to follow-up [27, 75] or condition-specific diagnostic procedures [26, 39]. Similarly, sensor systems grapple with incomplete data streams due to strategic intermittent deactivation for energy conservation, environmental noise, sensor obstruction, or hardware malfunctions [22, 7, 18]. Missing data for wearable mobile health sensors is especially prevalent and problematic. In addition to the aforementioned causes, user compliance issues (e.g. improper/insecure device attachment) or mobile-specific challenges (e.g. data transmission failures, battery charging periods) further exacerbate the problem [50, 68].

Self-Supervised Learning (SSL) has emerged as a powerful paradigm for learning transferable representations by exploiting inherent structures within unlabeled data [24]. When scaled to large pre-training datasets with sufficient compute, these approaches yield foundation models capable of strong generalization across diverse downstream tasks [45, 60]. This is especially promising for wearable sensors, where physiological signals contain rich information predictive of diverse health outcomes, with several recent large-scale data collection initiatives, such as UK Biobank [35], All of Us [33], and the Apple Heart and Movement Study [62]. This has enabled the development of wearable

Corresponding author(s): maxu@illinois.edu, girishvn@uw.edu, {xliucs,dmcduff}@google.com

sensor foundation models that generalize across multiple health prediction tasks [42, 70, 52, 1].

Unfortunately, state-of-the-art time-series SSL approaches typically assume fully-observed data inputs. As such, prior wearable sensor foundation models have handled missingness by modeling short context windows (i.e. <60s [1], 2.56s [70], 10s [47]), where incomplete instances can easily be filtered out. However, many clinically relevant physiological patterns (e.g. circadian rhythms [76], heart rate variability [15], and daily activity profiles [32]) require analyzing day-long recordings. Unfortunately, day-long data inevitably contains missingness due to wearable sensor limitations (e.g. battery drain necessitating strategic sensor deactivation, motion artifacts corrupting signals). As detailed in Section 3, our dataset exhibits pervasive missingness: 0% of records are complete. While prior work with similar data employed imputation methods in order to train their SSL model [42], such approaches risk introducing biases that can propagate to downstream models [34].

Figure 1 | **LSM-2 Models Incomplete Data.** Our method uses a learned mask token to represent existing missingness during inference. Then, if sensors are missing, it can directly reconstruct them [L] or classify directly on the incomplete data [R].

In this paper, we introduce the second generation of Large Sensor Model (LSM-2) based on Adaptive and Inherited Masking, `AIM`, a self-supervised learning approach that learns a representation directly from incomplete data with diverse missingness patterns. To the best of our knowledge, this is the first work to address representation learning directly on incomplete wearable sensor data. Building on masked autoencoder (MAE) pre-training [29], `AIM` uses a shared learnable mask token to represent both inherited and artificial masks. *Inherited masks* are derived from existing missingness in raw data, thereby masking incomplete data and avoiding the need for imputation. *Artificial masks*, are randomly applied on observed tokens, providing a ground truth for the reconstruction pre-training objective. Via `AIM`'s introduction of inherited masks, mask tokens are learned to represent real-world missingness. During evaluation, missingness still occurs in the raw data. Here, the inherited mask allows for missingness-aware embeddings. Like real missingness, the number of inherited mask tokens may vary, violating the naive MAE's assumption of a fixed number of masked tokens [29]. As such, the *adaptive* component of `AIM` is able to suppress any additional missing tokens from contributing to the final encoder output, ensuring that the encoding is a learned representation of the non-missing data solely. This encoding can then be used in conjunction with a linear probe to predict a variety of downstream classification and regression tasks, as well as being fed back into the decoder for downstream generative tasks.

The key contributions of our work are:

1. *We introduce LSM-2 and propose a novel training strategy, Adaptive and Inherited Masking,* `AIM`, that uses adaptive masking to jointly model artificial and inherited missingness and learn a strong, generalizable representation, directly on incomplete data. By incorporating adaptive masking during pre-training and inference, our method enables a single model to robustly support a variety of downstream tasks under real-world missingness conditions without requiring any explicit imputation.
2. *We demonstrate that our LSM-2 w/* `AIM` *pre-trained foundation model achieves state-of-the-art*

performance across diverse set of tasks (3x classification, 4x generative, 3x regression) that cover a wide range of semantics (clinical, mental health, wearables, demographics) after large-scale pre-training on 40 million hours of day-long multimodal sensor data. Our model also demonstrates superior scaling performance as compared to our prior LSM-1 model [42].

3. *We evaluate the robustness of our LSM-2 across a wide range of targeted missing scenarios*, dropping out specific sensors or time windows, and we demonstrate much less performance degradation compared to the baseline method that is pre-trained with imputed data. The missingness scenarios in which our model does express sensitivity is reflective of physiological domain knowledge, providing interesting insights into the nature of a given prediction target.

2. Related Work

Self-Supervised Learning for Time-Series Foundation Models. Our LSM-2 model utilizes `AIM`, an MAE [29] SSL framework that combines an artificial mask with an inherited mask from real-world sensor data. This differs from LSM-1 [42], the most closely-related work, which performs MAE pre-training with just an artificial mask and uses naive imputation to fill in pre-existing missingness, both of which negatively impacts downstream performance (see Section 6). Other MAE-style methods for time series data are limited in that they either: (a) focus exclusively on complete univariate signals [20, 37, 14], (b) work with highly correlated channels from a single modality [41], or (c) focus on task-specific forecasting without learning general-purpose embeddings [5, 44, 17]. Notably, none of these approaches handle the missing data patterns inherent in real-world multivariate sensor data. Alternatively, contrastive SSL methods learn representations by attracting positives and repelling negatives in embedding space. Positives are generated via augmentations [59] or sampling using temporal proximity [61], subject labels [1], domain knowledge [47], or motif similarity [69, 70]. However, these require strong assumptions, either carefully designed augmentations or reliable positive selection strategies and are unable to do reconstruction out-of-box unlike the MAE methods.

Learning from Incomplete Multimodal Data. Our model learns general-purpose embeddings directly from incomplete multimodal time-series data through self-supervised pre-training, enabling effective transfer to diverse downstream tasks via simple linear probes. Existing representation learning works for incomplete data have focused primarily on either tabular data [12] or irregularly-sampled event time-series [8], both of which differ fundamentally from wearable sensors. Tabular missingness consists of simple, scattered, point-wise missingness, unlike the complex structured patterns in wearables, in which sensor groups across a time window will be missing and not at random (Figure 2). While the irregularly-sampled domain shares some similarities, they have fundamentally different data characteristics. Irregularly sampled time-series such as ICU lab testing [56] are collected at arbitrary intervals with all other modalities typically missing, whereas wearables produce regularly-sampled data where some modalities will drop out in structured groups (Figure 2).

Alternatively, a separate body of incomplete multimodal data work has focused on learning imputation methods. The most relevant work is ReMasker [23], which combines inherited and artificial masking in an MAE framework. Our approach differs in three fundamental aspects: (1) we optimize for representation learning rather than imputation, (2) we handle the complex missingness patterns characteristic of multimodal time-series (see Fig. 2), as opposed to the simpler point-wise missingness in tabular data, and (3) we scale efficiently to long sequences (N=3744 tokens) compared to their limited context (N<20 tokens), representing a 35000x increase in compute (see Section 4 for details). Another approach, [65], similarly uses both inherited and artificial masks but limits attention to handcrafted time points (N=206) and uses self-attention blocks. While numerous deep learning methods exist for multivariate time-series imputation [72, 10, 49, 16], these approaches focus solely on reconstruction quality and fail to produce general-purpose embeddings necessary

Figure 2 | **The Fragmented Nature of Sensor Data.** Multimodal time-series sensor data frequently contains missing observations. Missingness can take several modes. In wearable data, these modes take the form of temporary periods in which a sensor(s) are off, periods in which the device is not warn, and measurements that are filtered out because they are clearly spurious/out of range.

for foundation models. [34] investigate various imputation methods and train classifiers on the reconstructed data, but do not learn representations for multiple downstream tasks. In contrast, our work handles real-world missingness patterns within a scalable representation learning framework.

3. Large Scale Incomplete Wearable Data

Data Summary. A primary contribution of our work is in modeling incomplete data during pre-training, post-training, and inference. To validate our method we curate a large, unlabeled, pre-training dataset in addition to two labeled datasets for downstream tasks. Each data sample contains 26 minutely aggregated features from a set of 5 sensors (photoplethysmography, accelerometer, skin conductance, altimeter, and temperature) for a time span of 1440 minutes (1 day). A core property of these data is that they have complex, structured missingness patterns. A representative example of sensor data with missingness can be seen in Fig. 2, along with the missingness distribution and statistics shown in Fig. 3. Missing data is ubiquitous in long-duration wearable sensor recordings, with 0% of samples over our entire dataset of 1.6 million instances of 1 day data. All pre-trained and downstream datasets utilize similar devices and thus are subject to similar missingness patterns. Please refer to the Appendix for further data descriptions and statistics.

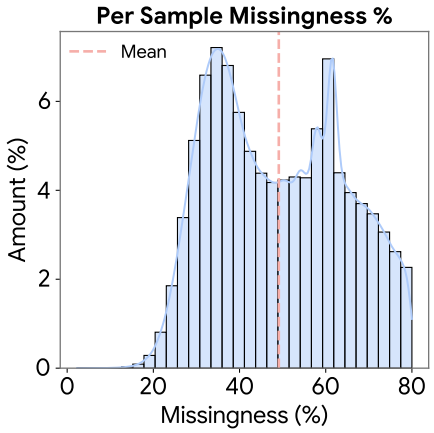

Figure 3 | **Distribution of Missingness % Per Sample.** Mean 49%, Median 48%, Std Dev 15%, Minimum 2%, Maximum 80%. Samples with > 80% missingness are discarded.

Pre-training Data. For pre-training, we used a de-identified dataset collected between March 1st and May 1st 2024 inclusive. The dataset included 3,581,748 person-days (or 40 million hours) sampled at minutely resolution from 60,440 people (37,352 men, 23,041 women, 47 unspecified). A mean of 59 days (min: 1, max: 93) were contributed per person with standard deviation of 32 days. All data used in this study were collected with the informed consent of research participants. This consent permits the use of data to generate findings for publication in scientific journals and other

outlets, contributing to general knowledge about health and science. The mean reported participant age was 42.5 years (min: 18, max: 96 years, st.dev.: 12.6). The population reflects a wide range of body-mass index (BMI) values with 37% healthy, 34% overweight and 25% obese in the training set and a similar cross-section in the validation set.

