DNA-PPG: A Foundation Model for Photoplethysmography via Dual Neighborhood Alignment

Yizhang Yang[1] , **Jinshi Cui**[1,†] , **Wenlong Wu**[2] , **Chunlong Tu**[2] , **Yang Zhang**[1] , **Junshi Lu**[3] , **Li Wang**[3] , **Guosong Gao**[2,†]

[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

[2]Xiaomi Laboratory of Sports and Health, Xiaomi Inc., Beijing, China

[3]School of Psychological and Cognitive Sciences, Peking University

1900012201@pku.edu.cn, cjs@cis.pku.edu.cn, gaoguosong@xiaomi.com

Abstract

Existing physiological foundation models face two limitations: rigid hard-negative sampling indiscriminately repels morphologically similar samples, distorting the natural manifold; and coarse discretization strategies sever the intrinsic continuity of physiological states, inducing precision loss. To address these challenges, we propose DNA-PPG[1], a novel pre-training framework anchored in Dual Neighborhood Alignment. DNA-PPG integrates the Morphology-Aware Self-Supervised Branch using Time-Frequency Soft Weighting to capture universal signal dynamics shared among diverse subjects, with the Physiological Semantic Alignment Branch that projects physiological indicators into continuous semantic space to explicitly embed precise physiological priors into the representation space. We scale the pre-training to 10.7 million PPG segments from over 8,400 subjects to ensure robust generalization. Extensive evaluations on six downstream benchmarks demonstrate that DNA-PPG significantly outperforms state-of-the-art baselines, achieving an 18% reduction in regression error and an 11.5% improvement in classification performance. These results validate DNA-PPG as a robust, universal feature extractor for diverse photoplethysmography applications.

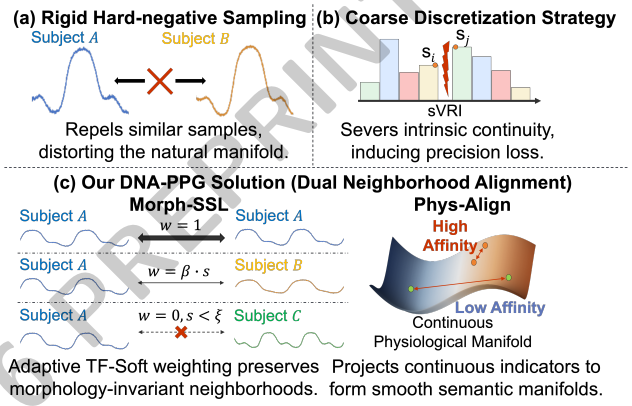

Figure 1: **Motivation and Proposed Strategy. (a-b) Existing Limitations.** Rigid hard-negative sampling indiscriminately repels pseudo-negatives, distorting the natural manifold (a), while coarse discretization strategies sever the intrinsic continuity of physiological states (b). **(c) Our DNA-PPG Solution. Left (Morph-SSL):** Employs TF-Soft to calculate alignment weight w based on the cosine similarity s of time-frequency dynamics. We introduce a threshold ξ to filter out noisy pairs and a scaling factor β to prevent signal dilution from negative dominance. **Right (Phys-Align):** Treats physiological indicators (BP, HR, SpO_2) as continuous semantic manifold. This forces the model to capture the intrinsic continuity of physiological states, ensuring that signals reflecting similar health conditions are naturally clustered in the embedding space.

1 Introduction

The proliferation of telemedicine, wearable devices, and intensive care technologies has yielded massive clinical physiological time series. However, the deployment of deep learning in this domain is hindered by a severe "annotation bottleneck": high-quality clinical labeling is costly, expert-dependent, and prone to inter-annotator noise. Consequently, supervised paradigms fail to capitalize on the vast archives of accumulated unlabeled data. To bridge this gap, Self-Supervised Learning (SSL) has emerged as a dominant paradigm, leveraging the intrinsic structure of data to learn representations without supervision. This has catalyzed the development of physiological foundation models, which aim to overcome the inherent fragility of traditional task-specific approaches by establishing robust, transferable representations across heterogeneous devices and patient populations.

To realize this vision, Contrastive Learning (CL) has become the standard for physiological foundation models due to its robust unsupervised learning capabilities. Representative frameworks, such as SimCLR [Chen *et al.*, 2020], established the instance discrimination paradigm by treating intra-batch samples as hard negatives. This paradigm was subsequently adapted for PPG pre-training [Abbaspourazad

[1]This work was done during an internship at Xiaomi Corporation. Code and pre-trained checkpoints are provided at https://github.com/yyzyyz001/DNA-PPG.

et al., 2023], achieving patient-level feature disentanglement by strictly defining signals from different subjects as negatives. Building upon this, recent works like PaPaGei [Pillai *et al.*, 2024] have attempted to integrate domain knowledge into the contrastive objective. Specifically, the PaPaGei-S variant employs a discretization (binning) strategy, where continuous hand-crafted morphological features are quantized into finite intervals to guide feature learning.

However, these paradigms face fundamental limitations. First, instance discrimination approaches are constrained by the rigid hard-negative assumption. By indiscriminately repelling pseudo-negative samples from different subjects, they overlook intrinsic morphological similarities—where distinct individuals often exhibit highly coherent time-frequency dynamics. This rigidity distorts the natural manifold structure of the feature space. Second, while approaches like PaPaGei-S incorporate domain priors, the discretization strategy disrupts the continuity of physiological states, inducing precision loss; furthermore, reliance on hand-crafted indices hinders transferability to other modalities or acquisition conditions.

