Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Time-series foundation models excel at tasks like forecasting across diverse data types by leveraging informative waveform representations. Wearable sensing data, however, pose unique challenges due to their variability in patterns and frequency bands, especially for healthcare-related outcomes. The main obstacle lies in crafting generalizable representations that adapt efficiently across heterogeneous sensing configurations and applications. To address this, we propose NORMWEAR, the first multi-modal and ubiquitous foundation model designed to extract generalized and informative representations from wearable sensing data. Specifically, we design a channel-aware attention mechanism with a shared special liaison [CLS] token to detect signal patterns in both intra-sensor and inter-sensors. This helps the model to extract more meaningful information considering both time series themselves and the relationships between input sensors. This helps the model to be widely compatible with various sensors settings. NORMWEAR is pretrained on a diverse set of physiological signals, including PPG, ECG, EEG, GSR, and IMU, from various public datasets. Our model shows exceptional generalizability across 11 public wearable sensing datasets, spanning 18 applications in mental health, body state inference, vital sign estimation, and disease risk evaluation. It consistently outperforms competitive baselines under zero-shot, partial-shot, and full-shot settings, indicating broad applicability in real-world health applications.

1 Introduction

Mobile and wearable sensors have been shown to be valuable for the field of healthcare by passively and continuously tracking physiological signals such as photoplethysmography (PPG) for pulse, electrocardiography (ECG) for heart activity, galvanic skin response (GSR), and electroencephalography (EEG) for brain activity. These time series signals are beneficial for early diagnosis, personalized health insights, and remote patient monitoring (Zhang et al., 2024a).

Recently, several foundation models have emerged for time series modeling, including Ansari et al. (2024); Abbaspourazad et al. (2023); Woo et al. (2024); Foumani et al. (2024). Another common approach for signal modeling involves converting raw signal series into 2D images or spectrograms, using fixed-size sliding windows, followed by the use of visual encoders like Vision Transformers (ViT) to extract representations for making inferences (Semenoglou et al., 2023; Wimmer & Rekabsaz, 2023; Vishnupriya & Meenakshi, 2018; Chun et al., 2016; Krishnan et al., 2020; Dosovitskiy et al., 2020). These works have significantly advanced the field and provided valuable insights, yet two main issues still exists which need further exploration to fully understand their potential in wearable scenarios. First, contrastive learning-based foundation models (Abbaspourazad et al., 2023) rely on a predefined set of input signal types, making them unsuitable when transferring to scenarios with different types and numbers of sensors. Second, while both time series foundation models (Ansari et al., 2024; Zhang et al., 2022; Woo et al., 2024) and spectral-based approaches (Semenoglou et al., 2023; Wimmer & Rekabsaz, 2023) attempt to address this issue by training a generic encoder that

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

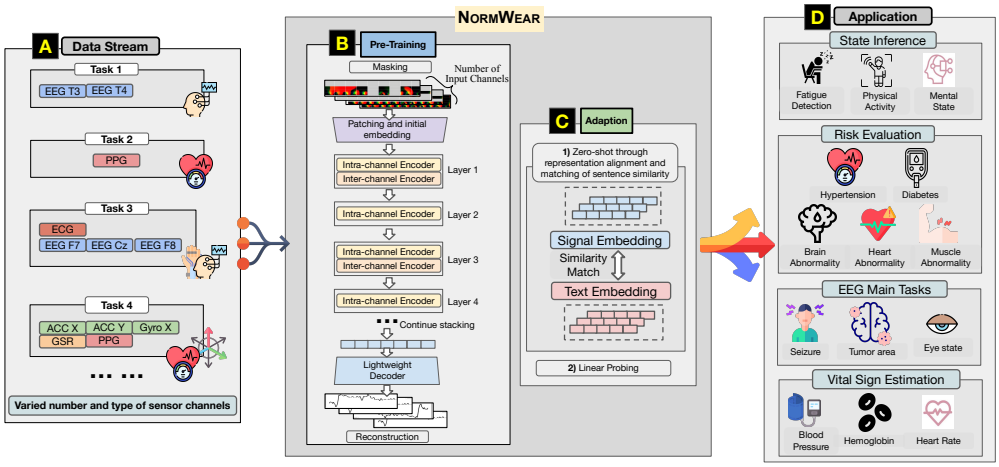


Figure 1: The role of our framework. Several icons from Freepik (n.d.); Zhang et al. (2024a).)

can handle type-agnostic series, they remain limited to processing only univariate series. Because of this constraint, these previous works fail to account for the heterogeneity of multivariate input data; specifically, they do not capture the complex relationships between signals from sensors located on different body parts. These two limitations of recent approaches hinder their generalization and usefulness for wearable health monitoring.

Moreover, Wearable-based multimodal physiological signals present unique challenges that distinguish them from general time series data, such as stock prices or weather patterns. Wearable signal modalities, such as PPG and EEG, vary in characteristics like dimensionality, sampling rate, and resolution, often requiring modality-specific preprocessing. Existing methods tokenize raw signals (Ansari et al., 2024; Zhang et al., 2022) or convert them into image or spectral representations (Wu et al., 2023; Mathew et al., 2024; Vaid et al., 2023). While effective for specific tasks, these approaches lack generalizability and fail to provide a consistent preprocessing pipeline across multiple modalities. A consistent framework that accommodates diverse signal requirements is essential for training deep learning-based foundation models and advancing multimodal signal analysis.

In this work, we present NormWear, a normative foundation model, aiming to learn effective wearable sensing representations, addressing the above-discussed research gaps. NormWear has been pretrained on more than 2.5 million multivariate wearable sensing segments, comprising total of 14,943 hours of sensor signal series, using publicibly avaliable datasets. We evaluated NormWear on 18 public downstream tasks against competitive baselines across zero-shot, few-show, and full-shot settings. Overall, our contributions with the proposed NormWear healthcare modeling framework can be summarized as follows:

- To our knowledge, we are the first to develop a foundation model specifically designed for wearable sensing data, capable of processing arbitrary configuration of multivariate signals from sources such as the heart, skin, brain, and physical body.
- NormWear comprises novel methodologies built upon the advanced practice in both the fields of signal processing and deep learning, including (a) continuous wavelet transform (CWT) based multi-scale representations for modality- and number-agnostic tokenization, (b) channel-aware attention layer that enables the model to process arbitrary multivariate inputs, and (c) a human sensing adapted fusion mechanism that enabled NormWear to achieve zero-shot inference on health related wearable sensing tasks.
- We are also the first to integrate and process a comprehensive wearable signals dataset with varied number of input channels for training self-supervised learning algorithms, with thorough downstream evaluation. These datasets cover key health applications, including mental and physical state inference, vital sign estimation, and disease risk evaluation.

Our proposed NormWear aims to provide a generalized data representation solution for smart health monitoring, benefiting the general public, and serving as a fundamental tool for researchers and professionals to address future healthcare challenges. We made the code and cleaned data to be publicly available to spur reproducible research.

2 Related Work

Foundation models have emerged as a transformative paradigm in machine learning, enabling generalizable and reusable representations across diverse downstream tasks (Bommasani et al., 2022).

In the time series domain, recent works (Ansari et al., 2024; Foumani et al., 2024; Abbaspourazad et al., 2023; Narayanswamy et al., 2024) have demonstrated success in tasks such as forecasting, classification, and anomaly detection. However, their generalizability to health-related wearable signals remains limited due to the lack of in-depth evaluation, reliance on specific sensor types (Wang et al., 2025; Jiang et al., 2024; Yang et al., 2023) and univariate data (Pillai et al., 2024; McKeen et al., 2024), as well as the inability to handle the heterogeneity of multivariate wearable signals. In contrast, NORMWEAR builds upon these principles by introducing a modeling framework that is agnostic to the sensor modality and number of input channels, as stated in section 1, and is presented in details in section 3. NORMWEAR has been evaluated on 18 digital healthcare tasks and demonstrate peak performance against solid time series modeling baselines, including common statistical approach, SoTA model in time series with self-supervised learning (Zhang et al., 2022), SoTA spectrum based modeling approach (Wu et al., 2023), and SoTA time series forecasting model (Ansari et al., 2024). Our work not only generalizes to arbitrary sensor configurations but also ensures compatibility across multivariate data, addressing key limitations of earlier approaches.

3 Method

3.1 Dataset construction for model pretraining and downstream evaluation

We curated a collection of 9 publicly available datasets (Table 5) exclusively for model pretraining, resulting in approximately 230,962 multivariate time series segments, comprising 4,294 hours of total sensor signal series, across various modalities, including PPG, ECG, EEG, GSR, PCG, and inertial measurement unit (IMU) data. To address the dataset size limitation, we then applied herustic data augmentation (algorithm 1) to expand the pretrain dataset to 2.5 million segments, comprising 14,943 hours of total sensor signal series. Notably, each sample segment may contain a variable number of input channels depending on the sensor signals provided by the respective datasets. This input configuration aligns seamlessly with our model's design, which is optimized to flexibly handle arbitrary numbers and configurations of sensor signal inputs.

To prevent potential data leakage in downstream tasks, we evaluate our model's transferability using an additional 11 publicly available datasets encompassing 18 modeling tasks, which include affective state classification, physical state recognition, biological estimation, and disease risk evaluation. Details about the datasets is presented in Table 4.

3.2 Tokenization

Tokenization is a fundamental term widely used in natural language processing. In the context of wearable sensing, we leverage this term to represent the stage of signal processing before sending the processed data to the deep learning-based encoder. Spectral methods, which utilize the short-time Fast Fourier Transform (FFT) (Brigham, 1988) with a sliding window to compute spectrograms, are widely regarded as the benchmark approach for tokenization. However, due to the inherent trade-off between time and frequency resolution, the spectral representation with a fixed window size cannot be generalized. This is because the window size has to be modulated accordingly when the modality varies. To enhance transferability, we propose a well-designed signal processing pipeline that preserves information in both the frequency and time domains across multiple scales. We begin by calculating the first and second derivatives for each single signal series, as suggested by Slapničar et al. (2019), followed by computing the continuous wavelet transform (CWT) on both the raw and derivative series, resulting in three scalograms. Then, we stack the three scalograms to form data in RGB-image-like format. The derivatives capture the rate of signal change at different moments, while the wavelet transform provides a multi-resolution encoding that preserves information from both the time and frequency domains Torrence & Compo (1998). For the wavelet transform, we use the Mexican Hat wavelet for signal convolution, as recommended by previous studies (Burke & Nasor, 2004; Hosni & Atef, 2023; Hassani, 2021; Negi et al., 2024; Nedorubova et al., 2021b). We apply scales ranging from 1 to 64, following the guidance of (Sengupta et al., 2022; Nedorubova et al., 2021a), which sufficiently covers most frequency bands of interest for physiological signals. Finally, this RGB-like scalogram is divided into patches, which is treated in the same way as tokens in an ViT (Dosovitskiy et al., 2020). In this way, this tokenization approach can be applied to various types of sensing signals without sensor-specific adjustments or reconfigurations.

3.3 Share-weighted Encoder

Rather than concatenating tokens from all channels into a single long sequence and processing them with a full attention transformer, we treat each channel of the multivariate signal as an independent

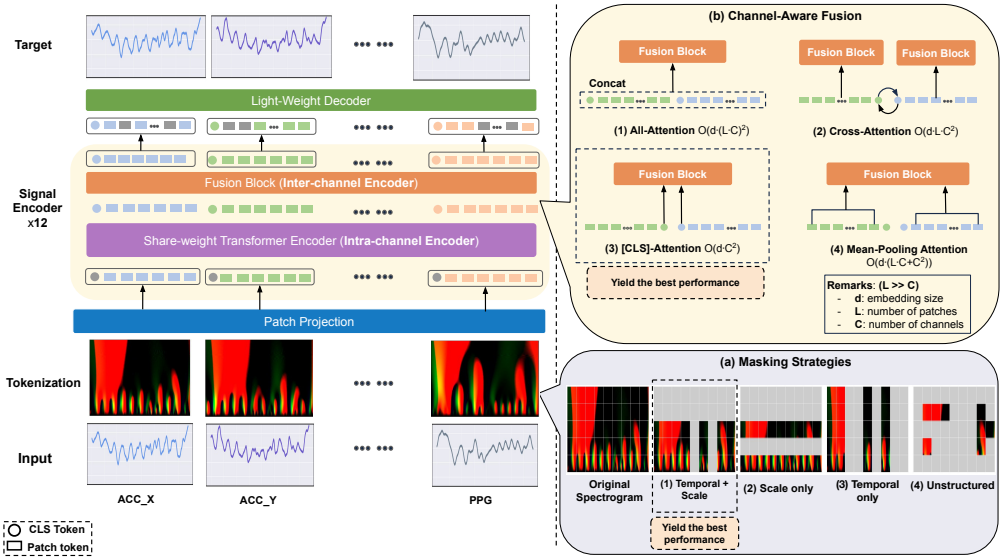


Figure 2: Overview of the pretrain pipeline.

input stream. Although all channels share the same transformer backbone, the forward pass is executed separately for each one. This design allows the model to first learn the temporal characteristics of each sensor without interference from others. It not only reduces computational cost but also increases flexibility. Because each channel is processed independently, the model can be pretrained on datasets with varying numbers or types of sensors and later fine-tuned on a target task with a different sensor configuration.

3.4 Channel-Aware Attention with Liaison Special Token

Following the tokenization step, we adopt common reconstruction-based pretraining strategies from Masked Auto Encoder (MAE) (He et al., 2021; Huang et al., 2023; Zhang et al., 2023), where input tokens are randomly masked and the model is trained to reconstruct the original time series using mean squared error (MSE) loss. Inspired by Huang et al. (2023), we experiment with four masking strategies, as shown in Figure 2 (a), including masking on (1) temporal and scale, (2) scale only, (3) temporal only, and (4) unstructured axes. We observe that the temporal and scalar masking yields the best performance for the downstream tasks. For the model architecture, we construct the backbone of our proposed framework with a convolutional patching layer followed by 12 standard Transformer blocks (Vaswani et al., 2023). For the same reason, NormWear uses a lightweight decoder consisting of 2 Transformer blocks, combined with a linear projection layer and a convolution layer to reconstruct the raw physiological signals both temporally and spatially. We also prepend a [CLS] token to each signal channel, following standard practice in transformer models, for learning a global representation of the input sequence for that channel.

Another important point to consider is that although empirical studies (Nie et al., 2023; Abbaspourazad et al., 2023) show that channel-independent structures effectively capture local patterns, they fail to account for relationships across channels. To address this, we use the [CLS] token from each signal channel as a liaison token, allowing them to exchange information through the channel-aware fusion layer afrer every other encoder block. We explore several fusion approaches and different design of liaison token as shown in Figure 2 (b), with each method described below:

(1) **All-Attention Fusion:** This approach involves concatenating all tokens from each modality without considering their individual properties and fusing the information through a self-attention module. However, this method requires quadratic computation time, as every token passes through the self-attention module, making it impractical for real-world applications.

(2) **Cross-Attention Fusion:** In addition to the cross-attention mechanism used in Cross-ViT (Chen et al., 2021), we introduce a slight modification to fit in our problem setting. We propose a symmetric fusion method, using the [CLS] token from each modality as an intermediary to exchange information between the patch tokens of another modality, then projecting the information back to its original modality in the subsequent Transformer layer. While this strategy is efficient, it restricts the model to handling only two time series signals or modalities, which deviates from our goal of building a general model capable of processing an arbitrary number of channels.

