HiMAE: Hierarchical Masked Autoencoders Discover Resolution-Specific Structure in Wearable Time Series

Anonymous authors
Paper under double-blind review

Abstract

Wearable sensors provide abundant physiological time series, yet the principles governing their predictive utility remain unclear. We hypothesize that temporal resolution is a fundamental axis of representation learning, with different clinical and behavioral outcomes relying on structure at distinct scales. To test this *resolution hypothesis*, we introduce *HiMAE* (Hierarchical Masked Autoencoder), a self-supervised framework that combines masked autoencoding with a hierarchical convolutional encoder–decoder. HiMAE produces multi-resolution embeddings that enable systematic evaluation of which temporal scales carry predictive signal, transforming resolution from a hyperparameter into a probe for interpretability. Across classification, regression, and generative benchmarks, HiMAE consistently outperforms state-of-the-art foundation models that collapse scale, while being orders of magnitude smaller. HiMAE is an efficient representation learner compact enough to run entirely on-watch, achieving sub-millisecond inference on smartwatch-class CPUs for true edge inference. Together, these contributions position HiMAE as both an efficient self supervised learning method and a discovery tool for scale-sensitive structure in wearable health.

1 Introduction

Wearable sensors have emerged as a primary modality for continuous health monitoring, providing access to rich physiological and behavioral signals in free-living settings (Erturk et al., 2025). Despite their ubiquity, the utility of wearable signals for machine learning in healthcare remains poorly understood. Unlike images (Dosovitskiy et al., 2021; Simonyan et al., 2014; Zhou et al., 2015; Petsiuk et al., 2018) or text (Brown et al., 2020; Li et al., 2016; Sundararajan et al., 2017; Arras et al., 2017), physiological time series rarely admit obvious visual cues that map cleanly to clinical outcomes, leaving open fundamental questions about which features carry predictive value. A particularly unresolved issue concerns temporal resolution: should models operate at a single universal resolution, or do different health outcomes depend on resolution-specific structure? Clinically actionable events can arise on second-level timescales, requiring representations that both capture fine-grained temporal patterns and support real-time inference under the computational constraints of wearable devices. We hypothesize that resolution is not a nuisance parameter but a fundamental axis of physiological representation learning. We refer to this as the *resolution hypothesis*, which posits that temporal granularity governs predictive performance in clinical and behavioral tasks. In this framing, "resolution" denotes the effective temporal context over which representations are formed—from fine-scale waveform morphology to coarse-scale dynamics spanning the whole sequence.

From an algorithmic perspective, much of the field defaults to transformer-based architectures (Vaswani et al., 2017), implicitly assuming that flexibility and capacity outweigh inductive bias. Yet wearable signals, while long in sequence length, are often generated by a few latent processes driven by biological mechanisms and captured through only a handful of sensor modalities. In this sense they are low-dimensional and highly structured. This raises the possibility that transformers may not only overfit but also obscure resolution-specific structure, rather than expose it. By contrast, hierarchical convolutional biases offer a natural mechanism for aligning architectures with the resolution hypothesis, capturing both local detail and long-range dependencies in a principled

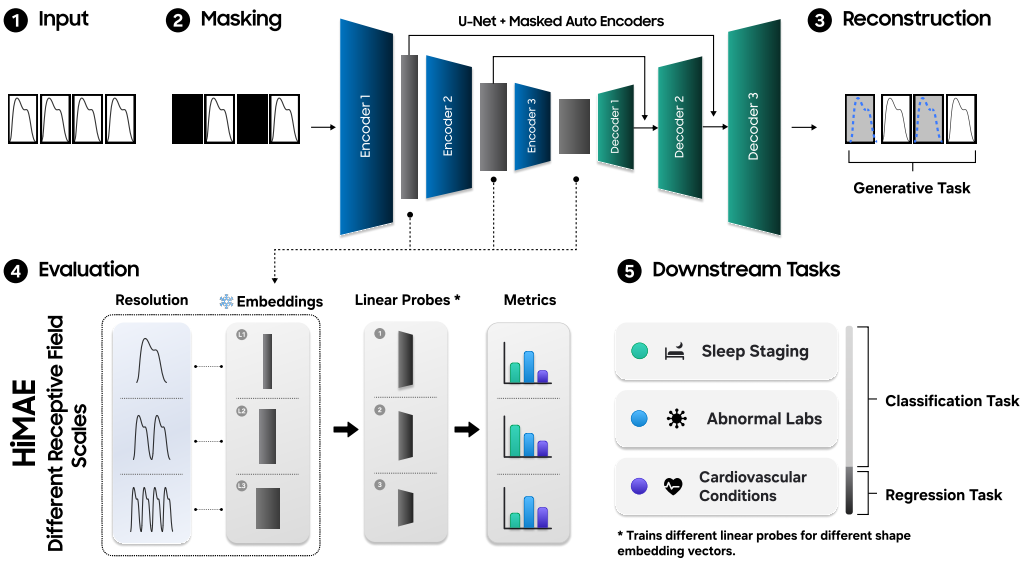


Figure 1: **HiMAE pre-training and evaluation pipeline.** (1) Physiological sequences are split into temporal patches. (2) Selected patches are masked randomly or contiguously. (3) A U-Net–style CNN encoder–decoder reconstructs missing values, with loss applied only to masked regions. (4) Multi-resolution embeddings feed linear probes for classification and regression benchmarking. (5) Three categorized task-lists are evaluated.

way. This motivates a re-examination of architectural design choices for self-supervised learning (SSL) on physiological time series.

In this work, we address these challenges by introducing *HiMAE* (Hierarchical Masked Autoencoder), a self-supervised pretraining framework for wearable time series that directly operationalizes the resolution hypothesis (Figure 1). HiMAE combines the masked autoencoding paradigm with 1D physiological signals by coupling patch-masking objectives (Wang et al., 2023) with a U-Net–inspired encoder–decoder (Ronneberger et al., 2015). Crucially, HiMAE produces multi-resolution embeddings, with each level of the hierarchy corresponding to a distinct temporal granularity. This design enables systematic interrogation of which resolutions carry predictive signal, while simultaneously yielding lightweight, efficient representations. Beyond its architectural advantages, HiMAE allows us to benchmark the resolution hypothesis across 14 classification and regression tasks. Our results reveal resolution-specific structure in wearable signals that is not readily identifiable by human experts, offering new insights into both representation learning and the interpretability of physiological time series in the time domain.

2 Related Work

Self-Supervised Pretraining Objectives for Wearable Signals Wearable devices equipped with photoplethysmography (PPG), electrocardiography (ECG), and accelerometry generate long, multi-channel time series encoding diverse physiological and behavioral phenomena, including cardiovascular dynamics (Castaneda et al., 2018), activity patterns (Yuan et al., 2024; Xu et al., 2025), sleep cycles (Li et al., 2021; Thapa et al., 2024; Logacjov et al., 2025), and other latent processes. These data streams are abundant, and predominantly unlabeled, making them well suited for large-scale self-supervised learning (Kaplan et al., 2020; Bommasani et al., 2021; Zhou et al., 2024; Liang et al., 2024).

SSL has become the dominant paradigm for wearable time-series representation learning, given the scarcity of labeled data and the ubiquity of unlabeled signals in free-living settings (Lee & Akamatsu, 2025). Among SSL strategies, masked autoencoding has emerged as a central approach, inspired by its success in vision (He et al., 2022; Vaid et al., 2023) and language modeling (De-

vlin et al., 2019). The method randomly occludes patches of the signal and tasks the model with reconstructing them, encouraging representations that capture latent physiological structure and temporal regularities (Zhang et al., 2022a; Kong et al., 2023). Recent large-scale efforts, most notably Google's LSM series (Narayanswamy et al., 2024; Xu et al., 2025), rely heavily on masked autoencoding, establishing it as a pretraining standard for multi-modal wearable datasets. Yet despite its effectiveness for local pattern recovery, vanilla masked autoencoding often struggles to capture multi-resolution features unless coupled with explicitly hierarchical architectures.

In parallel, contrastive learning enforces invariance by pulling semantically similar samples together in latent space while pushing dissimilar ones apart (Schmitt & Kuljanin, 2008; Jaiswal et al., 2020). The central challenge for wearables is defining positive and negative pairs without labels. One solution is participant-level contrastive training, where samples from the same individual are positives and samples from different individuals are negatives, an approach adopted in Apple's ECG and PPG foundation models (Abbaspourazad et al., 2023) and closely related to the SimCLR framework (Chen et al., 2020b). Other domain-specific innovations define pairs through physiological priors: PaPaGei leverages PPG morphology (Pillai et al., 2024), while SleepFM extends the paradigm across EEG, ECG, and EMG to enforce cross-modal consistency (Thapa et al., 2024). Additional embedding-level regularizers, such as differential entropy constraints (Jing et al., 2021; Abbaspourazad et al., 2023), further enrich learned representations. However, contrastive methods are highly sensitive to augmentation heuristics (which are rarely physilogically meaningful), computationally intensive, and limited in interpretability, providing little insight into which temporal structures are preserved.

HiMAE departs from both flat masked and contrastive approaches in two ways. First, instead of relying on a single-scale reconstruction or augmentation heuristics, HiMAE couples masked autoencoding with a hierarchical encoder–decoder that integrates information across resolutions, treating temporal scale as an explicit dimension of representation. Second, by extracting embeddings at multiple scales and probing them independently, HiMAE transforms SSL from a pretraining mechanism into a discovery tool: it directly tests which temporal resolutions carry predictive signal for downstream tasks. In doing so, HiMAE preserves the efficiency of masked autoencoding while introducing interpretability absent in contrastive or flat masked objectives.

Multi-scale Learning The emphasis on resolution awareness connects naturally to multi-scale learning, where modeling temporal signals across multiple granularities has emerged as a powerful inductive bias. In vision, multi-scale architectures such as pyramidal CNNs and hierarchical attention enable models to integrate fine-scale edges with coarse semantic structures, substantially improving recognition and generation in 2D (Wang et al., 2016; Yang et al., 2016; Liu et al., 2021a; Kusupati et al., 2024; Liu et al., 2024) and 3D (He et al., 2017; Ghadai et al., 2019; Zhang et al., 2022b).

In time series, multi-scale methods are fewer but increasingly influential. N-HiTS (Challu et al., 2022) improves long-horizon forecasting by allocating capacity across frequencies via hierarchical interpolation. Pyraformer (Liu et al., 2022) leverages pyramidal attention to capture dependencies over a tree of scales, while Scaleformer (Shabani et al., 2023) introduces iterative refinement across resolutions. Pathformer (Chen et al., 2024) further adapts pathways dynamically to match input-specific temporal dynamics. Together, these approaches highlight that temporal signals are inherently hierarchical and that resolution carries predictive structure rather than being a nuisance variable.

Prior multi-scale methods typically rely on fixed hierarchies or task-specific refinement stages (e.g., for forecasting), which constrains their generality. While HiMAE also inherits inductive biases from convolutional design choices (e.g., step size, padding, kernel width), these parameters define receptive fields rather than dictate which scales are salient. By coupling self-supervised reconstruction with these fields, HiMAE induces a hierarchy of temporal embeddings that can be probed independently.

3 METHODS

Hierarchical Masked Autoencoders (HiMAE) HiMAE combines masked autoencoding (Baldi, 2012; He et al., 2022) with 1-D physiological time series by coupling a patch-masking objective with

a U-Net–style convolutional encoder–decoder (Ronneberger et al., 2015). Given an input sequence $x \in \mathbb{R}^{C \times L}$, we partition it into $N = L/P$ non-overlapping patches of length P. A binary mask $m \in 0, 1^N$ is sampled from a Bernoulli distribution with parameter r, indicating the masking ratio. Masked indices are selected uniformly at random without replacement, expanded to match temporal resolution as $m' \in 0, 1^L$, and applied to the sequence, yielding $\tilde{x} = x \odot (1 - m')$. This masking procedure removes substantial context, forcing the model to infer higher-order dependencies. In addition to random masking, we also employ contiguous masking, in which adjacent patches are removed to mimic sensor dropout similar to recent protocols showing benefits (Xu et al., 2025). Both regimes are interleaved during pretraining to promote robustness across reconstruction settings.

The encoder f_θ is a hierarchical 1D CNN composed of residual convolutional blocks with stride-2 convolutions that downsample the temporal resolution by half at each stage, expanding the receptive field so that deeper layers capture long-range dependencies while shallow layers retain local detail. Each residual block consists of two convolutions with kernel size 5, batch normalization (Ioffe & Szegedy, 2015), and GELU activations (Hendrycks & Gimpel, 2023), along with a projection shortcut when input and output dimensions differ. The decoder g_ϕ mirrors this structure with transposed convolutions for upsampling and incorporates skip connections from encoder layers, concatenating intermediate features to restore fine-grained temporal structure. All convolutions are standard 1D operations defined over temporal windows, and striding handles subsampling directly. Intermediate activations use GELU, while the final layer applies a $\tanh$ nonlinearity so that outputs $\hat{x} \in \mathbb{R}^{C \times L}$ are bounded in $[-1, 1]$, matching the normalized input range.

We deliberately adopt a convolutional U-Net backbone rather than a transformer-based encoder for two reasons. First, physiological signals exhibit strong local dependencies governed by morphology (e.g., PPG waveform shape, ECG peaks), which are naturally modeled by finite receptive fields. Convolutions (O'Shea & Nash, 2015) encode this locality directly, whereas transformers must simulate it through restricted attention, often at higher parameter cost (Appendix H.2). Second, multi-resolution structure is intrinsic to physiology (e.g., heartbeats unfold over milliseconds, rhythms span seconds). A hierarchical CNN with skip connections provides an architectural bias toward such nested timescales, aligning directly with the resolution hypothesis and being orders of magnitude smaller than other proposed foundation models in this space (See Figure 2 for comparison). In contrast, transformers emphasize global mixing, which may obscure resolution-specific structure while consuming substantially more compute (Table 7). This rationale motivates HiMAE's design as not only efficient but also inductively aligned with the temporal statistics of wearable signals.

Multi-resolution embeddings extracted from different levels of the hierarchy are probed independently, with distinct linear classifiers trained per resolution (Alain & Bengio, 2018). This design enables us to systematically evaluate which temporal granularity carries predictive signal for downstream tasks, rather than collapsing embeddings into a single latent space. Finally, choices of patch length P and kernel size were guided by ablations (Appendix Section F.1), which confirmed that $P = 5$ and kernel size 5 yield the best balance between local fidelity and receptive field expansion when all other hyperparameters were fixed.

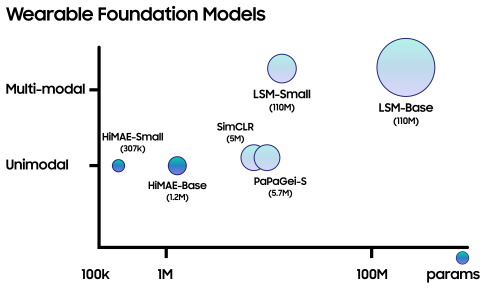


Figure 2: **HiMAE is lightweight**

Training minimizes a masked reconstruction loss restricted to occluded regions: $\mathcal{L}_{\text{MSE}}(\theta, \phi) = \frac{\|(\hat{x}-x)\odot m'\|_2^2}{\sum_{t=1}^{L} m'_t}$, where m' ensures that gradients are only computed on masked segments. This objective estimates $p(x_{\mathcal{M}}|x_{\mathcal{O}})$, with $\mathcal{M}$ and $\mathcal{O}$ denoting masked and observed indices, preventing trivial copying of visible inputs and promoting temporally coherent, multi-scale representations.