Downstream Metabolic Study Data. These data come from an IRB approved observational study of adults in the United States. We enrolled 4,416 participants, of which 1,250 had wearable data, labels and were included in our analysis. Demographics (age, BMI) and medical conditions (hypertension, anxiety) were collected via self-report.

Downstream Activity Study Data. These data come from the same source as our pretraining data. We randomly sampled approximately 5,000 examples for each of 20 activities for training and 1,000 examples of each activity for testing. The training and testing data were sampled in person-independent manner. The activities were from the following classes: *Walking*, *Bike*, *Playing Sports*, *Running*, *Aerobics*, *Elliptical*, *Spinning*, *Weightlifting*, *Swimming*, *Hiking*, *Playing Tennis*, *CrossFit*, *Pilates*, *Stairclimber*, *Dancing*, *Indoor climbing*, *Golf*, *Skiing*, *Snowboarding*, and *Kayaking*. In total, 104,086 activities were sampled from 46,199 people. The mean duration per activity was 66 minutes (min: 20 minutes, max: 360).

4. Learning to `AIM` with Adaptive Inherited Masking

Motivation. As sensor data frequently exhibits inherent missingness, our key idea is to inherit these missingness patterns to be used in conjunction with a masked pre-training framework [30]. These methods introduce an artificial mask on the present data and learn to reconstruct them. Artificial missingness sits in contrast to inherited missingness inherent to the data. Similar to the original MAE work [29], our method implements an transformer-based encoder-decoder structure.

Our method first takes an input matrix of sensor features, which are then tokenized to be $\mathbf{X} \in \mathbb{R}^{B \times N \times E}$ (B is batch size, N is number of tokens, and E is embedding dimension). We then define a binary vector mask, $\mathbf{M} \in \{0, 1\}^{B \times N}$ (where 1 is masked and 0 is non-masked) equal in length to the number of tokenized sensor inputs, where masked tokens are ignored by the encoder. Our method sets $\mathbf{M}$ as the union of the inherited and artificial masks such that:

$$\mathbf{M} = \mathbf{M}^{\text{inherited}} \vee \mathbf{M}^{\text{artificial}}$$

The inherited mask, $\mathbf{M}^{\text{inherited}}$, is the original, existing missingness present in the dataset. The artificial mask, $\mathbf{M}^{\text{artificial}}$, is a simulated missingness on observed data. Critically, this inclusion of the inherited mask ensures that the encoder exclusively learns representations from reliable sensor data without contamination from imputation artifacts.

Background. The original MAE work [30] implements masking through *dropout removal*, where masked tokens are not passed through the encoder. Specifically it assumes that a fixed number of tokens D are dropped for every sample, such that $\sum_{n=1}^{N} \mathbf{M}_{[b,n]} = D \; \forall \; b \in [1, B]$. The transformer encoder input can then be formulated as $\mathbf{X}_{[\mathbf{M},:]} \in \mathbb{R}^{B \times (N-D) \times E}$. This reduces the transformer's computational complexity from $O(N^2) \rightarrow O((N-D)^2)$, which translates to 25x less computation when masking 80% of tokens

Table 1 | **Capabilities of Different Masking Implementations.** We combine dropout removal's efficiency [30] with attention masking's flexibility [23] to allow us to process to long sequences with inherited masks that have varying mask %.

Masking	Complexity	Fixed Mask %	Dynamic Mask %	Allows Inherited Mask
Dropout Removal [30]	$O((N-D)^2)$	✓	✗	✗
Attention Mask [23]	$O(N^2)$	✓	✓	✓
`AIM` (ours)	$O((N-D)^2)$	✓	✓	✓

N: Number of tokens D: Number dropped

Figure 4 | **LSM-2 Pre-training with AIM [A-F] and Evaluation [G,H]**. Our mask is a union of **[A]** inherited missingness from real-world noise and **[B]** artificial masking of observed data. Both are modeled with identical, learnable tokens. Because the inherited mask introduces variable masking, **[C]** we first remove D (size of artificial mask) tokens and **[D]** then use an attention mask to remove the remaining. **[E]** Dropped tokens are reinserted before **[F]** the final reconstruction. **[G]** Reconstruction error is computed only on artificial masks with known ground truth. **[H]** For predictive evaluations, a linear probe is trained on a pooled representation of the non-missing data.

($D = 0.8N$). While efficient, this approach requires fixed masking amount D, in order to construct batched encoder input $\mathbf{X}_{[\mathbf{M},:]}$ with $B > 1$. The motivation of our AIM approach is to include inherited masking in the MAE procedure in order to model real-world missingness. However, we are unable to do so with dropout removal because the amount of pre-existing missingness will vary, and causing the D of the inherited mask also vary. Recent methods have attempted to handle variable masking [23] by utilizing the transformer's *attention masking* mechanism [64]. While flexible, these methods fail to use dropout removal, making them computationally prohibitive for long sequences and large scale pre-training.

Adaptive Attentive Masking Design. Our key insight is to combine both masking modes in a unified approach: we maintain dropout removal's efficiency while incorporating the flexibility of attention masking. This hybrid strategy is visualized in Figure 4. Dropout removal limits the number of tokens that must be encoded to the lower bound of artificially masked tokens. This is because the set of dropped tokens D is static. In scenarios where a sample has no inherent missing data, these dropped tokens must be entirely defined by the artificial mask. In practice, dropped-out tokens can be a mix of inherited and artificially masked tokens. Similarly, the remaining masked tokens, which are disregarded using the transformer's attention mask, can also be of either type. This fusion provides the benefits of both paradigms while mitigating their individual limitations.

Unified Framework for Pre-training and Evaluation. AIM provides a unified framework for LSM-2 that consistently handles missing data during both pre-training and inference. The full pre-training procedure can be seen in Figure 4 [A-G]. During pre-training, the adaptive masking not only enables the inclusion of varying inherited mask sizes, but also allows the artificial masking to include a mix of strategies with differing masking percentages. Our artificially masking mix seeks to model the real-world missingness patterns shown in Figure 2. The mix includes (1) 80% random imputation masking (to model noise), where a random patch is masked, (2) 50% temporal slice masking (to model off body), where all sensors at a random time point are masked, and (3) 50% signal slice masking (to model sensor off), where all time points for a random sensor are masked. Each instance uses a randomly selected masking strategy with equal probability. The specific masking

percentages were identified via an ablation study, reported within the Appendix. As such, we set $D = 0.5N$, boosting our computational efficiency by 4x.

Crucially, `AIM`'s adaptive masking is also used during evaluation, which can be seen in Figure 4 [G,H]. The pre-trained model is then able to operate directly on incomplete multimodal sensor data by dynamically attending only to observed segments. This eliminates the need for external preprocessing, such as imputing or discarding missing values, and ensures generalization from pre-training to downstream deployment in real-world settings.

5. Experiments

Pre-training Set-up. We pre-train LSM-2 on minutely multimodal wearable data ($\mathbf{A} \in \mathbb{R}^{T \times S}$) where $S = 26$ sensor features and $T = 1440$ minutes. Inputs are tokenized using a ViT-1D [21, 1] encoder with a 1D patch size of 10 minutes, resulting in 3744 tokens (144 tokens per signal). We apply a shared kernel across channels and use a 2D positional embedding to encode time and signal identity. The model has 25M parameters, 384-d hidden size, 12 encoder layers, and 4 decoder layers. Following Section 4, we apply a composite mask (80% random, 50% temporal, 50% signal slices) and optimize mean squared error over masked patches on reconstruction. Notably, we do not back-propagate on missing pixels for any of the SSL methods trained including baselines. Training is performed on 8x16 Google v5e TPUs with a batch size of 512 for 100K steps. SSL baselines—LSM-1 [42], SimCLR [13], DINO [11], and MSN [6]—are trained from scratch using the same setup unless otherwise noted. LSM-1 uses a ViT-2D with a (10,2) patch size and 0.8 random masking, while the contrastive methods rely on jittering, scaling, and time-flipping augmentations [59, 38, 74, 51]. All baselines use imputed data to meet their full-input requirement. See Appendix for further implementation details.

Downstream Evaluation. We evaluate LSM-2 across three downstream targets: generative, classification, and regression. For `generative`, we assess reconstruction under structured missingness patterns: (1) random imputation (30%, 50%, 80%), (2) temporal interpolation (contiguous masked windows of 10, 30, or 60 minutes), (3) temporal extrapolation (masked window at the end of the sequence), and (4) signal imputation (masking 2/26, 6/26, or 12/26 channels). Since contrastive baselines lack reconstruction objectives, we compare against LSM-1 [42] in addition to simple imputation methods used in practice—Linear Interpolation, Nearest Neighbors, and Mean Filling—under the same union masking scheme. We omit MICE [63] due to its missingness at random assumptions not holding and its lower performance in prior work [42]. For `classification`, we average embeddings over non-inherited-masked tokens and apply a trainable linear probe; LSM-1 pools across all tokens, and contrastive methods use the CLS token. We report F_1, Accuracy, Balanced Accuracy, and AUROC on targets including hypertension, anxiety (Metabolics dataset; see Section 3), and 20-class activity recognition (Activity dataset). For `regression`, we follow the same setup with a linear regression probe and report MAE and Pearson correlation on BMI and age (Metabolics dataset). See Appendix for further details.

6. Results and Discussion

Generalizability across classification, generative, and regression tasks. LSM-2 with `AIM` learns a strong generalizable representation, useful for classification, regression and generative tasks (Tables 2, 3, 4 respectively). This research presents preliminary findings and should not be interpreted as providing diagnostic tools or recommendations.

