# PPGPT: Transferring Next-Token Modeling from Language to PPG Signals

**Zexing Zhang, Huimin Lu*, Qingxin Zhao**
School of Computer Science and Engineering, Changchun University of Technology, China
Jilin Province Smart Health Joint Innovation Laboratory for the New Generation of AI, China
Jilin Province Science and Technology Innovation Center for Multimodal Cognitive Computing and Analysis of Medical Biometrics, China
{2202303094@stu., luhuimin@, 2202403139@stu.}ccut.edu.cn

## Abstract

The success of large language models (LLMs) in cognitive tasks prompts the question of whether their next-token prediction (NTP) paradigm can be adapted to model physiological signals from wearable devices. A key target for this adaptation is photoplethysmography (PPG), the most prevalent sensing modality in consumer wearables for non-invasive monitoring of diverse physiological conditions. Unlike in NLP, where NTP aligns with generative objectives, physiological signal analysis involves fundamentally different tasks, such as continuous parameter estimation (regression) and discrete state recognition (classification). This disparity creates a semantic mismatch between the pre-training paradigm and the downstream tasks. To bridge this gap, we propose PPGPT, the first foundation model that reformulates NTP into next-feature token prediction (NFTP), learning hierarchical feature transition probabilities to unify pre-training and downstream objectives. PPGPT features a novel dual-stream encoder that generates feature tokens by jointly modeling temporal dynamics and local-global morphological patterns. The model is developed using a two-stage training framework: it is first pre-trained on a large-scale mixed dataset of 1.6 billion data points and then validated on our newly released BioMTL benchmark, which includes data from 172 subjects over 285 days across seven different tasks. Extensive experiments show that PPGPT significantly outperforms competing methods, achieving a 16.5% improvement in F1-score and a 25.9% reduction in Mean Absolute Error (MAE). Furthermore, the model demonstrates robust few-shot learning capabilities.

**Code** — https://github.com/PPGPT/AAAI2026.git

## Introduction

The proliferation of wearable devices has established multi-parameter physiological monitoring via photoplethysmography (PPG) as a key area in non-invasive health sensing (Kim et al. 2025; Shah et al. 2025). A PPG signal comprises a pulsatile AC component, which carries information on cardiovascular blood flow dynamics (Ray et al. 2021), and a non-pulsatile DC component, which reflects basal blood volume, sympathetic nervous activity, and thermoregulation (Kim et al. 2025). Consequently, PPG allows for the estimation of diverse physiological parameters, such as heart rate (Ribeiro et al. 2023), blood glucose (Chen et al. 2024b), blood pressure (Wang et al. 2024), emotional states (Choi et al. 2023), mental stress (Feli et al. 2023), and sleep stages (Cajal et al. 2023). This capability supports a wide range of applications, from chronic disease management to affective computing. Nevertheless, existing methods that rely on handcrafted features or task-specific supervised models demonstrate poor adaptability, facing challenges in multi-task generalization, domain transfer, and data-scarce scenarios (Fig. 1 (i)).

In stark contrast, large language models (LLMs) are emerging as general-purpose cognitive systems (Yin et al. 2024), achieving unprecedented generalization in domains such as natural language processing, image generation, and protein structure prediction (Awais et al. 2025; Jumper et al. 2021). This revolutionary success is driven by their next-token prediction (NTP) mechanism (Fig. 1 (ii)–(iii)(a)). The power of this approach stems from the alignment between the pretraining proxy task and final downstream generation tasks, which facilitates effective representation alignment and semantic compression. However, this principle of task-aligned pretraining remains largely unexplored in physiological signal modeling. This gap prompts our central research question: *Can the NTP paradigm be adapted to wearable physiological signals to construct effective and generalizable foundation models?*

A primary challenge stems from a fundamental mismatch in task objectives. The next-token prediction (NTP) paradigm is inherently **generative**, designed to predict subsequent content. In contrast, the core tasks for wearable physiological signals are typically **discriminative**, framed as classification or regression problems that infer physiological states or parameters from temporal features (Ray et al. 2021). This disparity between the generative nature of NTP and the discriminative target tasks in physiological modeling creates a significant barrier to the paradigm's direct transfer and application.

Separately, extensive research has applied self-supervised learning (SSL) to physiological signal modeling, utilizing strategies from contrastive to generative learning (Zhang et al. 2024a; Abbaspourazad et al. 2024; Chen et al. 2024c; Saha et al. 2025; Ding et al. 2024; Zhang et al. 2024b; Jiang, Zhao, and Lu 2024; Luo et al. 2024; Pillai et al. 2025). To

