SPOTR: Spatio-temporal Pooling One-Token Reconstruction for Universal Physiological Signal Self-supervised Learning

Yiyu Gui[1,2], **Mingzhi Chen**[1,2], **Yuesheng Zhu**[2], **Guibo Luo**[2*] and **Yuchao Yang**[1*]
[1] New Cornerstone Science Laboratory, Guangdong Provincial Key Laboratory of In-Memory Computing Chips, Shenzhen Graduate School, Peking University, Shenzhen, China
[2] Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University, Shenzhen, China
{luogb, yuchaoyang}@pku.edu.cn

Abstract

Physiological signals such as EEG, ECG, and PPG are widely used in clinical monitoring. Recent self-supervised learning (SSL) methods offer an attractive way to leverage unlabeled recordings, yet they still fall short in practice. In particular, current SSL methods struggle across heterogeneous datasets, often distorting clinically meaningful structures or learning shortcuts from temporal and cross-channel redundancy. Consequently, existing SSL methods often deliver limited performance under linear probing, a lightweight adaptation setting that better matches real-world medical scenarios. Moreover, most Transformer-based SSL models encode a flattened spatiotemporal token sequence, incurring high computation and memory cost, and are typically developed within a single modality. To address these limitations, we present **SPOTR** (**S**patio-temporal **P**ooling **O**ne-**T**oken **R**econstruction), a compress-reconstruct pretraining framework that introduces a single-token global bottleneck for physiological signals. SPOTR compresses each waveform into a single-token representation and reconstructs the signal conditioned only on this representation. Meanwhile, SPOTR introduces an efficient spatio-temporal compaction module to reduce computation and memory cost. Pretrained on **20** datasets spanning EEG, iEEG, ECG, and PPG, SPOTR consistently outperforms the strongest baseline under linear probing, improving average AUC by **18.49%**, **21.71%**, **17.86%**, and **4.64%**, respectively. Compared with a representative general-purpose time-series foundation model, SPOTR achieves around **78%** lower latency and **52%** lower peak GPU memory on average. The code and supplementary material can be found at https://github.com/5GYYYYY/SPOTR.

1 Introduction

Physiological signals such as electroencephalography (EEG), electrocardiography (ECG), and photoplethysmography (PPG) are typically recorded as multi-channel, high-temporal-resolution waveforms, providing essential evidence for characterizing individual physiological states and health variations. Deep learning models have achieved remarkable progress in automated biosignal analysis [Wang *et al.*, 2024; Fan *et al.*, 2025] and have been widely applied to sleep analysis [Eldele *et al.*, 2021; Thapa *et al.*, 2024; Hu *et al.*, 2025], EEG-based seizure detection [Jing *et al.*, 2023; Hogan *et al.*, 2025; Chen *et al.*, 2026], ECG-driven arrhythmia detection [Guhdar *et al.*, 2025; Rahmani *et al.*, 2025], and PPG-based hypertension detection [Lin *et al.*, 2025; El-Dahshan *et al.*, 2024]. Despite their effectiveness, supervised methods often rely on large-scale datasets with high-quality annotations and expert review, which are costly and difficult to scale. To reduce reliance on expensive expert annotations, self-supervised learning (SSL) has made substantial advances leveraging unlabeled biosignals. Current SSL techniques for physiological signals can be broadly grouped into two categories: contrastive learning and masked reconstruction. Contrastive SSL (C-SSL) learns discriminative representations by bringing positives closer while separating negatives [Yang *et al.*, 2023; Pillai *et al.*, 2025; Saha *et al.*, 2025]. Masked reconstruction SSL (MR-SSL), in contrast, learns representations by predicting masked signal segments (or tokens) from unmasked ones [Jiang *et al.*, 2024; Wang *et al.*, 2025; Jin *et al.*, 2025; Na *et al.*, 2024].