Due to our improved pre-training reconstruction objective, LSM-2 obtains much stronger generative results compared to the prior state-of-the-art work - LSM-1 [42] which was limited in its masking

Table 2 | **Classification Task Results**

		Hypertension (2)				Anxiety (2)				Activity Recognition (20)			
	Method	↑F_1	↑Acc	↑BAcc	↑AUC	↑F_1	↑Acc	↑BAcc	↑AUC	↑F_1	↑Acc	↑BAcc	↑AUC
ST	ResNet	0.516	0.529	0.587	0.624	0.645	0.655	0.651	0.709	**0.729**	**0.721**	**0.734**	**0.965**
ST	ViT1D	0.481	0.516	0.509	0.520	0.583	0.597	0.586	0.620	0.351	0.367	0.374	0.863
LP	SimCLR	0.501	0.524	0.548	0.568	0.594	0.603	0.601	0.636	0.098	0.109	0.124	0.603
LP	DINO	0.487	0.536	0.504	0.510	0.551	0.557	0.562	0.582	0.102	0.110	0.124	0.635
LP	MSN	0.512	0.553	0.538	0.552	0.579	0.585	0.588	0.622	0.108	0.118	0.125	0.662
LP	LSM-1	0.640	0.676	0.682	0.739	0.670	0.678	0.678	0.743	0.470	0.470	0.489	**0.900**
LP	LSM-2	**0.651**	**0.687**	**0.693**	**0.754**	**0.683**	**0.690**	**0.692**	**0.758**	**0.474**	**0.472**	**0.493**	0.899
	Δ LSM-1	+1.7%	+1.6%	+1.6%	+2.0%	+1.9%	+1.8%	+2.1%	+2.0%	+0.8%	+0.4%	+0.8%	-0.1%

Metrics: F_1 Score, Accuracy, Balanced Accuracy, AUROC with Macro One-vs-Rest | Tasks: 20-class Activity Recognition, rest are binary | Methods: Supervised Training (ST), Linear Probe (LP).

Table 3 | **Generative Task Results**

	↓Random Imp.			↓Temporal Interp.			↓Temporal Extrap.			↓Signal Imp.		
Method	30%	50%	80%	10m	30m	60m	10m	30m	60m	2	6	12
Linear Int.	0.57	0.62	0.74	0.42	0.56	0.70	0.47	0.64	0.82	NA	NA	NA
NN Fill	0.70	0.76	0.90	0.52	0.69	0.84	0.47	0.64	0.82	NA	NA	NA
Mean Fill	0.92	0.96	0.93	0.79	0.80	0.84	0.78	0.80	0.83	1.28	1.30	1.29
LSM-1	0.21	0.24	0.30	0.49	0.55	0.60	0.45	0.52	0.56	0.73	0.58	0.45
LSM-2	**0.18**	**0.20**	**0.20**	**0.26**	**0.37**	**0.45**	**0.28**	**0.38**	**0.48**	**0.17**	**0.21**	**0.27**
Δ LSM-1	+14%	+17%	+33%	+47%	+31%	+25%	+38%	+27%	+14%	+77%	+64%	+40%

Metrics: Mean Squared Error | Tasks: Random Imputation (30%, 50%, 80% missing), Temporal Interpolation/Extrapolation (10, 30, 60 missing minutes), Signal Imputation (2, 6, or 12 out of 26 missing modalities) | Methods: Linear interpolation, Nearest neighbor fill, Mean filling

strategy (artificial random imputation masking). By introducing a mixture of artificial masking strategies with flexible missing ratios, as well as the inclusion of the inherited mask, not only do we achieve a **+33% performance increase** on the 80% random imputation evaluation, but we also achieve strong benefits across different generative tasks, with **+77% improvement** in 2 signal imputation and a **+47% improvement** in 10 minute temporal interpolation. This demonstrates that explicitly modeling diverse missingness patterns during pre-training leads to more robust representations that generalize better to real-world scenarios with complex data gaps.

Despite being pre-trained on with a reconstruction objective, LSM-2 achieves SOTA performance across classification tasks, **beating all other self-supervised learning baselines**. Even with a simple linear probe and frozen features, our model surpasses fully supervised baselines on hypertension and anxiety prediction — two challenging tasks that previously required hand-crafted features or custom architectures [55, 2]. This suggests that pre-training helps avoid overfitting and enables the model to capture subtle physiological cues that generalize across conditions. The strong results across both binary (hypertension/anxiety) and multi-class (activity recognition) tasks indicate that the model learns hierarchical

Table 4 | **Regression Task Results**

		Age		BMI	
	Method	↓MAE	↑Corr	↓MAE	↑Corr
ST	ResNet	7.43	0.618	5.07	0.515
ST	ViT1D	9.65	0.132	6.06	0.047
LP	SimCLR	9.21	0.345	5.85	0.235
LP	DINO	9.69	0.112	5.97	0.122
LP	MSN	9.42	0.255	5.84	0.250
LP	LSM-1	**6.41**	**0.728**	4.39	0.667
LP	LSM-2	6.49	0.722	**4.38**	**0.673**
	Δ LSM-1	-1.2%	-0.8%	+0.2%	+1.0%

Metrics: Mean Absolute Error, Pearson Correlation | Methods: Supervised Training (ST), Linear Probe (LP).

Figure 5 | **Scaling Performance of Our Model.** LSM-2 model achieves better scaling than LSM-1 across all dimensions: *subjects*, *data*, *compute*, and *model size*. LSM-2 uses a mixed masking strategy during pre-training, but here we report only random imputation loss to match LSM-1.

features suited to different levels of task complexity.

In regression tasks, LSM-2 improves correlation on BMI by +1.0%, while underperforming on age prediction by -0.8%. Since the absolute metric (e.g., mean absolute error) is affected by differing target scales (e.g., Age: 18–90 vs. BMI: 12–65), correlation offers a clearer view of model quality.

Strong scaling performance on 40 million hours of incomplete data. Figure 5 show that our `AIM` scales more effectively than the LSM-1 model across 4 different dimensions: subject, data, compute, and model. The LSM-1 model exhibits scaling saturation for the data and compute dimensions, but our model's trend indicates a more aggressive downwards slope that has not yet saturated. These results are promising as they suggest that continued investment in larger datasets and compute may yield further performance gains, indicating that our method has not yet reached its limits.

Strong Robustness to Targeted Missingness. LSM-2 with `AIM` demonstrates substantially greater resilience to targeted missingness compared to prior work, as seen in Figure 6. Across 11 out of 12 missingness scenarios, our model consistently maintains stronger performance. For example, when accelerometry is removed—a key sensor for activity recognition—our model's F_1 score drops from 0.47 to 0.20 (-57%), while LSM-1 degrades more severely from 0.47 to 0.14 (-71%). Notably, even in this degraded setting, our model still outperforms LSM-1 by **+47%** in absolute terms. A similar trend holds across other modalities: removing PPG during hypertension prediction leads to only a -6% drop for `AIM` (0.65 to 0.61), compared to -11% for LSM-1 (0.64 to 0.57).

Robustness also generalizes across temporal ablations. While both models reach similar peak activity recognition scores (~0.47 F_1), our model maintains an average F_1 of 0.43 across temporal ablations—substantially higher than LSM-1's 0.26 (**+65%** relative gain). Overall, these results validate the effectiveness of our adaptive masking strategy in modeling missingness patterns. Our model experiences **73% smaller performance drops** across all 12 ablation settings and retains **+15% higher** absolute performance in degraded states. This combination of robustness and accuracy makes `AIM` a more reliable choice for real-world deployment, where missing data is a reality.

Reflects Physiological Domain Knowledge and Other Real-world Implications. The targeted missingness experiments in Figure 6 also reveal clinically coherent patterns with real-world implications. LSM-2's hypertension and anxiety predictions show the expected nocturnal advantage, such that the removal of nighttime signals has 5% degradation in F_1 for both targets, compared to an average 0.4% and 0.01% degradation for the daytime windows for each target. This finding strongly aligns with clinical literature demonstrating the diagnostic value of nighttime biosignals for hypertension [71, 28] and stress prediction [36, 25], which are less affected by daily activity artifacts and better capture underlying pathophysiology.

Interestingly, LSM-2 also demonstrates a large 11% drop in performance for anxiety prediction

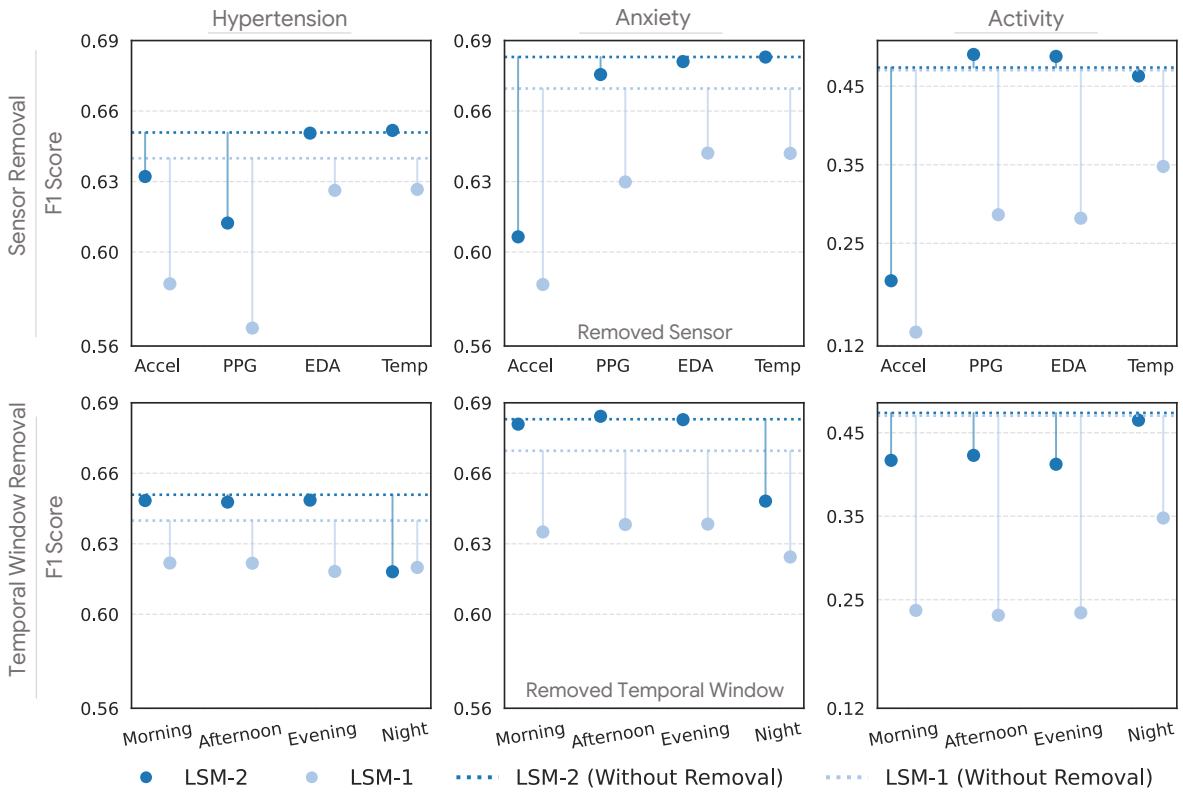

Figure 6 | **Robustness to Targeted Missingness.** In sensor removal, all signals derived from the specific sensor are removed. In temporal window removal, all signals are removed at a given timeframe (Morning [8am-12pm], Afternoon [12pm-4pm], Evening [4pm-8pm], Night [8pm-8am]). The dotted line denotes a model trained on all modalities. When evaluating with simulated sensor- or time-specific missingness, LSM-2 maintains consistent performance while LSM-1 degrades significantly. Where LSM-2 does show sensitivity, it aligns with domain knowledge. For example, nighttime BP's stronger predictive power of hypertension over daytime [28], accelerometry's role in distinguishing anxiety from physiological stress responses [54].

after removing the accelerometry sensor, whereas removing the other sensors only results in an average 0.5% drop. This suggests accelerometry provides unique signals for anxiety detection that are not captured by other modalities. There have been recent research works [54, 67] that demonstrate the importance of utilizing accelerometry sensors in stress prediction in order to distinguish anxiety and mental stress from physiological stress responses from physical activity.