Pretraining and Evaluation Protocol PPG Sequences were sampled at $f_s = 100$ Hz over fixed windows of $T = 10$s ($L = 1000$ timesteps). 10 second windows were selected due to clinically actionable events occurring in these time scales (ECG is collected at 10s intervals in clinical settings (Shuai et al., 2016; Elgendi, 2012)) and due to our interest in real-time monitoring on edge devices. Each signal was divided into non-overlapping patches of length $P = 5$ (200 patches total), and a masking ratio $r = 0.8$ was applied with patterns resampled per sequence and iteration to mitigate

overfitting (we empirically tested this masking ratio in Appendix Section F.1 with similar observations made in (Narayanswamy et al., 2024)). The encoder architecture employed channel widths $[16, 32, 64, 128]$, mirrored in the decoder. Optimization was performed with AdamW (Loshchilov & Hutter, 2019) (lr $= 10^{-3}$, weight decay $= 10^{-3}$) using a warmup–cosine schedule (10% linear warmup steps followed by cosine decay). Models trained up to 100k steps with batch size 2048 and early stopping triggered after 3 epochs without improvement similar to the protocols found in (Narayanswamy et al.). Data splits followed a 90/10 (train/validation) protocol across subjects, ensuring no identity overlap between pretraining and validation. Pretraining converged within 12 hours when distributing training across 4 Tesla T4 GPUs using PyTorch lightning (Paszke et al., 2019).

Pretraining datasets. We construct our pretraining corpus from approximately $80{,}000$ hours of wearable green PPG signals, drawn from seven large-scale free world studies conducted at REDACTED. These datasets include recordings from $47{,}644$ participants across seven distinct wearable devices, capturing broad demographic, behavioral, and hardware variability in a noisy environment (See Appendix Section B for ethics considerations). Although our modeling framework is modality-agnostic and can extend to other physiological signals such as electrocardiograms (see Appendix F.2), we focus here on PPG due to its prevalence and the scale of available data (we lack the same order of magnitude of ECG compared to PPG because ECG is not passively collected). To ensure reliability, we apply a standardized preprocessing pipeline that retains only high-quality segments, filtering by a Signal Quality Index (SQI). The retained signals are further refined using a bandpass filter of 0.5–8 Hz (Christiano & Fitzgerald, 2003), consistent across all pretraining and evaluation studies, to isolate physiologically relevant dynamics. Finally, signals are normalized to the range $[-1, 1]$ to match the output range of the $\tanh$ activation function used in our models.

4 Experimental Design

We follow the evaluation protocol of Narayanswamy et al. (2024) and extend it into a unified benchmark suite spanning generative, classification (and regression tasks in Appendix F.6), along with ablations to quantify how key architectural components interact with scaling. Across all experiments, our goal is not only to assess HiMAE's efficiency and transferability, but also to test the *resolution hypothesis*: whether predictive signal concentrates at specific levels of the hierarchical embeddings. Further analysis and results are displayed in full in Appendix Section F.

Model scaling and generative reconstruction. We first study HiMAE's scaling properties by measuring how reconstruction performance varies as a function of dataset size, number of participants, model capacity, and training compute capacity (batch size). For each axis, we systematically subsample or expand the relevant resource while holding others fixed, enabling us to isolate its contribution to representation quality. Scaling is assessed through mean squared error on masked reconstruction, which provides a direct measure of how model capacity and data availability govern loss reduction. We also squeeze in ablations in this experiment to assess how removing skip connections, and removing the hierarchal design affect scaling.

To complement this aggregate view, we also evaluate generative performance under three increasingly challenging reconstruction regimes defined in the LSM papers (Narayanswamy et al.; Xu et al., 2025): (i) random imputation, where patches are masked at random uniformly; (ii) temporal interpolation, where contiguous spans are removed to simulate sensor dropout; and (iii) temporal extrapolation, where future spans are occluded and predictions must rely solely on past context. We compute the mean squared error (MSE) for these evaluations.

Classification To assess downstream transferability and adaptability, we benchmark HiMAE on 12 binary classification tasks drawn from labeled datasets fully disjoint from our pretraining sources. We organize these into three groups: cardiovascular outcomes, sleep staging, and abnormal laboratory prediction. Cardiovascular outcomes, provide the most established benchmarks, with well-documented links between PPG and clinical endpoints (Shabaan et al., 2020). These include hypertension detection, estimating blood pressure (blood pressure regression pushed to Appendix 15 due to poor performance across all models), and arrhythmia-related events such as Premature Ventricular Contractions (PVCs), typically identified via electrocardiograms (ECGs). Sleep staging is another task we include which is of high interest, given the demand for wearables to track fine-grained sleep states despite the temporal and physiological complexity of the task (Imtiaz, 2021; Thapa et al.,

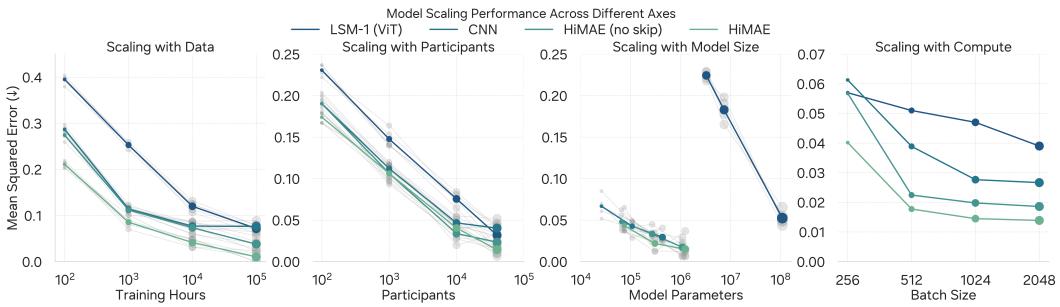


Figure 3: **HiMAE exhibits superior scaling across axes.** Mean squared error decreases most rapidly for HiMAE as data, participants, model size, and compute scale. Ablations without skip connections confirm that both the hierarchical design and skip pathways are helpful for generative pefromance. Grey lines indicate multiple runs whereas colored lines are average performance.

2024; Birrer et al., 2024). Laboratory predictions, on the other hand, serves as a discovery setting, testing whether PPG contains sufficient biomarker information to separate abnormal from healthy labs—an open question compared to patient-record benchmarks where such signals are more explicit (Kolo et al., 2024; McDermott et al., 2025). Together, these canonical and exploratory tasks form a spectrum that enables a comprehensive evaluation of representation quality across diverse digital health applications. All tasks are described in greater detail in Appendix Section D.

We compare HiMAE against state-of-the-art SSL methods adapted to the 1D setting for architectural comparability (More details on baselines in Appendix Section E). Specifically, we include SimCLR (Chen et al., 2020b), DINO (Caron et al., 2021), Masked Siamese Networks (MSN) (Assran et al., 2022), and a hierarchal Swin-Transformer (Liu et al., 2021b) as self-supervised baselines, along with the Large Signal Model (LSM) (Narayanswamy et al., 2024) and PaPaGei (Pillai et al., 2024) as established wearable foundation models. All models are evaluated under standard linear probing, in which the encoder is frozen and a linear classifier is trained on the resulting representations to measure AUROC as the main metric to measure discriminative abilities. For all architectures we use the full sequence embedding across the temporal dimension, without collapsing to a single summary token, to ensure that downstream probes have access to resolution-specific information. This setup allows us to test whether pretraining yields representations that are simultaneously transferable across tasks.

Resolution Hypothesis HiMAE produces embeddings at multiple temporal scales, and we probe each scale independently with linear classifiers. This allows us to test whether predictive information is concentrated at fine, intermediate, or coarse resolutions depending on the clinical endpoint. In this way, the classification tasks serve not only as benchmarks for transfer learning, but also as controlled tests of the resolution hypothesis (Receptive field lengths are described in Section C.1).

5 Results

5.1 Scaling and Generative Benchmark

Scaling: We first examine the scaling behavior in Figure 3 of HiMAE relative to baselines across data, participants, model parameters, and compute capacity (batch size). The overall scaling trends follow conventional expectations, error decreases monotonically with additional data, participants, or compute. However, scaling with model parameters reveals a interesting insight. HiMAE achieves substantially lower loss at smaller parameter capacities, while LSMs only begin to close the gap once scaled to orders of magnitude more parameters (we chose LSM parameter count based on their original paper (Narayanswamy et al., 2024)). This difference reflects an inductive bias. Transformer-based LSMs, which assume global receptive fields, appear to require considerably larger model capacity before capturing the local dynamics of the data (Further Mathematical Intuition is described in Appendix Section H). In contrast, HiMAE's hierarchical convolutional structure exploits spatial and temporal locality efficiently, yielding superior performance at modest scales. This observation reinforces the importance of architectural priors in low-capacity regimes.

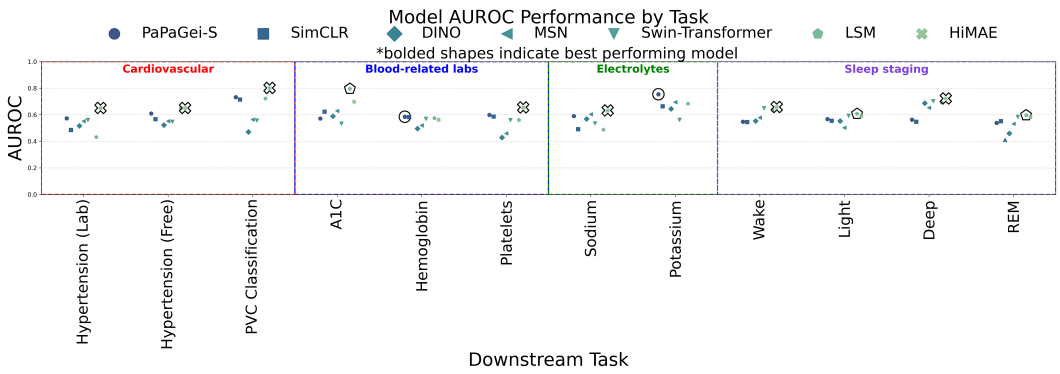


Figure 5: **AUROC across downstream tasks.** Highlighted shapes indicate best performing model. HiMAE consistently matches or outperforms foundation model baselines with far fewer parameters.

Generative: Turning to generative benchmarks, HiMAE consistently outperforms all baselines across random imputation, temporal interpolation, and temporal extrapolation tasks (Figure 4). In terms of mean squared error, HiMAE achieves the lowest reconstruction error in every setting, including cases with heavy missingness. This advantage persists when evaluated with R^2, where the mean-fill baseline serves as the reference. By achieving positive R^2 scores even in challenging extrapolation scenarios, HiMAE demonstrates reconstruction ability beyond naive heuristics (e.g., mean fill, nearest neighbor, or linear interpolation). Together, these results establish HiMAE as a strong generative model for missing data problems, with advantages that persist across scaling regimes and input corruption patterns.

Ablations: Ablation in Figure 3 and 4 further highlights the contributions of hierarchical design and skip connections in HiMAE. Removing either component results in increased error, indicating that both are crucial for effective representation learning. Nevertheless, even without these architectural elements, HiMAE variants remain competitive with larger LSM model, underscoring the robustness of the approach. More importantly, the full model exhibits improved generalization across scaling axes (Appendix Section F.4), suggesting that the combination of hierarchy and skip connections facilitates better transfer as data and compute grow.

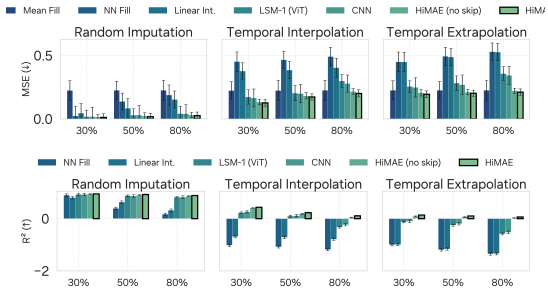


Figure 4: **Performance on generative benchmarks.** Mean squared error and R^2 for random imputation, temporal interpolation, and temporal extrapolation at varying missingness levels. Bold outline indicates best performing model.

5.2 Classification Benchmarking

Classification In Figure 5, HiMAE consistently secures the majority of wins, frequently outperforming or matching models that are considerably larger. This is particularly striking given that prior work has typically relied on heavy architectures to reach similar levels of performance, highlighting HiMAE's ability to capture a broad spectrum of physiological features with a compact design. These outcomes emphasize the model's robustness when applied to structured, temporally dependent problems that demand sensitivity to subtle variations in wearable signals.

Taken together, these results position HiMAE as the most consistently strong performer across the benchmark suite. In cases where HiMAE does not place first it is only $\sim$1-2% behind the winning model. Crucially, this level of performance is achieved with a substantially smaller model than competing approaches, demonstrating a favorable tradeoff between efficiency and predictive power. Rather than excelling only in isolated cases, HiMAE delivers broad, cross-domain competitiveness, suggesting that compact models, when designed with the right inductive biases, can rival or even surpass far larger architectures.

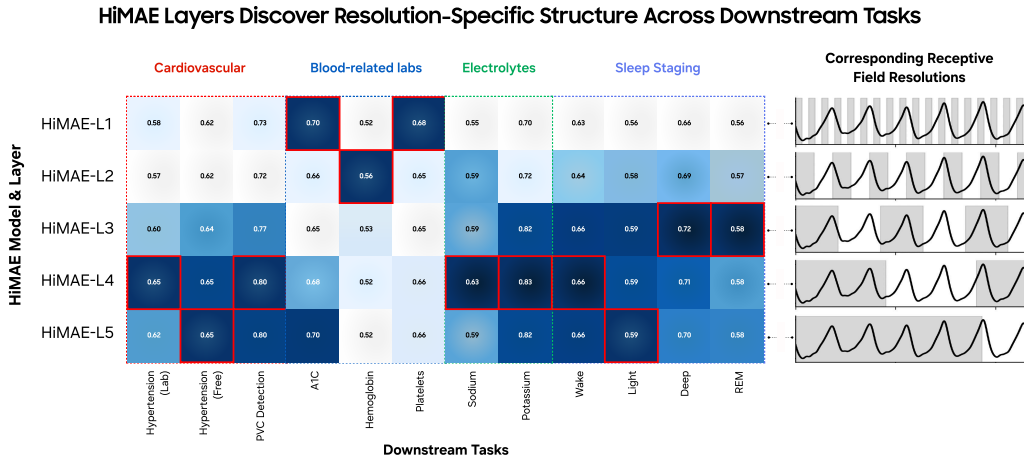
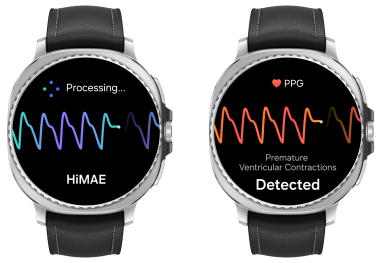


Figure 6: **HiMAE discovers task-specific structures for downstream tasks.** AUROC across layers shows that tasks rely on distinct temporal scales, highlighting HiMAE as a tool for discovering the most informative resolution in clinical machine learning.

5.3 Resolution Specific Clinical Interpretability

The resolution hypothesis predicts that different health outcomes depend on distinct temporal granularities. To test this, we analyze performance across HiMAE layers, where each layer corresponds to a progressively coarser resolution. Figure 6 reveals clear resolution-specific structure: individual downstream tasks achieve maximal AUROC at different layers, highlighted by the red boundaries.

This layer-task alignment underscores two key insights. First, temporal resolution is not a nuisance parameter but an axis of predictive structure: different outcomes are best represented at different scales (we show that collapsing an encoder decoder still has concordant results showing that our hierarchal model is not an artifact in Appendix Section F.5). Second, HiMAE naturally exposes this heterogeneity, functioning as a discovery tool for identifying the most informative resolution per task. This complements conventional interpretability methods (Amann et al., 2022; Xu et al., 2023; Lee et al., 2025) by shifting the focus from *which features* drive predictions to *which resolutions* matter. In doing so, HiMAE operationalizes the resolution hypothesis and provides insights to tasks where the resolution needed is not entirely clear.