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

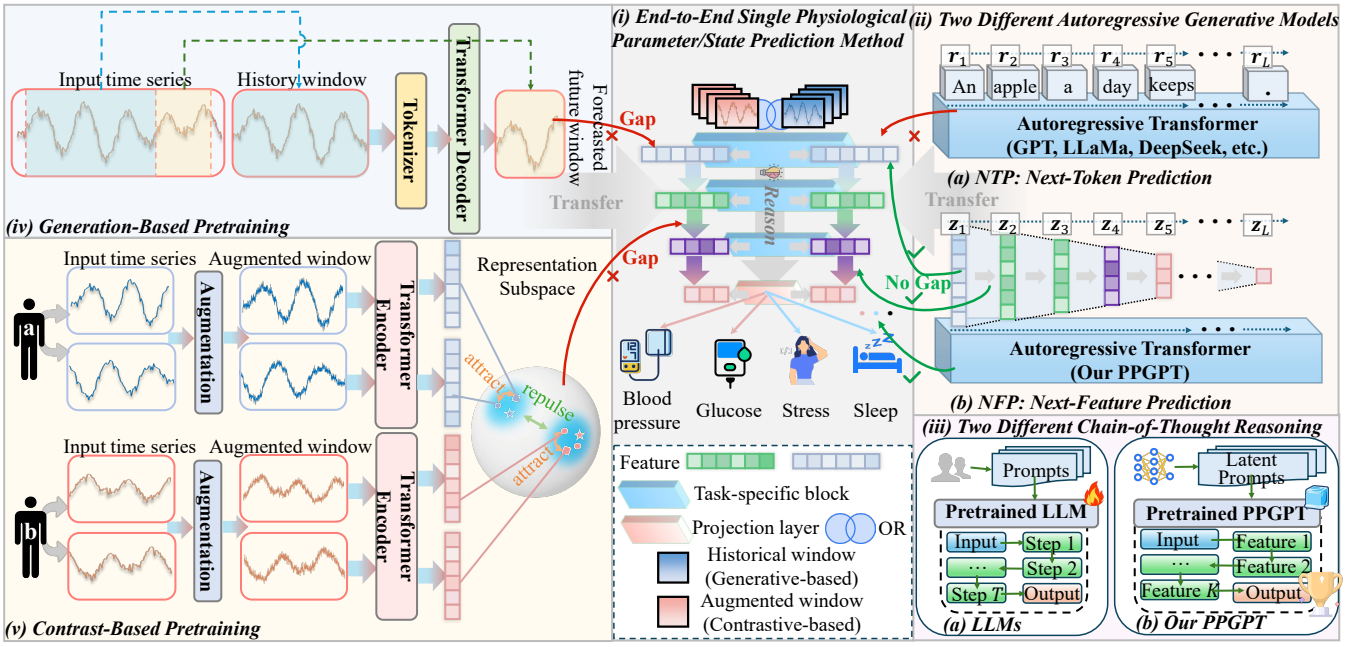



Figure 1: Conceptual comparison of pretraining paradigms for physiological signals. Prior methods suffer from a fundamental misalignment, as their pretraining proxy tasks are disconnected from the physiological semantics of downstream applications. In contrast, our PPGPT framework, using Next-Feature Token Prediction (NFTP), directly bridges this gap by learning physiologically meaningful feature transitions. This aligns the pretraining objective with downstream tasks, creating a unified and effective model. A detailed explanation of this paradigm is provided in the Introduction.

learn latent representations, these methods employ **structural proxy tasks**, such as signal reconstruction or segment contrast (Fig. 1 (iv)–(v)). However, a critical flaw in this approach is that these proxy tasks are defined by signal structure, not **semantic content**. This creates a fundamental misalignment with the objectives of downstream applications, a problem further detailed in Appendix A.

This leads to a central research question: *How can we design an "NTP-like" task for physiological signals that preserves the powerful "proxy-as-target" alignment of language models, while adapting it for the non-sequential, discriminative tasks for which autoregressive prediction is ill-suited?*

To address these challenges, we introduce PPGPT, a foundational model for physiological time series pretrained on 1.6 billion data points from large-scale PPG signals. Our framework redefines pretraining by learning feature transition probabilities, a paradigm designed for the complex, non-autoregressive nature of physiological monitoring (Fig. 1 (ii)–(iii)(b)). This work makes five key contributions:

- We construct and release **BioMTL**, the first multi-task benchmark for PPG-based physiological monitoring. It contains 285 days of longitudinal data from 172 subjects, with labels for seven tasks, including cuffless blood pressure, fingertip glucose, and mental stress.
- We propose **Next-Feature Token Prediction (NFTP)**, which learns transition probabilities between "Feature Tokens"—our novel, multi-level representation of signal dynamics and morphology. This directly aligns the pretraining objective with downstream discriminative tasks.
- Our work presents a new path for large-scale time-series pretraining that moves **beyond conventional point or patch-level modeling** and offers proven scalability.
- We design a dual-stream encoder that captures both temporal dynamics and local-to-global morphological features, providing a comprehensive representation of physiological signals.
- On the BioMTL benchmark, PPGPT outperforms all competing methods across all seven tasks and demonstrates exceptional few-shot learning capabilities.

## Methodology

### Overall Framework

We introduce a novel pretraining framework for wearable biosignals, centered on photoplethysmography (PPG), to address the absence of a unified and task-aligned paradigm. The core of our approach is to transform raw physiological signals into sequences of feature tokens, thereby enabling the model to learn the semantic evolution of their underlying multi-level patterns.

Our model processes a multi-source PPG dataset, $\mathcal{D}_{\text{raw}} = \{(\boldsymbol{x}_i^{\text{sig}}, \boldsymbol{x}_i^{\text{vis}})\}_{i=1}^{N}$. Each sample consists of a raw PPG time-series segment, $\boldsymbol{x}_i^{\text{sig}} \in \mathbb{R}^{1\times T}$, and its corresponding VisionPPG representation, $\boldsymbol{x}_i^{\text{vis}} \in \mathbb{R}^{H\times W\times 3}$. The VisionPPG