C-SSL: Performance Is Sensitive to View Design. Recent C-SSL methods mainly fall into two approaches: augmentation-based contrastive learning (e.g., BIOT [Yang *et al.*, 2023]) and domain-guided contrastive learning (e.g., PaPaGei [Pillai *et al.*, 2025] and Pulse-PPG [Saha *et al.*, 2025]). The former constructs multiple views via augmentations to encourage invariance in the learned representation, as shown in Fig. 1(a). The latter defines positive and negative pairs using task-specific or domain-specific rules and learns representations by pulling positives together while separating negatives, as shown in Fig. 1(b). While effective in specific settings, the learned invariances are highly dependent on pretraining design choices, and even slight changes in augmentation strength, view semantics, or pairing criteria can alter the supervision signal and induce unstable transfer behavior. In practice, this makes C-SSL harder to reliably generalize

*Corresponding authors.

across heterogeneous datasets, where no single augmentation or pairing rule consistently preserves clinically relevant structures.

MR-SSL: Masked Reconstruction Is Prone to Shortcut Learning. Current MR-SSL for biosignals includes reconstructing raw signals (e.g., ST-MEM [Na *et al.*, 2024], CBraMod [Wang *et al.*, 2025], and CSBrain [Zhou *et al.*, 2025]) and reconstructing discrete tokens (e.g., LaBraM [Jiang *et al.*, 2024] and HeartLang [Jin *et al.*, 2025]). Their difference lies in whether the reconstruction target is continuous waveforms (Fig. 1(c)) or tokenized representations (Fig. 1(d)). However, under the mask-then-reconstruct setup, models can often exploit temporal continuity and cross-channel redundancy to recover missing content, creating shortcuts that bypass compact global representation learning. Consequently, the reconstruction objective can be weakly aligned with learning task-relevant, modality-stable structures, allowing pretraining loss to drop quickly while downstream performance improves only marginally.

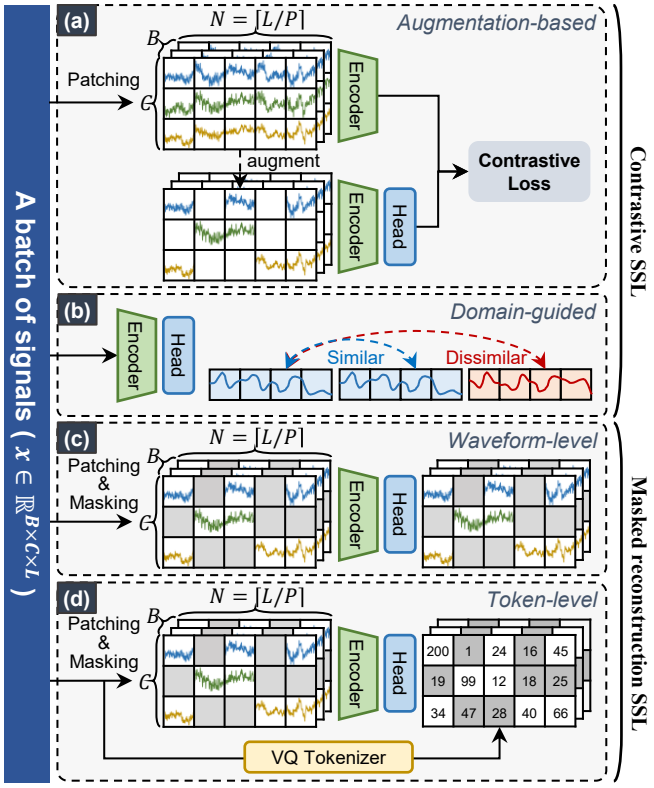

Figure 1: Representative SSL paradigms for physiological signals. (a) Augmentation-based contrastive learning (e.g., BIOT); (b) Domain-guided contrastive learning (e.g., PaPaGei, Pulse-PPG); (c) Waveform-level masked reconstruction (e.g., ST-MEM, CBraMod, CSBrain); (d) Token-level masked reconstruction (e.g., LaBraM, HeartLang).

Taken together, these limitations leave key gaps in generalization for building foundation models. As a result, pretrained representations are often weak under *linear probing*, a lightweight adaptation protocol that better matches real-world medical deployment where labeled data and compute are limited. In addition, existing architectures are frequently computationally heavy. As summarized in Table 1, most Transformer-based SSL models process a flattened spatiotemporal sequence with length on the order of $C\times N$, where C denotes the number of channels and N the number of temporal tokens per channel, leading to high attention cost and memory usage. Finally, most previous studies remain modality-specific in practice, typically tailoring pretraining and evaluation to a single signal type rather than learning a generalizable representation across EEG, ECG, and PPG within one universal model.