These results demonstrate three key advantages of our `AIM` adaptive masking approach: (1) performance degrades proportionally to a sensor's clinical importance, (2) cross-modal relationships are maintained when inputs are missing, and (3) known temporal biases in physiological data are preserved. This robustness is crucial for real-world deployment where missing data is inevitable, making `AIM` significantly more reliable in field settings.

Importance of Inheritance and Mask Mixing. `AIM` is composed of two main components: (1) inclusion of an Inherited Mask and (2) usage of a mix of artificial masking with randomly using either 80% random imputation, 50% temporal slices, or 50% signals slices. In Table 5, we show how removing inheritance leads to performance degradation across

Table 5 | **Ablation Study**

	Generative (↓MSE)		Classification (↑F_1)	
	80% R.Imp.	60m T.Interp.	Anxiety	Activity
`AIM`	0.20	0.45	0.683	0.474
w/o Inheritance	0.28	0.62	0.671	0.445
w/o Mixing	0.19	0.58	0.637	0.460

all of the various tasks. Without mixing, only an 80% random imputation pre-training task is used, matching prior work[42]. While the random imputation performance improves, all other tasks degrade, including the other generative task, temporal interpolation.

Limitations and Future Work. Our study has several important constraints. First, training and evaluation were limited to a specific private datasets, necessitating future work on exploring other datasets with complex missingness patterns, such as All of Us [33], and understanding missingness distribution shifts. Furthermore, we make use of minutely aggregated features, which is helpful for helping us model our 1-day longer time-scale day data, but uncommon in the broader wearable sensing space, which focuses primarily on raw high frequency sensor signal. Unfortunately, this is a practical limitation, as data is not stored in its raw form at such scale. Finally, although the focus of our work is on multimodal sensor data, our technique is broadly applicable and domain-agnostic requiring only that the data contains existing missingness. Therefore, future work can explore the application of our `AIM` across different missingness-afflicted domains.

7. Conclusion

In this work, we introduced the second generation of Large Sensor Model (LSM-2) with Adaptive and Inherited Masking, `AIM`, a novel self-supervised learning approach designed to learn robust representations directly from incomplete wearable sensor data. By integrating both inherited (real-world) and artificial masking strategies, AIM eliminates the need for explicit imputation while effectively modeling the pervasive missingness in real-world sensor data. Our experiments demonstrate that our foundation model LSM-2, pre-trained with `AIM`, achieves state-of-the-art performance and scaling capability across a diverse range of tasks across differing semantics. Our targeted missingness experiments reveal that LSM-2 maintains strong performance even when entire sensors are dropped, suggesting broad applicability to scenarios with varying sensor availability. Our model's strong performance under real-world missingness conditions demonstrates its practical applicability, and we hope the insights in our work will guide future work in machine learning methodologies for wearable sensors and health time-series.

References

[1] S. Abbaspourazad, O. Elachqar, A. C. Miller, S. Emrani, U. Nallasamy, and I. Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

[2] A. Abd-Alrazaq, R. AlSaad, S. Aziz, A. Ahmed, K. Denecke, M. Househ, F. Farooq, and J. Sheikh. Wearable artificial intelligence for anxiety and depression: scoping review. *Journal of Medical Internet Research*, 25:e42672, 2023.

[3] A. Afdala, N. Nuryani, and A. S. Nugroho. Automatic detection of atrial fibrillation using basic shannon entropy of rr interval feature. In *Journal of Physics: Conference Series*, volume 795, page 012038. IOP Publishing, 2017.

[4] M. Amiri and R. Jensen. Missing data imputation using fuzzy-rough methods. *Neurocomputing*, 205:152–164, 2016.

[5] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[6] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pages 456–473. Springer, 2022.

[7] S. Bähr, G.-C. Haas, F. Keusch, F. Kreuter, and M. Trappmann. Missing data and other measurement quality issues in mobile geolocation sensor data. *Social Science Computer Review*, 40(1):212–235, 2022.

[8] N. Beebe-Wang, S. Ebrahimi, J. Yoon, S. O. Arik, and T. Pfister. Paits: pretraining and augmentation for irregularly-sampled time series. *arXiv preprint arXiv:2308.13703*, 2023.

[9] G. Bleser, D. Steffen, A. Reiss, M. Weber, G. Hendeby, and L. Fradet. Personalized physical activity monitoring using wearable sensors. *Smart health: Open problems and future challenges*, pages 99–124, 2015.

[10] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[12] L.-W. Chang, C.-T. Li, C.-P. Yang, and S.-d. Lin. Learning on missing tabular data: Attention with self-supervision, not imputation, is all you need. *ACM Transactions on Intelligent Systems and Technology*.

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

[14] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

[15] H. ChuDuc, K. NguyenPhan, and D. NguyenViet. A review of heart rate variability and its applications. *APCBEE procedia*, 7:80–85, 2013.

[16] Z. Dai, E. Getzen, and Q. Long. Sadi: Similarity-aware diffusion model-based imputation for incomplete temporal ehr data. In *International Conference on Artificial Intelligence and Statistics*, pages 4195–4203. PMLR, 2024.

[17] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

[18] T. Decorte, S. Mortier, J. J. Lembrechts, F. J. Meysman, S. Latré, E. Mannens, and T. Verdonck. Missing value imputation of wireless sensor data for environmental monitoring. *Sensors*, 24(8):2416, 2024.

[19] C. M. DeGiorgio, P. Miller, S. Meymandi, A. Chin, J. Epps, S. Gordon, J. Gornbein, and R. M. Harper. Rmssd, a measure of vagus-mediated heart rate variability, is associated with risk factors for sudep: the sudep-7 inventory. *Epilepsy & behavior*, 19(1):78–81, 2010.

[20] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[22] J. Du, M. Hu, and W. Zhang. Missing data problem in the monitoring system: A review. *IEEE Sensors Journal*, 20(23):13984–13998, 2020.

[23] T. Du, L. Melis, and T. Wang. Remasker: Imputing tabular data with masked autoencoding. *arXiv preprint arXiv:2309.13793*, 2023.

[24] L. Ericsson, H. Gouk, and T. M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5414–5423, 2021.

[25] J. Fan, J. Mei, Y. Yang, J. Lu, Q. Wang, X. Yang, G. Chen, R. Wang, Y. Han, R. Sheng, et al. Sleep-phasic heart rate variability predicts stress severity: Building a machine learning-based stress prediction model. *Stress and Health*, 40(4):e3386, 2024.

[26] E. Ford, P. Rooney, P. Hurley, S. Oliver, S. Bremner, and J. Cassell. Can the use of bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? development of a novel method using simulated and real-life clinical data. *Frontiers in Public Health*, 8:54, 2020.

[27] S. Haneuse, D. Arterburn, and M. J. Daniels. Assessing missing data assumptions in ehr-based studies: a complex and underappreciated task. *JAMA Network Open*, 4(2):e210184–e210184, 2021.

[28] T. W. Hansen, Y. Li, J. Boggia, L. Thijs, T. Richart, and J. A. Staessen. Predictive role of the nighttime blood pressure. *Hypertension*, 57(1):3–10, 2011.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[32] A. Hecht, S. Ma, J. Porszasz, R. Casaburi, C. C. R. Network, et al. Methodology for using long-term accelerometry monitoring to describe daily activity patterns in copd. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 6(2):121–129, 2009.

[33] H. Jeong, A. Roghanizad, H. Master, and et al. Data from the All of Us research program reinforces existence of activity inequality. *npj Digital Medicine*, 8(8), 2025.

[34] J. Jungo, Y. Xiang, S. Gashi, and C. Holz. Representation learning for wearable-based applications in the case of missing data. *arXiv preprint arXiv:2401.05437*, 2024.

[35] M. Katori, S. Shi, K. Ode, Y. Tomita, and H. Ueda. The 103,200-arm acceleration dataset in the uk biobank revealed a landscape of human sleep phenotypes. *Proceedings National Academy of Science, U.S.A.*, 119(12), 2022.

[36] H. Kinnunen, A. Rantanen, T. Kenttä, and H. Koskimäki. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal hr and hrv assessed via ring ppg in comparison to medical grade ecg. *Physiological measurement*, 41(4):04NT01, 2020.

[37] Z. Li, Z. Rao, L. Pan, P. Wang, and Z. Xu. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023.

[38] Z. Liu, A. Alavi, M. Li, and X. Zhang. Guidelines for augmentation selection in contrastive learning for time series classification. *arXiv preprint arXiv:2407.09336*, 2024.

[39] M. McDermott, B. Nestor, E. Kim, W. Zhang, A. Goldenberg, P. Szolovits, and M. Ghassemi. A comprehensive ehr timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 257–278, 2021.

[40] S. Mekruksavanich, A. Jitpattanakul, K. Sitthithakerngkiet, P. Youplao, and P. Yupapin. Resnet-se: Channel attention-based deep residual network for complex activity recognition using wrist-worn wearable sensors. *IEEE Access*, 10:51142–51154, 2022.

[41] Y. Na, M. Park, Y. Tae, and S. Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

[42] G. Narayanswamy, X. Liu, K. Ayush, Y. Yang, X. Xu, S. Liao, J. Garrison, S. Tailor, J. Sunshine, Y. Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

[43] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, and S. Patel. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7914–7924, 2024.