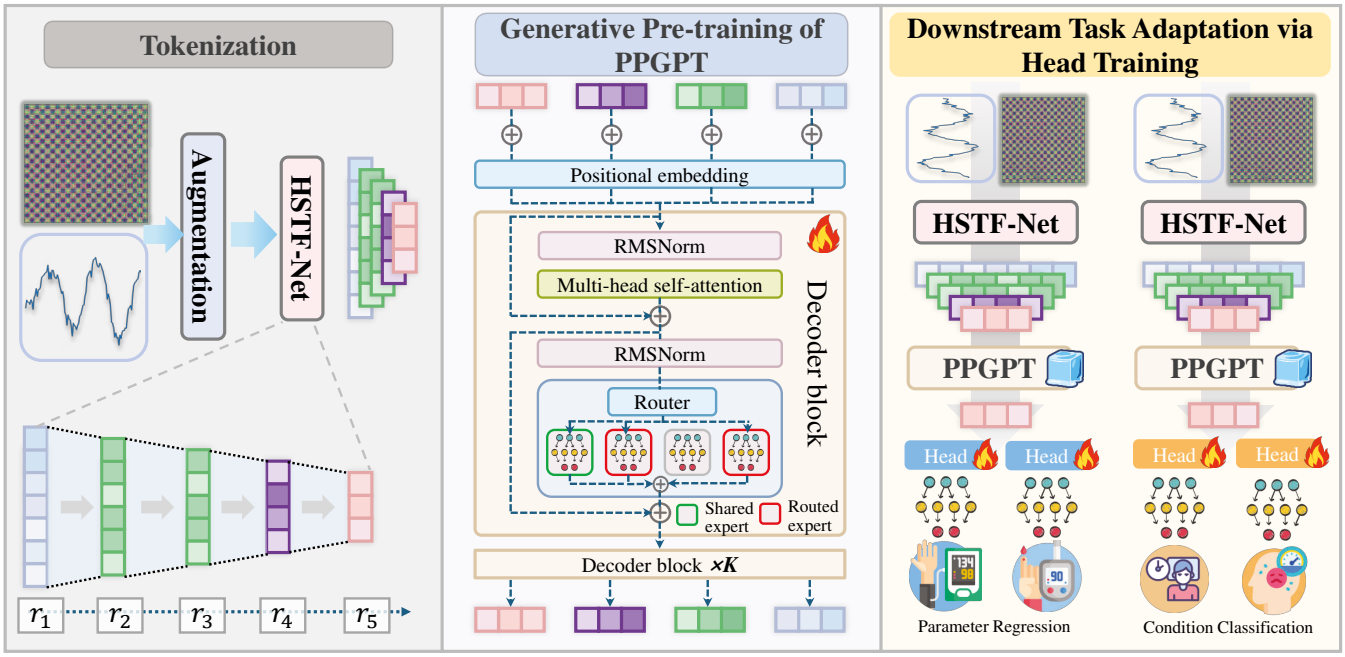



Figure 2: The three-stage architecture of PPGPT. (1) Tokenization: The HSTF encoder transforms a raw biosignal into a sequence of hierarchical feature tokens $\mathbf{z}^{(l)}$. (2) Generative Pre-training: A decoder is trained on the NFTP task to predict the next feature token from the preceding ones, learning the signal's intrinsic dynamics. (3) Downstream Adaptation: The pretrained encoder is frozen, and its feature tokens are aggregated to train a lightweight, task-specific head for final predictions.

transforms the signal into an image by encoding its spatiotemporal patterns. Appendix B provides the technical details for this transformation and for the HSTF architecture discussed next.

To unify these modalities, we designed the Hierarchical Spatio-Temporal Fusion Network (HSTF), denoted as $\mathcal{F}_{\text{HSTF}}$. This network maps the input pair $(\boldsymbol{x}_i^{\text{sig}}, \boldsymbol{x}_i^{\text{vis}})$ to a sequence of feature tokens $\mathcal{Z} = \mathbf{z}_1, \ldots, \mathbf{z}_L$, where each token $\mathbf{z}_l \in \mathbb{R}^d$ captures semantics at a distinct hierarchical level. This sequence $\mathcal{Z}$ then serves as the input to our core prediction model, $\mathcal{F}_{\text{PPGPT}}$.

Within each layer $l$ of the HSTF, a Temporal Contextualizer $(\phi_{\text{temp}}^{(l)})$ extracts features $\mathbf{F}_{PPG}^{(l)}$ from the time-series, while a Spatial Relation Aggregator $(\phi_{\text{spat}}^{(l)})$ extracts features $\mathbf{F}_{vis}^{(l)}$ from the VisionPPG image. These modality-specific features are integrated by a Bidirectional Attention Fusion module $(\phi_{\text{bi}}^{(l)})$. The resulting fused output is then projected through a normalization layer to produce the final token $\mathbf{z}_l$, ensuring a consistent feature dimension across all hierarchical levels.

Our framework is pretrained in two distinct stages on a mixed dataset of 1.6 billion time points. The first stage focuses on the HSTF encoder. To ensure training stability, we pretrain it on a high-quality subset, $\mathcal{D}_{\text{hq}} \subset \mathcal{D}_{\text{raw}}$. This process utilizes two auxiliary objectives: a contrastive loss $(\mathcal{L}_{\text{CL}})$ to maximize inter-subject variance, and a bidirectional jump prediction loss $(\mathcal{L}_{\text{BJP}})$ to preserve intra-subject temporal coherence. In the second stage, the pretrained HSTF generates feature tokens, $\mathcal{Z}$, from the entire raw dataset, $\mathcal{D}\text{raw}$. These tokens then serve as input for pretraining the $\mathcal{F}_{\text{PPGPT}}$ model via a Next Feature Token Prediction (NFTP) objective.

Following pretraining, the complete model is fine-tuned end-to-end for various downstream tasks. The entire workflow is illustrated in Fig. 2, and details on the dataset construction and filtering criteria are available in Appendix C.

## Hierarchical Spatio-Temporal Fusion Network

The HSTF encoder, $\mathcal{F}_{\text{HSTF}}$, is designed to process an input pair comprising a time-series signal $(\mathbf{x}_i^{\text{sig}})$ and its visual representation $(\mathbf{x}_i^{\text{vis}})$. Its core function is to generate a set of feature tokens, $\mathcal{Z} \in \mathbb{R}^{B\times L\times 2d}$, that encapsulates hierarchical, cross-modal semantics. The architecture operates through a sequential, multi-layer process. First, a modality-specific feature extraction module uses convolutional networks to independently map the temporal and spatial inputs into initial feature vectors, $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$. These initial vectors are then refined through $L$ hierarchical layers of cross-modal interaction. At each layer $l$, the network fuses the two feature streams to produce a dedicated feature token, $\mathbf{z}^{(l)}$. The final output, $\mathcal{Z}$, is the complete set of these tokens, representing the input data at multiple semantic levels.

**Temporal Contextualizer** The Temporal Contextualizer, $\psi_{\text{temp}}^{(l)}$, decomposes the input temporal feature $\mathbf{u}^{(l-1)}$ into task-relevant and sample-specific components using a dynamic gating mechanism. First, a gate vector $\mathbf{g}^{(l)}$ is com-

puted from the input. This gate then partitions the feature into a task-relevant component, $\mathbf{u}_{\text{task}}^{(l)} = \mathbf{g}^{(l)} \odot \mathbf{u}^{(l-1)}$, and a sample-specific component, $\mathbf{u}_{\text{sample}}^{(l)} = (1 - \mathbf{g}^{(l)}) \odot \mathbf{u}^{(l-1)}$, where $\odot$ denotes element-wise multiplication. Finally, each component is processed by a separate MLP, and the results are integrated with the original input via a residual connection to produce the refined feature vector: $\mathbf{u}^{(l)} = \mathbf{u}^{(l-1)} + \text{MLP}_{\text{task}}(\mathbf{u}_{\text{task}}^{(l)}) + \text{MLP}_{\text{sample}}(\mathbf{u}_{\text{sample}}^{(l)})$.