Our Approach. To address these gaps, we propose **SPOTR** (**S**patio-temporal **P**ooling **O**ne-**T**oken **R**econstruction), a universal SSL framework built around a *compress–reconstruct* pretraining scheme with a *single-token* global bottleneck. SPOTR compresses each waveform into a single-token representation and reconstructs the signal conditioned *only* on this representation, suppressing shortcut learning and encouraging globally organized, generalizable features that benefit linear probing. To improve efficiency, SPOTR introduces an efficient spatiotemporal compaction module (ST Compactor) that reduces the token sequence length from $C \times N$ to $C+N$ before sequence modeling, substantially lowering computation while preserving spatiotemporal structure. Finally, by adopting a modality-agnostic objective and architecture, we enable unified pretraining across multiple physiological signal types, targeting a universal foundation model that can generalize across modalities and downstream datasets.

We summarize our main contributions as follows:

(1) We introduce SPOTR, a universal SSL framework that learns generalizable representations across physiological signal modalities via a compress–reconstruct pretraining scheme.

(2) We design an efficient spatiotemporal compaction architecture that shortens the encoder token sequence from $C \times N$ to $C+N$, improving training and inference efficiency without sacrificing representation quality.

(3) We demonstrate consistent gains across diverse downstream tasks and modalities, supporting SPOTR as a universal pretraining framework for physiological signals.

2 Method

We present **SPOTR** (**S**patio-temporal **P**ooling **O**ne-**T**oken **R**econstruction), a compress-reconstruct pretraining framework that introduces a single-token global bottleneck for physiological signals. SPOTR comprises an **ST Compactor** $\mathcal{S}$ for spatiotemporal token compaction, a **Latent Aggregator** $\mathcal{A}$ for single-token global fusion, and a **Latent Renderer** $\mathcal{G}$ for bottleneck-conditioned signal reconstruction (Fig. 2).

2.1 SPOTR Training Objective

Physiological signals often exhibit substantial redundancy across both time and channel dimensions. Under masked reconstruction, a decoder may still leverage visible local tokens or neighboring context to predict masked regions, so the reconstruction loss can be dominated by local correlations and