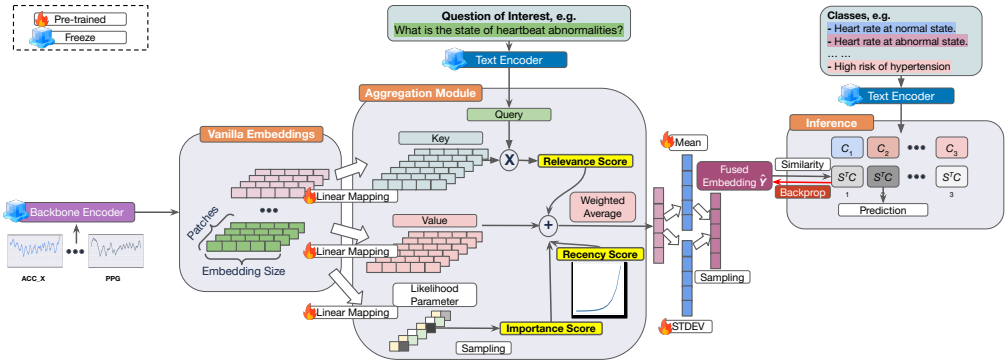


Figure 3: Memory stream inspired temporal fusion mechanism for representation alignment.

(3) **[CLS]-Attention Fusion** The [CLS] token serves as an abstract global representation for each signal modality. Here, we propose a hybrid fusion approach. We stack the [CLS] tokens from all signal modalities and perform feature fusion using a self-attention mechanism. The fused [CLS] token is then reattached to its original channel, enabling the newly learned information to be propagated to each patch token in subsequent transformer encoder layers.

(4) **Mean-Pooling Fusion** Similar to the [CLS]-Attention Fusion approach, we employ mean-pooling within each channel instead of using the [CLS] token as an abstract global representation.

Our empirical results show that [CLS]-attention fusion achieves the best downstreaming performance for our proposed NORMWEAR model. Details of all the ablation studies are reported in Appendix C. Beyond accuracy, we want to emphasize that the [CLS]-Attention Fusion design is highly flexible. This flexibility arises from the fact that self-attention is length-flexible and permutation-invariant (Vaswani et al., 2023). Consequently, it integrates naturally with our shared-weight encoder, allowing the model to accommodate a variable number of sensor channels presented in any order. We provide additional empirical evidence of NormWear's permutation invariance in Table 12, Appendix C.

3.5 Sensor-Semantic Representation Alignment

Zero-shot inference is an important aspect to evaluate foundation model. We evaluate our model in this setting by retrieving the closest text-derived label for each unseen task in the shared embedding space. Specifically, to unify information across diverse modalities, we incorporate a representation alignment objective that encourages the embeddings of physiological sensor data to reside in the same latent space as paired textual descriptions. Once this shared space is established, it naturally supports zero-shot inference by allowing unseen sensor inputs to be interpreted through their proximity to text-derived anchors, without additional task-specific training. Several important works in this direction focusing on domains of vision-language Radford et al. (2021), audio-language Wu et al. (2023), and motion-language (Zhang et al., 2024b). These works leverage end-to-end training to bind their modality of interest into semantic space. In this work, we extend this methodology to explore NORMWEAR's ability to generalize across unseen datasets and tasks.

Building on prior work in representation alignment, we notice that in healthcare-related tasks where flexible inference across diverse scenarios is often required, the ground truth labels often have substantial overlap. For instance, depression is inferred from stress levels (LeMoult, 2020), and running and cycling produce similar IMU signals (Li et al., 2019). Due to these nested relationships, it create potential challenge to representation alignment when using contrastive learning, which requires clearly defined positive and negative pairs. To address this, we first propose a novel way to fuse the signal representations together with improved qualities, then align the representation with vector distance as an auxiliary loss for contrastive learning method. In addition, to reduce computation cost and counteract the issue of catastrophic forgetting (Li et al., 2023), we use off-the-shelf frozen encoders for both signal and text modalities.

Human physiological signals are task-specific, dynamic, and often weakly labeled (He et al., 2018; Kim et al., 2022; Qian et al., 2021; Ma et al., 2021). To address these characteristics, we introduce three complementary scoring mechanisms during feature aggregation: *relevance scores* prioritize patches aligned with the task objective (e.g., IMU for activity recognition), guided by query sentences such as "*What activity is the subject doing?*"; *recency scores* emphasize recent segments to better reflect the current physiological or emotional state (Roelofs, 2017; Chowdhury et al., 2020; Chaudhury et al., 2021); and *importance scores* weigh signal segments that contain meaningful or transient

patterns often buried in weakly labeled sequences. Together, these scores guide the MSiTF fusion module to generate compact, task-aware representations. This design is inspired by memory-stream retrieval mechanisms (Park et al., 2023) and is tailored to the demands of human-centered sensing tasks such as risk assessment, affect detection, and activity recognition.

Memory Stream inspired Temporal Fusion (MSiTF). Our Aggregation or Fusion Module, MSiTF, is designed to addresses the above-discussed three challenges through three scores discussed below. Specifically, we denote f as the function that takes the semantic embedding of query sentence q and backbone output $H \in \mathbb{R}^{P \times E}$ as input, where P is the patch size and E is the embedding size, thus having the final fused representation $f(q, H) = \hat{Y} \in \mathbb{R}^{E}$.

We define the *Relevance* score as the cross attention between the key representations of each sensor time step and the query sentence embedding, obtained from a pretrained language model (Clinical TinyLlama (Muzammil, 2021)). This mechanism allows the model to identify distinct but contextually relevant segments in the sensor input. For the *Recency* score, we use an exponential decay function to reflect the intuition that recent time steps are more important than earlier ones. Finally, we consider the importance score IMP in this case to be whether to keep the representation at each time step or not. In order to achieve this, we assign binary parameters to each time step, denoted as $\theta_t = p(v_t) \in \mathbb{R}^2$ where $v_t \in \mathbb{R}^E$ is the representation vector at time step t and p is a trainable linear transformation function which will be optimized during pretraining. We then have the importance score for each patch defined as

$$W_{imp}(t) = \underset{i \in \{0,1\}}{\arg\max} \frac{\exp\left(\Big(\log(\theta_{t,i}) + \epsilon\Big)/\tau\right)}{\sum_{j \in \{0,1\}} \exp\left(\Big(\log(\theta_{t,j}) + \epsilon_j\Big)/\tau\right)} \tag{1}$$

where ϵ is the noise term sampled from Gumbel distribution (Jang et al., 2017), and τ is the temperature controlling the sharpness of the softmax function. Because $\arg\max$ is not a differentiable function, we will directly take the resulting probability corresponding to index at $j = 1$ to be the *importance* score, with τ being set to a small number to push the result closer to one hot vector from the softmax function. As a result, this logit function will determine to what extent to activate the gate during forward pass on each patch of the input signals. The final score for each patch is the summation of the three scores as described above. This score will be treated as the weight for aggregating the representations from all the patches to form the fixed length embedded output (vector with size of 768 in our case).

Once the signal embeddings are aggregated, we adopt a variational-inspired approach (Kingma & Welling, 2022). This design injects stochasticity into the representation, encouraging the model to explore and capture nuanced variations in semantic representations. Finally, we leverage contrastive learning with auxiliary loss on vector distance to train the MSiTF module with a projection layer to text representation on the pretraining datasets. The sentence template formation and training details are presented in Appendix B.5.

4 Experiments

NORMWEAR is pretrained exclusively on the data shown in Table 5. In this section, we present a comprehensive evaluation across 11 downstream publicly available datasets, focusing on 18 widely-recognized digital healthcare tasks. We evaluate the methods following order of zero-shot capability, partial-shot learning, and full-shot learning.

4.1 Selection of baselines covering representative modeling strategies

Modeling multivariate wearable signals with arbitrary input channels and sensor types, such as those capturing activities of heart, brain, and body physical motions, presents unique challenges, as no universally recognized open-source baseline or state-of-the-art (SoTA) model exists in this domain. To evaluate our approach, we selected diverse and representative baselines (as shown in Table 3).

In the literature, various modeling strategies have been proposed. Firstly, early approaches involved handcrafting statistical features, which was a widely adopted practice in signal processing (Yan et al., 2023a; Reyes-Ortiz et al., 2012; Mikelsons et al., 2017). We include this simple baseline as sanity check. Secondly, since sensory data can be naturally represented as time series (Woo et al., 2024; Semenoglou et al., 2023), we benchmarked our model against Chronos (Ansari et al., 2024) , as well

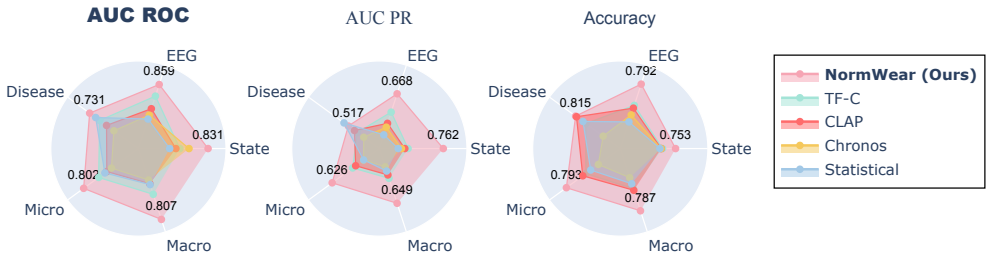


Figure 4: **Overview of performance trend of NORMWEAR, under full-shot linear probing, against competitive baselines in downstream tasks:** (1) Disease risk predictions. (2) EEG main tasks (mental and abnormal states prediction). (3) State recognition: physical and mental activities. (4) Macro: Average performance over types of tasks. (5) Micro: Average performance over each task.

as a self-supervised framework TF-C (Zhang et al., 2022). Finally, the spectrum-based modeling methods (Vishnupriya & Meenakshi, 2018; Chun et al., 2016; Krishnan et al., 2020) are widely used for signal modeling. Therefore, we incorporate CLAP (Wu et al., 2023) into baselines that has demonstrates SoTA performance in spectrogram-based modeling. Regarding the comparison with concurrent works proposing foundation models for a specific sensor modality, we leverage PaPaGei (Pillai et al., 2024) for PPG datasets, ECG-FM (McKeen et al., 2024) for ECG datasets, and CBraMod (Wang et al., 2025) for EEG datasets. These baselines span distinct paradigms, providing a solid foundation to demonstrate the strengths of our model in wearable signal tasks. For uni-modal baselines like Chronos and CLAP, we feed each signal separately into model and concatenate their representations after the forward pass. This ensures that all models have the same field of view, making the comparison fair.

4.2 Zero-shot Evaluation

We achieve zero-shot inference by pretraining our proposed novel temporal fusion module on the task of representation alignment. We include the SoTA spectral-based model CLAP Wu et al. (2023) as a baseline to provide a more comprehensive comparison of the results. For CLAP, we experimented with both Manhattan distance (MD) and dot product (DP) as similarity metrics during inference. We observe that there are no statistically significant differences in performance when using MD and DP for label retrieval in CLAP. From table 1, we could observe that overall, NORMWEAR equipped with MSiTF outperforms the baselines. We compare NORMWEAR with a few ablations by removing importance score (w/o IMP) and removing text augmentation (w/o text aug). We can observe that performance drop after removing each of the above components, verifying their respective importance in improving generalization across various downstream tasks. We present this outcome to demonstrate the zero-shot capability in the wearable signal domain, an aspect not present in recent studies. We also hope this outcome could potentially provide a new perspective that can help drive progress in this direction within the field.

Table 1: Zero-shot performance on the downstream datasets, with AUC ROC being reported. The last two columns show the average across the tasks and across group types respectively.

Model	*WESAD*	*UCI-HAR*	*DriverFatigue*	*GAMEEMO*	*Epilepsy (eye open state)*	*Epilepsy (eye relaxation)*	*Epilepsy (health area)*	*Epilepsy (tumor area)*	*Epilepsy (seizure)*	*PPG-BP (HTN)*	*PPG-BP (DM)*	*PPG-BP (CVA)*	*PPG-BP (CVD)*	*ECG-Abnormal*	*PhysioNet EMG*	*Micro Avg.*	*Macro Avg.*
CLAP - MD	45.3	62.8	58.5	53.1	44.9	45.1	47.6	30.5	**84.9**	**59.4**	41.8	46.0	57.4	22.9	55.4	50.4	51.2
CLAP - DP	50.7	52.3	**61.1**	51.6	54.4	41.9	58.6	46.4	74.3	52.2	41.4	50.6	58.9	42.7	38.3	51.7	52.2
before bind	44.1	48.2	52.1	48.4	54.1	62.6	53.9	52.5	24.6	48.8	49.6	46.3	56.8	54.3	48.2	49.6	49.4
NORMWEAR w/ MSiTF	55.8	71.2	57.2	51.0	**55.7**	61.3	**67.6**	55.8	66.0	57.1	**62.5**	**70.0**	59.0	63.1	70.1	**61.6**	**61.5**
- w/o IMP	56.2	70.3	55.4	49.8	54.0	56.5	66.9	**57.3**	52.9	56.5	54.3	61.7	**60.7**	**73.4**	65.2	59.4	59.6
- w/o text aug	54.8	65.8	55.2	49.2	31.0	58.4	58.6	32.8	58.1	50.2	52.6	50.8	50.6	47.7	33.6	50.0	51.4
- w/o refine	**59.5**	**72.8**	42.7	**57.3**	50.6	**69.0**	43.3	50.5	74.8	48.3	38.8	44.6	44.1	72.4	**75.7**	56.3	56.6

4.3 Partial-shot and Full-shot Evaluation

We evaluate the learned representations using linear probing through supervised training on each downstream dataset, and report performance on the corresponding held-out test set. To ensure

Table 2: **Detailed performance on various downstream wearable-signal-based health related applications under full-shot linear probing evaluation.**

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	Modality-Specific	NORMWEAR (Ours)
WESAD	66.213	71.489	72.383	69.865	56.656	**76.060**
UCI-HAR	95.784	91.593	96.420	96.892	-	**98.954**
DriverFatigue	63.249	76.722	61.889	66.882	**80.430**	74.292
Activity Recognition Avg.	75.082	79.935	76.897	77.880	-	**83.102**
Epilepsy (eye open state)	82.489	82.41	85.094	89.153	90.436	**92.743**
Epilepsy (eye relaxation)	87.457	88.218	89.867	94.416	**95.552**	94.828
Epilepsy (health area)	86.274	81.08	83.711	85.619	88.065	**88.541**
Epilepsy (tumor area)	82.816	81.034	83.644	86.348	**87.258**	87.197
Epilepsy (seizure)	88.272	97.572	**97.734**	93.998	94.616	97.053
GAMEEMO	51.009	53.747	52.551	**56.275**	55.420	54.937
EEG Main Tasks Avg.	79.720	80.677	82.100	84.302	85.225	**85.883**
ECG-Abnormal	97.092	98.585	97.23	98.275	89.898	**99.140**
PPG-BP (HTN)	59.499	52.425	56.757	**65.229**	61.839	62.341
PPG-BP (DM)	47.823	51.164	42.455	**57.883**	55.668	55.893
PPG-BP (CVA)	71.250	50.278	51.667	58.125	**73.125**	70.625
PPG-BP (CVD)	51.219	58.31	50.91	**58.674**	49.066	51.773
PhysioNet EMG	**99.309**	61.6	98.627	78.308	-	99.216
Risk Evaluation Avg.	71.032	62.060	66.274	69.416	-	**73.165**
Noninvasive-BP	92.310	91.79	91.922	87.481	90.596	**92.420**
PPG-Hgb	94.219	**95.005**	94.291	93.408	94.912	94.632
Fetal-fPCG	98.929	99.048	**99.195**	99.077	-	99.072
Vital Signs Avg.	95.153	95.281	95.136	93.322	-	**95.375**
Micro Avg.	78.623	76.782	78.130	79.773	-	**82.762**
Macro Avg.	80.247	79.488	80.103	81.230	-	**84.381**