**Spatial Relation Aggregator** The Spatial Relation Aggregator, $\psi_{\text{spat}}^{(l)}$, models intra-sample spatial dependencies within visual features $\mathbf{v}^{(l-1)}$. First, it partitions the features of each sample into a fixed number of chunks, $P$, creating localized spatial patches represented as $\mathbf{C}^{(l)} \in \mathbb{R}^{B \times P \times d}$. Next, a multi-head Graph Attention Network (GAT) operates on these chunks, constructing an implicit affinity graph based on cosine similarity and updating the chunk representations to $\mathbf{C}'^{(l)}$ via edge-aware propagation. Finally, to ensure dimensional consistency, the updated features are reshaped and projected back to the original feature dimension, yielding the output $\mathbf{v}^{(l)} = \text{Proj}(\text{Reshape}(\mathbf{C}'^{(l)})) \in \mathbb{R}^{B \times d}$. This mechanism embeds non-local spatial relationships, enhancing the model's topological context awareness.

**Bidirectional Attention Fusion** The Bidirectional Attention Fusion mechanism, $\psi_{\text{attn}}^{(l)}$, jointly refines the unimodal features $\mathbf{u}^{(l)}$ and $\mathbf{v}^{(l)}$. The process begins with a bidirectional cross-attention operation, computing attention outputs like $\mathbf{a}_{\text{uv}}^{(l)} = \text{Attn}(\mathbf{Q}_u^{(l)}, \mathbf{K}_v^{(l)}, \mathbf{V}_v^{(l)})$. The query, key, and value matrices are derived from linear projections of the input features (e.g., $\mathbf{Q}_u^{(l)} = \mathbf{u}^{(l)}\mathbf{W}_Q^u$). To enhance stability, these attention outputs are integrated with the original features using residual connections followed by layer normalization, yielding intermediate representations $\widetilde{\mathbf{u}}^{(l)} = \text{LayerNorm}(\mathbf{u}^{(l)} + \mathbf{a}_{\text{uv}}^{(l)})$ and a corresponding $\widetilde{\mathbf{v}}^{(l)}$. Finally, these are fed into modality-specific MLPs to produce the layer's final outputs, $\mathbf{u}_{\text{out}}^{(l)} = \text{MLP}_u(\widetilde{\mathbf{u}}^{(l)})$ and $\mathbf{v}_{\text{out}}^{(l)} = \text{MLP}_v(\widetilde{\mathbf{v}}^{(l)})$.

**Unified Representation Generator** First, the final layer features from the dual streams, $\mathbf{u}^{(L)}$ and $\mathbf{v}^{(L)}$, are concatenated into a joint vector $\mathbf{z} = \text{Concat}(\mathbf{u}^{(L)}, \mathbf{v}^{(L)})$. This vector is then mapped to a unified representation $\mathbf{p} \in \mathbb{R}^{B \times d}$ by an MLP-based projection head. To pre-train the feature extractor, we employ a dual-objective strategy. For each subject $i$, we sample two distinct, augmented temporal segments following (Abbaspourazad et al. 2024) and process them through our model $f_{\text{HSTF}}$ to obtain their embeddings, $\mathbf{p}_1$ and $\mathbf{p}_2$. The first objective is a contrastive loss, $\mathcal{L}_{\text{CL}} = -\log \frac{\exp(\text{sim}(\mathbf{p}_1, \mathbf{p}_2)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{p}_1, \mathbf{p}_j^-)/\tau)}$, which enforces semantic consistency. Here, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, $\tau$ is a temperature parameter, and $\mathbf{p}_j^-$ denotes a negative sample. The second objective, a bidirectional jump prediction loss $\mathcal{L}_{\text{BJP}} = \|\phi(\mathbf{p}_1) - \mathbf{x}_{i,2}^{\text{sig}}\|_2^2 + \|\phi(\mathbf{p}_2) - \mathbf{x}_{i,1}^{\text{sig}}\|_2^2$, preserves temporal dynamics. This task requires a decoder $\phi$ to reconstruct a non-consecutive signal segment from an embedding, encouraging the representation to capture slowly-varying attributes and preventing the model from learning trivial, identity-based shortcuts.

## Generative Pre-training of PPGPT

We introduce *Next Feature Token Prediction (NFTP)*, a generative pre-training method that adapts autoregressive principles from language modeling to continuous biosignal representations. The Hierarchical Spatio-Temporal Fusion (HSTF) encoder transforms an input signal $\mathbf{x}$ into a sequence of multi-level feature tokens, $\mathcal{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)}\}$. Each token $\mathbf{z}^{(l)}$ represents the signal at the $l$-th layer of semantic abstraction.

The core objective of NFTP is to predict the next feature token $\mathbf{z}^{(l+1)}$ given the sequence of all preceding tokens, $\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(l)}\}$. To achieve this, we employ a decoder-only Transformer, $\mathcal{D}_\theta$, which models the conditional probability of the next token. This decoder, based on the DeepSeekMoE architecture (Dai et al. 2024), is trained by minimizing the L2 reconstruction error across all layers:

$$\mathcal{L}_{\text{NFTP}}(\theta) = \sum_{l=1}^{L-1} \mathbb{E}_{\mathbf{x} \sim \mathcal{X}} \left\| \mathcal{D}_\theta(\mathbf{z}^{(1:l)}) - \mathbf{z}^{(l+1)} \right\|_2^2$$

where $\mathbf{z}^{(1:l)}$ denotes the concatenation of the first $l$ feature tokens.

Unlike conventional autoregressive models that predict future time steps, NFTP operates along the axis of semantic depth. It learns to predict a more abstract feature representation from a less abstract one. This process effectively teaches the model the transition dynamics between layers of abstraction, creating a powerful inductive bias for modeling how semantic features evolve. This learned dynamic representation is highly transferable, enhancing performance on downstream tasks. A theoretical analysis of NFTP's properties is provided in Appendix D.

## Downstream Task Adaptation via Head Training

For downstream tasks, we adopt an efficient adaptation strategy: we freeze the pretrained HSTF encoder and train only a lightweight, task-specific head (Pillai et al. 2025; Abbaspourazad et al. 2024; Saha et al. 2025). This head operates on the hierarchical feature tokens $\mathcal{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)}\}$ generated by the encoder.