To bridge these gaps, we propose **DNA-PPG**, a hybrid pre-training framework anchored in *Dual Neighborhood Alignment*. To mitigate the hard-negative bias, we design the **Morphology-Aware Self-Supervised Branch (Morph-SSL)**. Beyond standard subject consistency, it incorporates a *Time-Frequency Soft Weighting (TF-Soft)* mechanism that leverages spectral priors to measure dynamic affinity, adaptively relaxing the repulsion of morphologically similar false negatives to preserve time-frequency invariant neighborhoods. Simultaneously, to overcome discretization limitations, we introduce the **Physiological Semantic Alignment Branch (Phys-Align)**. Discarding coarse binning, this branch models physiological indicators (e.g., Blood Pressure, Heart Rate, and Oxygen Saturation) directly as *continuous manifold*, enforcing the learning of smooth semantic transitions consistent with continuous physiological variations. Extensive ablations validate this synergy, demonstrating that their joint optimization significantly outperforms single-branch variants.

Our contributions are summarized as follows:

- We propose **DNA-PPG**, a novel physiological foundation model framework anchored in *Dual Neighborhood Alignment*. By addressing the manifold distortion from rigid hard-negatives and the precision loss from discretization, it constructs a continuous, structure-preserving representation space superior to existing instance discrimination or binning-based paradigms.
- We design two synergistic mechanisms to align feature neighborhoods: **Morph-SSL** leverages the Time-Frequency Soft Weighting (TF-Soft) mechanism to maintain morphology-invariant neighborhoods, while **Phys-Align** exploits continuous physiological indicators to form smooth semantic manifolds. Their synergistic optimization ensures representations possess both structural robustness and semantic precision.
- Extensive experiments, leveraging a pre-training dataset of 10.7 million segments from over 8,400 subjects, demonstrate that DNA-PPG outperforms SOTA baselines, achieving an 18% MAE reduction in regression and an 11.5% improvement in classification performance. These results validate its efficacy as a robust universal feature extractor for PPG signals, evidenced by robust generalization across diverse healthcare domains.

2 Related Work

2.1 PPG-based Health Monitoring

Photoplethysmography (PPG) has emerged as a ubiquitous wearable sensing modality for non-invasive health monitoring [Nie *et al.*, 2024]. Deep learning has significantly advanced this field, showing remarkable efficacy in tasks such as blood pressure estimation [Wang *et al.*, 2020; Kim *et al.*, 2022], arrhythmia detection [Kwon *et al.*, 2019], and sleep apnea analysis [Papini *et al.*, 2020]. However, despite their performance, these fully supervised pipelines heavily rely on costly expert annotations and often suffer from poor generalization across varying devices and protocols. These limitations underscore the necessity of label-efficient pre-training for learning robust and transferable PPG representations.

2.2 Self-Supervised Learning for Time Series

Self-supervised learning (SSL) for time series has primarily categorized into discriminative and generative paradigms. Discriminative approaches, such as SimCLR [Chen *et al.*, 2020] and BYOL [Grill *et al.*, 2020], learn representations by maximizing similarity between augmented views of the same instance, utilizing strategies like negative sampling or asymmetric architectures to derive distinctive features. Conversely, generative frameworks like TiMAE [Li *et al.*, 2023] adopt masked modeling to capture temporal dependencies by reconstructing missing segments in the latent or signal space.

To enhance the robustness and generalization, recent studies have focused on refining training objectives. Addressing the false negative issue in rigid contrastive learning, SoftCLT [Lee *et al.*, 2023] employs soft assignments across instance and temporal dimensions, while SupCon [Khosla *et al.*, 2020] utilizes semantic labels to align intra-class samples. Furthermore, TF-C [Zhang *et al.*, 2022] incorporates spectral priors to ensure time-frequency consistency. Inspired by the scaling laws of large language models, the field is transitioning towards foundation models: Chronos [Ansari *et al.*, 2024] tokenizes time series for forecasting, while MOMENT [Goswami *et al.*, 2024] leverages massive pre-training on the "Time Series Pile" to derive universal features.

However, applying these generic paradigms to physiological signals remains suboptimal. Most methods treat time series as homogeneous data, overlooking the fine-grained morphological details critical for clinical analysis. Moreover, general foundation models typically ignore the continuous manifold structure of physiological states, failing to align representations with continuous clinical indicators. This necessitates a domain-specific framework that explicitly harmonizes morphological integrity with physiological continuity.

2.3 Foundation Models for Physiological Signals

Recent work explores foundation models for diverse physiological modalities to learn universal representations. For