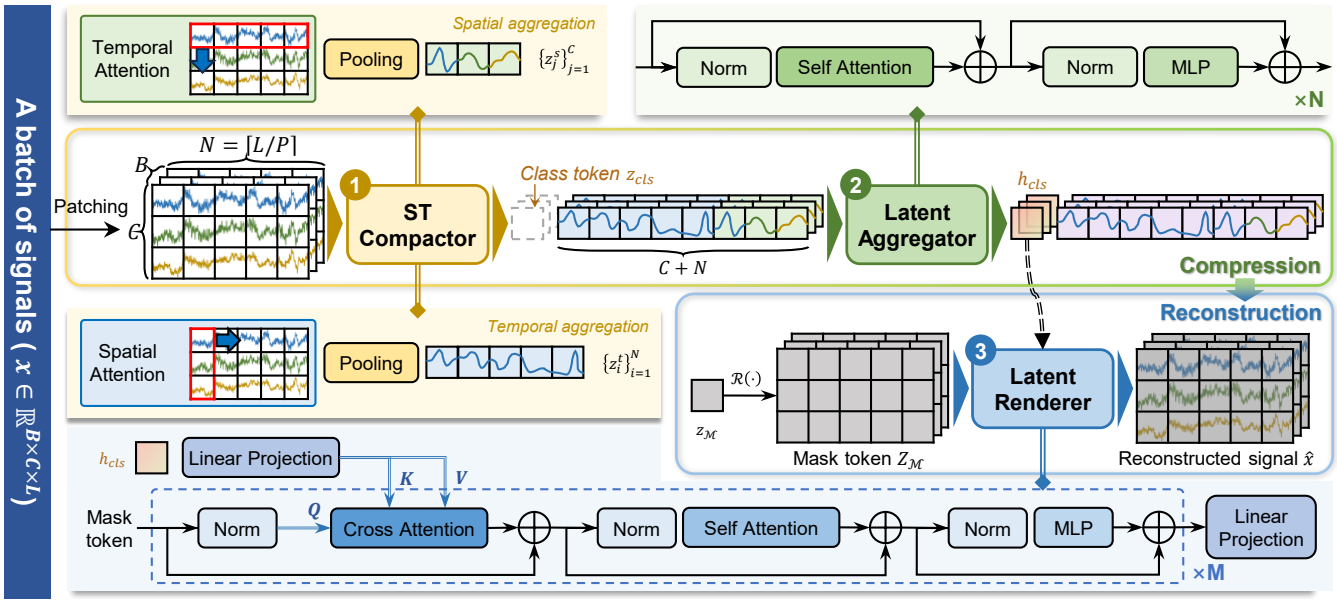

Figure 2: Overview of SPOTR. SPOTR performs compress–reconstruct pretraining with a single-token information bottleneck. The ST Compactor (1) compresses an input waveform into compact temporal tokens and spatial tokens. The Latent Aggregator (2) then fuses both streams into one global class token. For reconstruction, the Latent Renderer (3) starts from mask tokens and conditions the decoder on this single global token through cross-attention, forcing reconstruction to rely on global information rather than local visible context.

common-mode structure—yielding shortcut learning that bypass global representation learning [He *et al.*, 2021; Madan *et al.*, 2023; Wei *et al.*, 2023].

To address this issue and promote a more compact latent representation, SPOTR adopts a **compress–reconstruct pretraining scheme** with an explicit information bottleneck at both the encoder and decoder. The input is first compressed into two short sequences, then aggregated into a single class token as the global representation; reconstruction is conditioned only on this global token.

Let the physiological signal be $x \in \mathbb{R}^{B\times C\times L}$, where B is the batch size, C is the number of channels, and L is the number of sample points. We denote the ST Compactor, Latent Aggregator, and Latent Renderer by $\mathcal{S}(\cdot)$, $\mathcal{A}(\cdot)$, and $\mathcal{G}(\cdot)$, respectively. With patch size P, we define the number of temporal tokens as $N = \lceil \frac{L}{P} \rceil$.

(i) Spatiotemporal Compaction. The ST Compactor compresses the input into temporal tokens and spatial tokens, and uses learnable 2D positional encodings $PE_{2D}^{\mathcal{A}}$ to mark the relative position of each channel–time token:

$$Z = \mathcal{S}(x) + PE_{2D}^{\mathcal{A}} = [z_1^t,\ z_2^t, ...,\ z_N^t; z_1^s,\ z_2^s, ...,\ z_C^s] \in \mathbb{R}^{B\times(N+C)\times D_{\mathcal{A}}} \quad (1)$$

where $D_{\mathcal{A}}$ is the hidden dimension of the Latent Aggregator, z_i^t is the i-th temporal token, and z_j^s is the j-th spatial token.

(ii) Single-token Global Aggregation. We add a learnable class token z_{cls} and take the first output position as the global representation:

$$h_{\text{cls}} = A\big([z_{\text{cls}}; Z]\big)[:, 0:1] \in \mathbb{R}^{B\times 1\times D_{\mathcal{A}}} \quad (2)$$

Here h_{cls} serves as the only information bottleneck in SPOTR, responsible for capturing global structure across both time and channels. With this design, the reconstruction is no longer conditioned on a large set of visible local tokens; instead, it relies solely on a fixed-size global representation, imposing a stronger constraint on representation compression and information organization.