[44] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[45] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[46] Y.-C. Pan, B. Goodwin, E. Sabelhaus, K. M. Peters, K. F. Bjornson, K. L. Pham, W. Walker, and K. M. Steele. Feasibility of using acceleration-derived jerk to quantify bimanual arm use. *Journal of NeuroEngineering and Rehabilitation*, 17:1–8, 2020.

[47] A. Pillai, D. Spathis, F. Kawsar, and M. Malekzadeh. Papagei: Open foundation models for optical physiological signals. *International Conference on Learning Representations (ICLR)*, 2025.

[48] I. M. Pires, F. Hussain, N. M. Garcia, and E. Zdravevski. Improving human activity monitoring by imputation of missing sensory data: Experimental study. *Future Internet*, 12(9):155, 2020.

[49] R. Qin and Y. Wang. Imputegan: Generative adversarial network for multivariate time series imputation. *Entropy*, 25(1):137, 2023.

[50] M. M. Rahman, N. Ali, R. Bari, N. Saleheen, M. al'Absi, E. Ertin, A. Kennedy, K. L. Preston, and S. Kumar. mDebugger: Assessing and diagnosing the fidelity and yield of mobile sensor data. In *Mobile Health: Sensors, Analytic Methods, and Applications*, chapter 7, page 121–143. 2017.

[51] C. Rommel, J. Paillard, T. Moreau, and A. Gramfort. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020, 2022.

[52] M. Saha, M. A. Xu, W. Mao, S. Neupane, J. M. Rehg, and S. Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *arXiv preprint arXiv:2502.01108*, 2025.

[53] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[54] M. Sevil, M. Rashid, M. R. Askari, Z. Maloney, I. Hajizadeh, and A. Cinar. Detection and characterization of physical activity and psychological stress from wristband data. *Signals*, 1(2):188–208, 2020.

[55] G. F. Silva, T. P. Fagundes, B. C. Teixeira, and A. D. Chiavegatto Filho. Machine learning for hypertension prediction: a systematic review. *Current hypertension reports*, 24(11):523–533, 2022.

[56] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pages 245–248. IEEE, 2012.

[57] D. Spathis, I. Perez-Pozuelo, S. Brage, N. J. Wareham, and C. Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 69–78, 2021.

[58] B. Srimedha, R. N. Raj, and V. Mayya. A comprehensive machine learning based pipeline for an accurate early prediction of sepsis in icu. *Ieee Access*, 10:105120–105132, 2022.

[59] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[60] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[61] S. Tonekaboni, D. Eytan, and A. Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

[62] J. Truslow, A. Spillane, H. Lin, K. Cyr, A. Ullal, E. Arnold, R. Huang, L. Rhodes, J. Block, J. Stark, et al. Understanding activity and physiology at scale: The apple heart & movement study. *npj Digital Medicine*, 7(1):242, 2024.

[63] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[65] H. Wei, M. A. Xu, C. Samplawski, J. M. Rehg, S. Kumar, and B. M. Marlin. Temporally multi-scale sparse self-attention for physical activity data imputation. *Proceedings of machine learning research*, 248:137, 2024.

[66] J. M.-T. Wu, M.-H. Tsai, S.-H. Xiao, and Y.-P. Liaw. A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17, 2020.

[67] M. Wu, H. Cao, H.-L. Nguyen, K. Surmacz, and C. Hargrove. Modeling perceived stress via hrv and accelerometer sensor streams. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1625–1628. IEEE, 2015.

[68] M. Xu, A. Moreno, S. Nagesh, V. Aydemir, D. Wetter, S. Kumar, and J. M. Rehg. Pulseimpute: A novel benchmark task for pulsative physiological signal imputation. *Advances in Neural Information Processing Systems*, 35:26874–26888, 2022.

[69] M. A. Xu, A. Moreno, H. Wei, B. M. Marlin, and J. M. Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. *arXiv preprint arXiv:2311.00519*, 2023.

[70] M. A. Xu, J. Narain, G. Darnell, H. Hallgrimsson, H. Jeong, D. Forde, R. Fineman, K. J. Raghuram, J. M. Rehg, and S. Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. *arXiv preprint arXiv:2411.18822*, 2024.

[71] G. Yilmaz, X. Lyu, J. L. Ong, L. H. Ling, T. Penzel, B. T. Yeo, and M. W. Chee. Nocturnal blood pressure estimation from sleep plethysmography using machine learning. *Sensors*, 23(18):7931, 2023.

[72] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[73] H. Yuan, S. Chan, A. P. Creagh, C. Tong, A. Acquah, D. A. Clifton, and A. Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

[74] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

[75] Y. Zhou, J. Shi, R. Stein, X. Liu, R. N. Baldassano, C. B. Forrest, Y. Chen, and J. Huang. Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. *Journal of the American Medical Informatics Association*, 30(7):1246–1256, 2023.

[76] T. Zielinski, A. M. Moore, E. Troup, K. J. Halliday, and A. J. Millar. Strengths and limitations of period estimation methods for circadian data. *PloS one*, 9(5):e96462, 2014.

Appendix — LSM-2: Learning from Incomplete Sensor Data

Table of Contents

A.1. Data Details

A.1.1. Imputing Missingness for Non AIM Models

Although AIM is able to organically handle existing missing values using clever masking, the same cannot be said for our baseline methods. Furthermore, many standard deep learning frameworks (such as pytorch, jax, and tensorflow) are unable to handle nan values in model training and evaluation, causing value errors or propogating nans throughout the network during forward and backward passes. For this reason we impute missing (nan) values in our data. We use linear interpolation between gaps and then back and forward fill for missingness at the start and end of the sequence.

A.1.2. Device Details

There are many different types of smartwatches and fitness trackers. Fig. 7 shows the distribution of different trackers and smartwatches present in our pretraining dataset. Given the scale of our dataset we are able to train on examples of data from many different devices. Consequently, our model demonstrates robustness across diverse device types, handling their varying sensor technologies and differing inherent missingness patterns.

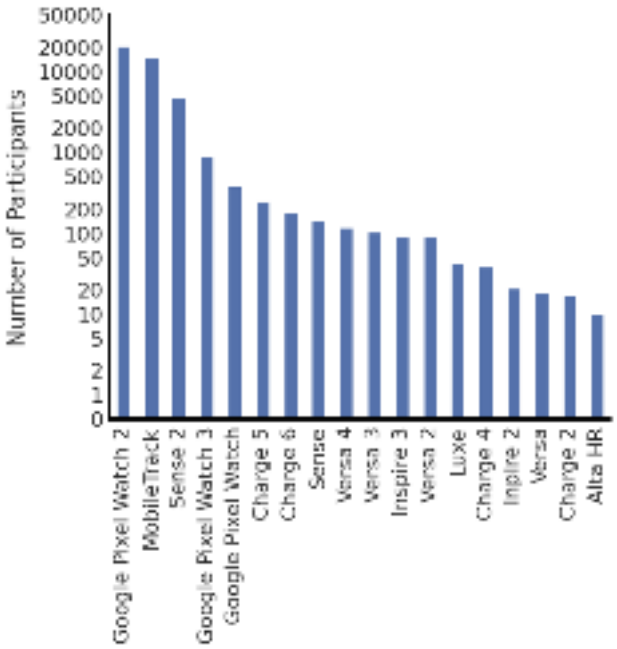

Figure 7 | **Device Distribution.** The count of each fitness tracker present in our pre-training dataset.

A.1.3. Sensor Derived Minutely Features

Our wearable devices utilize 5 different sensors: Photoplethysmography, Accelerometer, Skin Conductance (electrodermal activity or EDA), Temperature, and Altitude. Each of these sensors collects raw waveform signals at 100 Hz, 25 Hz, 200 Hz, 6 Hz, amd 10 Hz respectively, but we do not use the signals at this high resolution because (1) due to practical reasons (i.e. prohibitive storage costs and battery drain), data is not stored in this raw form at our scale, and (2) it is computationally impractical to learn models on raw waveforms across an entire day (i.e. 200 Hz for 1 day is T = 17million time-points, per instance). As such, various features are curated from the raw waveforms as minutely aggregrated features and saved to be used as inputs into our model. Each of these features are

grounded in the domain literature, based on prior work that has shown their clinical effectiveness. For example, heart rate variability metrics like RMSSD [19] or Shannon Entropy of RR intervals [3] have well-established prognostic value for cardiovascular health, while accelerometry features like jerk ratio [46] effectively characterize movement quality.

Each of the derived features, as well as their base sensor origin, can be found in Table 6 below. For the targeted sensor removal experiments, as well as any other descriptions of the sensor as a whole, *we refer to the sensor as all features derived from the sensor*. For example, when removing the PPG sensor in the targetted missingness experiment, we remove all PPG-derived features, from Heart Rate to Shannon Entropy RR Differences.

A.1.4. Demographic Breakdown

A statistical breakdown of our datasets, by demographic features can be found in Table 7. A subset of these, age and BMI, represent two of the regression tasks used to validate our method.

A.1.5. Discriminative Task Label Breakdown

Table 8 shows label and data breakdown of the discriminative tasks used to validate our method. These tasks include 20-class activity recognition (Table 8(a)) from the activity dataset, and binary anxiety and hypertension classification (Table 8(b.i)) from the metabolic dataset.

A.1.6. Acquisition and Approval

The data used for training in our analysis was curated from a large corpus of historical wearable data collected with consent from partcipants for these data to be used in research. Specifically, the consent language described use of the data for developing new health features and algorithms and being included in publications:

REDACTED will collect and use your data to research and develop new health and wellness products and services for you and others. This data includes your: Health and wellness data, such as steps, heart rate, and sleep data. Your data may also be used to generate findings that could be included in publications (such as scientific journals) to contribute to general knowledge about health and science. For example, activity, heart rate, and sleep data contributed to published findings that Fitbit devices could help detect flu outbreaks. None of the data used for these purposes will include your name, email, or other information that directly identifies you.