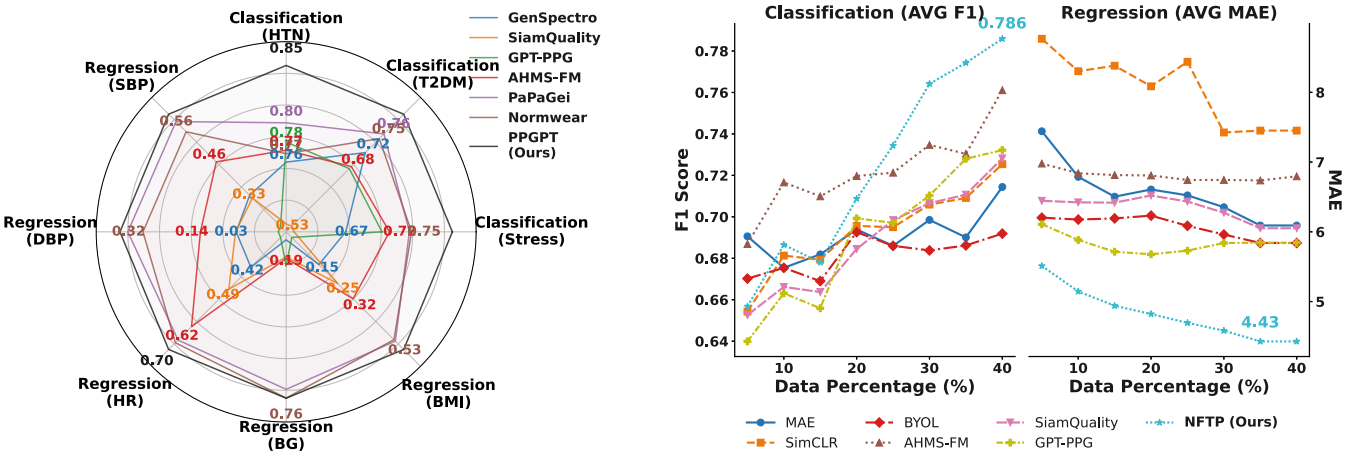
To create a single, powerful representation, we first aggregate these tokens using a weighted sum: $\mathbf{z}_{\text{agg}} = \sum_{l=1}^{L} \alpha_l \mathbf{z}^{(l)}$, where the weights $\alpha_l$ are learnable parameters constrained such that $\sum \alpha_l = 1$ and $\alpha_l \geq 0$. This aggregated feature vector, $\mathbf{z}_{\text{agg}}$, serves as the input to the task head.

Our primary task-specific head, $\mathcal{H}_\tau$, is a Multi-Layer Perceptron (MLP), which takes the aggregated feature vector $\mathbf{z}_{\text{agg}}$ as input to produce the final prediction $\hat{\mathbf{y}}_\tau = \text{MLP}_\tau(\mathbf{z}_{\text{agg}})$. In our analysis, we also explore simpler, non-parametric heads such as k-Nearest Neighbors (KNN) and Random Forests (RF) to benchmark the quality of the learned representations.

| Metric | Weight (kg) | Height (m) | Age (years) | BMI (kg/m$^2$) | HR (bp/m) | SBP (mm Hg) | DBP (mm Hg) | BG (mmol/L) |
|---|---|---|---|---|---|---|---|---|
| Mean | 66.02 | 1.71 | 48.9 | 22.38 | 80 | 119.50 | 72.39 | 6.4 |
| Minimum | 45 | 1.55 | 21 | 16.84 | 55 | 82 | 34 | 4.2 |
| Maximum | 98 | 1.87 | 96 | 36.73 | 116 | 164 | 105 | 16.3 |
| Standard Deviation | 12.93 | 0.09 | 25.5 | 3.29 | 13 | 17.99 | 10.63 | 2.1 |

BMI: Body mass index, SBP: Systolic blood pressure, DBP: Diastolic blood pressure, BG: Blood glucose.

Table 1: Demographic and physiological statistics of the BioMTL dataset.




(a) Task-wise Radar Plot

(b) Few-sample Performance

Figure 3: Multi-Task Evaluation Results

# Experiments

## Experimental Setup

**Datasets** We evaluate PPGPT on BioMTL, a novel photoplethysmography (PPG)-centric, multi-task learning benchmark for physiological signals. To foster reproducible research, we have made this dataset and code publicly available. BioMTL contains 1,808 high-quality samples from 172 subjects, aged 21 to 96, collected over 285 days. The data were gathered under an open-environment protocol, capturing natural signal variations and physiological drift. Each sample consists of a PPG signal synchronized with annotations for seven physiological tasks. These include four continuous parameters (cuffless blood pressure, fingertip blood glucose, heart rate, and BMI) and three binary states (mental stress, hypertension (HTN), and type-2 diabetes (T2DM)). As detailed in Appendix E, standardized acquisition procedures and rigorous quality assessments ensure inter-sample consistency. Tab. 1 summarizes the demographic and physiological statistics, highlighting the dataset's diversity (e.g., BMI: 16.84—36.73 kg/m$^2$; blood glucose: 4.2—16.3 mmol/L). A comprehensive benchmark evaluation is provided in Appendix O.

**Baselines and Metrics** We evaluate PPGPT against several baseline models, including general-purpose frameworks like PaPaGei (Pillai et al. 2025), task-specific architectures such as $\Delta$BP-Net (Wang et al. 2024), and self-supervised learning (SSL) strategies like SimCLR (Chen et al. 2020). To ensure fair comparisons, these SSL methods were adapted to our backbone architecture, with full details in Appendix F. We assess performance using MAE, RMSE, and $R^2$ for regression tasks (e.g., SBP, DBP, BG) and the F1-score and Accuracy for classification tasks (e.g., mental stress, HTN). The quality of the pre-trained representations is measured with RankMe (Garrido et al. 2023) and $\alpha$-ReQ (Agrawal et al. 2022), as detailed in Appendix G.

**Implementation Details** The HSTF encoder is first pre-trained for 50 epochs on $\mathcal{D}_{\text{hq}}$ using the $\mathcal{L}_{\text{CL}}$ loss with a temperature of 0.07, alongside the $\mathcal{L}_{\text{BJP}}$ loss. Next, the model undergoes 100 epochs of NFTP pre-training on $\mathcal{D}_{\text{raw}}$ with a 12-layer Transformer, as detailed in Appendix H. The process concludes with 50 epochs of head-only fine-tuning.