(iii) Mask-token Sequence and Conditional Rendering. To reconstruct the signal, we build a mask-token sequence aligned with the target spatiotemporal grid. Let $z_{\mathcal{M}} \in \mathbb{R}^{1\times 1\times D_{\mathcal{G}}}$ be a learnable mask token, and let $\mathcal{R}(\cdot)$ be a repeat operator that expands a single token to $B \times C \times N$ positions and flattens them into a sequence:

$$Z_{\mathcal{M}} = \mathcal{R}(z_{\mathcal{M}}; B, C, N) + \text{PE}_{\text{2D}}^{\mathcal{G}} \in \mathbb{R}^{B\times(CN)\times D_{\mathcal{G}}} \quad (3)$$

where $D_{\mathcal{G}}$ is the hidden dimension of the Latent Renderer and $PE_{2D}^{\mathcal{G}}$ is the decoder-side 2D positional encoding. The renderer conditions on h_{cls} and outputs the reconstruction:

$$\hat{x} = \mathcal{G}(Z_{\mathcal{M}}, h_{\text{cls}}) \in \mathbb{R}^{B\times C\times L} \quad (4)$$

(iv) Reconstruction Loss. Pretraining is optimized by minimizing the reconstruction error, and the objective can be formulated as the following empirical risk minimization problem:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\|\hat{x} - x\|_2^2\right] \quad (5)$$

where p_{data} is the pretraining data distribution and $\hat{x}$ is the reconstructed signal produced end-to-end by the ST Compactor, Latent Aggregator, and Latent Renderer. By minimizing the reconstruction loss under a single-token information bottleneck, this objective drives the model to learn global representations that are both compact and reconstructive. Unlike contrastive learning, which relies on constructing positive and negative pairs to form training signals, the reconstruction objective uses the input itself as supervision, thereby avoiding

the uncertainty introduced by negative-sample selection [Wei *et al.*, 2023].

2.2 ST Compactor: Compressing Spatiotemporal Structure

The ST Compactor $\mathcal{S}(\cdot)$ consists of a channel-independent convolutional patch embedding layer $\mathcal{C}$ followed by two orthogonal attention-based aggregation paths. The input signal $x \in \mathbb{R}^{B\times C\times L}$ is first segmented into temporal patches of length P and projected into patch tokens:

$$U = \mathcal{C}(x) \in \mathbb{R}^{B\times C\times N\times D_{\mathcal{A}}} \tag{6}$$

Two attention operations are then applied:

(i) Spatial Token Aggregation. For each channel j, temporal self-attention is applied over its N patches, followed by temporal pooling:

$$z_j^s = pool_t(MHA_t(U_{:,j,:,:})) \in \mathbb{R}^{B\times D_{\mathcal{A}}},\ j = 1, 2, ..., C \tag{7}$$

(ii) Temporal Token Aggregation. For each temporal token i, channel-wise self-attention aggregates information across channels, followed by spatial pooling:

$$z_i^t = pool_s(MHA_s(U_{:,:,i,:})) \in \mathbb{R}^{B\times D_{\mathcal{A}}},\ i = 1, 2, ..., N \tag{8}$$

This design reduces the subsequent sequence length from $C \times N$ to $C + N$, preserving spatiotemporal structure while significantly lowering computation.

2.3 Latent Aggregator: Aggregating into a Single Token

The Latent Aggregator $\mathcal{A}(\cdot)$ is stacked Transformer encoder blocks that jointly encode Z and aggregate it into the class token, yielding h_{cls}. By using the class token as the sole global representation throughout the network, $\mathcal{A}$ enforces a persistent single-token bottleneck, encouraging compact and transferable representations.

We adopt a pre-norm design with RMSNorm for training stability. The feed-forward network uses SwiGLU:

$$FFN_{SwiGLU}(x) = (Swish(W_g x) \odot (W_1 x))W_2 \tag{9}$$

where $\odot$ denotes the Hadamard product and W_g, W_1, W_2 are learnable projection matrices. This design effectively enhances the model's ability to conduct feature selection and knowledge encoding under limited parameters.

2.4 Latent Renderer: Rendering Signals from a Single Token Bottleneck

The Latent Renderer $\mathcal{G}$ reconstructs the full physiological signal by conditioning on the highly compressed global latent h_{cls} and progressively "rendering" the mask-token sequence $Z_{\mathcal{M}}$ into the signal domain. Concretely, $\mathcal{G}$ consists of an input projection $\mathcal{P}_{in} \in \mathbb{R}^{D_{\mathcal{A}}\times D_{\mathcal{G}}}$, an output projection $\mathcal{P}_{out} \in \mathbb{R}^{D_{\mathcal{G}}\times P}$, and a stack of Transformer decoder blocks. Here, $\mathcal{P}_{in}$ maps the conditioning vector to the decoder hidden space, while $\mathcal{P}_{out}$ regresses each token representation to a length-P local segment, which is then reassembled to form the reconstructed signal.