The use of data for pretraining in this manner was approved as exempt under 45 CFR § 46.104(d)(4) "*because the research involves the use of identifiable private information/biospecimens; and information, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects.*"

The Metabolic downstream dataset for anxiety and hypertension prediction came from an IRB approved study (protocol number removed for anonymization). The core objective of this study as described in the IRB protocol was to: "*Evaluate the feasibility of using the data provided by wrist-worn wearable devices to develop algorithms and scores to assess metabolic health.*"

In the consent for the observational study, participants were informed that data on up to 7,500 participants in the United States would be collected. We used a mobile study platform that allows participants to enroll, check eligibility and provide full informed consent. The same mobile application

Table 6 | **Sensor Feature Definitions and the Sensor they are Derived From.**

Feature	Unit	Definition
Photoplethysmography		
Heart Rate	Beats/Min	Mean of instantaneous heart rate.
Heart Rate at Rest	Beats/Min	Mean of heart rate at rest.
RR Percent Valid	%	% of 5-minute window with valid RR intervals.
RR 80^{th} Percentile	Msec	80^{th} percentile of 5-minute window of RR ints.
RR 20^{th} Percentile	Msec	20^{th} percentile of RR ints.
RR Median	Msec	Median RR interval.
RMSSD	Msec	Root mean squared st. dev. of RR ints.
SDNN	Msec	Standard deviation of RR intervals.
Shannon Ent. RR	Nats	Shannon entropy of the RR intervals.
Shannon Ent. RR Diffs	Nats	Shannon entropy of the RR interval differences.
Accelerometer		
Step Count	Steps	Number of steps.
Jerk Autocorrelation Ratio	a.u.	Ratio of lag=1 autocorrelation to energy in 1st 3-axis principal component.
Log Energy	a.u.	Log of sum of 3-axis root mean squared magnitude.
Covariance Condition	a.u.	Estimate of condition number for the 3-axis covariance.
Log Energy Ratio	a.u.	Log of ratio of sum of energy in 1st 3-axis principal component over energy of 3-axis root mean squared magnitude.
Zero Crossing St.Dev.	Seconds	Standard deviation of time between zero crossing of 1st 3-axis principal component.
Zero Crossing Average	Seconds	Mean of time between zero crossing of 1st 3-axis principal component.
Axis Mean	a.u.	Mean of 3-axis
Kurtosis	a.u.	Kurtosis of 3-axis root mean squared magnitude.
Sleep Coefficient	a.u.	Sum of 3-axis max-min range with 16 log-scaled bins.
Skin Conductance		
Skin Conductance Value	μSiemens	Center of linear tonic SCL value fit.
Skin Conductance Slope	μS/Min	Intraminute slope of SCL values.
Lead Contact Counts	Counts	Number of times sensor leads contacted the wrist in a minute.
Skin Temperature		
Skin Temperature Value	$^\circ$ C	Mean value of skin temperature.
Skin Temperature Slope	$^\circ$ C/Min	Slope of skin temperature.
Altimeter		
Altitude St.Dev. Norm	Hectopascals	Standard deviation of altimeter readings.

Table 7 | **Demographics of our Various Datasets.**

Category	Pre-training Train (%)	Pre-training Val (%)	Downstream Activity Train (%)	Downstream Activity Val (%)	Downstream Metabolic Train (%)	Downstream Metabolic Val (%)
Sex						
Male	37,352 (68.1)	3,657 (63.8)	27,653 (73.1)	6,092 (73.0)	551 (44.1)	258 (35.4)
Female	23,041 (38.1)	2,065 (36.0)	10,145 (26.8)	2,248 (26.9)	670 (53.6)	455 (62.4)
Not Specified	48 (0.1)	10 (0.2)	24 (0.1)	3 (0.1)	0 (0)	0 (0)
Age						
18–39	28,519 (47.2)	2,583 (45.1)	19,340 (51.1)	4,492 (53.8)	415 (33.2)	223 (30.6)
40–59	24,888 (41.2)	2,433 (42.4)	15,309 (40.5)	3,172 (38.0)	637 (51.0)	384 (52.7)
60–79	6,473 (10.7)	664 (11.6)	2,875 (7.6)	618 (7.4)	198 (15.8)	121 (16.6)
≥80	364 (0.6)	39 (0.7)	120 (0.3)	31 (0.4)	0 (0)	1 (0.1)
Not Specified	197 (0.3)	178 (0.5)	30 (0.4)	0 (0)	0 (0)	0 (0)
BMI						
Healthy (<25)	22,425 (37.1)	2,173 (37.9)	15,942 (42.2)	3,685 (44.2)	319 (25.5)	188 (25.8)
Overweight (25–30)	20,242 (33.5)	1,952 (34.1)	14,154 (37.4)	3,017 (36.2)	343 (27.4)	206 (28.6)
Obese (≥30)	14,799 (24.5)	1,330 (23.2)	6,131 (16.2)	1,316 (15.8)	481 (38.5)	274 (37.6)
Not Specified	230 (0.4)	14 (0.2)	81 (0.2)	18 (0.2)	49 (3.9)	28 (3.8)
Total	60,440 (100)	5,732 (100)	37,822 (100)	8,343 (100)	1,250 (100)	729 (100)

Table 8 | **Discriminative Task Dataset Distribution**

(a) **Activity Recognition Dataset**

Task / Label	Train (%)	Test (%)
Activity		
Walk	4,434 (6.0)	874 (5.8)
Bike	4,363 (5.9)	858 (5.6)
Sport	4,433 (6.0)	902 (5.9)
Run	4,023 (5.4)	790 (5.2)
Aerobics	4,417 (6.0)	906 (6.0)
Elliptical	4,402 (5.9)	879 (5.8)
Spinning	4,402 (5.9)	858 (5.6)
Weightlifting	4,335 (5.9)	841 (5.5)
Swim	4,280 (5.7)	867 (5.8)
Hike	4,062 (5.5)	841 (5.5)
Tennis	4,138 (5.6)	815 (5.4)
CrossFit	4,305 (5.8)	887 (5.8)
Pilates	4,365 (5.9)	846 (5.6)
Stairclimber	4,272 (5.8)	834 (5.5)
Dancing	4,288 (5.8)	826 (5.4)
Indoor climbing	3,520 (4.8)	853 (5.6)
Golf	3,003 (4.1)	710 (4.7)
Skiing	1,594 (2.1)	420 (2.8)
Snowboarding	662 (0.9)	167 (1.1)
Kayaking	732 (1.0)	212 (1.4)
Total	74,030 (100)	15,186 (100)

(b.i) **Metabolic Dataset** Classification Tasks

Task / Label	Train (%)	Test (%)
Anxiety		
Positive	55,030 (36.4)	34,749 (38.5)
Negative	96,316 (63.6)	55,437 (61.5)
Hypertension		
Positive	36,349 (24.0)	23,353 (25.9)
Negative	114,997 (76.0)	66,833 (74.1)
Total	151,346 (100)	90,186 (100)

enables the collection of Fitbit data using Fitbit devices or Pixel watches and allows participants to complete questionnaires. The participants reported their anxiety, depression and hypertension diagnoses through this app. Data was de-identified and stored in accordance with the approved IRB protocol. The participants were compensated with a free set of lab tests from Quest Diagnostics for participating in the study.

A.2. Missingness Visualizations

A core property of these data is that they are fragmented, and the missingness has several modal types. Three very common modes occur: 1) When the device is being charged or off all sensor stop recording data (device off), 2) when the device is in certain operation modes (e.g., when in sleep mode) certain signals stop being recorded (sensor off) and 3) when there is noise in the sensor data spurious values (e.g., values that are not physiologically possible - HR=0) are filtered out. The following sections demonstrate additional visualizations of the missingness patterns present from these mechanisms.

A.2.1. Additional Examples of Data with Existing Missingness

In order to demonstrate the ubiquity and broad range of missingness patterns found within the data, we randomly sample an additional 8 data examples, shown in Figure 8. These examples further demonstrate how some patterns are consistent across users, such as increased missingness during early morning hours (12am-6am) (reflecting device removal during sleep) or correlated missingness dropout across various sensor channels. However, it should be noted that all samples exhibit unique missingness signatures with no two patterns being identical with vastly differing missingness percentages (27-63%) and demonstrating the ubiquity of real-world missingness. These findings motivated our development of `AIM`'s flexible masking approach, which explicitly models such heterogeneous missingness patterns during pre-training.

A.2.2. Prevalence and Length of Missingness

In Figure 9, we demonstrate the prevalence of missingness as well as the length of the missingness, broken down across each sensor type across all 1.6 million instances of pre-training data. As we can see, each sensor has very different patterns of missingness, and across all sensors, their missingness presents as long extended gaps, making them non-trivial to reconstruct over. Notably, the accelerometry features in particular, have missingness in the form of these extended gaps, whereas most of the missingness for PPG sensors is of shorter length.

A.3. Pre-training Masking % Ablation Experiment

The adaptive component of our `AIM` methodologies allows for us to utilize a mix of artificial mask pre-training masking strategies. Each of these artificial masks are applied ontop of the existing, inherited mask. In order to model both dimensionalities of our data, across time and sensors, and the real-world missingness paradigms, we have a mix of 3 different artificial mask pre-training strategies:

1. Random Imputation Pre-training: Here we drop out a % of total tokens. This is useful for modeling sensor noise, in which random channels at random times will be missing.
2. Temporal Slice Pre-training: Here we drop out a % of total temporal slices, across all sensor channels. This is useful for modeling device off, in which, for a given period of time, all

Figure 8 | **Gallery of Data Examples with Real-world Missingness.** White designates missingness.