## Main Results

**Multi-Task Performance** We benchmarked PPGPT against eight state-of-the-art baselines on the BioMTL benchmark, which covers seven physiological tasks. These include regression for blood pressure, glucose, heart rate, and BMI, as well as classification for mental stress, hypertension, and type-2 diabetes. The baselines consist of AHMS-FM (Abbaspourazad et al. 2024), SimCLR (Chen et al. 2020), BYOL (Grill et al. 2020), SiamQuality (Ding et al. 2024), GPT-PPG (Chen et al. 2024c), PaPaGei (Pillai et al. 2025), Normwear (Luo et al. 2024), and MAE (He et al. 2022). As summarized in Tab. 2, PPGPT achieves an average F1-score gain of 16.5% and an MAE reduction of 25.9% across all baseline models. The performance gap

| Pretraining Method | Regression (AVG) | | Classification (AVG) | | Pretraining | |
|---|---|---|---|---|---|---|
| | MAE ↓ | $R^2$ ↑ | F1 ↑ | Accuracy ↑ | RankMe ↑ | $\alpha$-ReQ ↓ |
| AHMS-FM | 5.2683 (-23.8%) | 0.4089 (+49.1%) | 0.7133 (+14.9%) | 0.7856 (+9.9%) | 80.6230 (+55.2%) | 3.2040 (-29.6%) |
| SimCLR (our variation) | 5.5087 (-27.2%) | 0.3571 (+70.7%) | 0.7080 (+15.7%) | 0.7810 (+10.6%) | 28.8100 (+334.3%) | 3.0102 (-25.0%) |
| BYOL (our variation) | 5.5186 (-27.3%) | 0.3602 (+69.3%) | 0.6965 (+17.6%) | 0.7696 (+12.2%) | 16.2717 (+668.9%) | 2.6880 (-16.0%) |
| SiamQuality | 5.4810 (-26.8%) | 0.3590 (+69.9%) | 0.6826 (+20.0%) | 0.7605 (+13.6%) | 51.7325 (+141.8%) | 2.8255 (-20.1%) |
| GPT-PPG | 5.3744 (-25.3%) | 0.3759 (+62.2%) | 0.7092 (+15.5%) | 0.7828 (+10.3%) | 74.5020 (+67.9%) | 2.7738 (-18.6%) |
| MAE (our variation) | 5.3867 (-25.5%) | 0.3935 (+55.0%) | 0.7120 (+15.1%) | 0.7815 (+10.5%) | 87.7653 (+42.5%) | 2.9064 (-22.3%) |
| **NFTP (ours)** | **4.0123** | **0.6098** | **0.8193** | **0.8636** | **125.1079** | **2.2568** |

Table 2: Summary of the results for regression, classification, and pretraining.

| Model Variant | $\alpha$-ReQ↓ | RankMe↑ |
|---|---|---|
| w/o BAF | 3.202 | 62.568 |
| w/o TC | 2.940 | 47.400 |
| w/o SRA | 3.277 | 41.758 |
| Spatial-Only | 3.070 | 33.553 |
| Temporal-Only | 2.932 | 11.513 |
| **Full HSTF** | **2.560** | **75.428** |

Table 3: Ablation study of HSTF components. We evaluate the impact of the Temporal Contextualizer (TC), Spatial Relation Aggregator (SRA), and Bilinear Attention Fusion (BAF) modules on overall performance.

| Pre-training | Classification (AVG F1↑) | Regression (AVG MAE↓) |
|---|---|---|
| $(\mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{BJP}})$-Only | 0.56 | 7.82 |
| $\boldsymbol{\mathcal{L}_{\text{NFTP}}}$ | **0.81** | **4.01** |

Table 4: Results of the ablation study on our pre-training method.

remains substantial even against the strongest competitor, AHMS-FM, where PPGPT achieves a 23.8% lower MAE and a 14.9% higher F1-score. The task-specific radar plots in Fig. 3(a) further highlight these gains, particularly in blood pressure estimation, with an $R^2$ improvement from 0.46 to 0.62, and in mental stress classification, with an F1-score increase from 0.67 to 0.79. This demonstrates PPGPT's balanced and robust performance across diverse physiological tasks and underscores its ability to bridge the generative–discriminative semantic gap. Extended analyses are available in Appendix M.

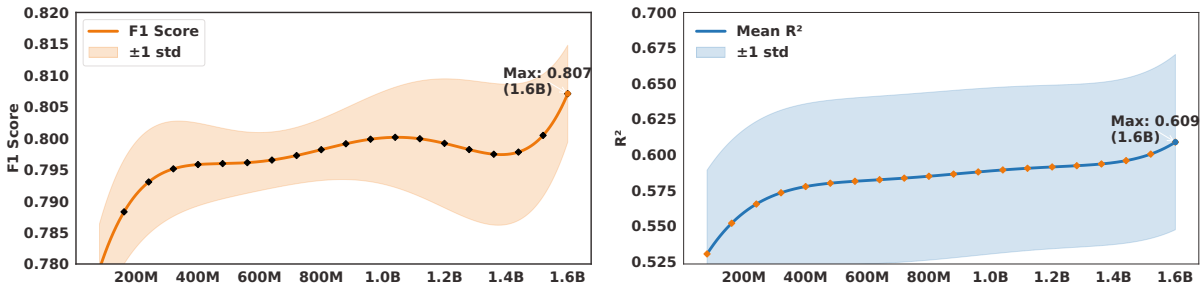
**Pretraining Effectiveness** To evaluate the quality of the learned representations, we use the RankMe and $\alpha$-ReQ pre-training metrics, with the results summarized in Tab. 3. PPGPT achieves a RankMe score of 125.1079 and an $\alpha$-ReQ of 2.2568. This marks a 42.5% improvement in RankMe over the top-performing baseline, MAE, and a 16.0% reduction in $\alpha$-ReQ compared to GPT-PPG. These findings suggest that our proposed Next-Feature Token Prediction (NFTP) paradigm produces more expressive and compact representations than alternative strategies based on contrastive learning, such as AHMS-FM and SimCLR, or reconstruction, like MAE and GPT-PPG. Furthermore, the hierarchical feature token modeling in NFTP enhances the alignment between the proxy task and downstream objectives. This strong alignment is evidenced by a high linear correlation between pre-training quality and downstream performance, yielding a Pearson's r of 0.89 between the RankMe and F1-scores. A complete analysis of the pre-training evaluation is detailed in Appendix I.