Model	Params (↓)	FLOPs (↓)	Memory (↓)	On-device Lat. (↓)
HiMAE	**1.2M**	**0.0647** gFLOPs	**4.8 MB**	**0.99 ms**
Efficient-Net B-1	7.8M	0.70 gFLOPs	31.1 MB	1.42 ms
Swin-Transformer	110.6M	11.89 gFLOPs	423.8 MB	2.95 ms
LSM-Base	110.6M	15.94 gFLOPs	441.3 MB	3.36 ms

Figure 7: **Model efficiency and on-device inference:** Sample on-device detections on REDACTED device. Size, compute cost, memory footprint, and CPU latency (ms per sample, batch size 2048) measured over a 10s sequence at 100Hz.

5.4 Case Studies

Case Study 1: On-Device Benchmarking A central novelty of HiMAE is that it is, to our knowledge, the first SSL method compact enough to run entirely *on-watch*, rather than on phone-class hardware. We evaluate on-device PVC detection on smartwatch-class CPUs sampled at 100 Hz (Figure 7). HiMAE is exceptionally lightweight (1.2M parameters, 0.0647 gFLOPs, 4.8 MB) and achieves 0.99 ms latency per sample, equivalent to processing ≈1,010 samples/s or ≈2.8 hours of signal per minute of wall time. By contrast it shows massive performance gains against transformer

baselines, Swin-Transformer (110M parameters, 11.9 gFLOPs, 423 MB) and LSM-Base (110M, 15.9 gFLOPs, 441 MB). HiMAE also outperforms optimized models like Efficient-Net B1 (Tan & Le, 2020) providing context to the latency and compactness of our model. HiMAE is thus $\sim$3–4$\times$ more efficient compared to transformers while fitting fully on-watch (without quantization (Jacob et al., 2017)), enabling continuous, private inference at the point of signal collection. *This prototype is strictly for research and is not deployed commercially.*

Case Study 2: HiMAE is adaptable in few shot settings

A central challenge in the wearable domain is that labels are scarce across tasks. Models that can adapt quickly from generic pretraining to specific detection tasks with limited supervision are therefore essential. Figure 8 illustrates this setting: HiMAE provides strong representations that can be adapted to diverse tasks such as PVC detection or hypertension monitoring with only a handful of labeled examples as reflected by the shape of the learning curves on the few-shot learning experiments. By reducing the supervision required to reach high performance, HiMAE enables new tasks to be supported on-device without the prohibitive cost of large curated datasets which help bolster its practical utility.

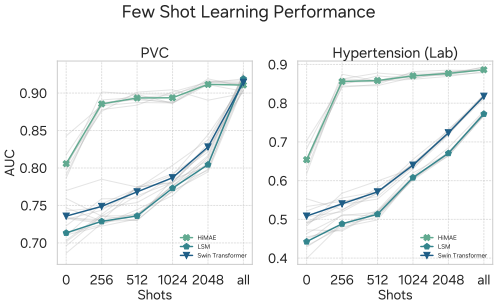


Figure 8: **Few-shot adaptation.** HiMAE adapts efficiently to new wearable tasks under sparse labels indicated by curve shape over transformer baselines.

6 Discussion

Summary. HiMAE advances wearable self supervised methods along three dimensions: (i) its flexible architecture is expressly designed for multi-resolution mapping, enabling seamless adaptation across heterogeneous tasks, (ii) by aligning task-dependent resolutions with model representations, it not only optimizes predictive performance but also offers a window into the temporal organization of physiological biomarkers, and (iii) by design of the compactness, it achieves the first demonstration of true *on-watch* inference, running entirely within smartwatch-class constraints while matching or surpassing performance on far larger models. These results position HiMAE as an efficient representation learner but also as a framework for interrogating which temporal resolutions carry signal.

Resolution as a structural prior. Our findings validate the resolution hypothesis and suggest a shift in how representation learning on wearables should be conceptualized. This reframing implies that representation learning for physiological signals should expose, rather than collapse, scale-specific embeddings. The layer-wise AUROC profiles in Figure 6 show that predictive performance peaks at different levels of the hierarchy depending on the task, with fine-scale embeddings capturing short-lived physiological events and coarse-scale embeddings capturing slower behavioral phenomena. By revealing this heterogeneity, HiMAE provides empirical evidence that resolution-specific representations are essential for wearable health modeling.

From "on-device" to "on-watch." HiMAE demonstrates that convolutional hierarchies can reduce model size by two orders of magnitude relative to transformer-based LSMs, enabling the first instance of true *on-watch* inference. This moves the deployment frontier from phone-class to watch-class processors, where inference occurs exactly at the point of sensing. Beyond efficiency, this shift has consequences for privacy (data never leave the device) and for clinical viability (continuous real-time monitoring becomes feasible).

Limitations and Future Works While we focus on PPG, the principles underlying HiMAE generalize to multimodal settings. Physiological signals are inherently multi-scale across modalities (e.g., ECG beats, accelerometer motion cycles, EEG rhythms), and resolution-aware architectures could expose complementary temporal signatures across them. Another limitation of our work is we don't handle sequences beyond 10 second windows which could unlock another breadth of tasks. Future works also warrants a clinical validation to the discoveries made by HiMAE which could be of significant interest to the health community.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pp. 265–283, USA, 2016. USENIX Association. ISBN 9781931971331.

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.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Malak Abdullah Almarshad, Md Saiful Islam, Saad Al-Ahmadi, and Ahmed S BaHammam. Diagnostic features and potential applications of ppg signal in healthcare: A systematic review. In *Healthcare*, volume 10, pp. 547. MDPI, 2022.

Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K Gilbert, Thilo Hagendorff, Sune Holm, Michelle Livne, et al. To explain or not to explain?—artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, 1(2):e0000016, 2022.

Ulzee An, Moonseong Jeong, Simon A Lee, Aditya Gorla, Yuzhe Yang, and Sriram Sankararaman. Raptor: Scalable train-free embeddings for 3d medical volumes leveraging pretrained 2d foundation models. *arXiv preprint arXiv:2507.08254*, 2025.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis, 2017. URL https://arxiv.org/abs/1706.07206.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49. JMLR Workshop and Conference Proceedings, 2012.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Vera Birrer, Mohamed Elgendi, Olivier Lambercy, and Carlo Menon. Evaluating reliability in wearable devices for sleep staging. *NPJ Digital Medicine*, 7(1):74, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Lucía Bouza, Aurélie Bugeau, and Loïc Lannelongue. How to estimate carbon footprint when training deep learning models? a guide and review. *Environmental Research Communications*, 5 (11):115014, November 2023. ISSN 2515-7620. doi: 10.1088/2515-7620/acf81b. URL http://dx.doi.org/10.1088/2515-7620/acf81b.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

M. Bridges, F. Feroz, M. P. Hobson, and A. N. Lasenby. Bayesian optimal reconstruction of the primordial power spectrum. *Monthly Notices of the Royal Astronomical Society*, 400(2): 1075–1084, December 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.15525.x. URL http://dx.doi.org/10.1111/j.1365-2966.2009.15525.x.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ebubekir Buber and DIRI Banu. Performance analysis and cpu vs gpu comparison for deep learning. In *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, pp. 1–6. IEEE, 2018.

Stephen Butterworth et al. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

Yong-Mei Cha, Glenn K Lee, Kyle W Klarich, and Martha Grogan. Premature ventricular contraction-induced cardiomyopathy: a treatable condition. *Circulation: Arrhythmia and Electrophysiology*, 5(1):229–236, 2012.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting, 2022. URL https://arxiv.org/abs/2201.12886.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1): 123–144, 2021.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. *arXiv preprint arXiv:2402.05956*, 2024.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020b. URL https://arxiv.org/abs/2002.05709.

Lawrence J Christiano and Terry J Fitzgerald. The band pass filter. *International economic review*, 44(2):435–465, 2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

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.

Chloë FitzGerald and Samia Hurst. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):19, 2017.

Sambit Ghadai, Xian Yeow Lee, Aditya Balu, Soumik Sarkar, and Adarsh Krishnamurthy. Multi-level 3d cnn for learning multi-scale spatial features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Thomas D Giles, Bradford C Berk, Henry R Black, Jay N Cohn, John B Kostis, Joseph L Izzo Jr, and Michael A Weber. Expanding the definition and classification of hypertension. *The Journal of Clinical Hypertension*, 7(9):505–512, 2005.

Thomas D Giles, Barry J Materson, Jay N Cohn, and John B Kostis. Definition and classification of hypertension: an update. *The journal of clinical hypertension*, 11(11):611–614, 2009.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon, 2023. URL `https://github.com/ml-explore`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Mingyi He, Bo Li, and Huahui Chen. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3904–3908. IEEE, 2017.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL `https://arxiv.org/abs/1606.08415`.

Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL `https://arxiv.org/abs/1502.03167`.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL `https://arxiv.org/abs/1712.05877`.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv*, 2021. doi: 10.48550/arxiv.2110.09348. URL https://arxiv.org/abs/2110.09348.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Yasin Kaya and Hüseyin Pehlivan. Classification of premature ventricular contraction in ecg. *International Journal of Advanced Computer Science and Applications*, 6(7), 2015.

Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14274–14285, 2020.

Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A Fries, Jeffrey N Chiang, Jungwoo Oh, Justin Xu, et al. Meds decentralized, extensible validation (meds-dev) benchmark: Establishing reproducibility and comparability in ml for health. *Machine Learning For Health Confernce*, 2024.

Lingjing Kong, Martin Q Ma, Guangyi Chen, Eric P Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7918–7928, 2023.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2024. URL https://arxiv.org/abs/2205.13147.

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

Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N Chiang. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1):394, 2025.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1082. URL https://aclanthology.org/N16-1082/.

Qiao Li, Qichen Li, Ayse S Cakmak, Giulia Da Poian, Donald L Bliwise, Viola Vaccarino, Amit J Shah, and Gari D Clifford. Transfer learning from ecg to ppg for improved sleep staging from wrist-worn wearables. *Physiological measurement*, 42(4):044004, 2021.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. Unraveling the hidden environmental impacts of ai solutions for environment life cycle assessment of ai solutions. *Sustainability*, 14(9):5172, April 2022. ISSN 2071-1050. doi: 10.3390/su14095172. URL http://dx.doi.org/10.3390/su14095172.

Yihan Lin, Zhirong Bella Yu, and Simon Lee. A case study exploring the current landscape of synthetic medical record generation with commercial llms. *arXiv preprint arXiv:2504.14657*, 2025.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0EXmFzUn5I.

Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Vision transformers with hierarchical attention. *Machine intelligence research*, 21(4):670–683, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021b. URL https://arxiv.org/abs/2103.14030.

Aleksej Logacjov, Kerstin Bach, and Paul Jarle Mork. Long-term self-supervised learning for accelerometer-based sleep–wake recognition. *Engineering Applications of Artificial Intelligence*, 141:109758, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Jianwen Luo, Kui Ying, and Jing Bai. Savitzky–golay smoothing and differentiation filter for even number data. *Signal processing*, 85(7):1429–1434, 2005.

Melissa D McCradden, Shalmali Joshi, James A Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, 27(12):2024–2027, 2020.

Matthew BA McDermott, Justin Xu, Teya S Bergamaschi, Hyewon Jeong, Simon A Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, et al. Meds: Building models and tools in a reproducible health ai ecosystem. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6243–6244, 2025.

Suril Mehta, Nipun Kwatra, Mohit Jain, and Daniel McDuff. Examining the challenges of blood pressure estimation via photoplethysmogram. *Scientific Reports*, 14(1):18318, 2024.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Jake Garrison, Shyam A Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, et al. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations*.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

Meir Nitzan and Zehava Ovadia-Blechman. Physical and physiological interpretations of the ppg signal. In *Photoplethysmography*, pp. 319–340. Elsevier, 2022.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. URL https://arxiv.org/abs/1511.08458.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Ignacio Perez-Pozuelo, Dimitris Spathis, Jordan Gifford-Moore, Jessica Morley, and Josh Cowls. Digital phenotyping and sensitive health data: Implications for data governance. *Journal of the American Medical Informatics Association*, 28(9):2002–2008, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

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

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals, 2025. URL https://arxiv.org/abs/2410.20542.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Neal Schmitt and Goran Kuljanin. Measurement invariance: Review of practice and implications. *Human resource management review*, 18(4):210–222, 2008.

Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of non-invasive blood pressure prediction from ppg and rppg signals using deep learning. *Sensors*, 21(18):6022, 2021a.

Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of deep learning based blood pressure prediction from ppg and rppg signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3820–3830, 2021b.

Muhammad Shabaan, Kaleem Arshid, Muhammad Yaqub, Feng Jinchao, M Sultan Zia, Giridhar Reddy Bojja, Muazzam Iftikhar, Usman Ghani, Loknath Sai Ambati, and Rizwan Munir. Survey: smartphone-based assessment of cardiovascular diseases using ecg and ppg analysis. *BMC medical informatics and decision making*, 20(1):177, 2020.

Mohammad Amin Shabani, Amir H. Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sCrnllCtjoE.

Wei Shuai, Xi-xing Wang, Kui Hong, Qiang Peng, Ju-xiang Li, Ping Li, Jing Chen, Xiao-shu Cheng, and Hai Su. Is 10-second electrocardiogram recording enough for accurately estimating heart rate in atrial fibrillation. *International journal of cardiology*, 215:175–178, 2016.

Gerald Simonneau, Nazzareno Galiè, Lewis J Rubin, David Langleben, Werner Seeger, Guido Domenighetti, Simon Gibbs, Didier Lebrec, Rudolf Speich, Maurice Beghetti, et al. Clinical classification of pulmonary hypertension. *Journal of the American College of Cardiology*, 43 (12S):S5–S12, 2004.

Gérald Simonneau, Ivan M Robbins, Maurice Beghetti, Richard N Channick, Marion Delcroix, Christopher P Denton, C Gregory Elliott, Sean P Gaine, Mark T Gladwin, Zhi-Cheng Jing, et al. Updated clinical classification of pulmonary hypertension. *Journal of the American college of cardiology*, 54(1_Supplement_S):S43–S54, 2009.

Gerald Simonneau, Michael A Gatzoulis, Ian Adatia, David Celermajer, Chris Denton, Ardeschir Ghofrani, Miguel Angel Gomez Sanchez, R Krishna Kumar, Michael Landzberg, Roberto F Machado, et al. Updated clinical classification of pulmonary hypertension. *Journal of the American College of Cardiology*, 62(25S):D34–D41, 2013.

Gérald Simonneau, David Montani, David S Celermajer, Christopher P Denton, Michael A Gatzoulis, Michael Krowka, Paul G Williams, and Rogerio Souza. Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *European respiratory journal*, 53(1), 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Benjamin A Teplitzky, Michael McRoberts, and Hamid Ghanbari. Deep learning for comprehensive ecg annotation. *Heart rhythm*, 17(5):881–888, 2020.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore Iv, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. In *International Conference on Machine Learning*, pp. 48019–48037. PMLR, 2024.

Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. Foundation models in healthcare: Opportunities, risks & strategies forward. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*, pp. 1–4, 2023.

Lan V Truong. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284*, 2022.

Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Jagat Narula, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digital Medicine*, 6(1):108, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling, 2023. URL https://arxiv.org/abs/2304.05919.

Ke Wang, Jiamu Yang, Ayush Shetty, and Jessilyn Dunn. Dreamt: Dataset for real-time sleep stage estimation using multisensor wearable technology. *PhysioNet https://doi. org/10.13026/62AN-CB28*, 2024.

Peng Wang, Yuanzhouhan Cao, Chunhua Shen, Lingqiao Liu, and Heng Tao Shen. Temporal pyramid pooling-based convolutional neural network for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2613–2622, 2016.

Lukasz Wesolowski, Bilge Acun, Valentin Andrei, Adnan Aziz, Gisle Dankel, Christopher Gregg, Xiaoqiao Meng, Cyril Meurillon, Denis Sheahan, Lei Tian, et al. Datacenter-scale analysis and optimization of gpu machine learning workloads. *IEEE Micro*, 41(5):101–112, 2021.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023.

Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 567–584, 2018.

Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5028–5037, 2017.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of healthcare engineering*, 2023(1):9919269, 2023.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094. PMLR, 2019.

Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022a.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. URL https://arxiv.org/abs/1512.04150.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pp. 1–65, 2024.

Wei Zhu, Qiang Qiu, Robert Calderbank, Guillermo Sapiro, and Xiuyuan Cheng. Scaling-translation-equivariant networks with decomposed convolutional filters, 2022. URL https://arxiv.org/abs/1909.11193.

APPENDIX

A FREQUENTLY ASKED QUESTIONS

What are the main conclusions from this work? We demonstrate that convolutional architectures benefit from inductive biases that remain advantageous for PPG signals. On our pre-training data, our model consistently outperforms alternative baselines. Furthermore, scaling experiments across model sizes reveal that brute-force scaling of generic architectures is possible, but less effective: our model achieves stronger performance and scales more gracefully due to a better initialization and inductive structure relative to other models. In addition to this inductive bias and compact design, our contributions are two fold in the sense that our model demonstrates the first on-device model which does not require phone level processors to run inference.

Is your pre-training dataset large enough? Our pre-training corpus was collected internally and is of comparable scale to recent public benchmarks such as PaPaGei and Apple's datasets. In terms of magnitude, we position our dataset as PaPaGei (Pillai et al., 2025) $<$ Ours $<$ Apple (Abbaspourazad et al., 2023) $<$ Google (Narayanswamy et al., 2024). Thus, while not the largest available, our dataset size is sufficiently large to validate the approach and lies within the range of accepted practice for self supervised learning wearable models.

Why do you model at 10-second windows? We deliberately adopt 10s windows sampled at 100Hz to balance physiological coverage with on-device feasibility. Many clinically actionable events, such as arrhythmic beats or premature ventricular contractions, unfold on the order of seconds and require rapid detection to enable continuous monitoring and real-time feedback. Shorter windows would impair the model's ability to capture meaningful temporal context, while much longer windows would hinder low-latency inference on watch-class hardware. By constraining the receptive field to 10s, HiMAE preserves second-level resolution while remaining efficient enough to process signals continuously under the hardware limits of edge devices. Additionally, 10-second window are a standard protocol that are adopted in the clinical setting where ECG for example is collected and interpreted at 10 second segments (Shuai et al., 2016).

What are the advantages of smaller models? From a research perspective, smaller models foster inclusivity by reducing reliance on brute-force scaling of transformer-based architectures that only industry-scale labs can realistically afford. From a deployment standpoint, compact models enable on-device inference on constrained hardware such as wearables. This dual benefit—lower research barriers and wider deployment potential—underscores the importance of investigating architectures that remain competitive at modest scale.

How large is too large to deploy on a smart watch? In principle, models up to approximately 50MB can be stored and executed on modern smart watches or larger models can be quantized (Jacob et al., 2017). In practice, however, latency and energy considerations suggest that models exceeding roughly 10MB may already hinder real-time inference and limit commercial viability. Additionally quantization does not do due dilligence to the original model and some level of the model's performance is lost. While smartphones relax these constraints, our contribution highlights that the proposed model remains sufficiently compact to fit within the computational and storage budgets of wearable devices such as watches, thereby supporting direct on-device deployment.

Can PPG predict abnormal laboratory results? We frame this as a binary classification task, testing whether photoplethysmography signal encodes biomarkers that separate "normal" from "abnormal" lab classes. Our investigation probes whether learned PPG representations capture biomarker signatures correlated with out-of-range labs, using lightweight classifiers on frozen embeddings with strict temporal alignment. Preliminary evidence suggests discriminative signal above chance, but these findings are designed to be exploratory and not clinically actionable.

B Ethics Considerations

B.1 Data Privacy and Consent

Wearable signals capture sensitive physiological and behavioral information (Erturk et al., 2025). Our study relies on both publicly available and proprietary (company-owned) datasets that have been carefully vetted. These datasets include transparent disclosure of data usage, explicit opt-in mechanisms, and the option for participants to withdraw (Perez-Pozuelo et al., 2021). Across the seven datasets used in this study, we obtained written consent—via paper or digital waivers—that clearly informed participants that their data may be used for commercial research purposes.

B.2 Bias and Representativeness

Physiological signals vary across age, gender, ethnicity, health status, and socioeconomic context, yet most existing datasets underrepresent key populations (FitzGerald & Hurst, 2017; McCradden et al., 2020; Chen et al., 2021). Such underrepresentation risks embedding biases into foundation models, leading to inequitable performance in downstream applications. Mitigation requires deliberate corpus curation, bias auditing, and systematic evaluation across diverse cohorts. In this study, we sought to mitigate bias by incorporating a pre-training corpus drawn from a wide range of wearable devices, collected across multiple regions of the world and over many years. However, patient-specific demographic information is not available. We do note that our data was collected across 4 countries including, USA, Brazil, Bangladesh, and South Korea.

B.3 Clinical Implications

Wearable foundation models are not substitutes for medical judgment. Their predictions require regulatory approval and clinical validation before integration into healthcare practice. Without safeguards, model misinterpretation could lead to misdiagnosis or inappropriate treatment. Development should involve clinical collaborators, real-world evaluations, and explicit positioning of models as decision-support rather than diagnostic systems. In our group, ongoing collaborations aim to evaluate where our foundation model performs well and how it may assist in forming clinical insights. We emphasize that no definitive clinical conclusions should be drawn from this work.

B.4 Environmental Impact

Training generative models entails substantial computational and environmental costs (Ligozat et al., 2022; Bender et al., 2021; Bouza et al., 2023). To minimize our footprint, we limited redundant runs, and reused checkpoints to avoid unnecessary GPU usage. All experiments were conducted on datacenter GPUs with efficient cooling systems and renewable energy credits to reduce carbon intensity. We emphasize that transparent reporting of compute usage and bounding resource allocation are necessary steps toward sustainable machine learning research.

C REPRODUCIBILITY STATEMENT

Table 1: HiMAE architecture components.

Encoder–Decoder

Layer	Output Shape
Input	[B, 1, T]
EncoderConvBlock(1→16)	[B, 16, T/2]
EncoderConvBlock(16→32)	[B, 32, T/4]
EncoderConvBlock(32→64)	[B, 64, T/8]
EncoderConvBlock(64→128)	[B, 128, T/16]
EncoderConvBlock(128→256)	[B, 256, T/32]
DecoderSkipBlock(256→128)	[B, 128, T/16]
DecoderSkipBlock(128→64)	[B, 64, T/8]
DecoderSkipBlock(64→32)	[B, 32, T/4]
DecoderSkipBlock(32→16)	[B, 16, T/2]
Final Deconv (16→1)	[B, 1, T]
Tanh	[B, 1, T]

EncoderConvBlock

Layer
Conv1d ($k = 5$, s=2, p=2)
BatchNorm
GELU
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
Conv1d ($k = 1$, s=2) + BN
GELU

DecoderSkipBlock

Layer
ConvTranspose1d ($k = 5$, s=2, p=2, op=1)
Concat skip connection
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
GELU
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
GELU

Due to restrictions around data licensing and industry policies, we are unable to release the full source code associated with HiMAE. To mitigate this limitation, we provide complete details of the model architecture, layer configurations, and hyperparameters in Table 1. This includes all encoder, decoder, and skip connection blocks, along with kernel sizes, strides, padding, activation functions, and normalization layers. Together, these descriptions are sufficient to re-implement the model faithfully in any modern deep learning framework (Paszke et al., 2019; Abadi et al., 2016; Bradbury et al., 2018; Hannun et al., 2023). In addition, we report all training settings (e.g., optimizer, learning rate schedule, and batch size) in the Appendix Section E to further support reproducibility. Our goal is to ensure that, while the exact implementation cannot be shared, independent researchers can replicate the methodology and validate the findings presented in this work.

C.1 TEMPORAL RESOLUTION AND RECEPTIVE FIELD

Let $x \in \mathbb{R}^T$ be a 1D input. Each `EncoderConvBlock` contains two convolutions on the main path, $\mathrm{Conv1d}(k{=}5, s{=}2)$ followed by $\mathrm{Conv1d}(k{=}5, s{=}1)$, and a $1{\times}1$ projection with stride 2 on the residual branch. The projection does not alter the main-path receptive field but aligns the skip in time and channel dimensions. For a stacked sequence of 1D convolutions with kernel sizes k_ℓ and strides s_ℓ (unit dilation), we define the effective input "jump" J_ℓ and receptive field R_ℓ after layer ℓ via

$$J_0 = 1, \quad R_0 = 1, \qquad J_\ell = J_{\ell-1}\, s_\ell, \qquad R_\ell = R_{\ell-1} + (k_\ell - 1)\, J_{\ell-1}.$$

Within one encoder block the first convolution halves the temporal resolution and expands the receptive field by $4J_{\mathrm{in}}$, and the second adds a further $8J_{\mathrm{in}}$ (because its stride is 1 but the jump has already doubled). Hence a block with effective stride 2 maps $(R_{\mathrm{in}}, J_{\mathrm{in}}) \mapsto (R_{\mathrm{out}}, J_{\mathrm{out}})$ with

$$J_{\mathrm{out}} = 2J_{\mathrm{in}}, \qquad R_{\mathrm{out}} = R_{\mathrm{in}} + 12\, J_{\mathrm{in}}.$$

After b encoder blocks this yields the closed form

$$J_b = 2^b, \qquad R_b = 1 + 12\sum_{i=0}^{b-1} 2^i = 1 + 12(2^b - 1).$$

Instantiating for our five encoder blocks ($b = 1, \ldots, 5$) gives the temporal resolutions $T/2, T/4, T/8, T/16, T/32$ and the cumulative receptive field at the end of the encoder of $R_5 = 373$ input samples per output position at stride $J_5 = 32$. Table 2 reports the resolution and cumulative receptive field after *every* main-path convolution; the 1 stride-2 projections on the residual branches are listed implicitly because they do not expand R_ℓ on the forward path.

Table 2: Temporal resolution and cumulative receptive field through the encoder. T denotes the input length in samples. R_ℓ is the receptive field after layer ℓ and J_ℓ the effective input stride ("jump").

Layer	Kernel k	Stride s	Output length	R_ℓ / J_ℓ
Enc1-conv1	5	2	$T/2$	5 / 2
Enc1-conv2	5	1	$T/2$	13 / 2
Enc2-conv1	5	2	$T/4$	21 / 4
Enc2-conv2	5	1	$T/4$	37 / 4
Enc3-conv1	5	2	$T/8$	53 / 8
Enc3-conv2	5	1	$T/8$	85 / 8
Enc4-conv1	5	2	$T/16$	117 / 16
Enc4-conv2	5	1	$T/16$	181 / 16
Enc5-conv1	5	2	$T/32$	245 / 32
Enc5-conv2	5	1	$T/32$	373 / 32

The figure of per-layer traces corroborates these counts: each block's first convolution visibly halves the temporal resolution, while the second refines features at the same scale; the cumulative growth of R_ℓ explains the progressive smoothing you observe at deeper layers, as each response aggregates over longer input windows. If your sampling rate is f_s Hz, multiply R_ℓ by $1/f_s$ to obtain the effective temporal support in seconds.

D DATASETS

D.1 AQUISTION AND APPROVAL

(IRB numbers have been excluded due to double blind review reviewing but will be included post-review)

All data analyzed in this study were collected under informed consent, with participants explicitly agreeing for their wearable-derived signals to be used in health-related research. The consent language stated that data could be used for developing new health features and algorithms and for inclusion in scientific publications. In particular, participants were informed that health and wellness data such as steps, heart rate, sleep, and photoplethysmography (PPG) signals could contribute to findings aimed at advancing general knowledge of health and science. No data used in this study included personally identifying information such as names or email addresses. We attach a portion of the protocols defined in our user data agreements below:

The use of these de-identified data for data usage was reviewed and classified as exempt. In addition, because the supporting records constitute case histories and document exposure to devices, we complied with the recordkeeping requirements in 21 CFR § 812.140(a)(3), including obtaining written digital consent and dated information. Participants could withdraw at any time; such withdrawals were documented in the case history, and data collected up to the point of withdrawal were retained and used for the investigation in accordance with the consent and applicable regulations.

For downstream evaluations, we relied on a combination of institutional review board (IRB)-approved datasets and publicly available resources. For instance, the PVC detection task used paired PPG and ECG recordings to derive annotations of premature ventricular contractions, with ECG-based labels verified both algorithmically and manually. The hypertension classification tasks were drawn from the REDACTED and REDACTED studies, both of which collected wrist-based PPG alongside reference blood pressure measurements under IRB-approved protocols. Sleep staging was evaluated using the DREAMT dataset, which combines PPG with gold-standard polysomnography annotations in individuals with and without diagnosed sleep disorders. Finally, a range of abnormal lab test prediction tasks were derived from the REDACTED dataset, linking PPG from REDACTED devices with clinical laboratory values for biomarkers (More details in Appendix Section D).

Across all studies, participants consented to data collection through mobile platforms that supported eligibility screening and enrollment, provided full informed consent, and enabled seamless integration of REDACTED devices for continuous signal acquisition. Where appropriate, participants also

reported medical histories or completed questionnaires through these platforms. All data were de-identified and stored in accordance with the approved study protocols, ensuring compliance with ethical and regulatory standards.

This layered consent and governance framework ensures that the data underpinning our pretraining and evaluation tasks are both ethically sourced and scientifically robust, supporting the broader goal of advancing health monitoring through consumer wearables such as the REDACTED Watch.

D.2 Pre-training and Generative Datasets

D.2.1 Device Distribution

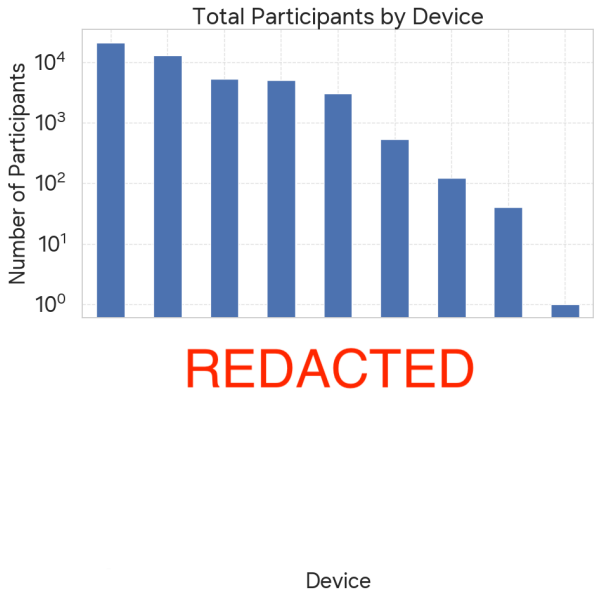


Figure 9: **Total Participants by Device.** The figure displays a bar chart illustrating the distribution of participants across different wearable devices used in the study. The y-axis is on a logarithmic scale to better show the wide range in the number of participants

The distribution of participants and data availability highlights both the diversity of collection devices and the heterogeneity of study contributions (Figure 9). At the device level, participation is primarily sourced from REDACTED DEVICE 1 REDACTED DEVICE 2 , and REDACTED DEVICE 3 , each contributing a lot of participants, while older models such as the REDACTED DEVICE 4 are represented by fewer users. This heterogeneity in devices provide us with a realistic and diverse set of raw wearable signals that can help us build generalizable foundation models. The presence of entries labeled as "NA" further reflects the mixture of collection devices and the occasional incompleteness of metadata. *We note that the devices used in our study are provided by two distributors limiting its generalizability and causing potential biases due to not having access to other consumer wearable devices.*

D.2.2 Participant Counts

In terms of study based segmentation, the dataset contains a handful of large-scale cohort studies, leading to diverse representation (Figure 10). Efforts were made to ensure representation across studies of varying sizes. This underscores the necessity of leveraging the vast scale of high-volume cohorts while simultaneously preserving the heterogeneity introduced by smaller studies, since both dimensions are essential for building foundation models that truly capture the variability and complexity of one-dimensional PPG signal modeling. Our data was collected across 4 countries (USA, South Korea, Brazil, Bangladesh) though most people specific demographic information is missing.