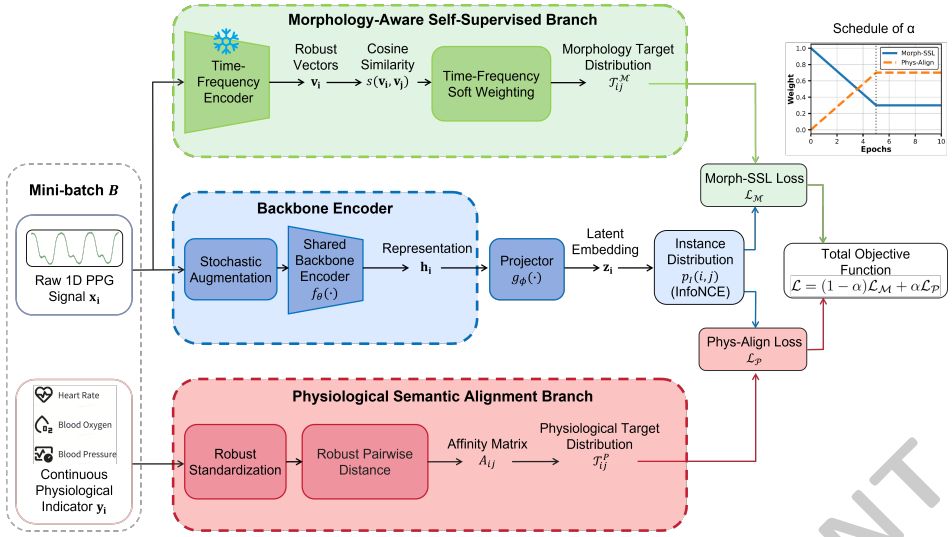

Figure 2: **Overview of the proposed DNA-PPG pre-training framework.** A shared **Backbone Encoder** (blue) extracts latent representations from augmented views. The framework implements the *Dual Neighborhood Alignment* strategy via two synergistic branches: (1) the **Morphology-Aware Self-Supervised Branch (Morph-SSL)** (green) utilizes *Time-Frequency Soft Weighting* to identify coherent signal dynamics across individuals, mitigating manifold distortion; and (2) the **Physiological Semantic Alignment Branch (Phys-Align)** (red) treats physiological indicators as auxiliary supervision to cluster signals reflecting similar health conditions. A schedule-based parameter α dynamically balances the joint optimization.

instance, RelCon [Xu *et al.*, 2024] leverages motif-based distances for accelerometry to learn fine-grained motion semantics, while Brant [Zhang *et al.*, 2023] employs dual-encoder Transformers to model spatiotemporal dependencies in EEG. Similarly, WBM [Erturk *et al.*, 2025] aggregates longitudinal metrics to encode high-level behavioral patterns.

In the specific domain of PPG, [Abbaspourazad *et al.*, 2023] established a benchmark by scaling contrastive learning to over 141,000 participants. However, this approach relies on a rigid instance discrimination paradigm, where signals from different subjects are strictly treated as negative pairs. This strategy indiscriminately repels morphologically similar samples solely due to subject identity, thereby distorting the natural feature manifold. To incorporate domain guidance, PaPaGei-S [Pillai *et al.*, 2024] aligns representations with hand-crafted metrics like the stress-induced vascular response index (sVRI). While incorporating priors, it relies on discretizing continuous features to construct positive pairs. Such quantization introduces hard decision boundaries that fragment the intrinsic continuity of physiological variations and induce precision loss. Moreover, the dependence on specific pre-defined indices limits the framework's adaptability. This highlights the urgent need for a unified approach that respects continuous physiological semantics while flexibly maintaining morphological neighborhoods.

3 Methodology

3.1 Overview

The proposed DNA-PPG framework serves as a foundation model designed to learn universal, transferable representations for photoplethysmography (PPG) data. The overall workflow is structured into two distinct phases: a hybrid pre-training phase and a downstream inference phase. Figure 2 specifically illustrates the architecture of the proposed hybrid pre-training phase. Formally, given a raw input sequence $\mathbf{x}$, we apply stochastic augmentations to generate a transformed view $\tilde{\mathbf{x}}$. The shared backbone encoder $f_\theta(\cdot)$ maps this view into a high-dimensional representation vector $\mathbf{h} = f_\theta(\tilde{\mathbf{x}})$. During pre-training, $\mathbf{h}$ is subsequently projected by a non-linear projection head $g_\phi(\cdot)$ into a latent embedding $\mathbf{z} = g_\phi(\mathbf{h})$ for contrastive optimization.

During the pre-training phase, the framework employs a *Dual Neighborhood Alignment* strategy to construct a semantically rich feature manifold. This is achieved by optimizing two complementary branches simultaneously. The **Morphology-Aware Self-Supervised Branch (Morph-SSL)** is designed to capture intrinsic signal dynamics and structural patterns. By employing a *Time-Frequency Soft Weighting* mechanism, this branch identifies samples with coherent time-frequency dynamics across different individuals, thereby mitigating the manifold distortion caused by rigid hard-negative sampling. Parallel to this, the **Physiological Semantic Alignment Branch (Phys-Align)** explicitly incorporates physiological priors into the representation learning process. By treating physiological indicators as auxiliary supervision signals, this branch forces the model to capture the intrinsic continuity of physiological states, ensuring that signals reflecting similar health conditions are naturally clustered in the embedding space. These two branches are jointly optimized via a progressive learning strategy, enabling the shared encoder to synergize morphological robustness with physiological semantic precision.

After pre-training, the projection head $g_\phi(\cdot)$ is discarded. The encoder $f_\theta(\cdot)$ is frozen and serves as a universal feature extractor. Lightweight, task-specific heads—such as linear

regressors or classifiers—are attached to the representation $\mathbf{h}$, enabling efficient adaptation to diverse healthcare applications with minimal overhead.

3.2 Morphology-Aware Self-Supervised Branch

In the Morph-SSL branch, we adopt a foundational contrastive assumption where two distinct PPG segments cropped from the same subject are treated as a natural positive pair (i, i'), while all other samples within the mini-batch of size B serve as candidate negatives. Given an anchor i and target j, the standard InfoNCE [Oord *et al.*, 2018] probability $p_I(i, j)$ is defined as:

$$p_I(i,j) = \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_j/\tau)}{\sum_{l\neq i}\exp(\mathbf{z}_i^\top \mathbf{z}_l/\tau)}, \tag{1}$$

where $\mathbf{z}$ is the normalized latent embedding and $\tau = 0.1$ is the temperature hyperparameter.

