PPG-DISTILL: Efficient Photoplethysmography Signals Analysis via Foundation Model Distillation

Juntong Ni[1], Saurabh Kataria[2], Shengpu Tang[1], Carl Yang[1], Xiao Hu[2], Wei Jin[1]
[1]Department of Computer Science, Emory University
[2]Nell Hodgson Woodruff School of Nursing, Emory University
{firstname.lastname}@emory.edu

Abstract

Photoplethysmography (PPG) is widely used in wearable health monitoring, yet large PPG foundation models remain difficult to deploy on resource-limited devices. We present **PPG-DISTILL**, a knowledge distillation framework that transfers both global and local knowledge through prediction-, feature-, and patch-level distillation. PPG-DISTILL incorporates *morphology distillation* to preserve local waveform patterns and *rhythm distillation* to capture inter-patch temporal structures. On heart rate estimation and atrial fibrillation detection, PPG-DISTILL improves student performance by up to 21.8% while achieving $7\times$ faster inference and reducing memory usage by $19\times$, enabling efficient PPG analysis on wearables. Our code is available at `https://github.com/LingFengGold/PPG-Distill`.

1 Introduction

Wearable sensors that are unobtrusive, widely accessible, and cost-effective have demonstrated strong potential for real-time health monitoring. Among these, photoplethysmography (PPG), an inherently time-series signal that captures continuous variations in blood volume over time, has become a widely used modality in smartwatches [4, 30]. Its popularity arises from enabling non-invasive physiological assessment without requiring firm skin attachment [30, 33]. The rich information in PPG arises from its local waveform morphology, which reflects cardiovascular events, and its long-range structural rhythm, reflecting periodicity and autonomic regulation. These properties enable applications from cardiovascular monitoring [28, 32, 31, 36, 10, 2, 26], clinical diagnostics [34, 27, 5, 14, 38, 18], to mental state assessment [41, 13, 35].

Given its wide range of applications, it is crucial to develop models that can learn generalizable representations from PPG signals and perform reliably across multiple downstream tasks. Recent studies have therefore introduced foundation models tailored to PPG signals [15, 25, 4, 30, 7]. Although these models demonstrate strong performance, deploying them on edge devices such as wearables remains difficult due to constraints on inference speed and memory usage. A natural solution is to leverage knowledge distillation (*KD*) [11, 9] to compress large teacher models into a smaller, more efficient student models (Figure 1). However, the primary challenge lies in knowledge preservation, since vanilla *KD* techniques may fail to transfer the nuanced understanding of PPG's unique characteristics. This raises a critical question: *What specific structural and temporal knowledge is essential for a PPG model, and how can it be effectively distilled from a teacher to a student?*

Most existing distillation methods concentrate on aligning output predictions [11] or intermediate feature [29] between a teacher and a student, namely **Global *KD***. Such approaches risk overlooking the local structural information that is central to PPG. In particular, waveform morphology within short temporal windows (patches) and structural rhythm between patches are essential for capturing both cardiovascular events and autonomic dynamics, yet these fine-grained patterns can be lost when only global prediction- or feature-level alignment is enforced. Moreover, recent PPG foundation

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

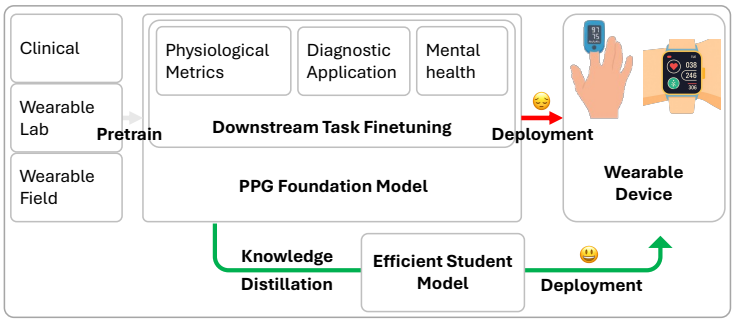

Figure 1: Illustration of our motivation. PPG foundation models are pretrained and finetuned for downstream tasks, but direct deployment on wearables is costly. *KD* produces efficient student models suitable for wearable deployment.