Figure 10: **Segment Count by Study.** This bar chart shows the number of data segments collected for each study, with the y-axis on a logarithmic scale to account for the large differences in segment counts.

D.2.3 PRE-PROCESSING PIPELINE

We operate on fixed-length windows (10 s) of raw PPG sampled at device-specific rates f_s. Each window is standardized via per-window z-scoring, $\tilde{x}_t = (x_t - \mu)/\sigma$, to remove level and scale effects that confound morphology-based quality metrics. To suppress gross amplitude artifacts (e.g., motion bursts), we compute the skewness of $|\tilde{x}|$, denoted $\gamma = \mathrm{skew}(|\tilde{x}|)$. Windows with heavy-tailed amplitude distributions ($\gamma > 2$) undergo an iterative trimming procedure that discards high-percentile excursions and recomputes γ until the distribution regularizes or a conservative floor is reached. This stage intentionally trades recall for precision: if trimming fails to regularize the distribution, the window is rejected.

For windows that pass amplitude checks, we impose a regularity prior using the sample autocorrelation $r[k] = \sum_t \tilde{x}_t \tilde{x}_{t+k}$. We locate zero-crossings of $r[k]$ near the origin and compute the dispersion of consecutive intervals, $\sigma_{\mathrm{zc}} = \mathrm{std}(\Delta k)/f_s$. Physiologically plausible pulsatile signals exhibit near-periodic structure; we therefore require a small timing dispersion to proceed. This criterion rejects segments whose periodicity is unstable, a signature of motion or sensor decoupling, and eliminates short or degenerate traces by enforcing a minimum number of intervals.

Surviving windows are band-limited with a low-order Butterworth filter to the cardiac band $[0.1, 2]$ Hz, which removes drift and high-frequency noise without distorting pulse morphology. We then quantify morphology via template matching against a canonical PPG waveform. Let $q_t \in [0, 1]$ denote the per-sample similarity score. We define a stringent acceptance mask $m_t = \mathbf{1}\{q_t > \tau\}$ with $\tau \in \{0.90, 0.95\}$ depending on whether the amplitude distribution was already regular ($\gamma \leq 2$). Two complementary statistics summarize quality: a "coverage" term $p = \frac{1}{T}\sum_t m_t$, measuring the fraction of the window that is confidently PPG-like, and an "agreement" term $a = \frac{1}{\max(1, \sum_t m_t)} \sum_t q_t m_t$, measuring how well accepted samples match the template. To penalize windows that have high agreement on vanishing coverage (or vice versa), we aggregate with the harmonic mean $H(a, p) = \frac{2ap}{a+p}$, yielding a continuous signal-quality index. A small additive term encodes whether the amplitude distribution was regular at entry, prioritizing windows that never required trimming. Windows that fail any upstream gate (amplitude regularization, periodicity stability, or template evaluation) are assigned null quality and excluded from downstream training.

At corpus scale, we apply this scoring in parallel and retain only windows with high composite quality. The resulting pretraining set emphasizes clean, consistent, periodic, and band-pass filtered signals harmonizing across devices and sampling rates, reducing the prevalence of motion artifacts and non-physiologic segments without relying on patient-level demographics or labels.

D.3 Downstream Evaluation Data

We evaluate HiMAE across diverse downstream tasks to assess the generality of wearable PPG representations. Rather than assuming a fixed mapping between PPG and outcomes, we exploit HiMAE's ability to learn hierarchical temporal features and adaptively resolve signal segments at scales most informative for prediction. This design allows us to probe the representational value of optical physiological signals across clinically and behaviorally relevant applications.

D.3.1 PVC Detection

Table 3: Stratified 80/20 Train/Test splits for PVC tasks (with per-task totals).

Task	Split	Negative	Positive	Total
PVC Detection	train	369987 (91.8%)	32832 (8.2%)	402819
	test	69880 (89.7%)	8019 (10.3%)	77899
	totals	439767 (91.4%)	40950 (8.6%)	480717

Premature Ventricular Contractions (PVCs) (Number Breakdowns in Table 3) are abnormal beats arising in the ventricles (Cha et al., 2012; Kaya & Pehlivan, 2015). We use paired PPG–ECG data, with ECG annotations generated using BeatLogic (Teplitzky et al., 2020) and manually verified. PPG inputs are 10s non-overlapping wrist segments, pre-processed with a Savitzky–Golay filter (Luo et al., 2005), a 0.5–4.0 Hz bandpass, normalization to $[-1, 1]$, and exclusion of segments with motion artifacts or disruptions > 1 s. This task evaluates whether ubiquitous PPG can approximate arrhythmia detection typically restricted to ECG.

D.3.2 Hypertension Classification

Table 4: Stratified 80/20 Train/Test splits for Hypertension tasks collected in a laboratory setting.

Task	Split	Negative	Positive	Total
Hypertension Classification (Lab)	train	2964 (86.7%)	454 (13.3%)	3418
	test	631 (76.7%)	192 (23.3%)	823
	totals	3595 (84.8%)	646 (15.2%)	4241

Table 5: Stratified 80/20 Train/Test splits for Hypertension tasks collected in a free-world setting.

Task	Split	Negative	Positive	Total
Hypertension Classification (Free World)	train	3959 (58.5%)	2812 (41.5%)	6771
	test	1042 (58.8%)	731 (41.2%)	1773
	totals	5001 (58.5%)	3543 (41.5%)	8544

Hypertension classification (Number Breakdowns in Tables 4, 5) relies on cuff-based references (Simonneau et al., 2004; Giles et al., 2005; 2009; Simonneau et al., 2009; 2013; 2019). Subjects within ± 8 mmHg of the diagnostic cutoff are excluded to reduce label noise, with remaining individuals labeled hypertensive or normotensive. Each 10s PPG segment undergoes Savitzky–Golay smoothing, 0.5–4.0 Hz bandpass filtering, normalization to $[-1, 1]$, and artifact removal. Unlike PVC detection, which is event-based, this task leverages PPG morphology and temporal dynamics to reflect vascular state. These evaluations contain both hypertension data collected in a naturalistic free world environment and within a controlled lab environment for both the hypertensive and blood pressure regression tasks.

D.3.3 SLEEP STAGING

Table 6: Stratified 80/20 Train/Test splits for Sleep Staging.

Task	Split	Wake	Light	Deep	REM	Total
Sleep Staging (4-class)	train	44829 (23.9%)	115932 (61.8%)	6696 (3.6%)	20214 (10.8%)	187671
	test	11298 (23.6%)	30153 (63.1%)	1416 (3.0%)	4881 (10.2%)	47748
	totals	56127 (23.8%)	146085 (61.9%)	8112 (3.4%)	25095 (10.6%)	235419

Sleep staging (Number Breakdowns in Tables 6) is evaluated on the DREAMT dataset (Wang et al., 2024) hosted on PhysioNet (Goldberger et al., 2000), which includes overnight wristband data with simultaneous PSG. Annotations follow AASM standards into wake, REM, NREM1, NREM2, and NREM3, excluding missing and preparation segments. PPG is bandpass filtered (0.5–12 Hz) (Butterworth et al., 1930), segmented into 10s windows, and normalized to zero mean and unit variance. Performance is measured with five-fold subject-independent cross-validation. This task examines whether PPG encodes temporal patterns sufficient for sleep stage classification. *We note that sleep staging has canonically been designed by leveraging the whole sleep cycle but we are assessing the ability to monitor real time sleep staging from much shorter PPG segments.*

D.3.4 ABNORMAL LAB TESTS

Table 7: Stratified 80/20 Train/Test splits for REDACTED tasks (with per-task totals).

Task	Split	Negative	Positive	Total
A1C	train	255 (31.6%)	553 (68.4%)	808
	test	64 (31.7%)	138 (68.3%)	202
	totals	319	691	1010
Hematocrit	train	1271 (77.0%)	380 (23.0%)	1651
	test	305 (77.0%)	91 (23.0%)	396
	totals	1576	471	2047
Hemoglobin	train	867 (81.2%)	201 (18.8%)	1068
	test	208 (81.3%)	48 (18.8%)	256
	totals	1075	249	1324
Platelets	train	622 (35.5%)	1129 (64.5%)	1751
	test	143 (35.7%)	258 (64.3%)	401
	totals	765	1387	2152
Potassium	train	731 (33.1%)	1476 (66.9%)	2207
	test	167 (33.1%)	338 (66.9%)	505
	totals	898	1814	2712
Sodium	train	203 (17.6%)	951 (82.4%)	1154
	test	48 (17.7%)	223 (82.3%)	271
	totals	251	1174	1425
WBC	train	247 (18.6%)	1082 (81.4%)	1329
	test	62 (18.7%)	270 (81.3%)	332
	totals	309	1352	1661

For abnormal lab test prediction, we use REDACTED Watch PPG collected at REDACTED University paired with clinical laboratory results. Each test is framed as a binary classification task: outcomes are labeled negative if within the 25th percentile of lab values and the positive labels are anything above the 75th percentile. All other labels are excluded. Preprocessing matches other tasks. Targets include A1C, hemoglobin, hematocrit, platelets, potassium, sodium, and WBC, each selected for established clinical relevance. This task extends evaluation beyond cardiovascular and behavioral endpoints to systemic markers of metabolic, renal, and hematologic health. *We note that it is unclear whether PPG can predict abnormal from healthy lab values based on the PPG alone.*

Despite this, REDACTED univeristy presents us with an opportunity to discover if PPG signal can provide digital signatures making this an exploratory task in our benchmark.

Clinical Relevance of Lab Tests Each lab test used for this analysis provides critical information about a patient's health status. Their inclusion in this study is based on their established role in diagnosing or monitoring chronic conditions and acute health issues.

- **A1C (Glycated Hemoglobin):** Measures average blood glucose levels over the past 2–3 months. It is the primary diagnostic tool for diabetes and a key indicator for managing long-term blood sugar control. Elevated A1C levels are linked to increased risk of cardiovascular disease, kidney damage, and other complications.
- **Hemoglobin:** Oxygen-carrying protein in red blood cells. Low levels indicate anemia, while elevated levels may suggest polycythemia vera.
- **Hematocrit:** Percentage of blood volume occupied by red blood cells. Used alongside hemoglobin to assess anemia or polycythemia.
- **Platelets:** Critical for clotting. Low count (thrombocytopenia) increases bleeding risk; high count (thrombocytosis) increases clot risk.
- **Potassium:** Essential electrolyte for nerve and muscle function. Both hypokalemia ($<$3.5 mEq/L) and hyperkalemia ($>$5.0 mEq/L) can trigger cardiac arrhythmias.
- **Sodium:** Regulates fluid balance and blood pressure. Abnormalities can indicate dehydration, renal disease, or endocrine disorders.
- **WBC (White Blood Cells):** Immune system cells. Leukocytosis ($>11\times10^9$/L) indicates infection, inflammation, or hematologic disease.

E BASELINES AND MODEL CONFIGURATION

Self Supervised Pre-trained methods have become a dominanat paradigm for health and wellness to study a variety of applications (Wornow et al., 2023; Thieme et al., 2023; He et al., 2024; An et al., 2025; Lin et al., 2025). Foundation models for one-dimensional signals are predominantly repurposed from architectures designed for vision, with adaptations that reinterpret temporal structure as a flattened analogue of spatial correlation. In this section we highlight our baseline models and model configurations

E.1 BASELINES

LSM (Narayanswamy et al., 2024) introduces a large-scale foundation model trained on multimodal wearable sensor data. The approach emphasizes scaling laws for wearable representation learning, leveraging a transformer-based backbone to capture temporal and cross-modal dependencies. Specifically, it adopts a vision transformer architecture trained via masked autoencoding with random masking. The model is designed as a general-purpose foundation, transferring effectively across a range of downstream tasks in physiological sensing and human activity recognition. In our work, we do not replicate the full multimodal design; instead, we adapt and constrain the model to a unimodal setting.

Swin-Transformer (Liu et al., 2021a) is a hierarchical Transformer that forms multi-scale representations by restricting self-attention to non-overlapping windows and alternating partitions with a shifted-window scheme, which enables cross-window communication while keeping computation near-linear in sequence length. We use this baseline as this is a direct comparison and counterpart to our proposed hierarchical HiMAE model. For wearable sensing, we adopt a 1D adaptation that tokenizes temporal patches and applies windowed attention along time, capturing both fine-grained waveform morphology and longer-range dependencies.

Masked Siamese Networks (MSN) (Assran et al., 2022) learn label-efficient representations by combining masked signal modeling with Siamese-style contrastive objectives. Instead of relying on class labels, MSN masks portions of the input and enforces consistency between augmented views. Architecturally, it employs a Vision Transformer encoder shared across views, while leveraging a predictor network to stabilize training. The key idea is to couple self-distillation with masked reconstruction to reduce sample complexity.

DINO (Caron et al., 2021) is a self-supervised framework that leverages knowledge distillation without labels. Using a teacher-student setup, the student network is trained to match the output distribution of the teacher under different data augmentations. Both networks are 1D-ViTs, and the method induces cluster-like emergent properties in the learned embedding space, enabling strong transfer performance without explicit contrastive pairs or handcrafted pretext tasks.

SimCLR (Chen et al., 2020b) establishes contrastive learning as a competitive self-supervised paradigm. The core idea is to maximize agreement between augmented views of the same signal in a latent space while pushing apart representations of different images. This is implemented using a ResNET encoder (He et al., 2015), a projection head, and a contrastive loss (NT-Xent (Chen et al., 2020a)).

PaPaGei (Pillai et al., 2024) is a domain-specific foundation model designed for optical physiological sensing, particularly photoplethysmography (PPG). It adapts ResNET-style CNN architectures to learn robust, generalizable representations from large-scale optical physiological datasets. PaPaGei releases both model weights and datasets to support reproducibility and broader adoption in physiological signal analysis. In our work, we used their source code to benchmark their method by pre-training on our volume of data to ensure fair comparison.

E.2 HYPERPARAMETERS FOR HIMAE AND BASELINES

To ensure a fair comparison across models, we aligned the training setup as closely as possible to the original implementations while maintaining consistency in optimizer choice and scheduling. All the methods trained from scratch (HiMAE, LSM, Swin-Transformer, MSN, DINO, SimCLR) were trained under identical optimization regimes, while PaPaGei follows its released open source training protocol. Table 8 summarizes the key hyperparameters for all models.

Table 8: Hyperparameter Configurations for Different Models

Configuration	HiMAE	LSM	Swin-Transformer	MSN	DINO	SimCLR	PaPaGei
Training Steps	50000						15000
Warmup Steps	2500						—
Optimizer	AdamW (Loshchilov & Hutter (2017))						
Opt. momentum $[\beta_1, \beta_2]$	[0.9, 0.95]	[0.9, 0.95]	[0.9, 0.95]	[0.9, 0.99]	[0.9, 0.99]	[0.9, 0.99]	—
Base learning rate	0.001	0.005	0.005	0.001	0.004	0.001	0.0001
Batch size	2048						256
Weight decay	0.0001						—
Gradient clipping	1.0	1.0	1.0	3.0	3.0	3.0	—
Dropout	0.0						—
Learning rate schedule	Linear Warmup & Cosine Decay						—
Loss Function	Mean Squared Error				Cross Entropy	Contrastive Loss	
Data resolution	1 (signal) - 100 Hz (Sampling Rate) $\times$ 10 (seconds)						
Augmentation	Flip, Time-Warping, Noise						

F Additional Results

F.1 Model Configurations Ablations

We conducted a comprehensive ablation study of HiMAE on a 100 Hz dataset comprising ten million segments (roughly 30k hours). The experiments systematically varied architecture and hyperparameters to understand their effect on reconstruction quality (Extrapolation task from our generative benchmark in tables where it is not explicitly stated as previously done in (Narayanswamy et al., 2024)), with multiple independent training runs averaged to reduce variance from stochastic initialization and data sampling. Unless otherwise noted, all training employed AdamW with a learning rate of 3×10^{-4}, cosine decay scheduling, and a batch size of 512.