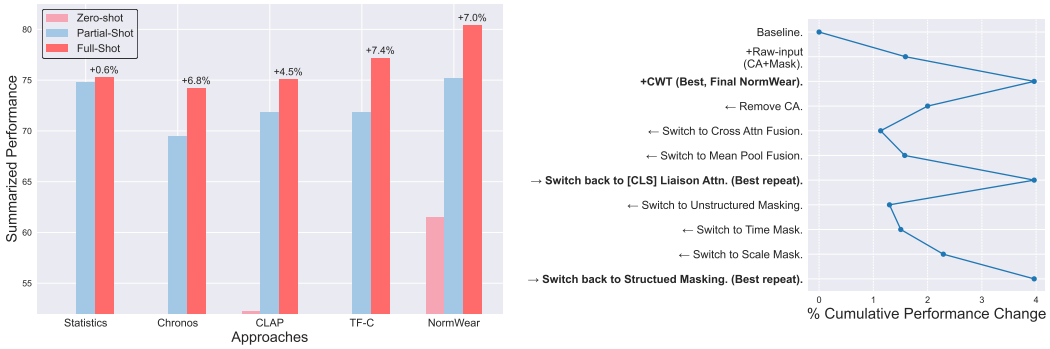
Table 3: **Baselines**

Baseline Methods	Modeling Strategies
Modality Specific (Zhang et al., 2022)	PaPaGei (Pillai et al., 2024), ECG-FM (McKeen et al., 2024), CBraMod (Wang et al., 2025).
TF-C (Zhang et al., 2022)	SoTA in TS SSL; modeling time and frequency domain information at same time.
CLAP (Wu et al., 2023)	SoTA in audio modeling; process signal as spectrogram
Chronos (Ansari et al., 2024)	SoTA in TS forecasting, leverage LLM for modeling
Statistical approach	Reserve full interpretability

fair comparison, we use a unified evaluation protocol with identical hyperparameter settings and implementation across all models and the dataset (Yuan et al., 2024). This design ensures that performance differences are not due to variations in learning rate, regularization, or data augmentation (Oliver et al., 2018). Specifically, the classification tasks, using logistic regression, are solved by Newton's method with conjugate gradient, with AUC ROC being reported as main metric. The regression (vital signs) tasks, using ridge regression, are solved by Cholesky's method with closed form solution, with relative accuracy being reported. For partial-shot evaluation, we leverage 10% of the training data for the linear probing, and detailed performance result is presented in Table 11. The full-shot evaluation results is presented in Table 2. All scores are the higher the better.

From Figure 4, Table 2, and Table 15, we observe that NORMWEAR consistently achieves peak performance across all task groups, including activity recognition, EEG signal analysis, disease risk evaluation, and vital sign estimation. Furthermore, its leading performance remains consistent across various evaluation metrics. Based on the macro-averaged total score across task groups, NORMWEAR delivers a 3.9% improvement over the state-of-the-art (SoTA) time-series self-supervised learning framework (Zhang et al., 2022), a 5.3% improvement over the SoTA spectrum-based modeling method (Wu et al., 2023), a 6.1% improvement over SoTA time-series forecasting models with LLM backbones (Ansari et al., 2024), and a 5.2% improvement over standard statistical baselines. On larger datasets, NORMWEAR significantly outperforms the statistical baseline by 9.0% and 7.5% for activity recognition and EEG brain activity monitoring tasks, respectively. On smaller datasets, it still achieves peak performance in disease risk evaluation. For vital sign estimation, all methods yield comparable results, suggesting inherent challenges in these regression tasks that warrant further investigation but are beyond the scope of this study.

When comparing with recent modality specific foundation models, NormWear's main benefit is that it capture cross-modal relationships, making it more versatile for wearable health tasks. While it sacrifices modality-specific optimization for adaptability, this may slightly reduce performance in highly specialized tasks. Single-signal models excel in their domains due to deeper modality-focused training. Instead of maximizing single-modality data, we prioritize signal diversity for better generalization. Benchmarking shows that NormWear, trained on a smaller dataset than EEG-only models, still achieves competitive results, highlighting the effectiveness of our pre-training approach. These findings illustrate NORMWEAR's capacity to balance consistency and adaptability across a diverse range of tasks and conditions. By excelling across standard benchmarks while addressing the intricacies of varied applications, NORMWEAR exemplifies the philosophy of a foundation model: a reliable generalist capable of performing robustly across both typical and challenging scenarios.



(a) Adaptation summary

(b) Inspect performance contribution. 'CA' refers to channel-aware attention.

Figure 5: Summary of adaptation performance and module-level performance contributions. Details of ablation study results are presented in Appendix C.

5 Conclusion and Discussion

Conclusion. In this work, we mainly propose a foundation model for wearable physiological signals. NORMWEAR is a practical tool that could serve as a starting point for researchers and clinicians when tackling a problem with wearable based signal data. Our proposed model could extract informative representations from raw signal series, which can be leveraged for further machine learning modeling, clustering, embedding vector-based information retrieval, and deployment of real-time health states monitoring with minimal tuning. We've justified the utilizability and generalization of NORMWEAR through an extensive evaluation of various ubiquitous health applications. As for future works, it is important to leverage our framework on larger scale clinical applications and explore the applicability of embedding vectors as state representations for intervention modeling problems that comprise the decision-making process.

Limitation and Future Work. We acknowledge several limitations to be addressed in future work. (1) The representation alignment component is currently trained on a limited set of healthcare-related objectives, and expanding the pretraining corpus with more diverse semantic labels may improve generalization. (2) While our design supports classification tasks well, adapting the framework for regression remains an open challenge, and future work may explore alternative formulations beyond label discretization. (3) NormWear currently focuses on physiological signals with relatively narrow frequency bands; extending its applicability to higher-frequency modalities such as audio or lower-resolution clinical summaries is a promising direction.

Broad Impact. NORMWEAR is the first foundation model tailored for multivariate physiological signals that supports a wide range of wearable health tasks across sensor modalities, device types, and clinical applications. Through a unified CWT-based tokenization pipeline and a channel-aware fusion mechanism, it enables robust, modality-agnostic representation learning. Our extensive evaluation across zero-shot, partial-shot, and full-shot settings demonstrates NormWear's strong generalizability and practical relevance. We believe NormWear provides a valuable resource for advancing foundation modeling in digital health and promoting more unified benchmarks in the community.

Ethics Statement

This study contains applications in the field of healthcare. We ensured that all the data being used during pretraining and evaluations were made publicly available by the original authors, and all these works were cited properly.

Reproducibility Statement

The full code base is submitted in supplementary material referred to as *NormWear_main.zip*, comprising all the scripts for exploratory data analysis and preprocessing, model construction, pretraining, downstream evaluation, result analysis, and all the visualizations that are described in this paper. The GitHub repository containing all the documentation will be published simultaneously with the paper.

References

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

Abuzairi, T., Vinia, E., Yudhistira, M. A., Rizkinia, M., and Eriska, W. A dataset of hemoglobin blood value and photoplethysmography signal for machine learning-based non-invasive hemoglobin measurement. *Data in Brief*, 52:109823, 2024. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2023.109823.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alakus, T. B., Gonen, M., and Turkoglu, I. Database for an emotion recognition system based on eeg signals and various computer games–gameemo. *Biomedical Signal Processing and Control*, 60: 101951, 2020.

Alzahab, N. A., Di Iorio, A., Apollonio, L., Alshalak, M., Gravina, A., Antognoli, L., Baldi, M., Scalise, L., and Alchalabi, B. Auditory evoked potential eeg-biometric dataset, 2022.

Andrzejak, R. G., Lehnertz, K., Rieke, C., Mormann, F., David, P., and Elger, C. E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state [dataset]. *Physical Review E*, 2023. doi: 10.34810/data490. URL `https://doi.org/10.34810/data490`.

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

Bajaj, N., Carrión, J. R., and Bellotti, F. Phyaat: Physiology of auditory attention to speech dataset. *arXiv preprint arXiv:2005.11577*, 2020.

Beh, W.-K., Wu, Y.-H., and Wu, A.-Y. A. Maus: A dataset for mental workload assessment on n-back task using wearable sensor, 2021. URL `https://dx.doi.org/10.21227/q4td-yd35`.

Bhaskaran, A., J, S. K., George, S., and Arora, M. Heart rate estimation and validation algorithm for fetal phonocardiography. *Physiological Measurement*, 43(7):075008, jul 2022. doi: 10.1088/1361-6579/ac7a8c. URL `https://dx.doi.org/10.1088/1361-6579/ac7a8c`.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D.,

Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Bousseljot, R., Kreiseler, D., and Schnabel, A. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. In *PTB-XL, a large publicly available electrocardiography dataset*, 2009. URL https://api.semanticscholar.org/CorpusID:111121953.

Brigham, E. O. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

Burke, M. and Nasor, M. Wavelet based analysis and characterization of the ecg signal. *Journal of Medical Engineering & Technology*, 28(2):47–55, 2004.

Carmona, C. U., Aubet, F.-X., Flunkert, V., and Gasthaus, J. Neural contextual anomaly detection for time series, 2021. URL https://arxiv.org/abs/2107.07702.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments, 2021. URL https://arxiv.org/abs/2006.09882.

Chaudhury, S., Yu, C., Liu, R., Kumar, K., Hornby, S., Duplessis, C., Sklar, J. M., Epstein, J. E., and Reifman, J. Wearables detect malaria early in a controlled human-infection study. *IEEE Transactions on Biomedical Engineering*, 69(6):2119–2129, 2021.

Chen, C.-F., Fan, Q., and Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

Chowdhury, M. H., Shuzan, M. N. I., Chowdhury, M. E., Mahbub, Z. B., Uddin, M. M., Khandakar, A., and Reaz, M. B. I. Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques. *Sensors*, 20(11):3127, 2020.

Chun, S. Y., Kang, J.-H., Kim, H., Lee, C., Oakley, I., and Kim, S.-P. Ecg based user authentication for wearable devices using short time fourier transform. In *2016 39th international conference on telecommunications and signal processing (tsp)*, pp. 656–659. IEEE, 2016.

Dar, M. N., Rahim, A., Akram, M. U., Khawaja, S. G., and Rahim, A. Yaad: young adult's affective data using wearable ecg and gsr sensors. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1–7. IEEE, 2022.

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

Esmaili, A., Kachuee, M., and Shabany, M. Nonlinear cuffless blood pressure estimation of healthy subjects using pulse transit time and arrival time. *IEEE Transactions on Instrumentation and Measurement*, 66(12):3299–3308, 2017.

Fekri Azgomi, H., Branco, L. R. F., Amin, M. R., et al. Regulation of brain cognitive states through auditory, gustatory, and olfactory stimulation with wearable monitoring. *Scientific Reports*, 13:12399, 2023. doi: 10.1038/s41598-023-37829-z. URL https://doi.org/10.1038/s41598-023-37829-z.

Foumani, N. M., Tan, C. W., Webb, G. I., and Salehi, M. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024.

Freepik. Hypertension; blood pressure gauge; motion sensor; student sleeping in class; diabetes; blood cells; edge computing; galvanic skin response; motion sensor; accelerometer sensor; eeg, n.d. URL prefix: https://www.flaticon.com/free-icon/ , IDs: hypertension_4939229; blood-pressure-gauge_3184052; motion-sensor_2818201; student-sleeping-in-class_43739; diabetes_2750352; blood-cells_3400003; edge-computing_11068838;galvanic-skin-response_11228469; motion-sensor_17881894; accelerometer-sensor_11330476; eeg_9851782.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Hassani, T. Federated emotion recognition with physiological signals-gsr, 2021.

He, J., Zhang, Q., Wang, L., and Pei, L. Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. *IEEE Sensors Journal*, 19(6):2287–2297, 2018.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners, 2021. URL `https://arxiv.org/abs/2111.06377`.

Hosni, A. and Atef, M. Remote real-time heart rate monitoring with recursive motion artifact removal using ppg signals from a smartphone camera. *Multimedia Tools and Applications*, 82(13): 20571–20588, 2023.

Hu, K., Ivanov, P. C., Chen, Z., Carpena, P., and Stanley, H. E. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1):011114, 2001.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen, 2023. URL `https://arxiv.org/abs/2207.06405`.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax, 2017.

Jiang, W.-B., Zhao, L.-M., and Lu, B.-L. Large brain model for learning generic representations with tremendous eeg data in bci, 2024. URL `https://arxiv.org/abs/2405.18765`.

Jolliffe, I. T. and Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Kachuee, M., Kiani, M. M., Mohammadzade, H., and Shabany, M. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Transactions on Biomedical Engineering*, 64(4):859–869, 2016.

Kazemnejad, A., Gordany, P., and Sameni, R. EPHNOGRAM: A Simultaneous Electrocardiogram and Phonocardiogram Database (version 1.0.0), 2021. URL `https://doi.org/10.13026/tjtq-5911`.

Kim, D., Lee, J., Cho, M., and Kwak, S. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20083–20093, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Krishnan, P., Yaacob, S., Krishnan, A. P., Rizon, M., and Ang, C. K. Eeg based drowsiness detection using relative band power and short-time fourier transform. *J. Robotics Netw. Artif. Life*, 7(3): 147–151, 2020.

LeMoult, J. From stress to depression: Bringing together cognitive and biological science. *Current Directions in Psychological Science*, 29(6):592–598, 2020.

Li, H., Derrode, S., and Pieczynski, W. An adaptive and on-line imu-based locomotion activity classification method using a triplet markov model. *Neurocomputing*, 362:94–105, 2019.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Liang, Y., Chen, Z., Liu, G., and Elgendi, M. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):1–7, 2018. doi: 10.6084/m9.figshare.5459299.v5.

Ma, H., Zhang, Z., Li, W., and Lu, S. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–25, 2021.

Mathew, G., Barbosa, D., Prince, J., and Venkatraman, S. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health*, 1(1):25, Oct 2024. ISSN 2948-2836. doi: 10.1038/s44325-024-00027-5. URL https://doi.org/10.1038/s44325-024-00027-5.

McKeen, K., Oliva, L., Masood, S., Toma, A., Rubin, B., and Wang, B. Ecg-fm: An open electrocardiogram foundation model, 2024. URL https://arxiv.org/abs/2408.05178.

Mikelsons, G., Smith, M., Mehrotra, A., and Musolesi, M. Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. In *ML4H Workshop at 31st Conference on Neural Information Processing Systems (NIPS)*, 2017. URL https://arxiv.org/abs/1711.06350.

Min, J., Wang, P., and Hu, J. The original EEG data for driver fatigue detection. *figshare.Dataset.*, 7 2017. doi: 10.6084/m9.figshare.5202739.v1.

Muzammil, M. Finetuning endevsols/tinyllama-2.5t-clinical model on clinical dataset., 2021. URL https://huggingface.co/muzammil-eds/tinyllama-2.5T-Clinical-v2.

Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Abdel-Ghaffar, S., and McDuff, D. Scaling wearable foundation models, 2024. URL https://arxiv.org/abs/2410.13638.