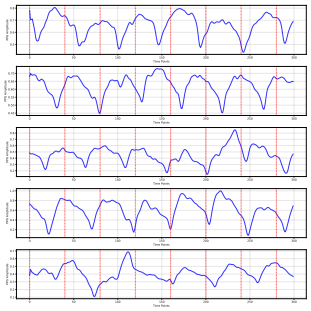

Figure 2: Real PPG signals from the StanfordAF dataset, segmented into patches by red lines (patch size = 40).

models [25, 4] already adopt a patch-based representation, which naturally encodes local dynamics but remains underutilized during distillation. To address this gap, we introduce **PPG-DISTILL**, a distillation framework that augments vanilla prediction- and feature-level transfer with two novel patch-level strategies: morphology distillation, which enforces discriminability among local segments, and rhythm distillation, which preserves structural dependencies across patches. By explicitly transferring both global knowledge and local morphology–rhythm patterns, PPG-DISTILL equips the student with richer PPG-specific representations. This design enables compact models that maintain strong task performance and are practical for on-device deployment. Across diverse benchmarks, PPG-DISTILL achieves up to 21.80% higher accuracy while reducing inference latency by up to $7\times$ and memory footprint by up to $19\times$ compared to the teacher, advancing the deployment of foundation-level PPG models in wearables. We discuss the related work in Appendix A

2 Methodology

We first introduce key notations. For PPG signal anslysis, given an input PPG signal $X \in \mathbb{R}^L$, where L represents the length of the PPG signal, the goal is to predict the value $Y \in \mathbb{R}^1$ for regression and the class $Y \in \mathbb{R}^C$ for classification, where C is the number of classes. Below, we propose and discuss several approaches to distill knowledge from a teacher PPG foundation model to a student. We start by adapting two Global *KD* methods: prediction-matching and feature-matching. Next, we motivate and introduce our proposed PPG-DISTILL, with patch-level matching strategies to distill additional patch-level local morphology-aware and structural rhythm information to the student.

2.1 Global *KD*

The student produces predictions $\hat{Y}_s$ and internal features $H_s \in \mathbb{R}^D$. The teacher produces predictions $\hat{Y}_t$ and internal features $H_t \in \mathbb{R}^D$. The objective of Global *KD* is:

$$\min_{\theta_s} \mathcal{L}_{sup}(Y, \hat{Y}_s) + \mathcal{L}_{KD}^{Y}(\hat{Y}_t, \hat{Y}_s) + \mathcal{L}_{KD}^{H}(H_t, H_s), \tag{1}$$

where θ_s is the parameter of the student; $\mathcal{L}_{sup}$ is the supervised loss (e.g., MAE for regression, cross-entropy for classification); $\mathcal{L}_{KD}^{Y}$ and $\mathcal{L}_{KD}^{H}$ are the distillation loss terms that encourage student model to learn knowledge from teacher on both **prediction level** [11] and **feature level** [29]. However, Global *KD* only matches the signal-level feature (i.e., $\mathcal{L}_{KD}^{H}$), making it less effective at preserving the local morphology within each PPG segment and the structural rhythm across segments (Figure 2).

2.2 PPG-DISTILL

In accordance with our intuition regarding preservation of local information of PPG signal, we propose a novel patch-level distillation framework, called PPG-DISTILL in Figure 3. Instead of focusing on matching global signal-level features, PPG-DISTILL focuses on distilling knowledge about local morphology and rhythm by patch-level morphology and rhythm distillation. We note that the term *morphology* here refers to data-driven local waveform representations within patches, rather than predefined or clinical morphological descriptors.

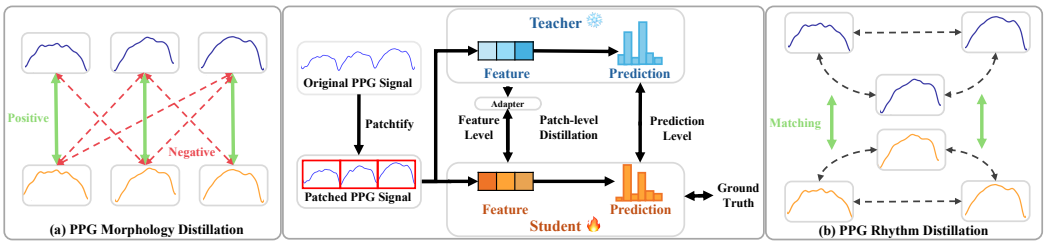

Figure 3: Overall framework of PPG-DISTILL.