Standard contrastive paradigms rely on a rigid, hard-negative assumption, indiscriminately repelling samples from different subjects. This ignores intrinsic morphological affinities and distorts the feature manifold. While direct similarity metrics (e.g., Euclidean distance or cosine similarity) could theoretically address this, they prove inadequate for raw PPG waveforms due to their susceptibility to phase misalignments and amplitude scaling, which induce high variance even among semantically similar signals. To address these challenges, we introduce the Time-Frequency Soft Weighting mechanism. This module integrates a dedicated auxiliary encoder based on the TF-C architecture [Zhang *et al.*, 2022], pre-trained on our unannotated PPG dataset. By performing joint time-frequency modeling, the encoder extracts robust representation vectors $\mathbf{v}_i$ that capture essential signal dynamics. Crucially, this auxiliary module remains frozen during the main training phase. Consequently, the additional computational overhead is limited to a single forward pass per sample, ensuring a negligible burden on the overall training efficiency. Specifically, we quantify the pairwise affinity in this auxiliary latent space using cosine similarity $s(\mathbf{v}_i, \mathbf{v}_j)$:

$$s(\mathbf{v}_i,\mathbf{v}_j) = \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\|_2\|\mathbf{v}_j\|_2}. \tag{2}$$

To reduce noise and prevent signal dilution from dominant negatives, we define the Morphology-Aware weight w_{ij} as:

$$w_{ij} = \mathbb{I}(s(\mathbf{v}_i,\mathbf{v}_j) > \xi)\cdot \beta \cdot s(\mathbf{v}_i,\mathbf{v}_j), \quad j \neq i, \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Here, ξ is a similarity threshold, set to 0.5 to filter out weakly correlated pairs. The scaling factor $\beta = 0.05$ is empirically chosen to attenuate the influence of soft negatives, thereby preserving the dominance of the primary positive pair (i, i').

By assigning soft weights to sequences with similar time-frequency morphologies, we preserve local neighborhood structures within the InfoNCE framework. This is formalized through a calibrated target distribution $\mathcal{T}_i^{\mathcal{M}}$, where the positive pair (i, i') is assigned a fixed weight of 1, and negative pairs utilize w_{ij}. These weights are normalized to yield a valid probability distribution:

$$\mathcal{T}_{ij}^{\mathcal{M}} = \frac{\mathbb{I}(j=i') + \mathbb{I}(j\neq i')w_{ij}}{1+\sum_{l\notin\{i,i'\}} w_{il}}. \tag{4}$$

The final Morphology-Aware loss is defined as the cross-entropy between the soft target distribution $\mathcal{T}_i^{\mathcal{M}}$ and the predicted probability distribution $p_I(i, \cdot)$, effectively mitigating the adverse effects of repelling morphologically similar samples:

$$\mathcal{L}_{\mathcal{M}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}\mathcal{T}_{ij}^{\mathcal{M}}\log p_I(i,j). \tag{5}$$

3.3 Physiological Semantic Alignment Branch

To structure the representation manifold consistently with physiological priors, we introduce the Physiological Semantic Alignment Branch (Phys-Align). Importantly, within the DNA-PPG framework, physiological indicators are not utilized as input features for the backbone encoder $f_\theta(\cdot)$. Instead, they serve strictly as auxiliary supervision signals to calibrate the embedding space, ensuring that signals reflecting similar health conditions are naturally clustered.

We treat the physiological state of the i-th sample as a continuous semantic vector $\mathbf{y}_i \in \mathbb{R}^K$, comprising key hemodynamic and physiological indicators: blood pressure (BP), oxygen saturation (SpO_2), and heart rate (HR), i.e., $\mathbf{y}_i = [\mathrm{BP}, \mathrm{SpO}_2, \mathrm{HR}]^\top$. Given physiological priors $\mathbf{y}_i \in \mathbb{R}^K$ and validity masks $\mathbf{m}_i \in \{0,1\}^K$, we first apply robust standardization to handle varying scales and potential outliers. For the k-th dimension, the standardized value $\tilde{y}_{i,k}$ is computed as:

$$\tilde{y}_{i,k} = \frac{y_{i,k}-\mu_k}{\sigma_k}, \tag{6}$$

where μ_k and σ_k denote the median and the Median Absolute Deviation (MAD) of the valid samples, respectively, serving as robust estimators for location and scale.

We define a robust pairwise distance metric $\delta_{ij,k}$ for missing data. Assuming missing entries follow $\mathcal{N}(0,1)$ in the standardized space (samples with all physiological labels missing are filtered out during preprocessing; see Sec. 4.2), the expected squared distance per dimension is:

$$\delta_{ij,k} = \begin{cases} (\tilde{y}_{i,k}-\tilde{y}_{j,k})^2, & m_{i,k}=1, m_{j,k}=1, \\ \tilde{y}_{i,k}^2+1, & m_{i,k}=1, m_{j,k}=0, \\ \tilde{y}_{j,k}^2+1, & m_{i,k}=0, m_{j,k}=1, \\ 2, & m_{i,k}=0, m_{j,k}=0, \end{cases}$$

which are aggregated across all K dimensions as $\mathcal{D}_{ij} = \frac{1}{K}\sum_{k=1}^{K}\delta_{ij,k}$.

Unlike prior works such as PaPaGei-S [Pillai *et al.*, 2024], which rely on rigid, interval-based discretization for hard contrastive sampling, our approach adopts a soft weighting mechanism. Such hard boundaries often lead to information loss and artificial manifold fragmentation. In contrast, we transform the Euclidean distance $\mathcal{D}_{ij}$ into an affinity matrix A, enabling the model to capture fine-grained degrees of physiological similarity. The diagonal is masked to zero to avoid self-reinforcement:

$$A_{ij} = \begin{cases} \exp\left(-\frac{\mathcal{D}_{ij}}{2\gamma^2}\right), & i\neq j \\ 0, & i=j \end{cases} \tag{7}$$

Dataset	Domain	# Subj.	# Segments	Fs (Hz)
VitalDB [Lee *et al.*, 2022]	Surgery	6,388	6,381,814	500
MESA [Zhang *et al.*, 2018]	Sleep	2,053	4,405,876	256