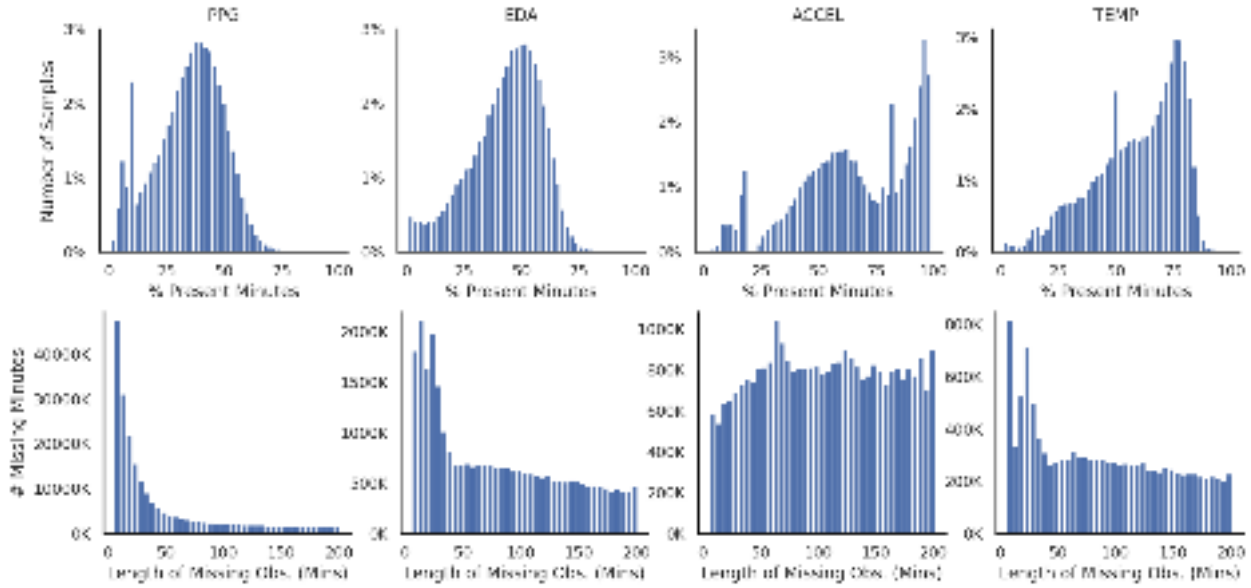

Figure 9 | **Distribution of Prevalence and Length of Missingness.**

sensors are off because the wearable device is off body. Here, we do not model it like temporal interpolation, in which the slices are necessarily contiguous. This is because, during pre-training, we would like to learn to reconstruction across a variable number of contiguous slices.

3. Sensor Slice Pre-training: Here we drop out a % of total sensor slices, across all time points. This is useful for modeling sensor off, in which a given sensor channel is off because of a non-random missingness mechanism that tells the device to turn off the channel (i.e. to save battery life).

Below in Tables 9, 10, 11, we see that an 80% random imputation mask %, 50% temporal slice %, and a 50% sensor slice % produce a good mix of reconstruction results across small and large amounts of evaluation masking, for each generative task. Note that when there is a tie, we would prefer higher masking %, in order to allow for a higher dropout removal ratio, and to produce a harder task for our model to pre-train with.

Table 9 | **Effect of Differing Pre-training Random Imputation Mask % on Random Imputation.**

	Random Imp. Eval Ratio		
PT Random Imp. Mask %	30%	50%	80%
90%	0.13	0.14	0.20
80%	0.10	0.12	0.19
70%	0.10	0.12	0.19
60%	0.10	0.12	0.19
50%	0.09	0.12	0.20

Table 10 | **Effect of Differing Pre-training Temporal Slice Mask % on Temporal Interpolation.**

	Temporal Interp. Eval Amount			
PT Temporal Slice %	10 min	30 min	60 min	180 min
70%	0.23	0.34	0.41	0.56
60%	0.26	0.36	0.42	0.57
50%	0.23	0.33	0.40	0.55
40%	0.22	0.33	0.40	0.56
30%	0.22	0.33	0.40	0.57

Table 11 | **Effect of Differing Pre-training Sensor Slice Mask % Ratios on Sensor Imputation.**

	Sensor Imp. Eval Amount			
PT Sensor Slice %	2/26	6/26	12/26	24/26
70%	0.19	0.23	0.28	0.43
60%	0.18	0.22	0.27	0.45
50%	0.17	0.21	0.27	0.48
40%	0.17	0.21	0.27	0.56
30%	0.16	0.21	0.30	0.63

A.4. Model Hyperparameter and Implementation Details

A.4.1. Pre-training Set-up.

We pre-train our models on a large set of wearable minutely sensor data described. The raw multimodal sensor data input can be denoted by $\mathbf{A} \in \mathbb{R}^{T \times S}$. $S = 26$, which is the full number of signals in our multimodal data. These signals are derived from 4 different wearable sensors: Accelerometry, PPG, EDA, and Temperature. In our setting, we set $T = 1440$, which is composed of all minutes from a full 24 hour day, from midnight to midnight local time. We use this window size as days normally have a consistent structure, allowing for a more meaningful absolute positional embedding than if an arbitrary window size was set (e.g. 300 minutes [42]).

Our model was pre-trained with a ViT-1D [21, 1] encoder backbone by using a 1D patch size of 10 time-steps (i.e. 10 minutes). This results in a total of 3744 tokens (the 1440 minutes are reduced to 144 tokens per signal. With 26 signals, 26*144=3744 is the final number of tokens). Similar to prior work [41], each signal channel is patched with a shared kernel, and we utilize a 2D positional embedding to encode information about the temporal position and signal channel. The ViT model had 25 million parameters with an encoding dimensionality of 384, 12 encoder layers, and 4 decoder layers. Our mask is a union of the inherited mask with an artificial masking mix of 80% random imputation, 50% temporal slices, and 50% signal slices. Our primary pre-training objective is to optimize the signal reconstruction loss (i.e. mean squared error), averaged over the artificially masked patches. The model was pre-trained on 8x16 Google v5e TPUs with a total batch size of 512 across 100,000 training steps. The training process uses the AdamW optimizer with a base learning rate of $5e-3$, weight decay set to $1e-4$, and betas set to 0.9 and 0.95. Gradients were clipped at 1.0. A linear warm-up schedule is applied for the first 5% of total steps, followed by a cosine learning rate decay to zero.

Our SSL baselines include LSM [42], SimCLR [13], DINO [11], and a Masked Siamese Network (MSN) [6]. LSM is an MAE [29] approach with 0.8 random masking ratio with no inherited masking. SimCLR, DINO, and MSN are augmentation-based contrastive approaches, and we utilize a set of common time-series augmentations [59, 38, 74, 51]: jittering, scaling, and time flipping. Each augmentation has a 0.5 probability of being applied. Jittering was implemented as a random sample from a gaussian distribution with zero-mean and a uniformly randomly sampled standard deviation frp, 0 to 0.5, per value in the time-series. Scaling was implemented by multiplying all of the data input with a scale, uniformly sampled from 1.1 to 1.5. For DINO, we omit scaling as the model was unable to converge.

Each of these baselines were all pre-trained from scratch, following the same previously stated training conditions, unless stated otherwise. All baselines expect full, complete data as input, and as such, they utilize the imputed version of our sensor dataset. LSM was trained with a ViT-2D with a 2D patch size of (10,2), in order to match their image-based encoding approach, and all other ViT parameters remain constant.

A.4.2. Downstream Evaluation

We group our downstream evaluation into three sections based on the target: generative, classification, and regression.

In our **Generative Evaluation**, we evaluate how well our model is able to reconstruct different types of structured missingness patterns that mimic real-world missingness patterns: (1) Random Imputation, where a [30%, 50%, 80%] of tokens is masked out, (2) Temporal Interpolation, where all signals in a contiguous temporal window of length [10, 30, 60 minutes] is completely masked out,

(3) Temporal Extrapolation, which is similar to interpolation, but the window is necessary at the end of the time-series, and (4) Signal Imputation, where all time points for a random set of [2/26, 6/26, 12/26] signal channels is masked. Reconstruction performance was calculated with mean squared error (MSE) on the artificially masked tokens, averaging only over the data points that have a ground truth.

Our deep learning baselines include the LSM model [42], another MAE-based model, which can be used to evaluate these generative tasks out-of-box by setting the artificial masking procedure to match the proposed tasks. Our `AIM` model is done in the same way, but the full encoder mask includes the inherited mask as well. Unfortunately, the contrastive SSL baselines are unable to provide generative performance metrics because they do not utilize a reconstruction objective. Instead, we use alternative simple generative baselines, which match practical applications. Many application-focused biosensor algorithms will employ simple imputation methods [48, 68, 58, 66, 4] as quick data preprocessing methods. Thus, we choose to include these additional methods as baselines: Linear Interpolation, K-Nearest Neigbhors, and Mean Filling. Similar to our method, we run these baselines with a union mask of the mask inherited from existing missingness and the artificial mask. MICE [63] is another popular, simple baseline designed for multivariate data, but we opted to not include it due to our existing missingness patterns violating the Missingness At Random assumption, and prior work demonstrate a relative poorer performance compared to nearest neighbor and linear interpolation [42].

In our **Classification Evaluation**, we evaluate how well our model's embedding representation is able to capture discriminative features. During evaluation, our model calculates the embedding on all non-inherited-masked tokens and uses an average pooling followed by a trainable linear probe to classify each of the prediction targets. For the LSM model, because it is unable to represent the inherited mask, the embedding for all tokens is pooled, such that tokens that were part of the existing missingness but have been filled with imputation will be included. For the contrastive methods, the learned CLS token is used as the pooled representation. We report performance with F1 score as it balances precision and recall for class-imbalanced targets, Accuracy as a straightforward measure of overall correctness, Balanced Accuracy to account for potential class imbalance, and AUROC to evaluate the model's ranking capability across all classification thresholds. The prediction targets are hypertension, anxiety, which originate from the Metabolics dataset and 20-class activity recognition, which originates from the Activity dataset.

The linear probe was trained by freezing the learned ViT backbone, averaging over the entire embedding and training a logistic regression head ontop of it. For our `AIM` model specifically, with the inherited mask, the average was only done over the non-masked tokens. Training was done with a batch size of 512, across 500 training steps with an AdamW optimizer with a base learning rate of $5e-3$, weight decay set to $1e-4$, and betas set to 0.9 and 0.95. Gradients were clipped at 1.0. For activity specifically, training steps and learning rate were increased to 1000 and $1e-1$ to achieve better convergence.

Additionally, we include two extra supervised baselines, ViT-1D [21] and a ResNet [31], that are trained end-to-end for each of our tasks. ViT-1D is a transformer-based architecture that follows the same architecture as our `AIM` with 25 million parameters, but with randomly initialized weights, trained end-to-end. ResNet is a strong CNN-based architecture that has seen broad success throughout the health biosignal time-series domain [70, 47, 1, 40]. This model was a ResNet-50 [31] with 25 million parameters, in order to match the ViT model. Specifically, it contains 50 layers, with 64 filters that double after each residual block, with a final average pooling and logistic regression head. Both models are trained with a batch size of 512, across 500 training steps with an AdamW optimizer with a base learning rate of $5e-3$, weight decay set to $1e-4$, and betas set to 0.9 and 0.95. Gradients

were clipped at 1.0. A linear warm-up schedule is applied for the first 5% of total steps, followed by a cosine learning rate decay to zero. Because these models do not handle missingness, they were trained directly on the imputed data.