Patchtify, for most PPG foundation models [25, 4], is the first step to process the original PPG signal X to non-overlapping patches [21, 16, 8]. Denote the patch length as P, then the patchifying process will generate a sequence of patches $X_p \in \mathbb{R}^{P\times N}$ where N is the number of patches, $N = L/P$.

PPG Morphology Distillation Let the student and teacher produce features for a PPG patch sequence X_p as $H_s^p \in \mathbb{R}^{N\times d_s}$ and $H_t^p \in \mathbb{R}^{N\times d_t}$. Because d_s and d_t can differ, we introduce a shared learnable linear adapter $A \in \mathbb{R}^{d_t\times d_s}$ and form $\tilde{H}_t^p = H_t^p A$. We then ℓ_2-normalize patch vectors row-wise, $\hat{H}_{s/t}^p = \mathrm{norm}(H_{s/t}^p)$. We align the i-th student patch to the i-th teacher patch and treat all other teacher patches as negatives. The similarity matrix is $Z = \frac{\hat{H}_s^p(\hat{H}_t^p)^\top}{\tau} \in \mathbb{R}^{N\times N}$, where τ is temperature. We use InfoNCE-style [22] loss with one positive per row:

$$\mathcal{L}_{mor} = \frac{1}{N}\sum_{i=1}^{N}\left(-\log\frac{\exp(Z_{ii})}{\sum_{j=1}^{N}\exp(Z_{ij})}\right).$$

This objective encourages one-to-one alignment of local morphology across patches, allowing the student to preserve the teacher's patch-level morphology feature.

PPG Rhythm Distillation To keep the PPG rhythm (beat-to-beat periodicity and timing regularity), we transfer the teacher's *inter-patch relations* to the student rather than only aligning individual patch features. We form pairwise Euclidean distance matrices with normalization $[D_t]_{ij} = \left\|\phi(H_{t,i}^p) - \phi(H_{t,j}^p)\right\|_2, [D_s]_{ij} = \left\|H_{s,i}^p - H_{s,j}^p\right\|_2$, The relational distillation loss matches these normalized structures with a smooth L1 penalty [23]:

$$\mathcal{L}_{rhy} = \frac{1}{N(N-1)}\sum_{i\neq j}\mathrm{smoothL1}\Big([\tilde{D}_s]_{ij},\ [\tilde{D}_t]_{ij}\Big). \tag{2}$$

This term penalizes discrepancies in relative inter-patch distances, thereby transferring the teacher's structural knowledge of rhythm to the student.

Joint Optimization While training PPG-DISTILL, we jointly optimize both the PPG morphology and rhythm distillation losses in addition to the Global *KD* losses. Therefore, the overall training loss that PPG-DISTILL adopts for the student is $\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{KD}^{Y} + \beta\mathcal{L}_{KD}^{H} + \gamma(\mathcal{L}_{mor} + \mathcal{L}_{rhy})$, where α, β, and γ are hyper-parameters which mediate the strengths of each loss term.

3 Experiment

Experimental Setting To evaluate the effectiveness of PPG-DISTILL, we benchmark it on both regression and classification tasks in PPG analysis, following GPT-PPG [4]. For regression, we use the DaLiA dataset [28], where the model is required to estimate patients' heart rates from PPG signals. For classification, we use the StanfordAF dataset [34], which targets atrial fibrillation (AF) detection. We adopt two PPG foundation models, GPT-PPG-19m [4] and PaPaGei [25], as teachers, and consider MLP as well as the lightweight GPT-PPG-1m variant of GPT-PPG as students. For regression, we report mean squared error (MSE) and mean absolute error (MAE) [20]. For classification, we report accuracy (Acc.) and F1 score. Further implementation details are provided in Appendix B.

Results Table 1 reports the effectiveness of the proposed PPG-DISTILL compared with Global *KD* on GPT-PPG-1m [4]. Since MLP does not patchify PPG signals, only Global *KD* can be applied to it.

Table 1: Performance comparison on DaLiA and StanfordAF. "+xx%" values indicate the relative improvement in student performance after distillation.