Architecture. We evaluated HiMAE alongside CNN baselines across increasing network depths, defined by the sequence of hidden channel dimensions $[16, 32, 64]$, $[16, 32, 64, 128]$, and $[16, 32, 64, 128, 256]$. Table 9 lists the parameter counts, showing a modest growth for HiMAE compared to CNN baselines, with the skip-connected HiMAE exhibiting slightly higher capacity than its no-skip variant.

Table 9: Model Parameters (in K or M)

Model *Depth*	**HiMAE-tiny** [16,32,64]	**HiMAE-small** [16,32,64,128]	**HiMAE-Base** [16,32,64,128,256]
CNN	26.2 K	108 K	437 K
HiMAE-no skip	66.1 K	271 K	1.10 M
HiMAE	75.3 K	309 K	1.25 M

The impact of network depth on mean absolute error (MAE) and mean squared error (MSE) is summarized in Table 10. Increasing depth consistently reduced both MAE and MSE for HiMAE, with the deepest configuration yielding the lowest reconstruction error. Skip connections were critical, as HiMAE consistently outperformed its no-skip variant across all depths.

Table 10: MAE and MSE for Different Network Depths

Model *Depth*	**HiMAE-tiny** [16,32,64]		**HiMAE-small** [16,32,64,128]		**HiMAE-Base** [16,32,64,128,256]	
	MAE ↓	**MSE** ↓	**MAE** ↓	**MSE** ↓	**MAE** ↓	**MSE** ↓
CNN	0.4052	0.2345	0.4177	0.2491	0.4008	0.2315
HiMAE-noskip	0.4031	0.2365	0.4006	0.2465	0.3975	0.2339
HiMAE	0.4008	0.2309	0.3892	0.2232	**0.3827**	**0.2210**

Patch Size. We varied the spatial-temporal patch sizes over 1, 5, 10, and 20. The results in Table 12 indicate that 5 provided the best trade-off between local resolution and generative performance. Smaller patches increased flexibility but slightly degraded performance due to reduced receptive field per token, while overly large patches caused loss of fine-grained structure.

Table 12: Model Performance for Different Patch Sizes

Model	**1**		**5**		**10**		**20**	
	MAE ↓	**MSE** ↓	**MAE** ↓	**MSE** ↓	**MAE** ↓	**MSE** ↓	**MAE** ↓	**MSE** ↓
CNN	0.4140	0.2391	0.4008	0.2315	0.4122	0.2449	0.4274	0.2613
HiMAE-noskip	0.4069	0.2398	0.3976	0.2339	0.4037	0.2462	0.4195	0.2629
HiMAE	0.3899	0.2268	**0.3827**	**0.2210**	0.3861	0.2312	0.4039	0.2479

Convolution Kernel Size. Kernel size was varied over $\{1, 5, 10, 20\}$. Table 13 shows that 5 yielded the lowest errors across all models, suggesting moderate receptive fields match the temporal and

spatial scales of our data. Very small kernels restricted context aggregation, while very large kernels oversmoothed latent features.

Table 13: Model Performance Across Convolution Kernel Sizes

Model	1		5		10		20	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	0.4162	0.2413	0.4010	0.2309	0.4103	0.2418	0.4241	0.2576
HiMAE-noskip	0.4090	0.2427	0.3959	0.2331	0.4032	0.2440	0.4208	0.2591
HiMAE	0.3921	0.2283	**0.3821**	**0.2206**	0.3885	0.2316	0.4047	0.2485

Stride. We evaluated stride values of 2, 4, and 8 (Table 14). Smaller strides yielded the best performance, particularly for HiMAE, by preserving high temporal resolution in early feature maps. Performance degraded monotonically with stride increases.

Table 14: Model Performance Across Stride Values

Model	2		4		8	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	**0.4016**	**0.2312**	0.4139	0.2445	0.4318	0.2678
HiMAE-noskip	**0.3976**	**0.2334**	0.4098	0.2471	0.4272	0.2702
HiMAE	**0.3829**	**0.2209**	0.3928	0.2325	0.4103	0.2504

Masking Ratio. Finally, we explored the effect of varying the latent masking ratio in the masked autoencoding objective for generative tasks, with ratios from 0.5 to 0.9. As shown in Table 15, interpolation and extrapolation both improved when increasing the ratio up to 0.8, after which performance degraded for interpolation and collapsed for extrapolation.

Table 15: MAE and MSE for HiMAE Across Different Masking Ratios Evaluated on Generative Tasks

HiMAE Masking Ratio	Temporal Interpolation		Temporal Extrapolation	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓
0.5	0.3972	0.2292	0.4077	0.2519
0.6	0.3889	0.2223	0.3975	0.2294
0.7	0.3848	0.2207	0.3963	0.2278
0.8	**0.3796**	**0.2183**	**0.3879**	**0.2217**
0.9	0.3818	0.2219	0.2881	0.2216

Final Selection. These controlled experiments informed the final HiMAE configuration: the deepest architecture $[16, 32, 64, 128, 256]$ with skip connections, patch size 5, kernel size 5, stride 2, and a masking ratio of 0.8, which jointly achieved the best trade-off between reconstruction fidelity and parameter efficiency.

F.2 ECG Pre-training

HiMAE attains the lowest masked-reconstruction error on ECG (Table 16), indicating that its hierarchical masking and reconstruction inductive biases capture reconstruction capacity beyond PPG. LSM-1 (ViT) is a close second, while the ablated HiMAE and CNN trail, reinforcing that the full HiMAE design transfers effectively to the ECG domain.

Table 16: Masked-reconstruction loss on ECG masked auto encoding task.

Model	MSE (↓)
HiMAE	**0.148**
LSM-1 (ViT)	0.162
HiMAE (no skip)	0.184
CNN	0.207

F.3 VISUALIZATION OF RECONSTRUCTIONS

We provide sample reconstructions on both ECG (Figure 11) and PPG (Figure 12) signal showcasing that our framework works across signal modalities. In our work, we limit our analysis to PPG, since ECG is not passively collected and obtaining paired PPG and ECG data was not attainable at scale.

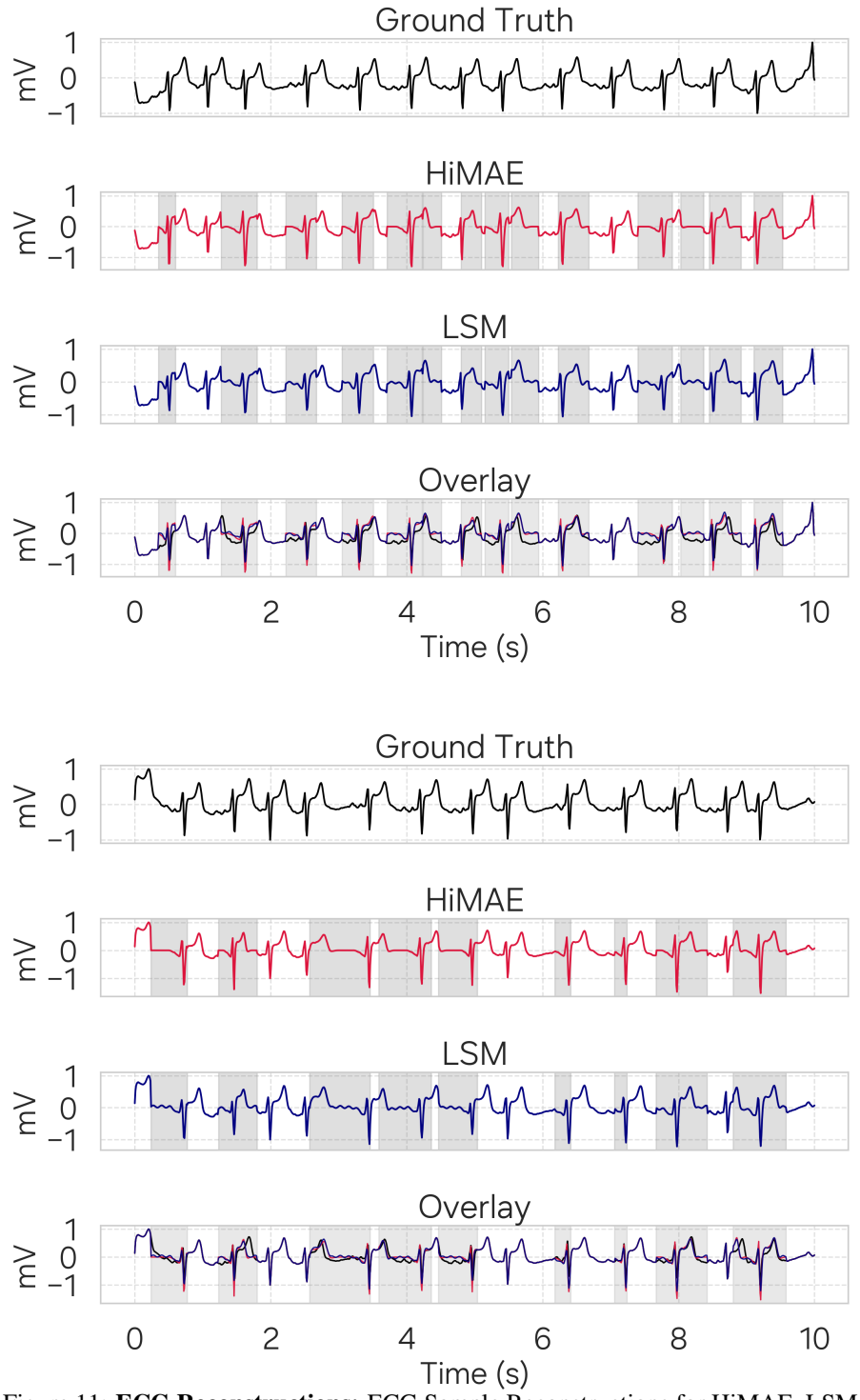


Figure 11: **ECG Reconstructions:** ECG Sample Reconstructions for HiMAE, LSM

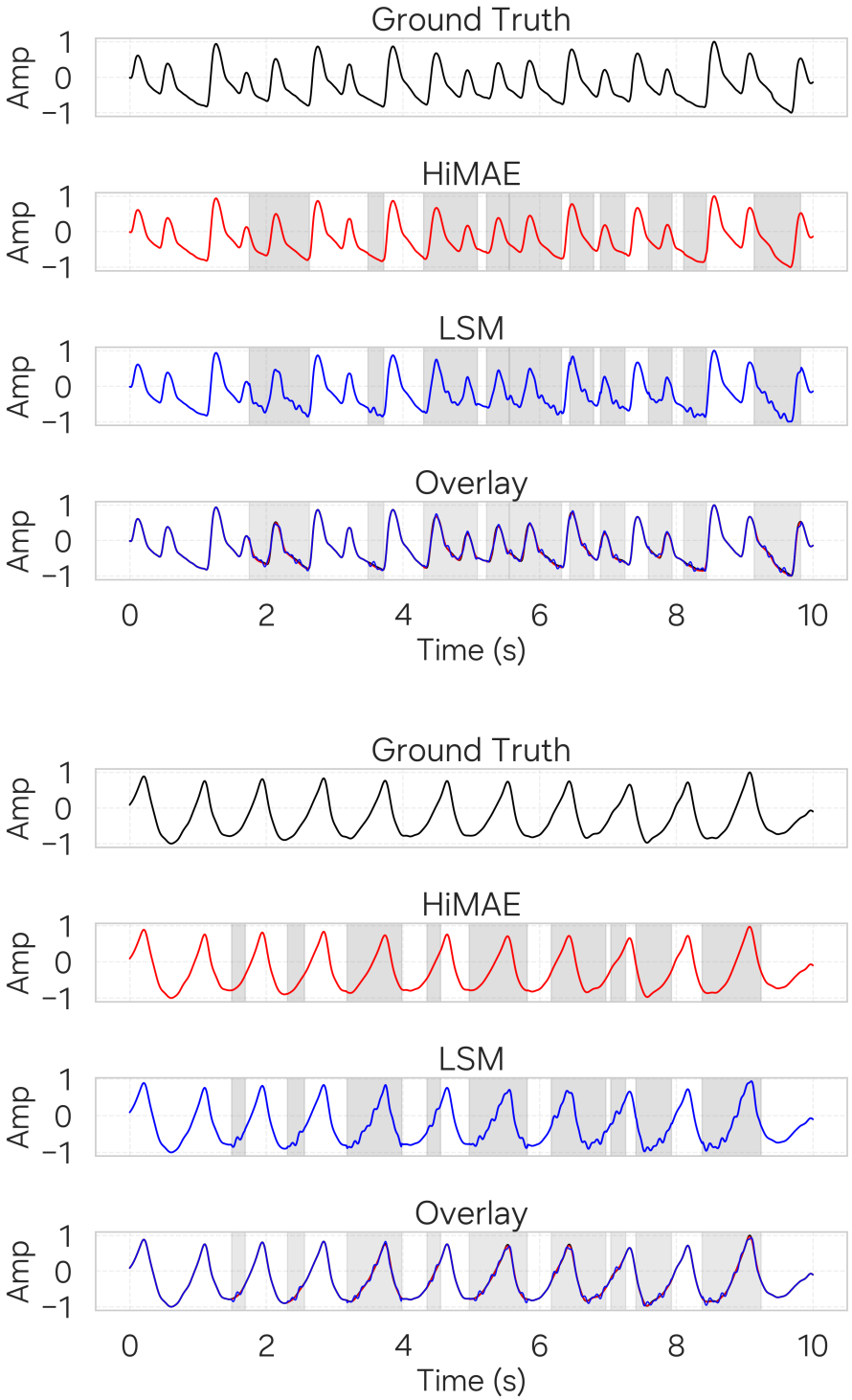


Figure 12: **PPG Reconstructions:** PPG Sample Reconstructions for HiMAE, LSM

F.4 Scaling Results for Generative Tasks

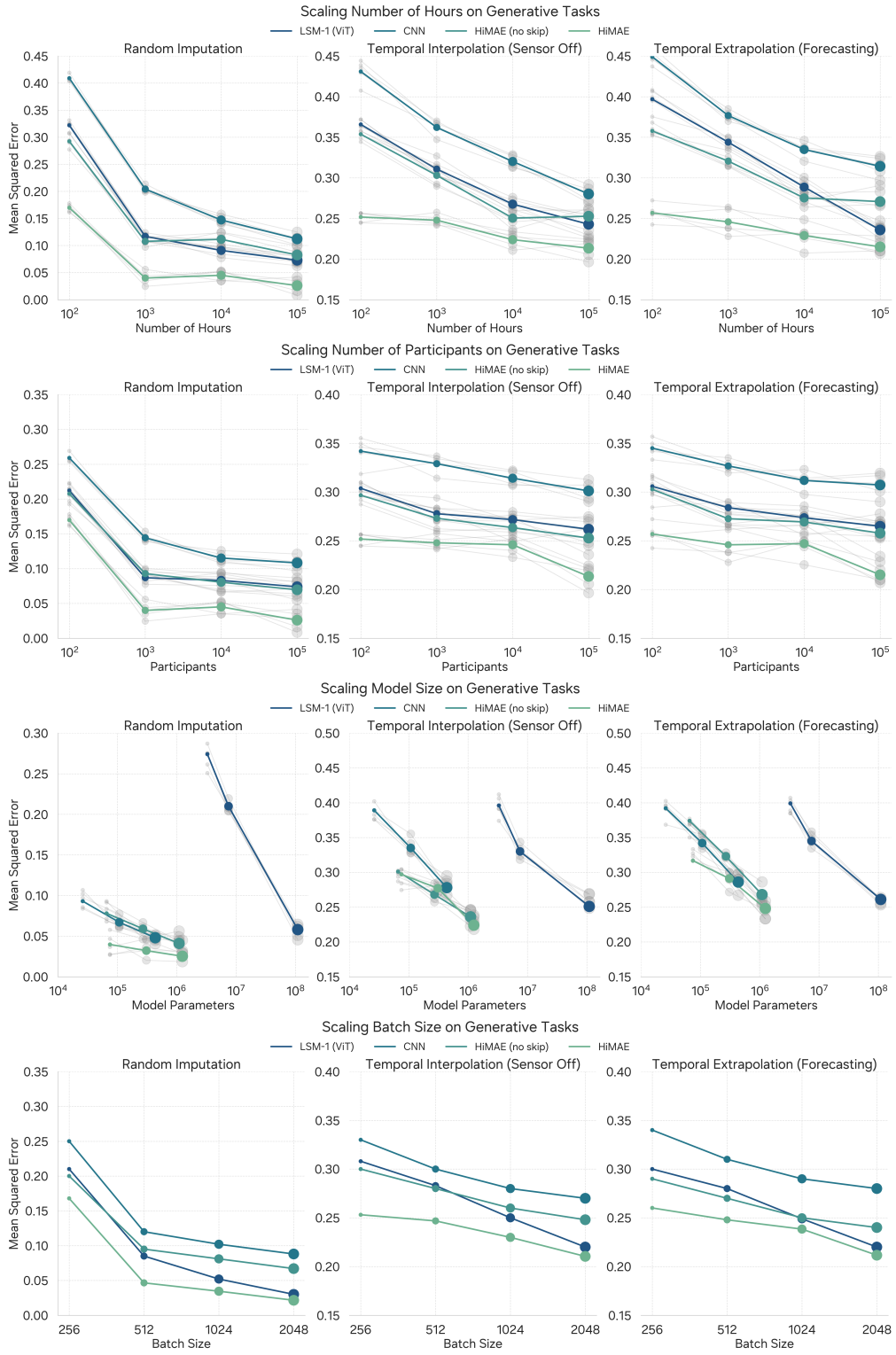


Figure 13: **Scaling Experiments on Generative Tasks:** Evaluation on the three generative tasks. HiMAE consistenly outperforms all model at our scale of data