Nedorubova, A., Kadyrova, A., and Khlyupin, A. Human activity recognition using continuous wavelet transform and convolutional neural networks. *arXiv preprint arXiv:2106.12666*, 2021a.

Nedorubova, A., Kadyrova, A., and Khlyupin, A. Human activity recognition using continuous wavelet transform and convolutional neural networks. *arXiv preprint arXiv:2106.12666*, 2021b.

Negi, P. C., Giri, H., Sharma, S., Sharma, N., et al. A comparative study of scalograms for human activity classification. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pp. 1–5. IEEE, 2024.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/c1fea270c48e8079d8ddf7d06d26ab52-Paper.pdf.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior, 2023.

Pillai, A., Spathis, D., Kawsar, F., and Malekzadeh, M. Papagei: Open foundation models for optical physiological signals, 2024. URL https://arxiv.org/abs/2410.20542.

Pimentel, M. A. F., Johnson, A. E. W., Charlton, P. H., Birrenkott, D., Watkinson, P. J., Tarassenko, L., and Clifton, D. A. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017. doi: 10.1109/TBME.2016.2613124.

Qian, B. and Rasheed, K. Hurst exponent and financial market predictability. In *IASTED conference on Financial Engineering and Applications*, pp. 203–209. Proceedings of the IASTED International Conference Cambridge, MA, 2004.

Qian, H., Pan, S. J., and Miao, C. Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions. *Artificial Intelligence*, 292:103429, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Reiss Attila, Indlekofer Ina, S. P. PPG-DaLiA. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C53890.

Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C54S4K.

Roelofs, K. Freeze for action: neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1718):20160206, 2017.

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

Semenoglou, A.-A., Spiliotis, E., and Assimakopoulos, V. Image-based time series forecasting: A deep convolutional neural network approach. *Neural Networks*, 157:39–53, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.10.006. URL https://www.sciencedirect.com/science/article/pii/S0893608022003902.

Sengupta, R., Polian, I., and Hayes, J. P. Wavelet transform assisted neural networks for human activity recognition. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1254–1258. IEEE, 2022.

Slapničar, G., Mlakar, N., and Luštrek, M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19(15):3420, 2019.

Thompson, J. M. T., Stewart, H. B., and Turner, R. Nonlinear dynamics and chaos. *Computers in Physics*, 4(5):562–563, 1990.

Torrence, C. and Compo, G. P. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

Vaid, A., Jiang, J., Sawant, A., Lerakis, S., Argulian, E., Ahuja, Y., Lampert, J., Charney, A., Greenspan, H., Narula, J., Glicksberg, B., and Nadkarni, G. N. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6(1):108, Jun 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00840-9. URL https://doi.org/10.1038/s41746-023-00840-9.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Vishnupriya, S. and Meenakshi, K. Automatic music genre classification using convolution neural network. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4, 2018. doi: 10.1109/ICCCI.2018.8441340. URL https://ieeexplore.ieee.org/document/8441340.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. Cbramod: A criss-cross brain foundation model for eeg decoding, 2025. URL https://arxiv.org/abs/2412.07236.

Wimmer, C. and Rekabsaz, N. Leveraging vision-language models for granular market change prediction, 2023. URL https://arxiv.org/abs/2301.10166.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. Determining lyapunov exponents from a time series. *Physica D: nonlinear phenomena*, 16(3):285–317, 1985.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

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

Yan, Y., Huang, Y.-C., Zhao, J., Liu, Y.-S., Ma, L., Yang, J., Yan, X.-D., Xiong, J., and Wang, L. Topological nonlinear analysis of dynamical systems in wearable sensor-based human physical activity inference. *IEEE Transactions on Human-Machine Systems*, 53(4):792–801, 2023a. doi: 10.1109/THMS.2023.3275774.

Yan, Y., Huang, Y.-C., Zhao, J., Liu, Y.-S., Ma, L., Yang, J., Yan, X.-D., Xiong, J., and Wang, L. Topological nonlinear analysis of dynamical systems in wearable sensor-based human physical activity inference. *IEEE Transactions on Human-Machine Systems*, 53(4):792–801, 2023b. doi: 10.1109/THMS.2023.3275774.

Yang, C., Westover, M., and Sun, J. Biot: Biosignal transformer for cross-data learning in the wild. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b30f3e2dd9cb53bbf2024402d02295-Paper-Conference.pdf.

Yuan, H., Chan, S., Creagh, A. P., Tong, C., Acquah, A., Clifton, D. A., and Doherty, A. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 7(1), April 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01062-3. URL http://dx.doi.org/10.1038/s41746-024-01062-3.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2017.

Zhang, W., Yang, L., Geng, S., and Hong, S. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35: 3988–4003, 2022.

Zhang, X., Chowdhury, R. R., Gupta, R. K., and Shang, J. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*, 2024a.

Zhang, X., Teng, D., Chowdhury, R. R., Li, S., Hong, D., Gupta, R. K., and Shang, J. Unimts: Unified pre-training for motion time series, 2024b. URL https://arxiv.org/abs/2410.19818.

A Datasets

Few openly accessible multi-channel or multi-device datasets for physiological signals exist, limiting advancements in this field. To address this gap, we curated a dataset comprising approximately 385 hours of recordings. Using the augmentation algorithm described below, we expanded this dataset to 4294 hours. The distribution of the pretraining dataset, as shown in Figure 6, reflects the inherent diversity of the original recordings, ensuring balanced representation across channels and devices. This curated and augmented dataset provides a critical resource for developing robust models, facilitating progress in multi-channel physiological signal research.

Table 4: **Downstream evaluation data that are unseen during pretraining.**

Downstream Dataset	Sensor	# Channels	Tasks	#Samp. (#Subj.)
WESAD (Schmidt et al., 2018)	IMU, PPG, ECG, GSR	10	Stress Detection	11050(15)
UCI-HAR (Reyes-Ortiz et al., 2012)	IMU	6	HAR	10299(30)
DriverFatigue (Min et al., 2017)	EEG	4	Fatigue Detection	2400(12)
Activity Recognition Total	-	-	-	**23749(57)**
Epilepsy (Andrzejak et al., 2023)	EEG	1	State Recognize	11500(500)
GAMEEMO (Alakus et al., 2020)	EEG	4	Valence-Arousal	5600(28)
EEG Main Tasks Total	-	-	-	**17100(528)**
ECG-Abnormal (Bousseljot et al., 2009)	ECG	1	Abnormal Detection	11640(249)
PPG-BP (Liang et al., 2018)	PPG	1	Risk of Diseases	657(219)
PhysioNet EMG (Goldberger et al., 2000)	EMG	1	Muscular Diseases	163(3)
Risk Evaluation Total	-	-	-	**12460(471)**
Noninvasive-BP (Esmaili et al., 2017)	PPG	3	BP Estimate	125(26)
PPG-Hgb (Esmaili et al., 2017)	PPG	2	Hgb Estimate	68(68)
Fetal-fPCG (Bhaskaran et al., 2022)	PCG	1	Fetal HR Estimate	47(47)
Vital Signs Total	-	-	-	**240(141)**
Total All	-	-	-	**53549(1197)**

Table 5: **Pretraining data.**

Pretrain Dataset	Sensors	#Samp (hours).
Cuff-Less-BP (Kachuee et al., 2016)	ECG, PPG	42934(72)
PPG-Dalia (Reiss Attila, 2019)	ECG, PPG IMU, GSR	42606(71)
Auditory-EEG (Alzahab et al., 2022)	EEG	13601(23)
PhyAAt (Bajaj et al., 2020)	EEG	19550(33)
MAUS (Beh et al., 2021)	ECG, PPG GSR	13068(22)
Mendeley-YAAD (Dar et al., 2022)	ECG, GSR	2964(5)
Brain-Cognitive (Fekri Azgomi et al., 2023)	EEG	51201(85)
EPHNOGRAM (Kazemnejad et al., 2021)	ECG, PCG	36611(61)
BIDMC (Pimentel et al., 2017)	ECG, PPG	8427(14)
Num Segments (# Segm.)	-	230,962(385)
# Segm. w/ Augment	-	2,576,418(4,294)
Num Sensor Signals (# Sign.)	-	**802,019(1,337)**
# Sign. w/ Augment	-	**8,965,538(14,943)**

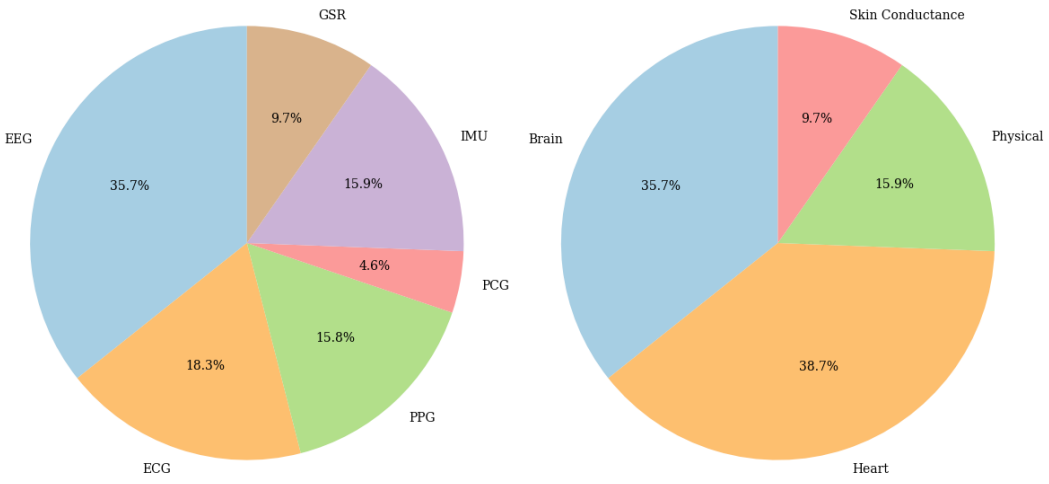


Figure 6: **Distribution of sensor signals used for pretraining.** *Left:* Distribution by sensor modality. *Right:* Distribution by type of physiological information.

Table 4 overviews used dataset in our experiement along with the modality and task type. We will gives further details for each dataset below:

WESAD (Schmidt et al., 2018) is a publicly available multimodal dataset used for wearable stress and affect detection, formulated as a classification task with labels: neutral, stress, and amusement. The dataset includes physiological and motion data collected from 15 subjects during a lab study,

using a chest-worn RespiBAN device and a wrist-worn Empatica E4 device. From the chest device, we use electrocardiogram (ECG), galvanic skin response (GSR), and triaxial acceleration (ACC-X, ACC-Y, ACC-Z), all sampled at 700 Hz. From the wrist device, we use photoplethysmogram (PPG), galvanic skin response (GSR, 4 Hz), and triaxial acceleration (ACC-X, ACC-Y, ACC-Z, 32 Hz). The selected channels span multiple physiological and motion modalities from both chest and wrist sensors. Each data segment is labeled with one of the three affective states, serving as the target output for classification tasks.

UCI-HAR (Reyes-Ortiz et al., 2012) dataset is publicly available and is used for classifying human activities based on sensor data. It comprises data from 30 volunteers, aged 19 to 48, each performing six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. During these activities, participants carried a waist-mounted smartphone equipped with embedded accelerometer and gyroscope sensors. The input channels consist of triaxial linear acceleration and triaxial angular velocity, totaling six channels. Each data segment is labeled with one of the six activities, serving as the target output for classification tasks. The sensors recorded data at a constant rate of 50 Hz.

Driver Fatigue EEG Dataset (Min et al., 2017) is a publicly available dataset used for detecting driver fatigue based on electroencephalogram (EEG) signals. EEG data were collected using a 40-channel Neuroscan amplifier. The recordings include EEG data corresponding to two states: alert and fatigued. Each data segment is labeled with one of these states, serving as the target output for classification tasks.

Epileptic Seizure Recognition (Andrzejak et al., 2023) dataset is publicly available and is used for classifying neurological and physiological states based on EEG signals. It comprises data from 500 subjects, each recorded for 23.6 seconds using a single EEG channel at a sampling rate of 178 Hz. Each sample is labeled with one of five brain states, allowing for the construction of multiple binary classification tasks that target different aspects of neurological assessment. Specifically, we formulated five tasks:

- *Eye Relaxation*: Detects eye fatigue by distinguishing between relaxed and alert states based on EEG changes related to eye closure.
- *Health Area*: Classifies brain regions as healthy or affected by neurological abnormalities.
- *Tumor Area*: Detects EEG patterns indicative of tumor presence in specific brain regions.
- *Seizure*: Identifies seizure activity from non-seizure states.
- *Eyes Open vs. Closed*: Differentiates EEG signals associated with visual input states.

GAMEEMO (Alakus et al., 2020) is a publicly available dataset used for emotion recognition based on EEG signals. It comprises data from 28 subjects, each playing four emotion-inducing computer games (boring, calm, horror, and funny) for five minutes per game, totaling 20 minutes of EEG data per subject. EEG signals were recorded using the EMOTIV EPOC+ headset, which includes 14 channels (AF3, AF4, F3, F4, F7, F8, FC5, FC6, O1, O2, P7, P8, T7, and T8) positioned according to the 10–20 system. The signals were sampled at 128 Hz. After each gameplay session, subjects rated their emotional response using the Self-Assessment Manikin (SAM) form, providing continuous scores for arousal and valence. These scores were quantized into binary values using subject-specific median thresholds: arousal and valence ratings above the median were labeled as high, and those below or equal to the median as low. Combining the binarized arousal and valence ratings yields four discrete emotional classes: low arousal and low valence, low arousal and high valence, high arousal and low valence, and high arousal and high valence. Each data segment is labeled with one of these four classes, serving as the target output for four-class emotion classification tasks.

ECG Heartbeat Categorization (Bousseljot et al., 2009) is a publicly available dataset used for classifying heartbeat signals based on electrocardiogram (ECG) recordings. It comprises two collections of heartbeat signals derived from PhysioNet's MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database. The first collection includes 109,446 samples categorized into five classes: normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and unknown (Q), with ECG signals sampled at 125 Hz. The second collection consists of 14,552 samples categorized into two classes: normal and abnormal, also sampled at 125 Hz. For our analysis, we restructured the dataset into a binary classification framework by consolidating the original categories into two classes: normal and abnormal heartbeats.

PPG-China (Liang et al., 2018) is a publicly available dataset used for classifying cardiovascular and metabolic conditions based on photoplethysmography (PPG) signals. It comprises 657 data records from 219 subjects, aged 20 to 89 years, including individuals with conditions such as hypertension and diabetes. PPG signals were recorded using a single channel at a sampling rate of 125 Hz. Each subject's data includes PPG waveforms and corresponding clinical information, facilitating the construction of multiple classification tasks focused on cardiovascular and systemic health monitoring. Specifically, we formulated four tasks:

- *PPG-HTN*: Identifies stages of hypotension severity by classifying PPG signals into four levels.
- *PPG-DM*: Detects diabetes by distinguishing between diabetic and non-diabetic individuals.
- *PPG-CVA*: Identifies the presence or absence of cerebrovascular accidents (strokes) based on PPG patterns.
- *PPG-CVD*: Assesses cardiovascular disease by classifying PPG signals into three cardiovascular health categories.