Teacher Models		**GPT-PPG-19m** [4]		**PaPaGei** [25]	
Metric		MSE (↓)	MAE (↓)	MSE (↓)	MAE (↓)
DaLiA	Teacher	221.78	8.82	160.39	6.81
	MLP	581.77	17.87	581.77	17.87
	+Global *KD*	230.59+60.36%	10.74+39.89%	575.40+1.10%	17.84+0.14%
	GPT-PPG-1m [4]	255.07	10.08	255.07	10.08
	+Global *KD*	234.16+8.20%	9.44+6.37%	220.26+13.65%	8.38+16.89%
	+PPG-DISTILL	**215.36+15.57%**	**8.34+17.32%**	**202.31+20.68%**	**7.90+21.62%**
Metric		Acc. (↑)	F1 (↑)	Acc. (↑)	F1 (↑)
StanfordAF	Teacher	0.93	0.88	0.83	0.70
	MLP	0.76	0.42	0.76	0.42
	+Global *KD*	0.76-0.09%	0.54+29.17%	0.73-4.31%	0.41-1.15%
	GPT-PPG-1m [4]	0.81	0.64	0.81	0.64
	+Global *KD*	0.82+0.80%	0.65+2.73%	0.83+1.83%	0.67+5.69%
	+PPG-DISTILL	**0.87+6.73%**	**0.77+21.80%**	**0.88+7.68%**	**0.77+21.35%**

Several key observations can be drawn from the results. **First**, PPG-DISTILL consistently improves the performance of GPT-PPG-1m across both regression (DaLiA) and classification (StanfordAF) tasks. In particular, PPG-DISTILL achieves up to a **+21.8% relative F1 improvement** on StanfordAF and a **+13.7% relative MSE improvement** on DaLiA, highlighting its strong and consistent gains across tasks. Notably, on the DaLiA dataset with GPT-PPG-19m as the teacher, GPT-PPG-1m trained with PPG-DISTILL even outperforms its teacher while using $19\times$ fewer parameters, demonstrating that structural *KD* can close, and even invert, the capacity gap between teacher and student. **Second**, MLP, even with Global *KD*, fails to surpass GPT-PPG-1m, highlighting the limitation of its shallow architecture in modeling complex PPG dynamics. **Third**, PPG-DISTILL consistently yields stronger performance than Global *KD* when applied to GPT-PPG-1m, confirming that PPG-DISTILL is more effective than Global *KD*, particularly in transferring fine-grained rhythm and morphological cues that are crucial for PPG signal analysis. **Fourth**, on the DaLiA dataset, stronger teachers (e.g., PaPaGei) generally lead to better students, suggesting that high-quality teacher representations provide richer relational structure for distillation. However, this trend does not hold for the StanfordAF dataset, where the performance gap between teachers is smaller, and dataset-specific factors likely play a larger role. We conduct an ablation study and hyperparameter sensitivity analysis in Appendix C.

Efficiency Analysis To further evaluate the efficiency of PPG-DISTILL, we compare throughput (measured in Batch/s) and model size (measured in number of parameters) across different models, as shown in Figure 4.The results highlight two points. First, foundation models such as GPT-PPG-19m and PaPaGei provide strong accuracy but suffer from low throughput and high memory cost, making them unsuitable for wearables. Second, GPT-PPG-1m distilled with PPG-DISTILL achieves the highest throughput with nearly $19\times$ fewer parameters, showing that compact students can retain strong performance while enabling efficient on-device inference. We provide detailed efficiency results in Appendix D.

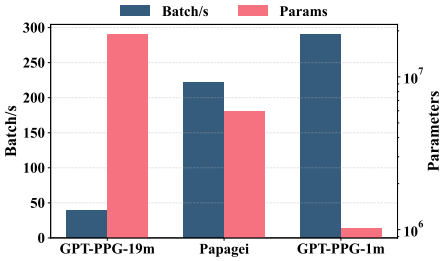

Figure 4: Inference throughput (Batch/s) and parameter size comparison across GPT-PPG-19m, PaPaGei, and GPT-PPG-1m.

4 Conclusion and Future Work

We proposed PPG-DISTILL, a distillation framework that combines prediction-, feature-, and patch-level strategies to transfer both global and local knowledge from large PPG foundation models to lightweight students. Experiments on heart rate estimation and atrial fibrillation detection show notable performance gains with much higher efficiency, enhancing the feasibility of real-world deployment of these models. Future work includes extending to more tasks and datasets, deeper analysis of the framework, and exploring diverse teacher models beyond foundation models.