**Few-Sample Learning** To evaluate robustness in data-scarce scenarios, we fine-tune the model on BioMTL subsets ranging from 40% down to 5%. We assess generalization performance using the average F1-score across classification tasks and the Mean Absolute Error (MAE) across regression tasks. As illustrated in Fig. 3(b), PPGPT consistently outperforms the majority of baselines across all data scales. This robust performance is attributable to its semantically aligned NFTP pre-training and effective cross-modal feature fusion. These findings confirm PPGPT's effectiveness in few-sample settings, highlighting its suitability for real-world applications with limited labeled data.

## Ablation Studies

**Architectural Components** As detailed in Tab. 3, the complete HSTF model achieves the best performance, recording an $\alpha$-ReQ of 2.560 and a RankMe score of 75.428. Removing the BAF module causes the most significant performance degradation, increasing the $\alpha$-ReQ to 3.202 and reducing the RankMe score to 62.568. This result underscores the module's critical role in integrating temporal and spatial features. Similarly, excluding either the TC or SRA module substantially impairs representation quality, causing RankMe scores to drop to 47.400 and 41.758, respectively. Furthermore, the single-stream variants, namely the TC-only and SRA-only models, perform poorly and yield RankMe scores of just 11.513 and 33.553. These findings confirm that all components are integral and collectively demonstrate the synergistic advantage of our dual-stream design for learning expressive representations. Crucially, optimal performance is achieved only when HSTF pre-trained representations are subsequently refined by the PPGPT phase, a finding that validates the NFTP methodology. Furthermore, the HSTF encoder's architectural modularity ensures extensibility, permitting its replacement with alternative architectures.

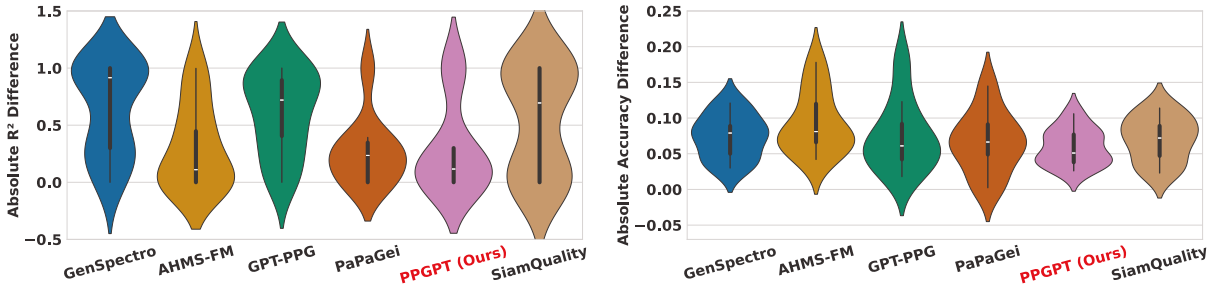
Table 4 shows that the performance gain primarily comes from the NFTP-based feature transfer, not from the struc-




(a) Classification F1 Score vs. Data Scale

(b) Regression $R^2$ vs. Data Scale

Figure 4: Performance of PPGPT trained on varying fractions (5%–100%) of the 1.6B-point dataset, showing decreasing $R^2$ and F1-score with smaller data scales, and diminishing returns beyond 80%.




(a) $R^2$ Differences (Regression)

(b) Accuracy Differences (Classification)

Figure 5: Violin plots showing absolute performance differences across demographic splits for six models: NFTP (ours) and five baselines (GenSpectro, AHMS-FM, GPT-PPG, SiamQuality, PaPaGei). Comparisons are made across three subgroup pairs: adult vs. elderly (AvE–EvA), female vs. male (FvM–MvF), and normal vs. abnormal BMI (NvA–AvN). (a) $R^2$ differences for regression tasks (BG, DBP, SBP); (b) Accuracy differences for classification tasks (HTN, Stress, T2DM).

tured proxy tasks, confirming NFTP's ability to learn broadly transferable representations.

**Impact of Pretraining Data Scale** As shown in Fig. 4, representation quality scales positively with the pretraining data volume. PPGPT exhibits notable data efficiency by maintaining robust performance even when trained on a moderate 50% data subset, confirming its viability for resource-constrained settings. Further details are provided in Appendix N.

## Qualitative Analysis of Demographic Robustness

The qualitative analysis in Fig. 5 confirms the demographic robustness of NFTP-based models. They consistently maintain low prediction error variance across subgroups, a stark contrast to the significant instability observed in baseline models. This superior robustness underscores NFTP's generalizability for physiological modeling in heterogeneous real-world populations.

## Related Work

Monitoring physiological parameters such as blood pressure and stress from wearable signals like PPG is a critical task, yet models often fail to generalize across diverse individuals and conditions (Zhang et al. 2024b; Chen et al. 2025). Traditional approaches depend on handcrafted features or task-specific architectures, which limits their scalability and adaptability (El-Dahshan et al. 2024; Chen et al. 2024a). While modern self-supervised learning (SSL) methods improve robustness (Ding et al. 2024; Chen et al. 2024c; Zhang et al. 2024b), they suffer from a fundamental misalignment: their proxy tasks, like masked reconstruction or segment contrast, are disconnected from the physiological semantics needed for downstream applications (Hou et al. 2025; Tian et al. 2025). In contrast, our NFTP paradigm directly bridges this gap by aligning the pretraining objective with these semantics, creating a unified and effective model for diverse physiological monitoring tasks. (More related work is included in Appendix A)

## Conclusion

We introduce PPGPT, a foundation model that utilizes Next-Feature Token Prediction (NFTP) to unify physiological representations. Pretrained on a 1.6-billion-point corpus via a dual-stream HSTF encoder, PPGPT achieves a 25.9% MAE reduction and 16.5% F1-score improvement on the BioMTL benchmark. The model demonstrates robust generalization in few-shot and diverse demographic scenarios. Furthermore, we release the BioMTL benchmark to facilitate scalable research in multi-modal health applications.

## Acknowledgments

This research was supported by the 2023 Jilin Provincial Development and Reform Commission Industrial Technology Research and Development Project [No.2023C042-6].

The authors also acknowledge the Jilin Province Science and Technology Innovation Center for Multimodal Cognitive Computing and Analysis of Medical Biometrics, as well as the Smart Health Joint Innovation Laboratory for the Next Generation of AI, for their valuable contributions to this research.

## References

Abbaspourazad, S.; Elachqar, O.; Miller, A.; Emrani, S.; Nallasamy, U.; and Shapiro, I. 2024. Large-scale Training of Foundation Models for Wearable Biosignals. In *The Twelfth International Conference on Learning Representations*.