PhysioNetEMG (Goldberger et al., 2000) is a publicly available dataset used for classifying neuromuscular conditions based on electromyography (EMG) signals. It comprises single-channel EMG recordings from the tibialis anterior muscle of three subjects: one healthy, one with neuropathy, and one with myopathy. The EMG signals were recorded at a sampling rate of 4,000 Hz. Each recording was segmented into time series samples using a fixed-length window of 6 second. Each segment is labeled according to the subject's condition—healthy, neuropathy, or myopathy—serving as the target output for classification tasks.

Non-invasive Blood Pressure Estimation (Esmaili et al., 2017) is a publicly available dataset used for cuff-less blood pressure (BP) estimation. It comprises data from 26 subjects, each with recorded electrocardiogram (ECG) and photoplethysmogram (PPG) signals, sampled at 1,000 Hz. Reference BP measurements were taken during signal acquisition. Each subject's data also includes demographic information such as age, weight, and height. The dataset is structured to facilitate regression tasks aimed at predicting systolic and diastolic BP values.

PPG-HGB (Abuzairi et al., 2024) is a publicly available dataset used for non-invasive hemoglobin (Hb) measurement based on photoplethysmography (PPG) signals. It comprises data from 68 subjects, aged 18 to 65 years, with a gender distribution of 56% female and 44% male. PPG signals were recorded using the MAX30102 sensor, which emits red and infrared light. The sensor's analog-to-digital converter (ADC) output data rate can be programmed from 50 samples per second (sps) to 3200 sps. Each subject contributed 12 sets of PPG signals, totaling 816 data records. We formulate regression tasks aimed at predicting Hb concertration levels.

Fetal-fPCG (Bhaskaran et al., 2022) is a publicly available dataset designed for estimating fetal heart rate (FHR) using fetal phonocardiography (fPCG) signals. It includes recordings from 60 pregnant women, aged 18 to 37 years, with gestational ages between 31 and 40 weeks. The recordings were collected at St. John's Hospital in Bangalore using an electronic stethoscope (SS30LA) connected to a Biopac MP36 data acquisition system. The stethoscope was placed on the lower abdomen of each subject to capture the fPCG signal, which was sampled at 2,000 Hz. The dataset supports regression tasks, where the goal is to predict continuous FHR values directly from the fPCG waveforms.

B Implementation Detail

B.1 Data Preprocess.

For the data preparation, we set the uniform sampling rate and interval length to 65 HZ and 6 seconds respectively. In our case, 65 Hz covers most of the frequency bands of interest such as heart activity, physical motions, and neuron activity up to the beginning of Gamma power (above 30 Hz). And a great amount of samples are less than 6 seconds such as (Reyes-Ortiz et al., 2012; Liang et al., 2018; Bousseljot et al., 2009). We conduct basic pre-processing for each signal with identical setting: (1) de-trended by subtract the result of a linear least-squares fit to series data from the raw time series, and (2) Gaussian smoothed with standard deviation of 1.3 (0.02 seconds), ensuring a highly consistent dataset for training.

Since the Transformer's computational requirements scale quadratically with input length, to release the full potential of our self-supervised algorithm, we segment our multivariate time series into

intervals with a uniform length and pad shorter samples with zeros. This approach not only enables parallel processing of samples in large minibatches but also addresses variation in the length of individual samples.

For the downstream task, we split the data into train and test sets for linear probing evaluation with portion of 80% and 20% correspondingly. The split is stratified on the anonymized subject ID if this information is provided by the dataset.

B.2 Data Augmentation.

Since there are very few publicly available datasets containing multiple devices or modalities, we aim to expand our curated training set to fully leverage the potential of self-supervised learning. Inspired by data augmentation techniques in computer vision and natural language processing (Zhang et al., 2017; Carmona et al., 2021), we adopt a heuristic approach to augment the dataset. Specifically, we augment each sub-dataset by a factor of 10. For each dataset, we sample two time series, randomly extract a segment from one, and substitute it with a transformed counterpart, as outlined in the pseudocode in Algorithm 1. As a result, our training set is expanded to 2,586,404 segments, corresponding to 4,294 hours of data.

Algorithm 1 Time Series Mixup Augmentation

Input: Time series dataset $\mathcal{X}$, number of augmentations n
Output: Augmented Dataset $\tilde{\mathcal{X}}$

1: **for** $i = 1$ to n **do**
2: Sample two time series $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \sim \mathcal{X}$
3: Sample a chunk size $\lambda \sim \mathcal{U}(0, l)$
4: Sample start indices $s_1, s_2 \sim \mathcal{U}(0, l - \lambda)$
5: Swap chunk from $\mathbf{x}^{(2)}$ into $\mathbf{x}^{(1)}$:

$$\mathbf{x}^{(1)}_{s_1:s_1+\lambda} \leftarrow \mathbf{x}^{(2)}_{s_2:s_2+\lambda}$$

6: Append $\mathbf{x}^{(1)}$ into $\tilde{\mathcal{X}}$
7: **end for**
8: **return** $\tilde{\mathcal{X}}$

B.3 Pretraining Framework.

Normwear is derived from the Masked Autoencoder (MAE) (He et al., 2021). The detailed hyper-parameter choice is descibe in 6. We use a Conv2D layer with a kernel size of (9, 5) and a stride of (9, 5), ensuring no overlapping patches. This layer takes input with 3 channels and projects it to 768 channels, matching the hidden size of our encoders. In Normwear, we apply structured masking independently to each variate along both the frequency and time axes, with respective masking ratios of 0.6 and 0.5. This results in an expected overall masking ratio of 0.8 for each variate. Only the unmasked tokens are passed to the encoder, reducing computational complexity. To enhance representation learning, we introduce six additional transformer blocks as fusion layers, interleaved with the original 12 encoder blocks, creating a total of 18 blocks. Each transformer block has a hidden dimension of 768 and uses LayerNorm as in the original MAE. The latent embeddings obtained from the encoder are projected from 768 to 512 dimensions. Learnable masked tokens are reinserted at their original positions, and positional embeddings are added to guide the decoder in reconstructing the input series. The lightweight decoder consists of two transformer blocks with a hidden dimension of 512, followed by two Conv1D layers. The first Conv1D layer maps from the flattened multivariate signal embedding to an intermediate dimension, and the second Conv1D layer maps from this intermediate dimension back to the original multivariate signal space. A GELU activation function is used between these layers, with BatchNorm applied to the input. The decoder reconstructs the original input series, and the model is trained using Mean Squared Error (MSE) loss on all data points. Our models are pre-trained for 45,000 steps with a batch size of 256, using the AdamW optimizer with a learning rate of 10^{-4}. We did not perform on-the-fly data augmentation,

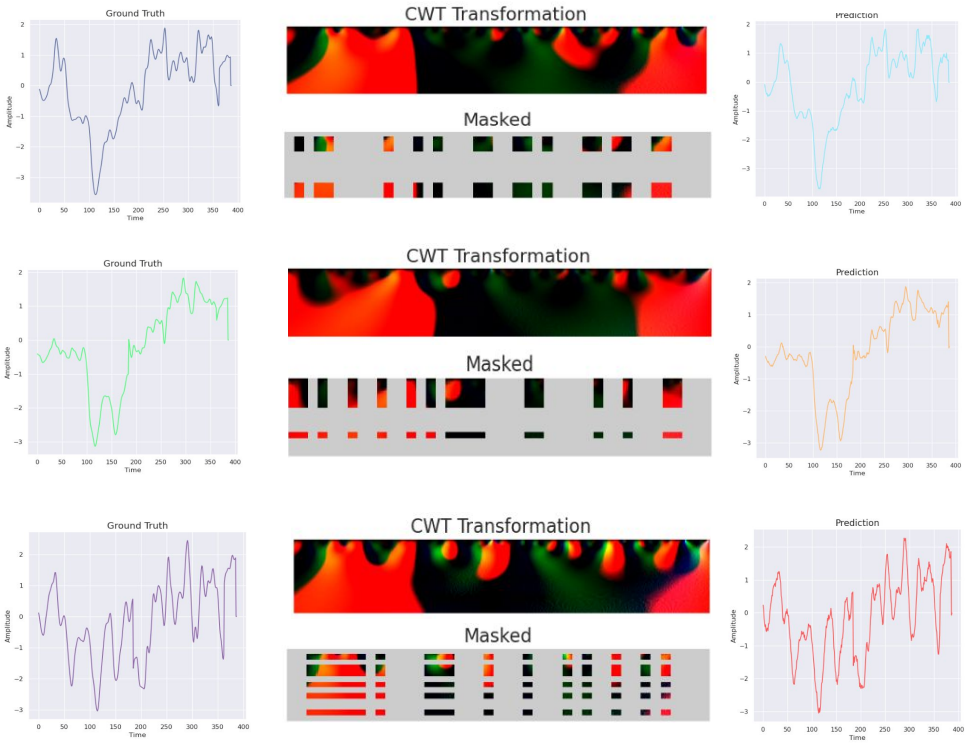


Figure 7: **Visualization of original time series (left), CWT transformation image with structured masking (middle), and reconstructed time series (right).**

as suggested in the MAE framework, due to the high masking ratio. (An end-to-end example of the input and output of this pretraining pipeline is illustrated in Fig. 7)

All the models are pretrained on 4 NVIDIA RTX 3090 graphical computing unit (GPU), with 24GB of GPU memory on each card.

B.4 MSiTF.

For pretraining the representation alignment module, we have the training hyper-parameters in Table 7.

Table 6: **NormWear Pretraining Hyper-parameters.**

Hyper-parameter	Value
# cross-patches Transformer Encoder	12
# cross-channels Transformer Encoder	6
# Transformer Decoder	2
# Attention Heads	12
Encoder Latent Size	768
Decoder Latent Size	512
Feedforward Latent Size	3072
Normalization	LayerNorm
Patch size (time axis)	9
Patch size (scale axis)	5
Optimizer	AdamW
Loss Scalar	NativeScaler
Base Learning Rate (blr)	1e-3
Epochs	140
Batch size	192

Table 7: **MSiTF Hyper-parameter**

Hyper-parameter	Value
Learning rate (lr)	1e-3
Epochs	40
Batch size	32
L2 regularization	5e-6
lr decay rate	0.997
λ	0.5
τ	0.5

B.5 Aligner Module, Objective Function, and Pretraining.

The Aligner Module matches two vectors: the fused representation $f(q, H) = \hat{Y} \in \mathbb{R}^E$ with the semantic embedding (Y) of ground truth sentence, which is obtained from prompting the ground truth label using a template, for example, *"The subject is presently {activity_label}"*. In the same manner as the query embedding, the ground truth sentence is encoded using the same pre-trained language model (Muzammil, 2021). At this stage, Y is leveraged to supervise the fused output $\hat{Y}$. The vanilla contrastive learning loss formula following Zhang et al. (2024b) is:

$$Loss_{ctl}(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\hat{Y_i}^T Y_i)^{\frac{1}{\gamma}}}{\sum_{k=1}^{N}\exp(\hat{Y_i}^T Y_k)^{\frac{1}{\gamma}}} \tag{2}$$

where N is the batch size and γ is the learnable temperature parameter. We denote this loss function as contrastive loss with batch normalizer. We also leverage a refine process after contrastive learning using simlarity loss with per sample normalizer, which is essentially cosine similarity loss, with vector distance as supplemental penalty:

$$Loss_{refine}(\hat{Y}, Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\left(1 - \frac{\hat{Y}_i^T Y_i}{\|\hat{Y}_i\|\|Y_i\|}\right) + \lambda|\hat{Y}_i - Y_i|\right) \tag{3}$$

where λ is hyper-parameters controlling the weight of the supplemental loss components.

B.6 Sentence template example for signal-sext alignment.

To enhance the expressiveness and diversity of supervision signals for our MSiTF alignment module, we convert categorical labels into natural language descriptions using varied prompt templates. We apply this strategy to several pretraining tasks. We present example sentence templates below for emotion recognition and activity recognition to demonstrate the general idea of how we derive text modality from the raw label:

For the **emotion recognition** task, we use:

- " The emotion detected is {}."
- " This subject is feeling {}."
- " The emotional state is {}."
- " The identified emotion is {}."

For the **activity recognition** task, we use:

- " This subject is currently {}."
- " The subject is engaged in {}."
- " Current activity is {}."
- " Subject's activity is {}."

By exposing the model to multiple phrasings for the same label, this design helps it learn modality-invariant representations that are more robust to linguistic variation and better aligned across modalities. Specifically, to increase the diversity of semantic representations of query and ground truth sentences in the pretraining signal corpus, we utilize large language models (GPT-3.5) (Achiam et al., 2023) to generate 20 alternative variations for each sentence, from which only one is randomly sampled during pre-training. During test-time inference on downstream datasets, each ground truth label is converted into a sentence (details in appendix B.6), which is transformed into a semantic embedding using a frozen text encoder. The sentence with the closest distance with the embedding from our foundation model is used as the final inferential result.

B.7 Statistical Feature List.

Our statistical baseline includes features extracted from both the time and frequency domains. In the time domain, we compute the mean, standard deviation, maximum, minimum, skewness, kurtosis,

25% quantile, median, and 75% quantile. In the frequency domain, we extract the spectral centroid, spectral spread, mean frequency, peak frequency, as well as the 25%, 50% (median), and 75% quantile frequencies.

B.8 Radar Plot or Performance Trend.

To enhance the visual contrast between model performances across tasks, we applied the Softmax function to the raw performance scores. This transformation rescales the scores to a range between 0 and 1, accentuating relative differences between models. While the Softmax transformation emphasizes the relative improvement of our model over others, we note that the absolute scores may differ from those originally reported.

C Ablation Study

Due to computational constraints, we will conduct the ablation study on our smaller dataset (37k samples) to train and evaluate the model, establishing a proof of concept and demonstrating the effectiveness of our approach in a controlled setting.

Fusion Schemes. Table 8 shows the performance of different fusion schemes, including (1) no fusion, (2) cross-attention fusion, (3) [CLS]-attention fusion, and (4) mean-pooling fusion. We excluded all-attention fusion in our ablation study because it is computationally prohibitible. Among all the compared strategies, the [CLS] token fusion generally achieves the best accuracy with a minor increase in parameters.

Table 8: **Performance Comparison of Various Fusion Schemes**

Downstream Tasks	No fusion	Cross-Attention fusion	Mean pooling fusion	[CLS] Token fusion
WESAD	72.209	74.165	71.99	**75.390**
UCI-HAR	97.793	96.908	97.566	**98.928**
DriverFatigue	73.252	60.308	72.552	**75.167**
Activity Recognition Avg.	81.085	77.127	80.703	**83.162**
Epilepsy (eye open state)	90.966	84.075	89.817	**92.203**
Epilepsy (eye relaxation)	94.399	93.589	93.912	**94.908**
Epilepsy (health area)	87.866	86.899	87.248	**88.117**
Epilepsy (tumor area)	86.599	86.861	**87.152**	86.888
Epilepsy (seizure)	**97.477**	96.351	96.719	96.638
GAMEEMO	**57.695**	56.724	58.079	56.532
EEG Main Tasks Avg.	85.834	84.083	85.488	**85.881**
ECG-Abnormal	99.429	**99.441**	99.268	99.041
PPG-BP (HTN)	61.850	60.983	**63.577**	60.344
PPG-BP (DM)	58.333	**62.800**	62.200	59.459
PPG-BP (CVA)	61.319	61.458	59.236	**70.278**
PPG-BP (CVD)	48.417	**53.585**	46.961	52.596
PhysioNet EMG	93.715	95.49	86.749	**98.184**
Risk Evaluation Avg.	70.511	72.293	69.665	**73.317**
Noninvasive-BP	88.356	**92.759**	88.719	92.470
PPG-Hgb	95.031	93.413	**95.086**	94.766
Fetal-fPCG	98.582	**99.145**	98.771	99.088
Vital Signs Avg.	93.990	95.106	94.192	**95.441**
Micro Avg.	81.294	80.831	80.867	**82.833**
Macro Avg.	82.855	82.152	82.512	**84.450**

Masking Strategies in Pre-training. We ablated our masking strategy introduced in Section 3.4. Using a consistent mask ratio of 0.8 in all strategies, we found that applying masking along the scale and time axes produced the best performance (details in Table 9).