In our **Regression Evaluation**, we utilize the same evaluation procedure described in classification, only instead the linear probe is specifically a linear regression. We report performance with MAE as it provides an interpretable deviation from the correct value, as well as Pearson Correlation Coeffecient, as it is a common metric for evaluating how well a regressor is able to capture the trend of the target [70, 73]. The prediction targets are BMI and Age.

The linear probe was trained by freezing the learned ViT backbone, averaging over the entire embedding and fit a linear regression head ontop of it using Scikit-Learn's LinearRegression implementation out-of-box. The supervised baselines were trained in an identical way as done in the classification evaluation, but using a linear regression head instead of logistic regression.

A.5. Additional Results

A.5.1. Confusion Matrices

Figure 10 illustrates the utility of AIM learned embeddings for downstream applications. Specifically, this confusion matrix shows the performance of AIM, post-trained on the 20-class activity recognition task using a linear probe. It is clear that the embedding are useful in discriminating between a large number of activities, even those which may be semantically clustered, such as skiing and snowboarding. Future work may explore how to expand to even more activities and behavioral events, and investigate the utility of large-scale pre-training in address long-tail task labels.

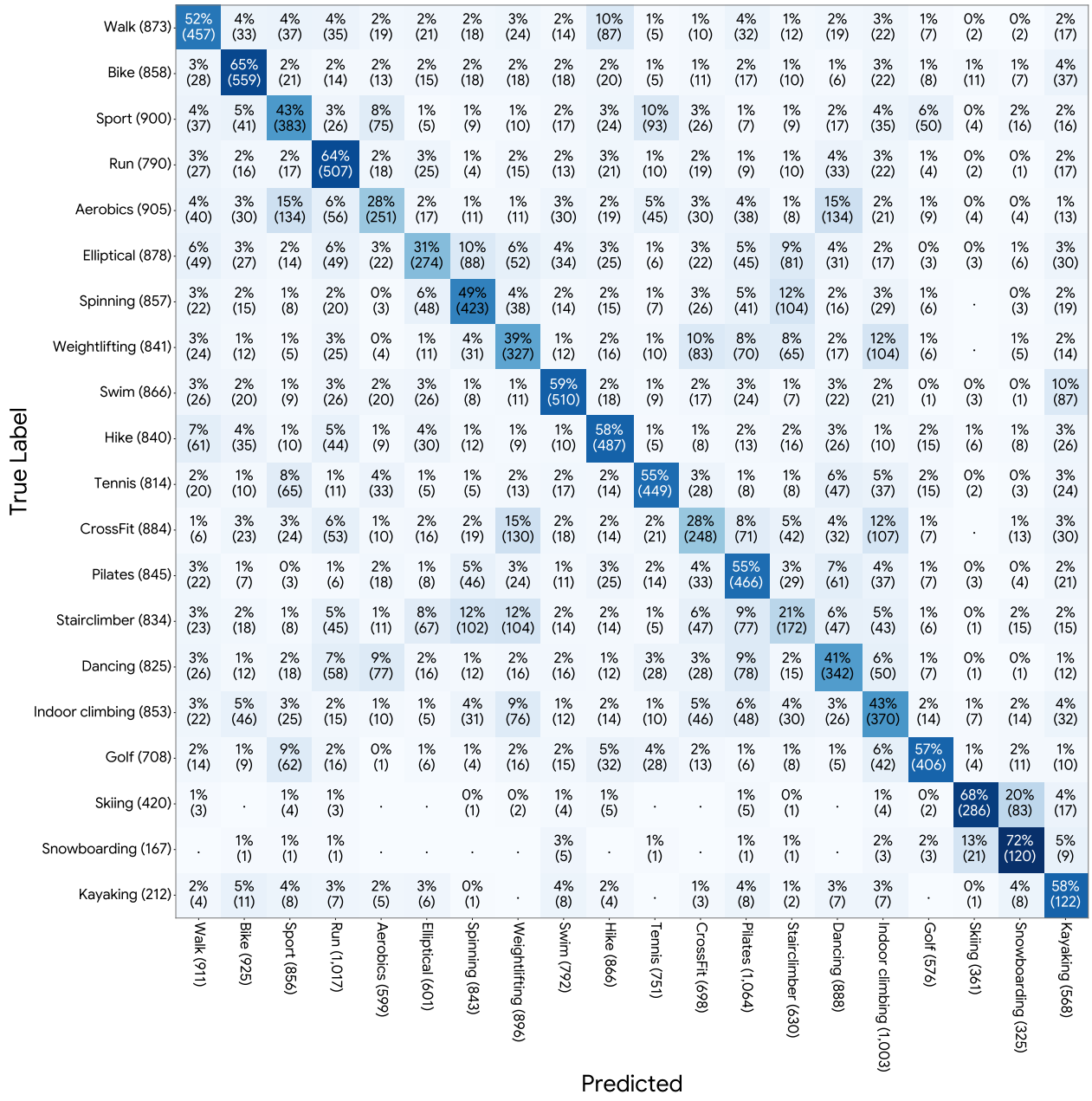

Figure 10 | **Activity Recognition Confusion Matrix.** The results of a linear probe applied to AIM for the 20-class activity recognition task. Rows add up to 100%.

A.5.2. Reconstruction Examples

Figure 11 shows various reconstruction examples for a specific sensor signal. Here we can clearly see Our `AIM` approach leads to much stronger performance, across different generative tasks.

Figure 11 | **Reconstruction Examples for 2/26 Sensor Signal Imputation (Row 1), 3 Hour Temporal Interpolation (Row 2), 3 Hour Temporal Extrapolation (Row 3).** Red highlighted regions demonstrate regions of artificial masking. Orange shows original data with imputation (i.e. the first 400-500 steps of the each row were originally missing, then imputed, as demonstrated by the straight line) and blue shows the reconstructed data.

A.6. Additional Discussions

A.6.1. The Utility of Day-Level Features

Traditionally, generalist methods for time-series health signals have focused on small windowed segments of data on the order of seconds or sub-seconds [1, 70, 43, 73]. Such methods allow for fine-grain activity and physiological tracking. An adjacent body of work has explored the utility of longer observations, on the order of hours [57, 42], enabling more complex person-level insights. In this work seek to expand the observation window to encode a high-level of context. Day level features allow models to learn relationships not possible from shorter spans, for example, how a person's activity during the day may affect their night-time resting heart rate. Looking forward, we intend to continue exploring how best to encode large context windows to include known week, seasonal, and year level periodicities.

A.6.2. Person-Level versus Event-Level Performance

Analysis of the discriminative results (classification and regression) presented in the main body of the paper, raise an interesting question: how do generative pre-training affect performance on person-

level and event-level tasks. For person-level tasks (hypertension, anxiety, age, BMI) we find that AIM consistently outperforms supervised baselines while only using a simple linear probe. In contrast, we find for the event-level task (20-class activity recognition), ResNet50, a supervised baseline performs extremely well, and likely a fully-finetuned AIM model is needed to surpass it. This suggest that while supervised methods easily capture event-level features (e.g., sudden heart rate changes due to activity), they struggle to learn slow-changing, near-constant day-level features more-relevant to person-level tasks. This highlights how method, like are own, learn a more complex representation of the data via generative pre-training. We further concede that our contrastive SSL baselines fail to fully realize the gains of pre-training. We hypothesizes that more complex time-series augmentations are needed to leverage their effect.

A.6.3. Limitations and Future Work

Here we expand upon the limitations and future work introduced in the main body of the paper.

Generalizing to New Devices. Though many commodity wearables host a similar suite of sensors there are inevitable differences between these software-hardware systems. We acknowledge that our methods focuses on a small subset of such devices. Future work will explore the generalizability of our methods to additional devices and datasets, and investigate the extent to which device specific missingness patterns result in a distribution shift.

Generalizing to Open Data. Most publicly available wearable datasets (e.g. WESAD [53], PAMAP2 [9]) are composed of high-frequency raw signals that are very limited in their temporal context with only a subset of the sensors we have available. Thus, they are unable to shown to be used in our setting of day-level context. All of Us [33] demonstrates an interesting avenue to apply our work. Although limited to only the Heart Rate and Step Count channels (compared to our 26 channels), the dataset contains with long context windows and minutely data, and presents an interesting direction in future work to apply our AIM method.

Data and Feature Scales. Time-series analysis often requires explicit assumptions regarding data scale. As such, our method focuses on day-long samples. We acknowledge that such data disregards known periodicities (e.g., weekly, seasonal, etc.). Future work will explore combining our fine-grained behavioral and physiological modeling with insights from longer windows. Furthermore, our method utilizes minutely aggregated features as opposed to the raw sensor feeds common in sensing research. This is a practical limitation, as data is not stored in its raw form at this scale.

Handling Sensor Feature. Our method utilizes 26 features derived from a set of 5 sensors, and regards each feature as independent in the modeling. In reality there are significant correlations between features from the same sensor (e.g., heart rate and heart rate variability). More work can be done to explore how best to combine these multimodal features – potentially sensor-specific encoders, cross-attention, or special class tokens per-sensor feed.

A.6.4. Broader Impact

Personal and ubiquitous health technologies, including smart phones and wearables, have the potential to scale to billions of individuals. Such devices allow for significant self- and longitudinal tracking, and in so doing may augment the current paradigm of clinical healthcare. To-date, consumer health technologies focus on low-level insights, such as steps, resting heart-rate, and sleep staging, which allow users to reason on personal higher-level insights (e.g., "my resting heart-rate has been elevated ever since I fell sick").

In contrast, our method, trained on day-level samples, learns behavioral and physiological patterns

useful in deriving more complex insights. For example, our method shows the potential to predict anxiety and hypertension, insights that humans and commercial algorithms would struggle to derive given only sensor data. We believe this line of work will one day enable people to make the most of their tracked wearable data, better understand their behavior and physiology, and in so doing receive more proactive and better informed care.

A.6.5. Ethics Statement

While consumer health research holds potential for significant positive impact, with so many possible stake holders, such research must be performed intentionally to ensure that it is safe and fair. Additionally, there exists the unfortunate possibility that bad-actor may attempt to leverage methods, such as our own, in negligent ways. As researchers in the field, the burden falls to us to consider the implications of this research, and act to fulfill the positive impacts and mitigate the associated risks.

Building upon this, we concede that training our methods on closed (non-public) data, prevents the scientific community from fully replicating our work. We acknowledge this as a limitation and attest our support for open science and open data. However, due to the sensitive nature of health data, these considerations must be balanced by with the privacy and protection of our participants.