Scaling analysis. We evaluate HiMAE's reconstruction error under participant, recording hour, batch size, and model size scaling, following the regimes of Narayanswamy et al. (2024); Xu et al. (2025): random imputation, temporal interpolation, and temporal extrapolation. Across all settings HiMAE follows clean scaling law trends (Kaplan et al., 2020) and maintains a margin over LSM-1 (ViT) and CNN baselines.

The most pronounced effect is model size. At small capacities HiMAE achieves lower error than much larger transformer baselines, highlighting the advantage of hierarchical inductive bias over sheer parameter count. LSM-1 only begins to close the gap at orders of magnitude more parameters. The transformer could surpass our HiMAE model when given a larger capacity but this again highlights the effectiveness of the inductive bias that we are conveying.

Participant, hour, and batch size scaling follow canonical patterns. More participants and longer recordings steadily reduce error, with HiMAE continuing to improve where baselines saturate, especially on interpolation and extrapolation.

Ablations confirm the mechanism: removing skip connections or collapsing the hierarchy to a single scale uniformly degrades performance, with gaps widening as data or model size grow. Task difficulty follows the expected order (imputation $<$ interpolation $<$ extrapolation), with the largest relative gaps in extrapolation, where hierarchical structure effectively lengthens usable context. Overall, HiMAE reaches lower error at smaller scales, showing that efficiency derives from inductive bias rather than brute force capacity.

F.5 Hierarchal Concordance

Layer concordance across depths. We further assess the stability of the resolution hypothesis by comparing HiMAE trained with four versus five encoder–decoder stages (Figure 14). The resulting heatmaps reveal that the alignment between downstream tasks and temporal resolutions is largely preserved across depths. Cardiovascular endpoints such as PVC detection and hypertension consistently achieve their best performance at finer layers, while blood related labs benefits from coarser layers. Although minor fluctuations appear in intermediate levels, the overall hierarchy of predictive resolutions is concordant. This suggests that the resolution–task mapping uncovered by HiMAE is not an artifact of architectural depth, but a robust property of the representations themselves.

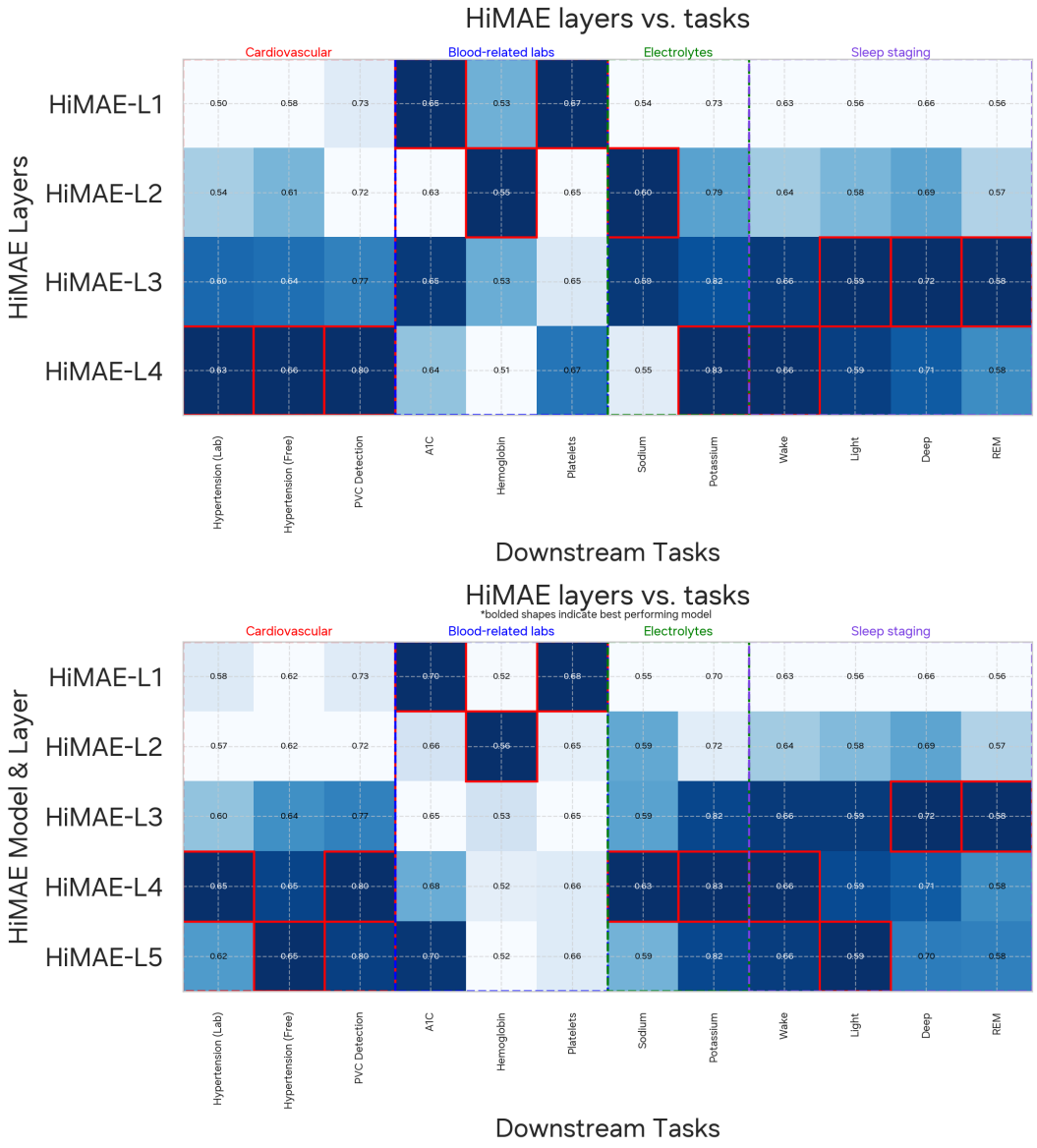


Figure 14: **HiMAE layer concordance across encoder depths.** Heatmaps compare downstream AUROC when probing HiMAE at 4 layers (top) versus 5 layers (bottom). Despite the removal of an encoder–decoder stage, the resolution–task alignment remains highly concordant: tasks such as PVC detection and hypertension consistently peak at similar layers, while sleep staging benefits from coarser representations. Minor deviations appear in intermediate layers, but the overall hierarchy of predictive resolutions is preserved, indicating robustness of the resolution hypothesis to architectural depth.

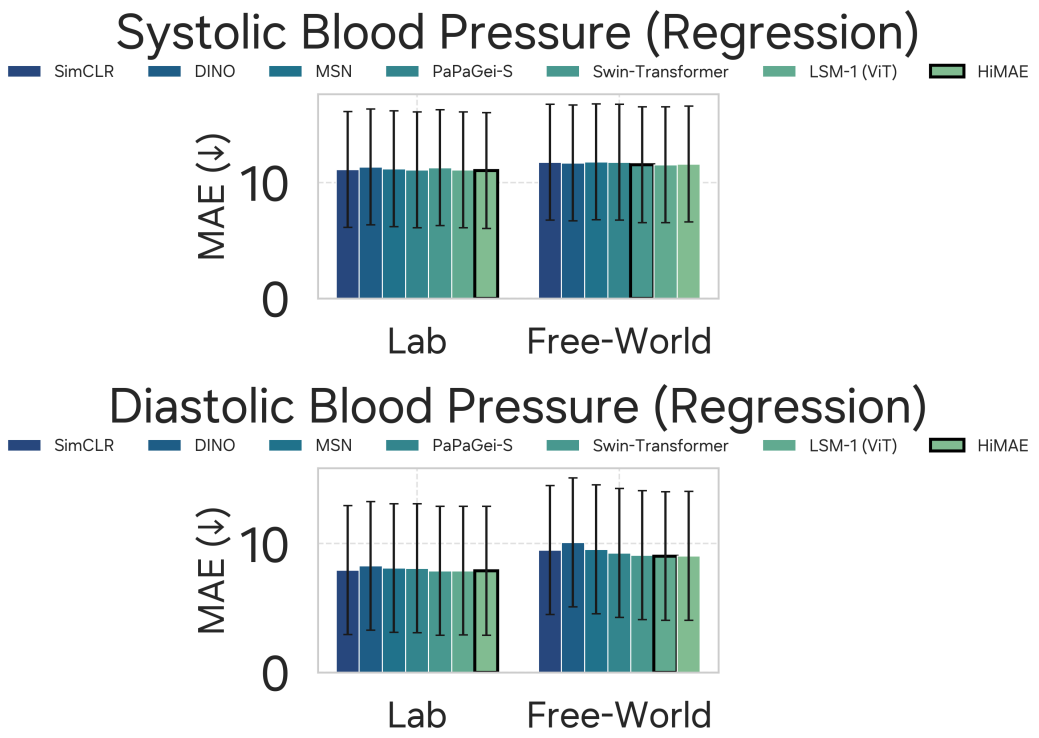


Figure 15: **Performance on regression benchmarks.** Mean absolute error (↓) for regressing systolic and diastolic blood pressure.

F.6 Regression

Continuous regression of blood pressure from wearable signals represents a canonical benchmark for physiological monitoring, yet the task remains highly challenging (Schrumpf et al., 2021a;b; Mehta et al., 2024). The objective is to recover systolic and diastolic pressures directly from sensor data, a setting where accuracy demands are clinically stringent but input signals are noisy and weakly correlated with the target (Figure 15). On the diastolic task, all approaches converge to errors on the order of 10 mmHg across both the REDACTED and REDACTED datasets. All Foundation Models yield similar performance, with HiMAE and LSM-1 providing marginal improvements but no decisive advantage. The systolic task exhibits a similar profile. Across datasets, performance saturates at errors slightly around 10 mmHg, with self-supervised approaches again clustered closely together. Despite this performance, our model does achieve the lowest mean absolute error across 2 out of the 4 comparisons showing that the model design does achieve better performance under the majority of scenarios. However, despite methodological advances, the achievable error floor has yet to approach clinically useful levels (Mehta et al., 2024).

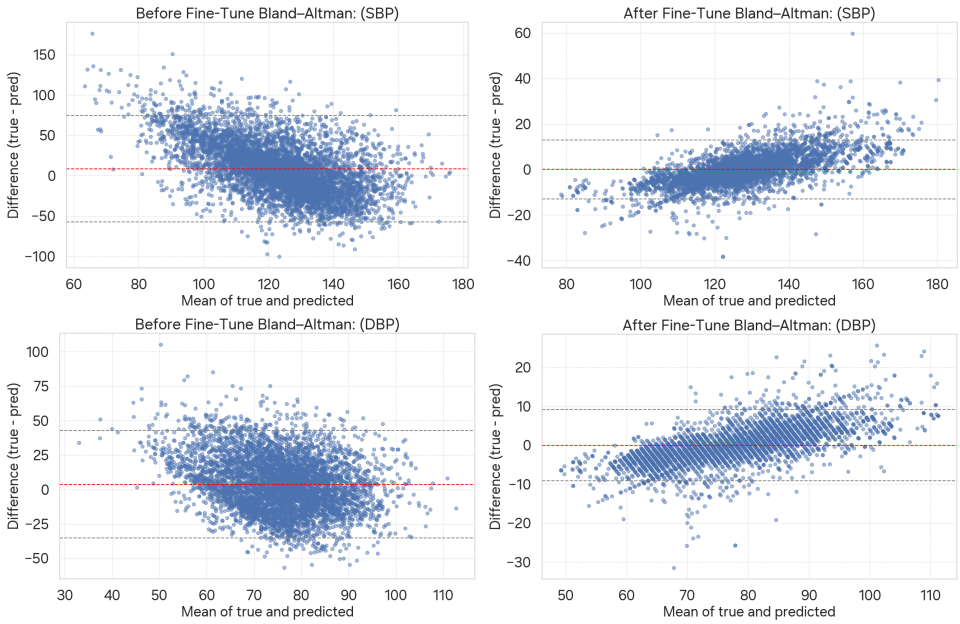


Figure 16: **Bland–Altman plot before and after fine-tuning on blood pressure regression.** The plots illustrate the agreement between predicted and reference blood pressure values, with mean bias (solid line) and 95% limits of agreement (dashed lines). Fine-tuning substantially reduces systematic bias and narrows the limits of agreement, indicating improved calibration and reliability of HiMAE-derived representations for regression.

F.7 Finetuning Improves Regression Performance

Fine-tuning substantially improves the regression behavior of our blood pressure estimators, as evidenced by the Bland–Altman plots in Figure 16. Prior to fine-tuning, both systolic and diastolic predictions exhibit large variance and systematic deviations, with wide limits of agreement and bias patterns that suggest poor calibration. After fine-tuning, the error distributions contract markedly: variance is reduced, biases approach zero, and the limits of agreement narrow considerably. These shifts indicate that fine-tuning not only enhances point prediction accuracy but also improves the overall reliability of the regression component, yielding estimates that are more clinically consistent with reference values. Despite this improvement, the model also indicates errors exceeding +/- 20mmHg which again highlight a limitation in these approaches to do well on estimating blood pressure.