References

[1] Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Haider Ali, Imran Khan Niazi, David White, Malik Naveed Akhter, and Samaneh Madanian. Comparison of machine learning models for predicting interstitial glucose using smart watch and food log. *Electronics*, 13(16):3192, 2024.

[3] David Campos, Miao Zhang, Bin Yang, Tung Kieu, Chenjuan Guo, and Christian S Jensen. Lightts: Lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1(2):1–27, 2023.

[4] Zhaoliang Chen, Cheng Ding, Saurabh Kataria, Runze Yan, Minxiao Wang, Randall Lee, and Xiao Hu. Gpt-ppg: a gpt-based foundation model for photoplethysmography signals. *Physiological Measurement*, 46(5):055004, 2025.

[5] Gari D Clifford, Ikaro Silva, Benjamin Moody, Qiao Li, Danesh Kella, Abdullah Shahin, Tristan Kooistra, Diane Perry, and Roger G Mark. The physionet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the icu. In *2015 Computing in Cardiology Conference (CinC)*, pages 273–276. IEEE, 2015.

[6] Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J Lee, Cynthia Rudin, and Xiao Hu. Siamquality: a convnet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, 2024.

[7] 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. In *Forty-second International Conference on Machine Learning*, 2025.

[8] Tianning Feng, Juntong Ni, Ezequiel Gleichgerrcht, and Wei Jin. Seizureformer: A transformer model for iea-based seizure risk forecasting. *arXiv preprint arXiv:2504.16098*, 2025.

[9] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

[10] Jiayu He, Jianlin Ou, An He, Lin Shu, Tao Liu, Ruowen Qu, Xiangmin Xu, Zhuoming Chen, and Yifeng Yan. A new approach for daily life blood-pressure estimation using smart watch. *Biomedical Signal Processing and Control*, 75:103616, 2022.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

[13] Spyridon Kontaxis, Eduardo Gil, Vaidotas Marozas, Jesus Lazaro, Esther Garcia, Mar Posadas-de Miguel, Sara Siddi, Maria Luisa Bernal, Jordi Aguilo, Josep Maria Haro, et al. Photoplethysmographic waveform analysis for autonomic reactivity assessment in depression. *IEEE Transactions on Biomedical Engineering*, 68(4):1273–1281, 2020.

[14] Remo Lazazzera, Margot Deviaene, Carolina Varon, Bertien Buyse, Dries Testelmans, Pablo Laguna, Eduardo Gil, and Guy Carrault. Detection and classification of sleep apnea and hypopnea using ppg and spo _2 signals. *IEEE Transactions on Biomedical Engineering*, 68(5):1496–1506, 2020.

[15] Simon A Lee and Kai Akamatsu. Foundation models for physiological signals: Opportunities and challenges. 2025.

[16] Zewen Liu, Juntong Ni, Max SY Lau, and Wei Jin. Cape: Covariate-adjusted pre-training for epidemic time series forecasting. *arXiv preprint arXiv:2502.03393*, 2025.

[17] Zewen Liu, Juntong Ni, Xianfeng Tang, Max SY Lau, and Wei Jin. Can large language models adequately perform symbolic reasoning over time series? *arXiv preprint arXiv:2508.03963*, 2025.

[18] Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. Graph odes and beyond: A comprehensive survey on integrating differential equations with graph neural networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6118–6128, 2025.

[19] Juntong Ni, Zewen Liu, Shiyu Wang, Ming Jin, and Wei Jin. Timedistill: Efficient long-term time series forecasting with mlp via cross-architecture distillation. *arXiv preprint arXiv:2502.15016*, 2025.

[20] Juntong Ni, Shiyu Wang, Zewen Liu, Xiaoming Shi, Xinyue Zhong, Zhou Ye, and Wei Jin. Are we overlooking the dimensions? learning latent hierarchical channel structure for high-dimensional time series forecasting. *arXiv preprint arXiv:2507.15119*, 2025.

[21] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

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

[23] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[25] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open Foundation Models for Optical Physiological Signals. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, Singapore, April 2025. Accepted. arXiv preprint arXiv:2410.20542.

[26] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2016.

[27] Ming-Zher Poh, Yukkee Cheung Poh, Pak-Hei Chan, Chun-Ka Wong, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Daniel Wai-Sing Chu, and Chung-Wah Siu. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart*, 104(23):1921–1928, 2018.

[28] 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.