Table 1: Statistics of the pre-training datasets. The dataset scale is measured by the number of subjects and 10-second signal segments.

where $\gamma = 1.0$ is the bandwidth parameter. The physiological target distribution is then obtained by row-normalization, $\mathcal{T}_{ij}^{\mathcal{P}} = A_{ij} / \sum_{l=1}^{B} A_{il}$, and the physiological alignment loss $\mathcal{L}_{\mathcal{P}}$ is the cross-entropy between this distribution and the contrastive probability $p_I(i, j)$ in Eq. (1):

$$\mathcal{L}_{\mathcal{P}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j \neq i} \mathcal{T}_{ij}^{\mathcal{P}} \log p_I(i, j).$$

3.4 Joint Optimization Strategy

To achieve holistic representation learning, we propose a joint objective that synergizes the intrinsic structural robustness captured by Morph-SSL with the precise clinical semantics encoded by Phys-Align. The total objective function $\mathcal{L}$ is defined as:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\mathcal{M}} + \alpha \mathcal{L}_{\mathcal{P}}, \tag{8}$$

where α denotes the dynamic coefficient that modulates the injection of physiological priors into the representation space.

We implement a progressive learning strategy, inspired by curriculum learning [Bengio *et al.*, 2009] principles. Specifically, α is initialized at 0 and linearly ramps up to 0.7 throughout the training process. This schedule allows the backbone encoder to first establish a stable manifold based on signal dynamics via Morph-SSL before gradually incorporating complex physiological semantic constraints. This strategy effectively prevents the high variance or noise inherent in physiological labels from disrupting feature alignment in the early stages.

4 Experiments

4.1 Datasets

We evaluate our framework using a large-scale pre-training dataset and six downstream benchmarks. Following the PaPaGei [Pillai *et al.*, 2024] benchmark, we align evaluation settings for fair comparison while utilizing a refined preprocessing pipeline. Statistics are summarized in Tables 1 and 2.

Pre-training Datasets

We leverage unlabeled PPG signals from VitalDB [Lee *et al.*, 2022] and MESA [Zhang *et al.*, 2018; Chen *et al.*, 2015] (see Table 1). To unify the format, raw signals were processed into non-overlapping 10-second segments. Unlike the original PaPaGei benchmark, we exclude MIMIC-III [Johnson *et al.*, 2016] to ensure full reproducibility, given its recent access restrictions.

Downstream Datasets

We evaluate the generalization capability of our foundation model across six downstream datasets summarized in Table 2.

Regression Tasks These tasks evaluate the model's precision in hemodynamic estimation using Mean Absolute Error (MAE). For **PPG-BP** [Liang *et al.*, 2018], we estimate three indicators: Systolic Blood Pressure (SysBP), Diastolic Blood Pressure (DiasBP), and Heart Rate (HR). Similarly, **VV** [Toye, 2023] is used for SysBP and DiasBP estimation. For **PPG-DaLiA** [Reiss *et al.*, 2019], which contains ambulatory data with motion artifacts, we perform HR estimation.

Type	Dataset	Target Task	Metric	# Subj.
Regression	PPG-BP [Liang *et al.*, 2018]	SysBP DiasBP Heart Rate (HR)	MAE MAE MAE	219
	PPG-DaLiA [Reiss *et al.*, 2019]	Heart Rate (HR)	MAE	15
	VV [Toye, 2023]	SysBP DiasBP	MAE MAE	231
Classification	SDB [Garde *et al.*, 2014]	AHI (Normal/Abn.)	AUROC	146
	WESAD [Schmidt *et al.*, 2018]	Valence (High/Low) Arousal (High/Low)	AUROC AUROC	15
	ECSMP [Gao *et al.*, 2021]	TMD (High/Low)	AUROC	89

Table 2: Overview of downstream tasks. Regression tasks are evaluated by MAE, and Classification tasks are evaluated by AUROC.

Classification Tasks All classification tasks are formulated as binary problems to assess the model's ability to capture high-level physiological semantics, evaluated using the Area Under the Receiver Operating Characteristic (AUROC). For sleep analysis, we utilize **SDB** [Garde *et al.*, 2014], where the original four-level Apnea-Hypopnea Index (AHI) is converted into a binary target by distinguishing between normal (0) and abnormal (> 0) cases. In the domain of affective computing, we address label subjectivity on **WESAD** [Schmidt *et al.*, 2018] and **ECSMP** [Gao *et al.*, 2021] by adopting a median-split protocol. Specifically, the continuous scores for Valence and Arousal in WESAD, as well as the Total Mood Disturbance (TMD) in ECSMP, are binarized based on their population medians to differentiate between high and low physiological states.

4.2 Implementation Details

Data Preprocessing Pipeline

We adopt a standardized pipeline for pre-training data to ensure signal quality, while employing task-specific strategies for downstream benchmarks. **Pre-training Processing.** For VitalDB, raw PPG signals are first resampled to 125 Hz and synchronized with physiological labels. We isolate continuous valid intervals, applying bandpass filtering (0.5-12 Hz) and smoothing to eliminate baseline wander and high-frequency noise, before segmenting into non-overlapping 10-second windows. To ensure data integrity, we discard segments where the PPG signal is entirely absent or where both HR and SpO_2 labels are missing. Additionally, flatline artifacts are filtered utilizing a dynamic threshold (0.01× the 5th-95th percentile amplitude range). MESA follows a similar procedure, processing raw signals in 20-minute slices. Finally, valid segments from both sources undergo Z-score normalization. **Downstream Processing.** Following pre-training, downstream signals undergo bandpass filtering, resampling (125 Hz), and Z-score normalization. Variable-length inputs are adjusted to a fixed 10 seconds via symmetric padding or center-cropping. For regression tasks involving discrete measurements (PPG-BP, VV), we aggregate feature representations by averaging multiple segments per subject. For continuous monitoring in PPG-DaLiA, a sliding window protocol (8s window, 2s stride) is applied. For long-term recordings (WESAD, SDB, ECSMP), we trim the initial