F.8 TSNE PLOTS ON LINEAR PROBES AND FINE-TUNED

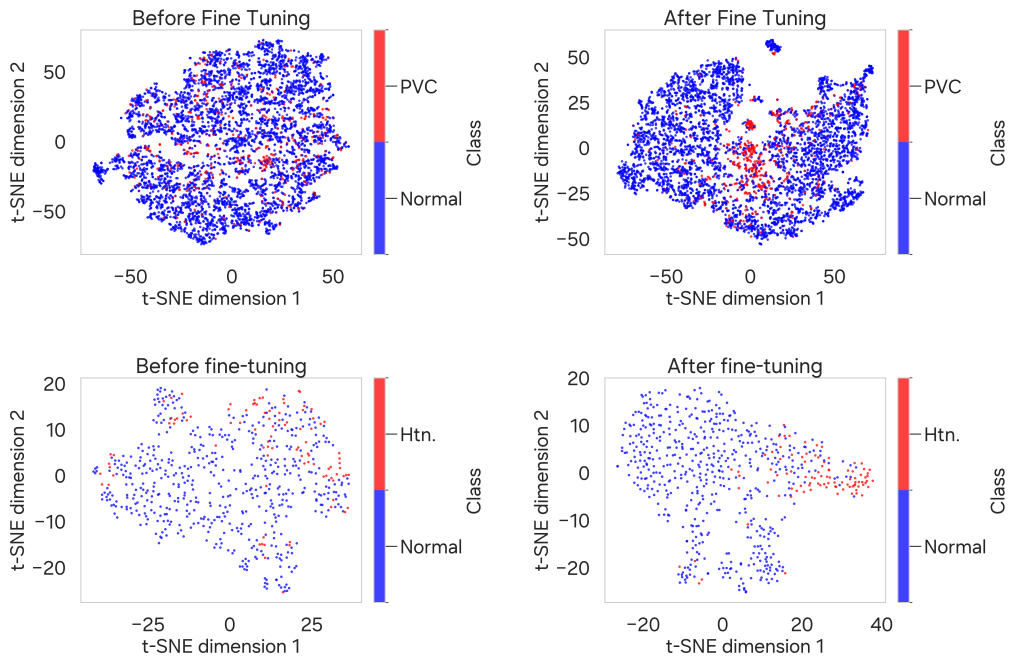


Figure 17: **t-SNE Visualization of Representations Before and After Fine-tuning.** Two representative tasks are shown: premature ventricular contraction (top) and hypertension detection (bottom). Each panel displays a 2D t-SNE projection of HiMAE embeddings colored by class label. Before fine-tuning, the clusters for normal and abnormal cases overlap substantially. After fine-tuning, the separation between classes becomes more pronounced, indicating that task-specific supervision sharpens decision boundaries in the learned representation space.

t-SNE analysis. Figure 17 visualizes embeddings using t-SNE before and after fine-tuning. Prior to fine-tuning, normal and abnormal samples form largely overlapping clusters, indicating that pretraining alone does not fully separate task-specific structure. After fine-tuning, separation between classes becomes more distinct, particularly for PVC detection, suggesting that lightweight task-specific adaptation sharpens decision boundaries while preserving the efficiency of the pretrained HiMAE representations. This confirms that HiMAE provides a strong initialization that benefits from minimal supervised refinement.

G On-device Experiments

G.1 Inference Efficency

We benchmarked the inference efficiency of our proposed HiMAE against the transformer baseline (LSM-Base), measuring three aspects: model footprint and computational complexity in terms of parameters, memory, and FLOPs per 10-second input window at 100 Hz (Table 17); latency, defined as mean per-sample forward-pass time at batch size 1; and throughput, defined as the maximum number of samples processed per second (Table 18). All experiments were run on a Samsung Watch Series 8. Benchmarks were run on-device, using Exynos W1000 CPUs. We also tested on a T4 GPU for potential mobile device deployment; although the T4 is a datacenter GPU, modern mobile processors like the Qualcomm Adreno 750 found on commercial phones are optimized for high-performance ML and can deliver comparable efficiency (Buber & Banu, 2018; Wesolowski et al., 2021), underscoring the practicality of on-device deployment.

Model	Params	FLOPs	Memory
HiMAE	1.2M	0.0647 gFLOPS	4.8 MB
Efficient-Net	7.8M	0.70 gFLOPS	31.1 MB
Swin-Transformer	110.6M	11.89 gFLOPS	423.8 MB
LSM-Base	110.6M	15.94 gFLOPS	441.3 MB

Table 17: **HiMAE is lightweight and efficient:** Model size and compute cost comparison between HiMAE and LSM. FLOPs measured per forward pass on a 10s sequence at 100Hz.

Results Despite being more than two orders of magnitude smaller in parameter count, the HiMAE consistently outperforms the transformer baseline across all efficiency metrics. Between Efficient-Net (Tan & Le, 2020), it remains marginally better which is encouraging due to the optimizations designed in this model.

Model	GPU Lat.	GPU Thr.	CPU Lat.	CPU Thr.
HiMAE	0.039 ms	25.8k/s	0.99 ms	1.2k/s
Efficient-Net	0.082 ms	12.2k/s	1.42 ms	0.704k/s
Swin-Transformer	0.704 ms	1.42k/s	2.95 ms	0.456k/s
LSM-Base	0.80 ms	1.24k/s	3.36 ms	0.298k/s

Table 18: **Inference Performance:** Latency (ms per sample, batch size 2048) and throughput (samples/sec) measured over 10 s windows.

Model footprint: HiMAE reduces parameters from 110M to 0.31M ($\sim 355\times$ fewer), FLOPs from 15.94G to 0.0647G ($\sim 246\times$ fewer), and memory from 441.3MB to 3.6MB ($\sim 123\times$ smaller). These reductions highlight that computational savings scale with the compactness of the model, without loss of representational capacity for the task.

Latency: HiMAE achieves substantially faster per-sample inference. On GPU, latency drops from 0.80ms to 0.039ms ($\sim 20\times$ faster), while on CPU it falls from 3.93ms to 0.99ms ($\sim 4\times$ faster). The reduction in latency follows directly from the smaller computational footprint, reflecting a consistent efficiency advantage.

Throughput: These improvements translate into higher throughput across hardware. On GPU, throughput increases from 1.24k to 25.8k samples/s ($\sim 21\times$ higher), while CPU throughput rises from 0.255k to 1.2k samples/s ($\sim 5\times$ higher). These results confirm that computational gains extend beyond memory and FLOPs, yielding end-to-end speedups at inference time.

In summary, HiMAE achieves a favorable tradeoff between compactness and efficiency, providing lower FLOPs, smaller memory footprint, and faster inference despite its reduced model size. It also outperforms Efficient-Net B1 which was specially designed and optimized for performance and compactness giving a comparison and context to our models performance.

H Additional Theoretical Motivation for HiMAE

H.1 Intuition and Derivation of HiMAE Design

Mathematical walk-through. Let $x \in \mathbb{R}^{C\times L}$ denote a PPG sequence of length $L = 1000$ samples (10 s at 100 Hz). HiMAE partitions x into $N = L/P$ non-overlapping patches of length P, applies a binary mask $m \in \{0,1\}^N$ with masking ratio r, and lifts it to the sample resolution $m' \in \{0,1\}^L$. Training minimizes a reconstruction loss restricted to masked indices M:

$$\mathcal{L}(\theta,\phi) \;=\; \frac{1}{|M|}\sum_{i\in M}\big\|x_i - g_\phi(f_\theta(x\odot m'))_i\big\|^2.$$

Receptive field growth. The encoder f_θ consists of D strided Conv1D stages, each with stride $s_\ell = 2$ and kernel size $k = 5$. Denoting by R_ℓ the receptive field at depth ℓ (measured in input samples), we obtain the recursion

$$R_0 = 1, \qquad R_\ell = R_{\ell-1} + (k-1)\prod_{j=1}^{\ell-1} s_j = R_{\ell-1} + 4\cdot 2^{\ell-1}.$$

Unrolling gives

$$R_\ell = 4\cdot 2^\ell - 3.$$

At $\ell = 4$, $R_4 = 61$ samples (≈ 0.61 s), illustrating that deeper features expand exponentially with $\Theta(2^\ell)$, thereby constructing a natural temporal hierarchy.

Masking scale. Because reconstruction operates on patches, the smallest imputation unit is P. Under i.i.d. masking, the expected run length of a masked region is

$$\mathbb{E}[\ell_{\text{mask}}] = P\cdot\frac{r}{1-r}.$$

For $P = 5$ and $r = 0.8$, this expectation is 20 samples (200 ms). Contiguous masking further increases ℓ_{mask}, shifting the training signal toward meso-scale temporal coherence.

Scale alignment. We define the "challenge band"

$$\mathcal{S} = [\underline{s}, \overline{s}],$$

where $\underline{s} \approx P$ and $\overline{s}$ is set by the typical masked run length. Since encoder receptive fields $\{R_\ell\}$ grow exponentially, effective representation learning occurs when some $R_\ell \in \mathcal{S}$, with shallower layers capturing sub-band morphology and deeper layers integrating slower rhythms. This condition ensures that different depths specialize to distinct physiological scales.

In HiMAE, the chosen hyperparameters ($P = 5$, $k = 5$, stride-2, $r = 0.8$) yield $\mathcal{S}$ on the order of 10^1–10^2 samples, which is naturally bracketed by intermediate layers ($R_3 = 29$, $R_4 = 61$). This alignment explains the resolution-specific probing optima observed empirically.

H.2 Why CNN/U-Net for Physiological Foundation Models?

Transformer-based architectures (Vaswani et al., 2017; Narayanswamy et al., 2024) have emerged as the dominant design choice for modern foundation models. Yet their application to physiological signals reveals critical limitations. Photoplethysmography (PPG), for instance, is highly nonstationary, with morphology that varies across participants. Its dynamics combine quasi-periodic rhythms with subtle aperiodic perturbations arising from arrhythmias, vascular tone, and motion artifacts (Nitzan & Ovadia-Blechman, 2022; Almarshad et al., 2022). Capturing such behavior requires sensitivity to both fine-scale temporal structure and long-range dependencies. Transformers, which lack built-in inductive biases for locality, often force an unfavorable trade-off: small models underfit, while large models rely on brute-force capacity and incur quadratic compute overhead. In contrast, convolutional encoder–decoders such as U-Nets directly encode locality and multiscale structure, providing a more natural choice for compact physiological foundation models.

Local sufficiency for masked prediction. Let $(x_t)_{t=1}^{L}$ denote a physiological time series. Under β-mixing with exponential decay, conditional dependence between x_t and observations outside a finite neighborhood $\mathcal{N}_R(t)$ vanishes rapidly:

$$I\big(x_t; x_{O\setminus\mathcal{N}_R(t)} \,\big|\, x_{O\cap\mathcal{N}_R(t)}\big) \le \varepsilon, \tag{1}$$

for all masked indices $t \in M$, with $R(\varepsilon) = O(\log(1/\varepsilon))$. Thus the Bayes-optimal reconstruction (Bridges et al., 2009) depends only on a finite receptive field. Convolutional architectures implement such translation-equivariant predictors directly as confirmed by many studies (Worrall et al., 2017; Worrall & Brostow, 2018; Kayhan & Gemert, 2020; Zhu et al., 2022). Transformers, by treating all L positions as globally coupled, must simulate locality through restricted attention, leading to inefficiency in both parameters and computation.

Generalization via hypothesis complexity. Denote by $\mathcal{H}_{\mathrm{conv}}$ the class of depth-D CNNs with kernel size k and stride s, and by $\mathcal{H}_{\mathrm{trf}}$ width-d, H-head Transformers. Under spectral norm constraints, convolutional Rademacher complexity (Yin et al., 2019; Truong, 2022) scales as

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{conv}}) \;\lesssim\; \frac{B \cdot kC \sum_{\ell=1}^{D} \prod_{j\le\ell} \|W_j\|_2}{\sqrt{n}}, \tag{2}$$

while self-attention contributes effective rank $\Omega(L)$, yielding

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{trf}}) \;\gtrsim\; \frac{B \cdot \sqrt{H}\, d \sqrt{\log L}}{\sqrt{n}}. \tag{3}$$

Since physiological signals admit local optimal predictors, convolutional models achieve the same approximation error with lower estimation error, giving smaller excess risk under low-data conditions. This explains why U-Net based HiMAE attains competitive or superior downstream accuracy relative to Transformer-based large sequence models, despite using orders of magnitude fewer parameters and FLOPs.

Approximation properties. Physiological signals exhibit multiscale, approximately shift-invariant structure, naturally modeled in Sobolev or Besov spaces. U-Nets implement a multiresolution analysis: downsampling by stride-s grows the receptive field as $R \asymp s^D$, while skip connections preserve fine-scale detail. By analogy with wavelet bases, a depth-D U-Net with $O(D)$ channels achieves approximation error $O(M^{-m})$ with M parameters for functions in $B^m_{2,2}$. In contrast, approximating a local Toeplitz operator with self-attention requires rank $r = \Omega(R)$, rendering Transformers parameter-inefficient for local FIR-like dynamics. This accounts for HiMAE's design choice of a U-Net backbone (Figure 1).

Resolution as an information axis. The resolution hypothesis posits that predictive information in physiological signals is stratified across temporal scales. Formally, let $\mathcal{X}^{(r)}$ denote a representation of the signal at resolution r (e.g., by subsampling or local averaging). The mutual information $I(\mathcal{X}^{(r)}; Y)$ with respect to a downstream target Y is generally non-monotonic in r: fine resolutions retain morphology useful for tasks, while coarse resolutions capture dynamics relevant for circadian rhythms. A model that collapses across resolutions risks discarding scale-specific sufficient statistics.

CNN/U-Nets operationalize this by producing a hierarchy $\{\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(D)}\}$ of embeddings, where each $\mathcal{Z}^{(d)}$ corresponds to receptive field $R_d \asymp s^d$. By the data-processing inequality,

$$I(\mathcal{X}; Y) \;\ge\; I(\mathcal{Z}^{(1)}; Y) \;\ge\; \ldots \;\ge\; I(\mathcal{Z}^{(D)}; Y),$$

but critically, different downstream tasks may maximize $I(\mathcal{Z}^{(d)}; Y)$ at different depths d. This creates a natural testing ground for the resolution hypothesis: if task-specific performance peaks at intermediate $d^\star$, then $I(\mathcal{Z}^{(d^\star)}; Y)$ is locally maximal, showing that neither the finest nor coarsest scale is universally optimal.

Transformers, in contrast, produce globally mixed representations $\mathcal{Z}^{\mathrm{trf}}$ where scale separation is implicit. While these embeddings may approximate $I(\mathcal{X}; Y)$ overall, they do not yield a structured decomposition across resolutions, limiting their utility as discovery tools. By inducing explicit scale-indexed embeddings, CNN/U-Nets make the information–resolution tradeoff observable and probe-able.

Compute–statistical efficiency. For sequence length L, convolutions require $O(Lkd)$ time and $O(Ld)$ memory, while self-attention demands $O(L^2d)$ time and $O(L^2)$ memory. Under fixed hardware, convolution allows larger batches or longer windows, which reduce gradient variance in masked autoencoding pretraining. This directly improves statistical efficiency. In practice, Hi-MAE's convolutional encoder–decoder attains strong downstream generalization while incurring a much smaller compute footprint than Transformer-based LSMs (Table 17).

Theoretical summary. Let $f_{\mathrm{conv}} \in \mathcal{H}_{\mathrm{conv}}$ and $f_{\mathrm{trf}} \in \mathcal{H}_{\mathrm{trf}}$ denote ERM solutions under equal parameter budgets M. Under locality,

$$\mathbb{E}[\mathcal{L}(f_{\mathrm{conv}}) - \mathcal{L}^\star] \leq c_1\varepsilon + c_2\frac{\sqrt{\log L}}{\sqrt{n}}, \tag{4}$$

$$\mathbb{E}[\mathcal{L}(f_{\mathrm{trf}}) - \mathcal{L}^\star] \geq c_3\varepsilon + c_4\frac{\sqrt{\log L}}{\sqrt{n}} + c_5\frac{\sqrt{Hd}}{\sqrt{n}}, \tag{5}$$

with $c_3 > c_1$ unless the number of heads grows with receptive field size R. This analysis suggests that CNN/U-Nets achieve strictly better compute and statistical efficiency for masked-reconstruction on physiological signals, while also aligning with the resolution hypothesis by producing embeddings that explicitly preserve scale-dependent information.