Agrawal, K. K.; Mondal, A. K.; Ghosh, A.; and Richards, B. 2022. $alpha$-ReQ: Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35: 17626–17638.

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2025. Foundation Models Defining a New Era in Vision: a Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cajal, D.; Gil, E.; Laguna, P.; Varon, C.; Testelmans, D.; Buyse, B.; Jensen, C.; Hoare, R.; Bailón, R.; and Lázaro, J. 2023. Obstructive sleep apnea screening by joint saturation signal analysis and PPG-derived pulse rate oscillations. *IEEE Journal of Biomedical and Health Informatics*, 28(1): 228–238.

Chen, Q.; Yang, X.; Chen, Y.; Han, X.; Gong, Z.; Wang, D.; and Zhang, J. 2024a. A blood pressure estimation approach based on single-channel photoplethysmography differential features. *Biomedical Signal Processing and Control*, 97: 106662.

Chen, S.; Fan, S.; Qiao, Z.; Wu, Z.; Lin, B.; Li, Z.; Riegler, M. A.; Wong, M. Y. H.; Opheim, A.; Korostynska, O.; et al. 2025. Transforming Healthcare: Intelligent Wearable Sensors Empowered by Smart Materials and Artificial Intelligence. *Advanced Materials*, 2500412.

Chen, S.; Qin, F.; Ma, X.; Wei, J.; Zhang, Y.-T.; Zhang, Y.; and Jovanov, E. 2024b. Multi-view cross-fusion transformer based on kinetic features for non-invasive blood glucose measurement using PPG signal. *IEEE Journal of Biomedical and Health Informatics*, 28(4): 1982–1992.

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

Chen, Z.; Ding, C.; Modhe, N.; Lu, J.; Yang, C.; and Hu, X. 2024c. Adapting a generative pretrained transformer achieves sota performance in assessing diverse physiological functions using only photoplethysmography signals: A gpt-ppg approach. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Choi, J.; Hwang, G.; Lee, J. S.; Ryu, M.; and Lee, S. J. 2023. Weighted knowledge distillation of attention-LRCN for recognizing affective states from PPG signals. *Expert Systems with Applications*, 233: 120883.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; Xie, Z.; Li, Y.; Huang, P.; Luo, F.; Ruan, C.; Sui, Z.; and Liang, W. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1280–1297. Bangkok, Thailand: Association for Computational Linguistics.

Ding, C.; Guo, Z.; Chen, Z.; Lee, R. J.; Rudin, C.; and Hu, X. 2024. SiamQuality: a ConvNet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8): 085004.

El-Dahshan, E.-S. A.; Bassiouni, M. M.; Khare, S. K.; Tan, R.-S.; and Acharya, U. R. 2024. ExHyptNet: An explainable diagnosis of hypertension using EfficientNet with PPG signals. *Expert Systems with Applications*, 239: 122388.

Feli, M.; Kazemi, K.; Azimi, I.; Wang, Y.; Rahmani, A. M.; and Liljeberg, P. 2023. End-to-end ppg processing pipeline for wearables: From quality assessment and motion artifacts removal to hr/hrv feature extraction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1895–1900. IEEE.

Garrido, Q.; Balestriero, R.; Najman, L.; and Lecun, Y. 2023. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, 10929–10974. PMLR.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hou, Y.; Wang, B.; Zhang, C.; Wang, Q.; Li, J.; Meng, P.; Zhang, Y.; Han, C.; Hong, F.; and Zhang, T. 2025. OSASformer: A transformer-based model for OSAS screening via multi-source representation fusion. *Knowledge-Based Systems*, 316: 113365.

Jiang, W.; Zhao, L.; and Lu, B.-l. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *The Twelfth International Conference on Learning Representations*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kim, C.; Lee, K.; Kim, J.; Yang, D.; Lee, H.; Moon, G.; Kim, Y.; Cho, D.; Bae, K. S.; Kim, G.; et al. 2025. Multipoint sensing organic light-emitting diode display based mobile cardiovascular monitor. *Nature Communications*, 16(1): 1666.

Luo, Y.; Chen, Y.; Salekin, A.; and Rahman, T. 2024. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals. *arXiv preprint arXiv:2412.09758*.

Pillai, A.; Spathis, D.; Kawsar, F.; and Malekzadeh, M. 2025. PaPaGei: Open Foundation Models for Optical Physiological Signals. In *The Thirteenth International Conference on Learning Representations*.

Ray, D.; Collins, T.; Woolley, S. I.; and Ponnapalli, P. V. 2021. A review of wearable multi-wavelength photoplethysmography. *IEEE Reviews in Biomedical Engineering*, 16: 136–151.

Ribeiro, L.; Hu, X.; Oliveira, H. P.; and Pereira, T. 2023. Ai-based models to predict the heart rate using ppg and accelerometer signals during physical exercise. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4398–4403. IEEE.

Saha, M.; Xu, M. A.; Mao, W.; Neupane, S.; Rehg, J. M.; and Kumar, S. 2025. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3): 1–35.

Shah, K.; Wang, A.; Chen, Y.; Munjal, J.; Chhabra, S.; Stange, A.; Wei, E.; Phan, T.; Giest, T.; Hawkins, B.; et al. 2025. Automated loss of pulse detection on a consumer smartwatch. *Nature*, 1–3.

Tian, Z.; Liu, A.; Zhu, G.; and Chen, X. 2025. A paralleled CNN and Transformer network for PPG-based cuffless blood pressure estimation. *Biomedical Signal Processing and Control*, 99: 106741.

Wang, W.; Mohseni, P.; Kilgore, K. L.; and Najafizadeh, L. 2024. $\Delta$ BP-Net: Monitoring "Changes" in Blood Pressure Using PPG with Self-contrastive Masking. *IEEE Journal of Biomedical and Health Informatics*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Zhang, K.; Wen, Q.; Zhang, C.; Cai, R.; Jin, M.; Liu, Y.; Zhang, J. Y.; Liang, Y.; Pang, G.; Song, D.; et al. 2024a. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE transactions on pattern analysis and machine intelligence*.

Zhang, Z.; Lu, H.; Ma, S.; Peng, J.; Lin, C.; Li, N.; and Dong, B. 2024b. A general framework for generative self-supervised learning in non-invasive estimation of physiological parameters using photoplethysmography. *Biomedical Signal Processing and Control*, 98: 106788.