Table 9: Performance Comparison of Different Masking Strategies

Downstream Tasks	**Unstructured Mask** ($P = 0.8$)	**Time Mask** ($P_t = 0.8, P_f = 0.0$)	**Scale Mask** ($P_t = 0.0, P_f = 0.8$)	**Structured Mask** ($P_t = 0.6, P_f = 0.5$)
WESAD	71.46	71.952	72.201	**75.390**
UCI-HAR	97.097	98.438	98.106	**98.928**
DriverFatigue	72.719	73.424	**78.354**	75.167
Activity Recognition Avg.	80.425	81.271	82.887	**83.162**
Epilepsy (eye open state)	89.521	91.895	89.407	**92.203**
Epilepsy (eye relaxation)	93.471	94.808	93.786	**94.908**
Epilepsy (health area)	86.812	**88.510**	87.317	88.117
Epilepsy (tumor area)	86.524	**88.254**	85.502	86.888
Epilepsy (seizure)	96.59	**97.791**	95.29	96.638
GAMEEMO	**58.043**	56.770	55.771	56.532
EEG Main Tasks Avg.	85.160	**86.338**	84.512	85.881
ECG-Abnormal	99.085	**99.316**	98.296	99.041
PPG-BP (HTN)	58.880	55.333	59.230	**60.344**
PPG-BP (DM)	61.074	48.386	58.896	**59.459**
PPG-BP (CVA)	56.389	58.472	64.167	**70.278**
PPG-BP (CVD)	52.572	46.557	**55.666**	52.596
PhysioNet EMG	85.160	95.490	83.922	**98.184**
Risk Evaluation Avg.	68.860	67.259	70.030	**73.317**
Noninvasive-BP	90.124	90.650	91.152	**92.470**
PPG-Hgb	**95.314**	95.055	94.713	94.766
Fetal-fPCG	98.630	**99.121**	98.926	99.088
Vital Signs Avg.	94.689	94.942	94.930	**95.441**
Micro Avg.	80.526	80.568	81.150	**82.833**
Macro Avg.	82.284	82.453	83.090	**84.450**

Input Representations. Table10 compares the performance of two input representations: (1) CWT scalogram and (2) raw time series. The CWT scalogram converts the time series into a time-frequency representation, while the raw time series retains the original sensor data. Among the two representations, the model trained on CWT scalograms demonstrates better performance, suggesting that the time-frequency features enhance model accuracy.

Table 10: Performance Comparison Between CWT Scalogram and Raw Time Series as Inputs.

Downstream Tasks	Raw Series Input	CWT Scalogram Input
WESAD	70.862	**75.390**
UCI-HAR	97.969	**98.928**
DriverFatigue	73.854	**75.167**
Activity Recognition Avg.	80.895	**83.162**
Epilepsy (eye open state)	91.978	**92.203**
Epilepsy (eye relaxation)	94.781	**94.908**
Epilepsy (health area)	88.045	**88.117**
Epilepsy (tumor area)	85.619	**86.888**
Epilepsy (seizure)	**97.722**	96.638
GAMEEMO	54.651	**56.532**
EEG Main Tasks Avg.	85.466	**85.881**
ECG-Abnormal	97.701	**99.041**
PPG-BP (HTN)	52.614	**60.344**
PPG-BP (DM)	**62.012**	59.459
PPG-BP (CVA)	56.181	**70.278**
PPG-BP (CVD)	**54.812**	52.596
PhysioNet EMG	93.756	**98.184**
Risk Evaluation Avg.	69.513	**73.317**
Noninvasive-BP	89.850	**92.470**
PPG-Hgb	93.832	**94.766**
Fetal-fPCG	98.977	**99.088**
Vital Signs Avg.	94.220	**95.441**
Micro Avg.	80.845	**82.833**
Macro Avg.	82.523	**84.450**

Semi-Supervised Learning (Partial-shot). To evaluate the generalizability and quality of learned representations, we conducted a semi-supervised learning evaluation following the protocol established by prior self-supervised methods (Caron et al., 2021). Specifically, we assessed performance on the NORMWEAR dataset using frozen features and a limited labeled subset (10%). We deliberately excluded the commonly used 1% label evaluation due to the inherently small sample size of our downstream medical dataset. A 1% labeling scenario would provide fewer than ten labeled instances, rendering the results statistically unreliable and scientifically unjustified. Instead, we sampled 10% of the training data while preserving the original label distribution, and then trained a linear classifier atop the frozen NORMWEAR features for classification tasks and regression tasks. The results, summarized in Table 11, demonstrate the effectiveness of our method under realistic semi-supervised constraints.

Table 11: **Semi-supervised learning on Downstream tasks**. We linear-prob the model with 10% labels and report AUCROC scores.

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	Modality-Specific	NORMWEAR (Ours)
WESAD	64.869	64.908	68.626	62.218	59.371	**70.25**
UCI-HAR	94.124	73.124	92.794	92.334	-	**98.355**
DriverFatigue	63.237	72.454	50.193	54.613	**69.004**	55.094
Activity Recognition Avg.	74.077	70.162	70.538	69.722	-	**74.566**
Epilepsy (eye open state)	82.186	80.082	84.103	88.02	**89.152**	85.456
Epilepsy (eye relaxation)	87.480	81.820	88.716	93.670	**95.191**	92.369
Epilepsy (health area)	86.096	77.682	82.651	84.940	**87.377**	85.471
Epilepsy (tumor area)	82.153	78.364	82.579	85.450	**86.962**	83.033
Epilepsy (seizure)	88.179	96.786	**97.386**	92.900	94.063	92.345
GAMEEMO	**54.527**	50.176	51.952	49.714	52.046	52.633
EEG Main Tasks Avg.	80.104	77.485	81.231	82.449	**84.132**	81.885
ECG-Abnormal	96.420	**97.613**	95.432	94.769	79.918	93.921
PPG-BP (HTN)	52.491	49.407	48.397	53.800	**57.544**	53.967
PPG-BP (DM)	41.254	48.574	38.664	45.383	56.532	**57.545**
PPG-BP (CVA)	**83.056**	51.944	48.125	51.667	64.792	66.597
PPG-BP (CVD)	55.753	47.547	**59.505**	55.651	47.586	54.614
PhysioNet EMG	**92.993**	70.248	92.415	79.412	-	87.503
Risk Evaluation Avg.	**70.328**	60.888	63.756	63.447	-	69.025
Noninvasive-BP	90.589	**93.783**	91.614	92.707	92.671	90.694
PPG-Hgb	**95.068**	94.999	94.712	94.981	94.916	94.633
Fetal-fPCG	**99.020**	99.153	98.889	98.902	-	98.813
Vital Signs Avg.	94.892	**95.978**	95.072	95.530	-	94.713
Micro Avg.	78.305	73.815	75.931	76.174	-	**78.516**
Macro Avg.	79.850	76.129	77.649	77.787	-	**80.047**

Permutation-Invariant Input Channel Analysis. In many multimodal or multichannel sensing tasks, the input channel order is typically fixed and determined by hardware or preprocessing pipelines, limiting flexibility during deployment. This constraint raises the question of whether *Normwear* relies on a specific channel ordering to perform well. To examine this, we conducted an experiment on datasets with multiple input channels by circularly shifting the channel order by one position and evaluating the resulting model performance. As shown in Table 12, the model performance remains consistent across different permutations. These results suggest that our model does not rely on a fixed input channel configuration and is robust to variations in channel ordering, making it more applicable in practical scenarios where such inconsistencies may occur.

k-fold Analysis. To evaluate whether Normwear maintains consistent performance on datasets with limited subject diversity, we conducted 5-fold cross-validation stratified by subject ID. We applied this protocol to all downstream tasks containing 30 or fewer subjects to ensure a robust assessment. As shown in Table 13, our model consistently outperformed the baselines across all tasks, demonstrating the robustness of our evaluation metric.

Table 12: **Performance of NormWear with original input channel order compared to random shuffling across tasks.**

Task	Original Order	Random Shuffle
WESAD (IMU, PPG, ECG, GSR)	0.761	0.763
UCI-HAR (IMU)	0.989	0.975
Drive Fatigue (EEG)	0.743	0.721
GAMEEMO (EEG)	0.549	0.530
Noninvasive-BP (PCG, PPG, ECG)	0.924	0.914
PPG-HGB (Red, IR)	0.946	0.948

Table 13: **Performance on downstream health-related tasks under linear probing using 5-fold subject-stratified cross-validation.** Classification reports AUC ROC; regression reports relative accuracy. All metrics are higher-is-better.

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	NormWear-L (Ours)
WESAD	79.992 ± 0.707	83.332 ± 0.841	87.824 ± 0.463	82.701 ± 0.536	**89.585 ± 0.683**
UCI-HAR	95.602 ± 0.148	91.956 ± 0.256	96.864 ± 0.175	97.382 ± 0.138	**98.179 ± 0.06**
DriverFatigue	69.614 ± 1.138	**72.48 ± 2.848**	66.251 ± 0.471	65.026 ± 1.198	68.971 ± 1.32
GAMEEMO	64.281 ± 1.292	56.694 ± 0.878	64.119 ± 0.543	62.925 ± 0.999	**67.863 ± 0.72**
Noninvasive-BP	92.83 ± 0.386	92.223 ± 0.356	92.612 ± 0.272	88.707 ± 0.622	**93.381 ± 0.516**
Avg.	80.464 ± 0.734	79.337 ± 1.036	81.534 ± 0.385	79.348 ± 0.699	**83.596 ± 0.660**

Table 14: **Checking reliance on demographic information.** Simple baseline: for regression tasks (yellow), the mean prediction is used; for classification tasks (blue and red), the mode prediction is used. NormWear-Medium and NormWear-Large refer to NormWear's pretrained checkpoints trained on 2.58 million and 8.97 million signal segments, respectively.

Downstream Tasks	Empirical Distribution	Demographic	NormWear-Medium	Demographic + NormWear-Medium	NormWear-Large	Demographic + NormWear-Large
WESAD	50.000	49.907	**74.227**	69.06	**76.06**	68.755
Noninvasive-BP	92.988	92.954	91.427	90.84	92.42	92.528
PPG-Hgb	94.816	95.634	94.911	95.835	94.632	96.384
Fetal-fPCG	99.033	99.039	98.997	99.001	99.072	99.097
Vital Signs Avg.	95.612	**95.876**	95.112	95.225	95.375	**96.003**
PPG-BP (HTN)	50.000	59.899	62.746	64.482	62.341	61.291
PPG-BP (DM)	50.000	47.297	62.613	47.86	55.893	60.135
PPG-BP (CVA)	50.000	81.875	67.639	83.681	70.625	77.847
PPG-BP (CVD)	50.000	71.011	51.504	70.37	51.773	67.466
Risk Evaluation Avg.	50.000	65.021	61.126	**66.598**	60.158	**66.685**
Micro Avg.	67.105	74.702	75.508	**77.641**	75.352	**77.938**
Macro Avg.	65.204	70.268	76.821	**76.961**	**77.198**	77.148

Demographic Anlysis. Several previous works (Abbaspourazad et al., 2023; Narayanswamy et al., 2024) have used learned representations to infer demographic labels. These results suggest that wearable signals do contain demographic information. In Table 14, we wanted to investigate that NormWear does not extract only demographic information (e.g. age, sex, height, etc. depending on what is available within each dataset), hence indicating that the representation that our proposed model extracted and the demographic could be used as complementary features to each other during downstream modeling. From Table 14, we observe that demographic information and wearable signal representations each excel at different tasks. In most cases, concatenating them improves overall performance. However, the occasional performance drop after concatenation suggests a confounding relationship between the two, implying that demographic data and NormWear's wearable representations capture different aspects.

D Statistical significance on the model comparison

We performed a statistical analysis to test the significance of the differences in model performance. First, we ran the downstream evaluations 100 times for each model on every task without fixing the random seed. The results remained consistent due to the stable optimization process.

Next, we applied a permutation test on the results from these 100 runs to determine whether NormWear's AUC ROC score is greater than that of the baselines. The reported p-value represents the probability of observing a test statistic as extreme or more extreme than the observed difference under the null hypothesis, which assumes that NormWear's score is not higher than the baseline. In nearly all cases, the p-value is less than 0.01, confirming the statistical significance and indicating the robustness and superiority of our approach. Table 8 presents the statistical test results across different task groups (as indicated by the color coding in the main tables) along with the overall average scores.

We also include a critical difference (CD) diagram to visually compare the performance of multiple models across datasets and highlight statistically significant differences. To generate the CD diagram, we first conducted a Friedman Chi-square test on the models' scores across all downstream tasks, which yielded a p-value of $P < 0.001$, confirming that the models' performances come from different distributions. We then applied the Conover post hoc test to examine pairwise differences between model performances; the p-values for NormWear compared with the baselines are shown in the last row of Table 8. Finally, based on these results, we generated the CD diagram displayed in Figure 9. In this diagram, our proposed model, NormWear, is well separated from the others, indicating its statistical superiority over the competitive baselines.

Ours/Baselines	Stats	Chronos	CLAP	TFC
NormWear - activity	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - eeg	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - risk	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - vital	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - micro avg.	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - macro avg.	$P < .01$	$P < .01$	$P < .01$	$P < .01$
Conover post hoc	$P < .001$	$P < .001$	$P < .001$	$P < .05$

Figure 8: Permutation test on models' performance.

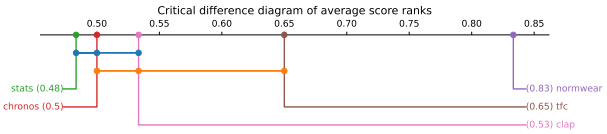


Figure 9: Critical Difference Diagram.

E Supplementary Metrics

Normwear's performance is summarized in Fig. 4 and detailed in Table 15. Normwear consistently exceeds the baseline models by a wide margin, demonstrating a clear advantage.