Dataset	Task	Hand-crafted (Stat. Feat.)	SimCLR [Chen *et al.*, 2020]	BYOL [Grill *et al.*, 2020]	TF-C [Zhang *et al.*, 2022]	Chronos [Ansari *et al.*, 2024]	Moment [Goswami *et al.*, 2024]	PaPaGei-S [Pillai *et al.*, 2024]	DNA-PPG (Ours)
Panel A: Regression (MAE ↓ – Lower is better)									
PPG-BP	SysBP	16.72 [12.9, 20.8]	14.25 [11.7, 17.4]	15.82 [12.4, 19.2]	14.14 [9.9, 16.5]	16.44 [12.9, 20.1]	14.85 [11.1, 18.8]	14.52 [11.0, 18.1]	**11.92 [8.8, 15.2]**
	DiasBP	8.23 [6.2, 10.1]	8.44 [6.8, 10.0]	8.69 [6.6, 11.2]	9.57 [7.8, 11.5]	8.46 [6.5, 10.5]	8.17 [6.5, 10.0]	7.34 [5.6, 9.1]	**6.62 [4.8, 8.7]**
	HR	6.07 [4.7, 7.5]	5.51 [4.1, 7.1]	7.40 [5.5, 9.3]	5.69 [4.5, 6.8]	6.55 [5.0, 8.1]	4.04 [3.1, 5.0]	5.29 [4.1, 6.4]	**3.90 [3.0, 4.8]**
PPG-DaLiA	HR	14.98 [14.8, 15.1]	14.17 [13.9, 14.3]	14.07 [13.8, 14.2]	13.92 [13.7, 14.1]	14.20 [14.0, 14.4]	14.21 [14.0, 14.3]	14.26 [14.0, 14.4]	**13.87 [13.6, 14.0]**
VV	SysBP	16.60 [14.5, 18.7]	15.97 [12.9, 19.1]	15.17 [12.1, 18.7]	17.70 [11.9, 25.3]	17.19 [13.9, 20.6]	14.48 [11.5, 18.0]	16.38 [12.3, 20.6]	**13.63 [10.6, 16.4]**
	DiasBP	8.62 [7.3, 9.8]	8.61 [7.0, 10.5]	8.62 [6.7, 10.8]	8.93 [6.9, 11.0]	8.76 [7.0, 10.6]	8.54 [6.5, 10.4]	9.24 [7.3, 11.2]	**7.03 [5.5, 8.5]**
Regression Avg.		11.87	11.16	11.63	11.66	11.93	10.72	11.17	**9.50**
Panel B: Classification (AUROC ↑ – Higher is better)									
SDB	AHI	0.536 [0.492, 0.579]	0.540 [0.298, 0.775]	0.595 [0.392, 0.795]	0.620 [0.397, 0.826]	0.571 [0.397, 0.755]	0.530 [0.297, 0.769]	**0.665 [0.406, 0.862]**	0.640 [0.409, 0.846]
WESAD	Valence	0.596 [0.478, 0.709]	0.538 [0.412, 0.664]	0.518 [0.400, 0.635]	0.565 [0.450, 0.668]	0.577 [0.456, 0.683]	0.616 [0.496, 0.717]	0.533 [0.363, 0.580]	**0.737 [0.646, 0.810]**
	Arousal	0.546 [0.440, 0.671]	0.502 [0.399, 0.594]	0.513 [0.392, 0.631]	0.574 [0.466, 0.673]	0.566 [0.456, 0.668]	0.598 [0.507, 0.704]	0.521 [0.400, 0.639]	**0.602 [0.499, 0.709]**
ECSMP	TMD	0.524 [0.435, 0.607]	0.554 [0.473, 0.633]	0.631 [0.559, 0.702]	0.612 [0.525, 0.700]	0.545 [0.447, 0.633]	0.589 [0.497, 0.660]	0.560 [0.483, 0.642]	**0.665 [0.572, 0.753]**
Classification Avg.		0.551	0.534	0.564	0.593	0.565	0.583	0.570	**0.661**

Table 3: Comparison with baselines and state-of-the-art methods. **Panel A** reports MAE (95% Confidence Interval in brackets) for regression tasks. **Panel B** reports AUROC (95% Confidence Interval in brackets) for classification tasks. Best results are highlighted in **bold**. Abbreviations: **SysBP**: Systolic Blood Pressure; **DiasBP**: Diastolic Blood Pressure; **HR**: Heart Rate; **AHI**: Apnea-Hypopnea Index; **TMD**: Total Mood Disturbance; **MAE**: Mean Absolute Error; **AUROC**: Area Under the Receiver Operating Characteristic curve.

and final transition periods to exclude noise, and subsequently sample representative 10-second clips from the remaining effective duration.

Backbone Architecture

The backbone encoder maps raw 1D PPG signals into a high-dimensional feature space. To determine the most suitable architecture for physiological feature extraction, we evaluated three representative networks: a ViT-1D [Dosovitskiy, 2020] with 12 Transformer blocks, which processes the sequence via patching and aggregates global information utilizing the [CLS] token, and convolutional baselines EfficientNet-1D [Tan and Le, 2019] and ResNet-1D [He *et al.*, 2016]. Empirically, we adopt **ResNet-1D**, configured for a 512-dimensional output, as our default backbone, as it yielded the superior performance in our contrastive learning framework (validated in Sec. 4.4).