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

[30] Mithun Saha, Maxwell A Xu, Wanting Mao, Sameer Neupane, James M Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *arXiv preprint arXiv:2502.01108*, 2025.

[31] Fatemeh Sarhaddi, Kianoosh Kazemi, Iman Azimi, Rui Cao, Hannakaisa Niela-Vilén, Anna Axelin, Pasi Liljeberg, and Amir M Rahmani. A comprehensive accuracy assessment of samsung smartwatch heart rate and heart rate variability. *PloS one*, 17(12):e0268361, 2022.

[32] 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.

[33] Ping Shi. *Photoplethysmography in noninvasive cardiovascular assessment*. PhD thesis, Loughborough University, 2009.

[34] Jessica Torres-Soto and Euan A Ashley. Multi-task deep learning for cardiac rhythm detection in wearable devices. *NPJ digital medicine*, 3(1):116, 2020.

[35] Jie Wang, Tuantuan Lu, Ruogu Huang, and Yongxiang Zhao. Classifying engagement in e-learning through gru-tcn model using photoplethysmography signals. *Biomedical Signal Processing and Control*, 90:105903, 2024.

[36] Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.

[37] Qing Xu, Zhenghua Chen, Mohamed Ragab, Chao Wang, Min Wu, and Xiaoli Li. Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks. *Neurocomputing*, 485:242–251, 2022.

[38] Yuhao Xu, Xiaoda Wang, Yi Wu, Wei Jin, Xiao Hu, and Carl Yang. Ecg-moe: Mixture-of-expert electrocardiogram foundation model. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*.

[39] 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.

[40] Zexing Zhang, Huimin Lu, Songzhe Ma, Jianzhong Peng, Chenglin Lin, Niya Li, and Bingwang Dong. A general framework for generative self-supervised learning in non-invasive estimation of physiological parameters using photoplethysmography. *Biomedical Signal Processing and Control*, 98:106788, 2024.

[41] Lili Zhu, Petros Spachos, Pai Chet Ng, Yuanhao Yu, Yang Wang, Konstantinos Plataniotis, and Dimitrios Hatzinakos. Stress detection through wrist-based electrodermal activity monitoring and machine learning. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2155–2165, 2023.

A Related Work

A.1 PPG Signal Analysis

PPG has been used to estimate key physiological metrics, including heart rate [28, 32], heart rate variability [31], blood glucose [2], respiration rate [26], and blood pressure [36, 10]. Beyond general monitoring, PPG contributes to diagnostic applications by supporting the detection of cardiovascular conditions such as atrial fibrillation [34, 27], reducing false arrhythmia alarms [5], and identifying hypoxia [14]. In addition, it is increasingly applied in mental health and wellness contexts, where it enables tracking of stress [41], emotion [13], and cognitive states such as focus [35].

A.2 Foundation Model for PPG Signal

A foundation model is a large pre-trained model that learns general representations transferable to many downstream tasks [17]. Recent advances in foundation models for PPG signals can be categorized by their pre-training data sources. **Clinical or lab PPG-based models** include *PaPaGei* [25], which leverages morphology-aware contrastive learning on 57,000 hours of clinical PPG and provides open-source weights, *SiamQuality* [6], which enforces robustness to signal quality variations using over 36 million clinical PPG pairs, and *GPT-PPG* [4], which adapts generative transformers to ICU-collected PPG and demonstrates both predictive and denoising capabilities. In addition, *REGLE* [39] employs autoencoders to extract disentangled embeddings from biobank-scale clinical PPG for genomic discovery and disease risk prediction, while *TS2TC* [40] introduces a generative self-supervised framework trained on the VitalDB dataset of surgical patients, aiming at physiological parameter estimation. **Field PPG-based models** directly address wearable applicability: *Apple-PPG* [1] is trained on data from more than 140K Apple Watch users and achieves strong generalization, though it remains closed-source, while *Pulse-PPG* [30] represents the first open-source foundation model trained exclusively on large-scale wearable field PPG, showing improved robustness to motion noise and free-living conditions.

A.3 Knowledge Distillation

Knowledge distillation (*KD*) [11] transfers knowledge from a larger, more complex model (teacher) to a smaller, simpler model (student) while maintaining comparable performance. By aligning the output distributions of teacher and student models, *KD* provides richer training signals than hard labels alone, enabling the student to capture subtle patterns that the teacher has learned. In the context of time series signal, CAKD [37] uses adversarial and contrastive learning for feature distillation without a specific design for time series, while LightTS [3] designs a *KD* framework for ensemble classifiers, limiting its generality. Unlike these, TimeDistill [19] targets time series-specific patterns, such as multi-scale and multi-period, pioneering cross-architecture *KD* for time series analysis. To the best of our knowledge, we are the first attempt to apply the *KD* technique to the PPG signal.