Table 15: **Details of Incidental Performance Metrics.**

Task Group	Methods	AUC ROC	AUC PR	Accuracy	Precision	Recall	F1 Score
Activity Recognition	Statistical	75.082	63.996	65.298	61.450	61.56	61.034
	Chronos	79.935	65.622	66.175	62.044	61.512	60.522
	CLAP	76.897	67.026	66.349	62.790	62.826	62.435
	TF-C	77.880	68.228	67.175	64.967	64.798	64.783
	NormWear (Ours)	**83.102**	**76.232**	**75.254**	**72.606**	**72.177**	**72.053**
EEG Main Tasks	Statistical	79.720	50.172	73.921	63.567	57.529	57.948
	Chronos	80.677	55.507	75.285	72.442	52.520	47.671
	CLAP	82.100	57.518	76.391	68.506	61.961	62.650
	TF-C	84.302	61.864	76.825	71.702	65.517	67.889
	NormWear (Ours)	**85.883**	**66.841**	**79.182**	**72.485**	**69.158**	**69.698**
Disease Risk Evaluation	Statistical	71.032	**53.783**	79.688	52.718	53.235	50.807
	Chronos	62.060	40.673	71.910	45.512	43.739	40.569
	CLAP	66.274	48.232	81.327	53.028	54.721	52.804
	TF-C	69.416	46.312	78.929	52.123	52.352	51.349
	NormWear (Ours)	**73.165**	51.666	**81.530**	**54.133**	**56.314**	**54.428**
Micro Average	Statistical	75.317	51.596	74.503	58.804	56.618	55.709
	Chronos	73.082	51.596	72.113	59.590	50.806	47.401
	CLAP	74.729	55.705	76.357	61.171	59.238	58.669
	TF-C	77.063	56.916	75.737	62.523	60.107	60.652
	NormWear (Ours)	**80.240**	**62.649**	**79.336**	**65.168**	**64.624**	**64.061**
Macro Average	Statistical	75.278	55.983	72.969	59.245	57.441	56.596
	Chronos	74.224	53.934	71.123	59.999	52.590	49.587
	CLAP	75.091	57.592	74.689	61.441	59.836	59.296
	TF-C	77.199	58.801	74.310	62.931	60.889	61.340
	NormWear (Ours)	**80.717**	**64.913**	**78.656**	**66.408**	**65.883**	**65.393**

F Scaling up the Pretraining Data Size

In addition to demonstrating that NormWear outperforms all strong baselines, we further investigate the effect of varying pretraining data size on the model's downstream performance to examine whether the scaling law applies to our proposed methodology. As shown in Figure 10, the overall performance (measured by accuracy) significantly improves as the pretraining data size increases from approximately 37k (62 hours) to nearly 2.5M (4000 hours) samples of wearable signal data. This observation indicates that our model adheres to the scaling law, highlighting its potential scalability and suitability for future large-scale applications.

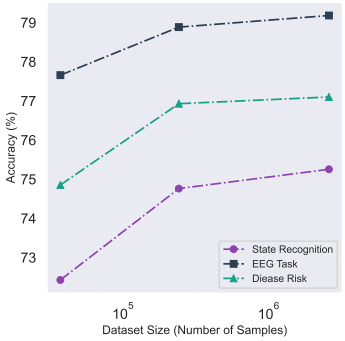
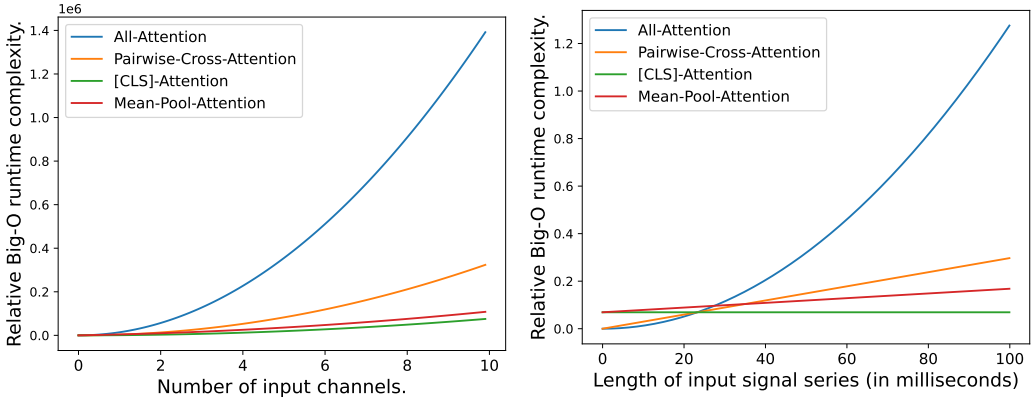


Figure 10: **Impact of scaling the pretraining dataset on downstream tasks.** The y-axis represents the average accuracy across tasks, while the x-axis denotes the size of the pretraining dataset in terms of the number of samples.

G Channel Fusion Complexity analysis

When conducting multi-channel modeling, for example, when the input comprises an arbitrary number of signals, a fusion operation needs to be conducted across all channels in order to let the model extract correlation information. Because we will deploy the model on an edge device like Jetson Nano, other than empirical evidence of the performance, we also have to consider the computation complexity of different approaches. A brief visualization of the runtime complexity of different approaches is presented in figure 11. The detailed derivation is presented in the following subsections.



(a) Runtime Big-O when increasing the number of channels

(b) Runtime Big-O when increasing the sequence length

Figure 11: **Visualization of runtime complexity when scaling up the number of channels or the sequence length.**

G.1 All-Attention

For the approach of conducting self-attention by concatenating all the patches, we arrive the Big-O complexity expression as follows:

- We denote C as the number of input channels, d as the embedding size, L as the number of patches convolved from the time series in each channel (proportional to sequence length), and $x \in \mathbb{R}^{C \times L \times d}$ as the input data before feeding into the fusion block. We have a total of $L \cdot C$ patches.
- When calculating the attention scores, dot products are computed for each pair of the patches, which results in the following calculation process:
  where "1), 2), 3)" represents the operations conducted at the first, second, and third rounds of entering the entire nested loops. The complexity for the first round of operation results in

Algorithm 2 All-Attention Complexity

for $i \in [1, 2, \ldots, C]$ **do**
  for $j \in [1, 2, \ldots, L]$ **do**
    $N \leftarrow \exp(\text{attn}(x_{i,j}))$ $\implies O(L \cdot C)$
    for $k \in [1, 2, \ldots, C]$ **do**
      for $l \in [1, 2, \ldots, L]$ **do**
        1) Calculate dot product: $\text{attn}(x_{i,j}, x_{k,l}) = x_{i,j}^T x_{k,l}$ $\implies O(2d)$
        2) Softmax over attention scores: $\frac{\exp(\text{attn}(x_{i,j}, x_{k,l}))}{N}$ $\implies O(1)$
        3) Weighted average: $x_{i,j} + \text{attn}(x_{i,j}, x_{k,l}) \cdot x_{k,l}$ $\implies O(2d)$
      end for
    end for
  end for
end for

a complexity of:

$$\sum_{i=1}^{C}\sum_{j=1}^{L}\sum_{k=1}^{C}\sum_{l=1}^{L} 2d = \sum_{i=1}^{C}\sum_{j=1}^{L}\sum_{k=1}^{C} L \cdot 2d = \sum_{i=1}^{C}\sum_{j=1}^{L} C \cdot L \cdot 2d = O(d \cdot (L \cdot C)^2) \quad (4)$$

where in the case of multi-head attention, the dot product still has the complexity of $O(2d)$, and because the number of heads is a constant, the final complexity is equivalent to the result in equation 4.

- Similarly, the softmax operation will result in a complexity of $O((L \cdot C)^2)$, and the final weighted average operation will also have a complexity of $O(d \cdot (L \cdot C)^2)$, which results in total complexity of

$$O(d \cdot (L \cdot C)^2) + O((L \cdot C)^2) + O(d \cdot (L \cdot C)^2) = O(d \cdot (L \cdot C)^2) \quad (5)$$

G.2 Cross-Attention

For the pairwise cross-attention approach following guidance of Chen et al. (2021), we have the operation defined as

Algorithm 3 Cross-Attention Complexity

for i in $[1, 2, ..., C-1]$ **do**
  for j in $[1, 2, ..., C]$ **do**
    2) $N = \exp(\text{attn}(x_{i,1}))$, $\implies O(L)$
    for k in $[2, 3, ..., L]$ **do**
      1) Calculate $\text{attn}(x_{i,1}, x_{j,k})$, $\implies O(2d)$
      2) Softmax over all-attention scores, $\frac{\exp(\text{attn}(x_{i,1}, x_{j,k}))}{N}$, $\implies O(1)$
      3) Weighted average: $x_{i,1} + x_{j,k}$, $\implies O(2d)$
    end for
  end for
end for

with the same notion in the previous subsection. The total complexity is

$$O(C^2 \cdot L \cdot 2d) + O(C^2 \cdot L) + O(C^2 \cdot L \cdot 2d) = O(d \cdot L \cdot C^2) \quad (6)$$

G.3 [CLS]-Attention

This is the approach that we adopted for the final version of our proposed foundation model. Only the embedding corresponding to the [CLS] token of each channel is involved during the self-attention operation. Therefore, the complexity is

$$O(d \cdot C^2) \quad (7)$$

G.4 Mean-pool Attention

For fusion with mean-pool attention, we first calculate the mean representation for each channel, resulting in a complexity of $O(C \cdot L \cdot d)$. And self-attention with Tese mean representations has the same complexity as [CLS]-attention, which is $O(d \cdot C^2)$. Thus, the total complexity is

$$O(C \cdot L \cdot d) + O(d \cdot C^2) = O(d \cdot (L \cdot C + C^2)) \tag{8}$$

H MSiTF Complexity analysis

Algorithm 4 MSiTF Runtime Complexity

key embedding $E_k = k(S) \in \mathbb{R}^{p \times d}$, $\implies O(d^2)$
value embedding $E_v = v(S) \in \mathbb{R}^{p \times d}$, $\implies O(d^2)$
Relevance score $Rel = E_k^T Q \in \mathbb{R}^p$, $\implies O(pd)$
likelihood parameter $E_l = l(S) \in \mathbb{R}^{p \times 2}$, $\implies O(d^2)$
Importance score sampling $W_{imp} \in \mathbb{R}^p$ (equation 1) $\implies O(p)$
Fused embedding $E_{final} = E_v^T(\alpha W_{imp} + \beta W_{rel} + \kappa W_{rec}) \in \mathbb{R}^d$, $\implies O(pd)$
Inference final score $c = \underset{i \in |C|}{\arg\max} \; C_i^T E_{final}$, $\implies O(cd)$

Where d being the latent size, p being the number of total patches, c being the number of available ground truth choice, k and v being the key and value linear mapping, $S \in \mathbb{R}^{p \times d}$ as the signal embeddings, Q as the query sentence embedding, and C as the list of available answer choice sentences. The total runtime complexity is $O(d^2 + pd + p + cd)$. Since d is constant, we have runtime complexity of $O(p + c)$.

Regarding memory complexity of MSiTF, with m being the size of text encoder, w being the size of normwear, we have (i) Signal representations: $O(pd)$; (ii) Text representations: $O(cd)$; (iii) Total: $O(m + w + d(p + c))$. Since m, w, and d are all constants, we have memory complexity of $O(p + c)$.

I Feature Visualization

Feature visualization serves as a tool to interpret and analyze the internal representations learned by the model. By examining activation patterns or embedding structures at various layers, we aim to understand how the model encodes input signals and whether these representations align with relevant semantic or structural information. This analysis provides insight into the effectiveness of the learned features and can inform architectural or training modifications to improve performance and generalization.

I.1 The model is agnostic to the input signals

This section investigates whether, without requiring the signal modality type information as input, NORMWEAR can effectively distinguish between different signal sources. We randomly sampled 500 samples for each sensor type and fed them into our pretrained model. We use t-SNE (Van der Maaten & Hinton, 2008), with PCA (Jolliffe & Cadima, 2016) initialization to visualize the learned representations corresponding to the [CLS] special token at the last layer. The PCA preserves the global structure, while t-SNE emphasizes local relationships in the data. From Figure 13(a), we observe that representations from sensors located at the same body position are clustered closely together, while representations from different body locations are clearly separated. This suggests that our model is signal-agnostic, as it can recognize the signal type differences, map their representations appropriately in the embedding space, and guide feature extraction within each Transformer block.

I.2 Waveform visualization

Figure 13 (b) under “Feature Associations" shows the features extracted by our model. Each patch corresponds to a representation with a vector size of $\mathbb{R}^{768}$. When ordered by time sequence, these representations form 768 waveforms per layer, representing the model’s extracted features. The figure

displays 64 randomly sampled waveforms from a selected layer. The features highlighted in purple and gray indicate the top 10 patterns positively and negatively associated with the target task (diabetes classification, in this example), with associations determined by linear regression parameters during linear probing. Additionally, our relevance-based fusion mechanism identifies the contribution of each time step during inference, highlighted by red dots in the “Time Step Relevance" section of Figure 13 (b).

Such a visualization pipeline can assist researchers and clinicians by offering insights into how the model reaches its final predictions. Given the millions of parameters and hundreds of waveform features per layer, visualizing these features individually is inefficient for understanding the overall behavior of the proposed foundation model. As a result, we use several techniques in nonlinear dynamic analysis (Thompson et al., 1990) to quantify the overall patterns of these extracted features, which are discussed in detail in section I.3.

I.3 Quantify the intrinsic behaviors: nonlinear dynamics analysis on the layer-wise waveforms

Understanding the representations extracted by intermediate layers is crucial to interpreting our model’s behavior. To quantify the meaningfulness of these representations, we conducted a nonlinear dynamics analysis inspired by chaos theory. This method analyzes the features’ intrinsic behaviors through metrics like the Lyapunov exponent (Wolf et al., 1985) (sensitivity to initial conditions), Hurst exponent (Qian & Rasheed, 2004) (self-correlation/seasonality), and persistence entropy (Yan et al., 2023b) (unpredictability in system states). We obtain the following key observations:

1. Deeper Layers Capture Higher-Order Complexity.

- For signals such as GSR, EEG, and ACC, deeper layers show lower self-correlation (DFA (Hu et al., 2001)) and higher unpredictability (persistence entropy), indicating a transition to representations that are less periodic and more chaotic.
- The decrease in the Lyapunov exponent across layers suggests reduced variation in extracted features, aligning with the idea that deeper layers capture more abstract, long-term patterns with broader receptive fields.

2. Modalities with Simpler Dynamics. In contrast, PPG and ECG signals, dominated by regular heart activity, exhibit more stable patterns across layers. This aligns with their simpler waveform structures and less complex dynamics compared to signals related to neural and physical activities.

These visualizations reveal that the model progressively transforms raw sensory data into representations aligned with the complexity of each signal. For GSR and EEG, deeper layers exhibit increased unpredictability and reduced periodicity, highlighting the extraction of nuanced, higher-order patterns critical for human sensing. In contrast, the stability of representations for PPG and ECG reflects their simpler dynamics, demonstrating the model’s adaptability to varying signal characteristics. This analysis confirms that the intermediate representations are purposefully optimized to capture the temporal and structural nuances of each modality, supporting the conclusion that the model learns meaningful features tailored to human sensing tasks.