Training Setup

Experiments were conducted on a single NVIDIA A800 GPU. Models were trained for 10 epochs (batch size 256) with fixed 10-second inputs (1,250 time steps). We employed the Adam optimizer [Kingma, 2014] (initial lr=1×10^{-3}) and a cosine annealing scheduler [Loshchilov and Hutter, 2016] featuring 5% linear warmup and subsequent decay to zero. During training, data augmentations—including Gaussian noise, random scaling, and time warping—were applied to construct diverse views for contrastive learning.

4.3 Comparison

We compare DNA-PPG against seven baselines: statistical features (**Hand-crafted**), general contrastive learning methods (SimCLR [Chen *et al.*, 2020], BYOL [Grill *et al.*, 2020]), time-series specific frameworks (TF-C [Zhang *et al.*, 2022]), and time-series foundation models (Chronos [Ansari *et al.*, 2024], Moment [Goswami *et al.*, 2024]). We also include PaPaGei-S [Pillai *et al.*, 2024] as a representative PPG-focused benchmark, reporting the better result between the official checkpoint and a version retrained on our pre-training data. Notably, REGLE [Yun *et al.*, 2024] is excluded because only its spirogram inference code is available; applying a spirogram-trained model to PPG creates a modality mismatch, rendering the comparison unreliable.

The quantitative results are summarized in Table 3. In the regression tasks (Panel A), DNA-PPG outperforms all baselines with the lowest average MAE of **9.50**. Our model shows a significant advantage in hemodynamic estimation, reducing the SysBP error on PPG-BP by approximately **18%** compared to PaPaGei-S. This improvement suggests that our Physiological Semantic Alignment Branch effectively captures fine-grained physiological semantics, which general contrastive methods often fail to model. Furthermore, on the noise-prone PPG-DaLiA dataset, DNA-PPG maintains superior performance (MAE 13.87) over foundation models like Chronos and Moment, indicating that the Morphology-Aware Self-Supervised Branch disentangles intrinsic physiological patterns from motion artifacts.

For classification tasks (Panel B), DNA-PPG achieves the highest average AUROC of **0.661**. In affective computing benchmarks (WESAD, ECSMP), our framework demonstrates exceptional sensitivity to subtle signal variations, surpassing Moment in WESAD Valence detection by a large margin (0.737 vs.0.616). While PaPaGei-S exhibits a slight edge in sleep apnea detection (SDB), DNA-PPG delivers a more balanced and robust performance across the full spectrum of cardiovascular, sleep, and emotion tasks, verifying its capability as a versatile foundation model.

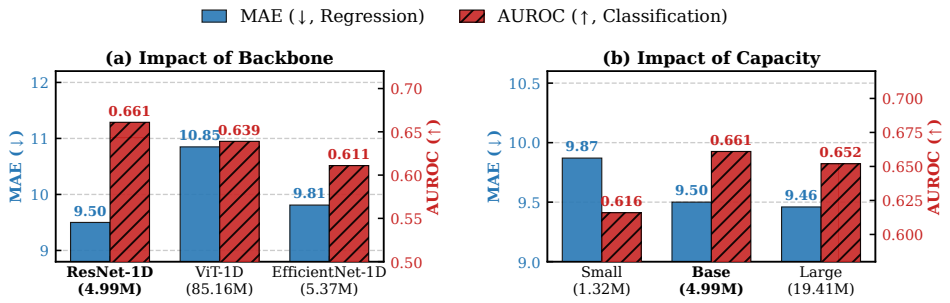

Figure 3: **Backbone ablation.** (Left) Comparison of architectures (ResNet-1D, ViT-1D, EfficientNet-1D). (Right) Impact of parameter size. The **Base** variant (4.99M) achieves the optimal trade-off, avoiding the diminishing returns observed in larger models.

Variant	Model Components			Downstream Tasks	
	Morph-SSL	TF-Soft	Phys-Align	Class. (AUROC ↑)	Reg. (MAE ↓)
Setting 1	√	–	–	0.611	10.33
Setting 2	√	√	–	0.624	10.22
Setting 3	–	–	√	0.603	10.40
Setting 4	√	–	√	0.632	10.02
DNA-PPG (Full)	√	√	√	**0.661**	**9.50**

Table 4: Component-wise ablation study. We evaluate the impact of the Morphology-Aware Self-Supervised Branch(**Morph-SSL**), Time-Frequency Soft Weighting mechanism(**TF-Soft**), and Physiological Semantic Alignment Branch(**Phys-Align**). Results report the average AUROC for classification and MAE for regression.

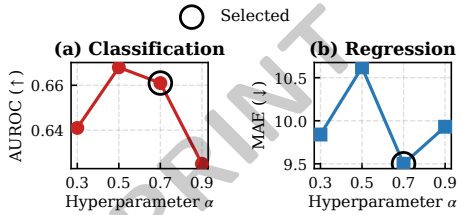

Figure 4: **Sensitivity analysis of α.** The plot illustrates the trade-off between classification (AUROC, left) and regression (MAE, right). We adopt $\alpha = 0.7$ as it minimizes MAE while maintaining competitive classification performance.

4.4 Ablation Study

Contribution of Key Components

Table 4 quantifies the contribution of each module. Single-branch variants (Settings 1 & 3) yield suboptimal generalization, whereas their combination (Setting 4) significantly improves performance, confirming the synergy between morphological learning and physiological alignment. Furthermore, incorporating Time-Frequency Soft Weighting (TF-Soft) consistently boosts results across all configurations, evidenced by the improvements of Setting 2 over 1 and DNA-PPG over Setting 4. The full framework achieves the best metrics (**0.661** AUROC, **9.50** MAE), verifying that refining sample interactions via time-frequency consistency is crucial for robust representation learning.