B Implementation Details

All experiments are implemented in PyTorch [24] and conducted on one NVIDIA L40S GPU. The teacher models are trained using their default configurations as reported in their respective papers. When using PPG-DISTILL for distillation, the teacher model is frozen, and only the student is trained. Following GPT-PPG [4], we set the patch size to 40. We use Adam [12] for optimization. The initial learning rate is set by lr_init=1e-5, and further adjustments are handled by the scheduler. A warmup and cosine annealing strategy is applied at the batch level with lr_max=1e-3, eta_min=1e-6, warm up ratio=25%. We apply early stopping with a patience value of 20 epochs. The batch size is set to 64. The temperature τ for patch-level contrastive distillation is set to $\tau = 2$. We perform a hyperparameter search for α, β and γ within the range $\{0.1, 0.5\}$.

C Ablation study and Hyperparameter sensitivity

We varied α, β, and γ in the joint objective $\mathcal{L}$ in Section 2.2 to examine the effect of each loss term. As shown in Figure 5, α strongly influences performance: small values improve learning while large values degrade it. β remains stable across settings, indicating feature-level distillation is less sensitive.

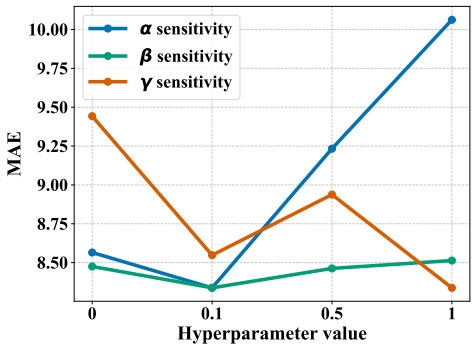

Figure 5: Effect of hyperparameters (α, β, γ) on MAE for the DaLia dataset (Teacher: GPT-PPG-19m, Student: GPT-PPG-1m).

γ shows a non-monotonic trend, with $\gamma = 1$ achieving the best MAE, confirming the importance of patch-level objectives for capturing morphology and rhythm.

D Full Results of Efficiency

Table 2: Comparison on DaLiA dataset.

DaLiA	GPT-PPG-19m	Papagei	MLP	GPT-PPG-1m
MAE	8.82	6.81	10.74	7.90
Batch/s	128.06	225.80	4248.70	291.50
Params	19,018,417	5,917,197	41,473	1,017,197
Memory (MB)	72.6	22.6	0.16	3.9

Table 3: Comparison on StanfordAF dataset.

StanfordAF	GPT-PPG-19m	Papagei	MLP	GPT-PPG-1m
F1	0.88	0.70	0.54	0.77
Batch/s	39.19	222.30	1546.70	290.00
Params	19,034,290	5,917,454	154,242	1,021,690
Memory (MB)	72.7	22.6	0.59	3.9

Tables 2 and 3 compare accuracy, inference throughput, and parameter efficiency across different models on the DaLiA and StanfordAF datasets. Several observations can be made. **First**, large foundation models such as GPT-PPG-19m achieve strong accuracy (MAE of 8.82 on DaLiA, F1 of 0.88 on StanfordAF) but come with high computational cost, processing fewer than 130 batches/s on DaLiA and fewer than 40 batches/s on StanfordAF. **Second**, PaPaGei provides a favorable trade-off, reducing parameters by about $3\times$ while maintaining competitive accuracy and substantially increasing throughput. **Third**, MLP achieves extremely high throughput (over 4000 batches/s on DaLiA), but its limited capacity results in a clear accuracy drop (MAE 10.74 on DaLiA, F1 0.54 on StanfordAF). **Finally**, GPT-PPG-1m, when distilled with PPG-DISTILL, offers the best balance: it achieves accuracy close to or surpassing its teachers with only around 1M parameters, while running an order of magnitude faster than GPT-PPG-19m. These results highlight that PPG-DISTILL enables lightweight models to approach the accuracy of large PPG foundation models while retaining significantly higher efficiency.