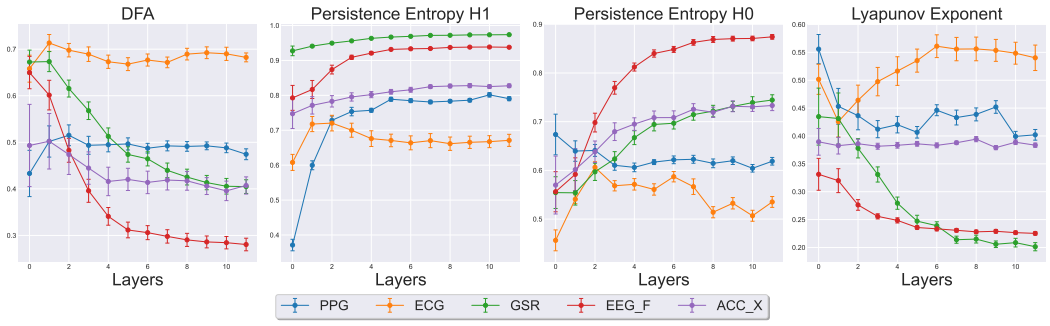
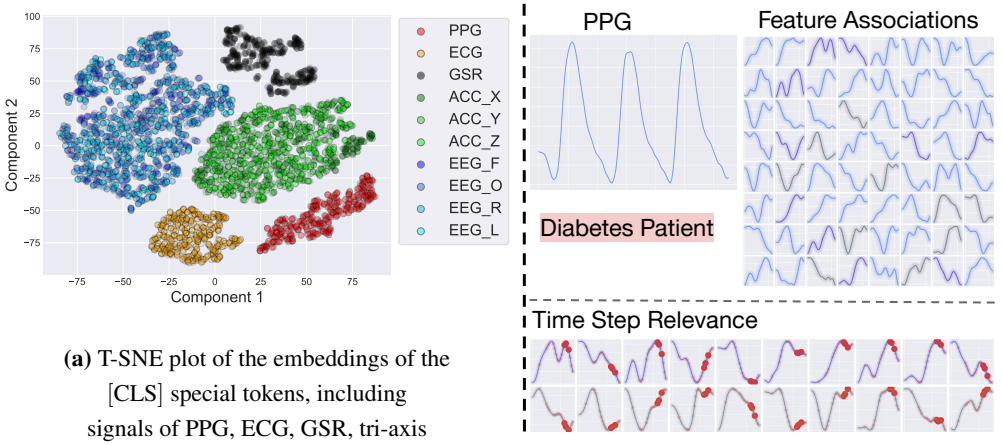


Figure 12: Nonlinear dynamic analysis on the waveforms extract at different layers of our model.



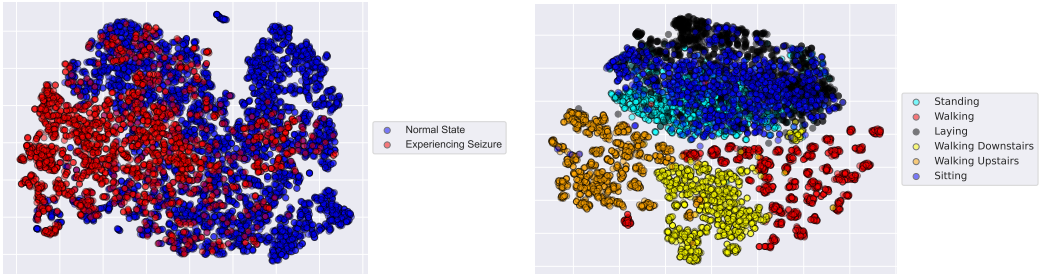
(a) T-SNE plot of the embeddings of the [CLS] special tokens, including signals of PPG, ECG, GSR, tri-axis accelerometer, and EEG at lobe of front, occiput, right, and left

(b) Visualization of features extracted by the intermediate layer

Figure 13: Feature visualization.

I.4 T-SNE plot among classes

In this section, we present T-SNE plots of NormWear's embeddings across different classes to provide insights into their structure and assess their suitability for sample similarity-based information retrieval. It is important to note that these plots are exploratory in nature and do not serve as a claim of the embeddings' superiority. As shown in Figures 14a and 14b, clear class separations can be observed in certain scenarios. For example, EEG samples from seizure subjects and normal subjects are distinctly separated, and physical activity types are well-clustered. For ECG data, abnormal heartbeats tend to form cohesive clusters. However, it is essential to recognize that these T-SNE plots reduce the latent representations into a 2D space, which may not fully capture the inherent properties of the embeddings in their original high-dimensional form.



(a) Visualization of embedding on EEG signals.

(b) Visualization of embedding on signals from IMU sensors.

Figure 14: Visualization of example signal embeddings.

I.5 Supplementary Qualitative Analysis of MSiTF

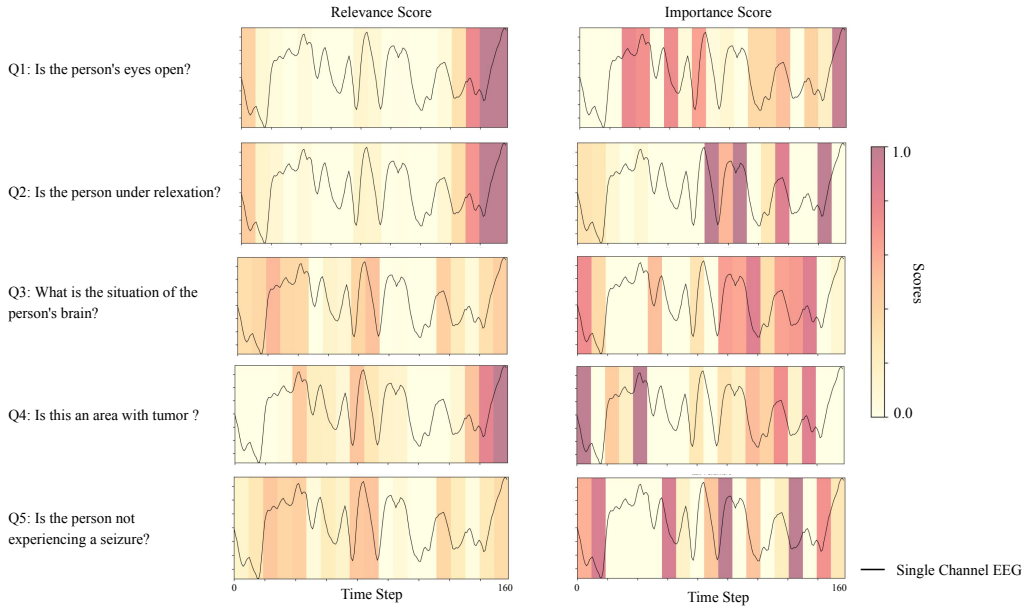


Figure 15: **Visualization of relevance scores(left) and importance scores (right) for a single channel EEG sample from the Epilepsy dataset under five task-specific questions.** The background color follows a yellow-to-red scale, where darker regions indicate higher scores.

To understand how each of our proposed gating modules in MSiTF—relevance, recency, and importance—select useful features for different tasks, we visualize the scores assigned to each time window. As shown qualitatively in Figure 15, the heatmaps reveal that both relevance and importance scores are sensitive to task differences. For example, in the eye closure detection task, the model focuses on the last few patches, whereas in the seizure detection task, it emphasizes patches with large fluctuations. A similar pattern is observed for the importance score, where patches are weighted differently across tasks. This suggests that our gating mechanism can adaptively select relevant features based on the task. We include a figure of the recency score (Figure 16) for completeness. Since the recency score is derived from a fixed decay function and is not learned, it remains the same across tasks.

To improve visualization, we aggregated token scores using a window size of 9, which matches our tokenization patch size. We then applied Z-score normalization to ensure comparability across tasks. The sample was selected from the Epilepsy dataset due to its multiple and diverse task types.

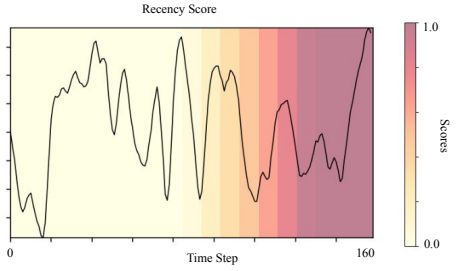


Figure 16: **Recency score generated by a decay function.** The sample is selected from the Epilepsy dataset.

J Reconstruction Example

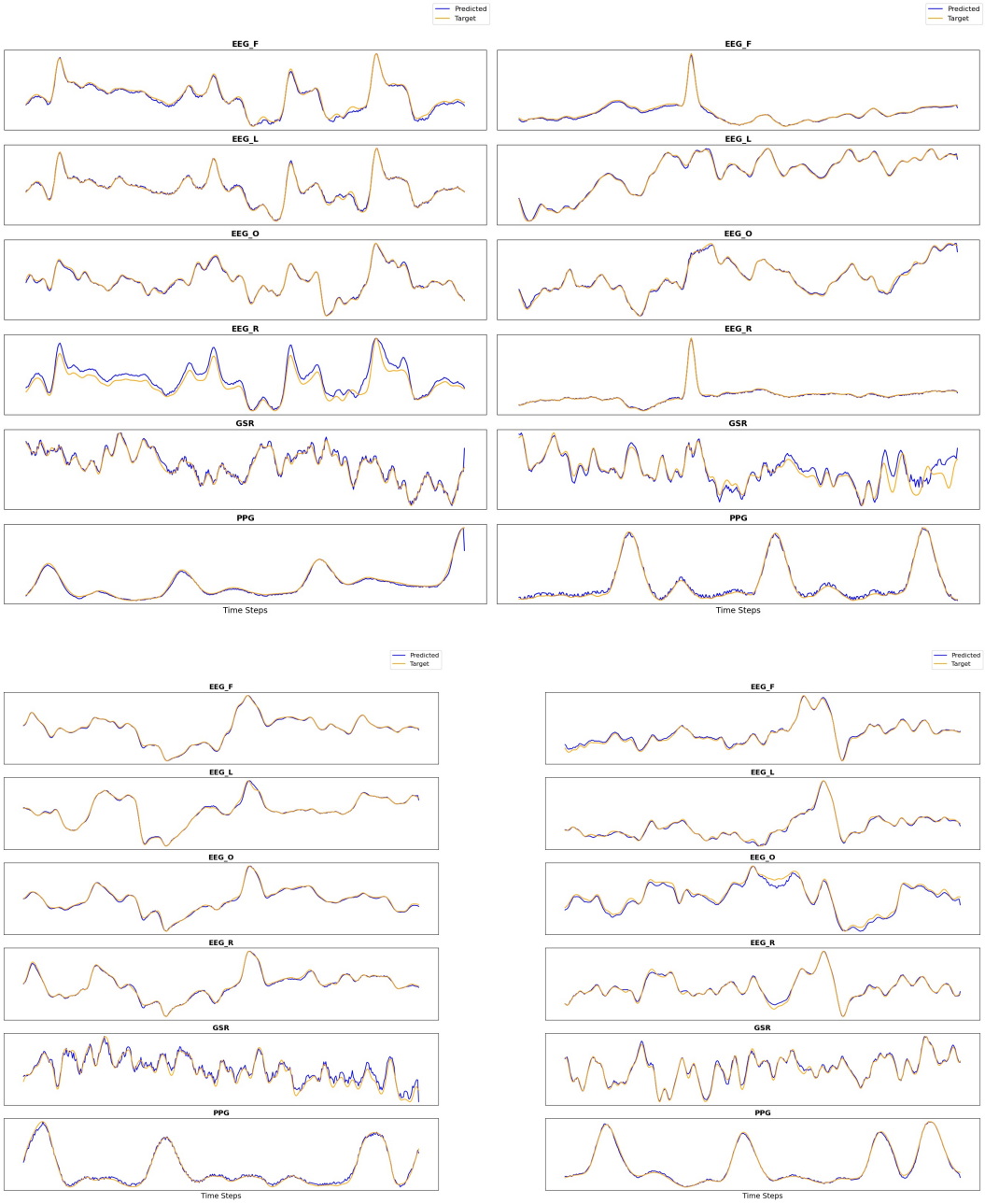


Figure 17: **Uncurated random samples** on Phyatt scalogram, using a NORMWEAR trained in our training set. The masking ratio is 80%.

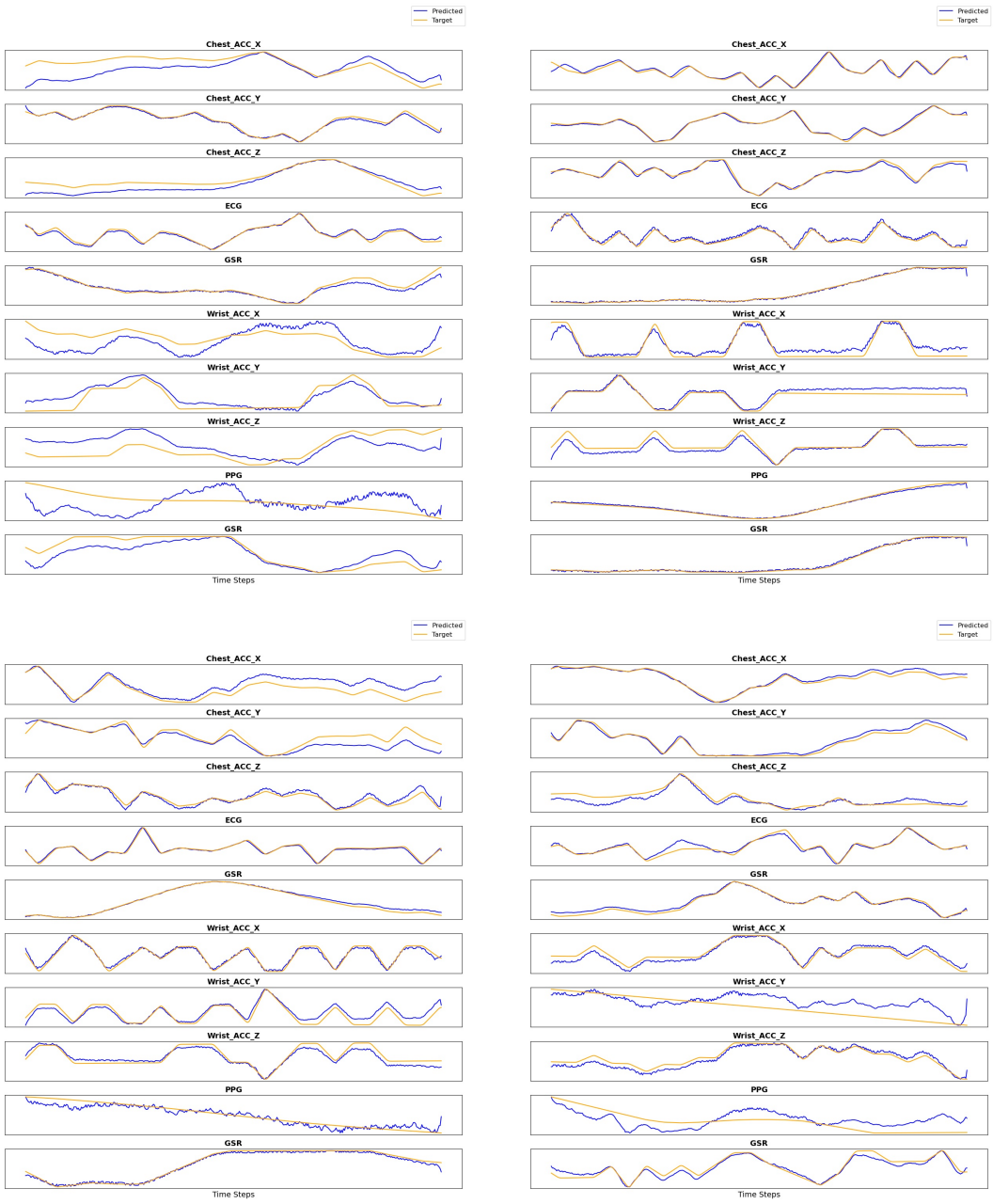


Figure 18: **Uncurated random samples** on WESAD scalogram, using a NORMWEAR trained in our training set. The masking ratio is 80%. Note that the IMU data are not in the training set and, in general, NORMWEAR is able to reconstruct this with high accuracy.

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 [NA] .
- [NA] 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 NA).

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 be asked to also 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 "[NA] " 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.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA 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 NA 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: [NA]

   Guidelines:

   - The answer NA 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: [NA]

   Justification: [NA]

   Guidelines:
   - The answer NA 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: [NA]

   Guidelines:
   - The answer NA 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: [NA]

   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA 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: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.