Backbone Analysis

To determine the optimal backbone encoder for our framework, we first evaluated three candidate architectures: ResNet-1D, ViT-1D, and EfficientNet-1D. As shown in the left panel of Fig. 3, **ResNet-1D** yields the best performance, attaining the lowest regression error (9.50 MAE) and highest classification score (0.661 AUROC). Based on this, we further investigate model scaling by varying the channel width of ResNet-1D (1.32M, 4.99M, and 19.41M parameters). As plotted in the right panel of Fig. 3, performance does not improve monotonically with scale. While the 4.99M variant yields significant gains over the 1.32M baseline, further scaling to 19.41M results in diminishing returns: regression error decreases negligibly (9.50 $\rightarrow$ 9.46 MAE) while classification performance degrades (0.661 $\rightarrow$ 0.652 AUROC). Consequently, we adopt the 4.99M ResNet-1D as the default backbone to balance performance and computational efficiency.

Sensitivity to Dual-Branch Balancing

Finally, we analyzed the sensitivity to the hyperparameter $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$, which controls the trade-off between the morphological and physiological objectives. As plotted in Fig. 4, the performance varies significantly with α. While $\alpha = 0.5$ achieves the highest classification score (0.668 AUROC), it incurs a substantial increase in regression error (10.62 MAE). Setting $\alpha = 0.7$ provides the most balanced performance, achieving the lowest MAE (9.50) while maintaining a competitive AUROC (0.661). Therefore, we adopt $\alpha = 0.7$ as the default setting.

5 Conclusion

In this paper, we presented DNA-PPG, a unified foundation model addressing manifold distortion and semantic precision loss in physiological representation learning. Through a *Dual Neighborhood Alignment* strategy, we bridge intrinsic signal morphology and clinical semantics. Specifically, Time-Frequency Soft Weighting captures inter-subject morphological affinities, while the Physiological Semantic Alignment branch exploits continuous targets for smooth health representations. Comprehensive experiments across cardiovascular, sleep, and affective domains verify that DNA-PPG significantly outperforms existing state-of-the-art frameworks. Future work will extend this paradigm to multi-modal signals and low-resource environments. Moreover, we plan to leverage the vast wearable-collected PPG signals lacking physiological indicators by using prediction-algorithm outputs as approximate gold standards, thereby further scaling pre-training with pseudo-supervision.

References

[Abbaspourazad *et al.*, 2023] Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

[Ansari *et al.*, 2024] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[Chen *et al.*, 2015] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Erturk *et al.*, 2025] Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. *arXiv preprint arXiv:2507.00191*, 2025.

[Gao *et al.*, 2021] Zhilin Gao, Xingran Cui, Wang Wan, Wenming Zheng, and Zhongze Gu. Ecsmp: A dataset on emotion, cognition, sleep, and multi-model physiological signals. *Data in Brief*, 39:107660, 2021.

[Garde *et al.*, 2014] Ainara Garde, Parastoo Dehkordi, Walter Karlen, David Wensley, J Mark Ansermino, and Guy A Dumont. Development of a screening tool for sleep disordered breathing in children using the phone oximeter™. *PloS one*, 9(11):e112959, 2014.

[Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[Kim *et al.*, 2022] Dong-Kyu Kim, Young-Tak Kim, Hakseung Kim, and Dong-Joo Kim. Deepcnap: A deep learning approach for continuous noninvasive arterial blood pressure monitoring using photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3697–3707, 2022.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kwon *et al.*, 2019] Soonil Kwon, Joonki Hong, Eue-Keun Choi, Euijae Lee, David Earl Hostallero, Wan Ju Kang, Byunghwan Lee, Eui-Rim Jeong, Bon-Kwon Koo, Seil Oh, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. *JMIR mHealth and uHealth*, 7(6):e12770, 2019.

[Lee *et al.*, 2022] Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

[Lee *et al.*, 2023] Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. *arXiv preprint arXiv:2312.16424*, 2023.

[Li *et al.*, 2023] Zhe Li, Zhongwen Rao, Lujia Pan, Pengyun Wang, and Zenglin Xu. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023.

[Liang *et al.*, 2018] Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):1–7, 2018.

[Loshchilov and Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[Nie *et al.*, 2024] Guangkun Nie, Jiabao Zhu, Gongzheng Tang, Deyun Zhang, Shijia Geng, Qinghao Zhao, and Shenda Hong. A review of deep learning methods for photoplethysmography data. *arXiv preprint arXiv:2401.12783*, 2024.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Papini *et al.*, 2020] Gabriele B Papini, Pedro Fonseca, Merel M van Gilst, Jan WM Bergmans, Rik Vullings, and Sebastiaan Overeem. Wearable monitoring of sleep-disordered breathing: estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography. *Scientific reports*, 10(1):13512, 2020.

[Pillai *et al.*, 2024] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

[Reiss *et al.*, 2019] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[Schmidt *et al.*, 2018] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[Toye, 2023] Pieter-Jan Toye. Vital videos: A dataset of face videos with ppg and blood pressure ground truths. *arXiv preprint arXiv:2306.11891*, 2023.

[Wang *et al.*, 2020] Dingliang Wang, Xuezhi Yang, Xuenan Liu, Shuai Fang, Likun Ma, and Longwei Li. Photoplethysmography based stratification of blood pressure using multi-information fusion artificial neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 276–277, 2020.

[Xu *et al.*, 2024] Maxwell A Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J Raghuram, James M Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. *arXiv preprint arXiv:2411.18822*, 2024.

[Yun *et al.*, 2024] Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, et al. Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. *Nature Genetics*, 56(8):1604–1613, 2024.

[Zhang *et al.*, 2018] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[Zhang *et al.*, 2022] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

[Zhang *et al.*, 2023] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023.