Different from standard Transformer decoders that apply self-attention before cross-attention, we adopt a cross-attention-first update order, so that each position reads from the global condition h_{cls} prior to any intra-sequence interaction. Specifically, let the input to the k-th decoder block be $Q^{(k-1)} \in \mathbb{R}^{B\times(CN)\times D_{\mathcal{G}}}$, $\widetilde{h} = \mathcal{P}_{in}(h_{cls}) \in \mathbb{R}^{B\times 1\times D_{\mathcal{G}}}$. We first inject the global condition via cross-attention:

$$\widetilde{Q}^{(k)} = Q^{(k-1)} + MHA(Q^{(k-1)}, \widetilde{h}, \widetilde{h}) \tag{10}$$

Next, we perform self-attention over the mask-token sequence to enable information exchange across positions:

$$\bar{Q}^{(k)} = \widetilde{Q}^{(k)} + MHA(\widetilde{Q}^{(k)}, \widetilde{Q}^{(k)}, \widetilde{Q}^{(k)}) \tag{11}$$

Finally, the resulting features are passed through a feed-forward network with a residual connection:

$$Q^{(k)} = \bar{Q}^{(k)} + FFN_{SwiGLU}(\bar{Q}^{(k)}) \tag{12}$$

After stacking K decoder blocks, we obtain $Q^{(K)}$. The final reconstruction is produced by projecting each position back to a length-P segment and reshaping tokens into the original spatiotemporal layout:

$$\hat{x} = Reshape(\mathcal{P}_{out}(Q^{(K)})) \in \mathbb{R}^{B\times C\times L} \tag{13}$$

This design forces all local reconstructions to be mediated by the same global condition h_{cls}, thereby preventing the mask tokens from completing reconstruction solely via local correlations.

3 Experimental Setup

We pretrain and evaluate models under the SPOTR framework on multi-modal physiological signal datasets, and refer to the resulting family of models as **SPOTR**.

3.1 Baselines

As summarized in Table 1, we compare **SPOTR** with representative SSL baselines. For each modality, we include (i) a general-purpose time-series SSL foundation model, **MOMENT**, and (ii) domain-specific SSL models. Specifically, for EEG/iEEG we use **LaBraM**, **CBraMod**, and **CSBrain**; for ECG we use **ST-MEM** and **HeartLang**; and for PPG we use **PaPaGei** and **Pulse-PPG**. We do not include **BIOT** [Yang *et al.*, 2023] in our comparison because, although it claims applicability to multiple physiological modalities, only pretrained EEG weights are publicly available. More details are provided in the Appendix.

3.2 Datasets

Pretraining Datasets

We curate a large-scale pretraining dataset covering 4 physiological modalities, including scalp EEG, intracranial EEG (iEEG), ECG, and PPG. Overall, the pretraining data collection contains **20 datasets**, including **6** EEG datasets [Miltiadous *et al.*, 2023; Kemp *et al.*, 2000; Quan *et al.*, 1997; Shoeb, 2009; Van Dijk *et al.*, 2022; Khan *et al.*, 2022], **2** iEEG datasets [Li *et al.*, 2021; Burrello *et al.*, 2019], **8** ECG datasets [Gow *et al.*, 2023; Zheng *et al.*, 2022; Liu *et al.*,

Model	Signal Type				Pretraining Paradigm		Model Architecture	Number of tokens	Adapt to any number of channels
	iEEG	EEG	ECG	PPG					
LaBraM (ICLR [Jiang *et al.*, 2024])	✓	✓			Fig. 1(d)	MR-SSL	Transformer	C×N	✓
ST-MEM (ICLR [Na *et al.*, 2024])			✓		Fig. 1(c)	MR-SSL	Transformer	C×N (C=12)	
MOMENT (ICML [Goswami *et al.*, 2024])	✓	✓	✓	✓	Fig. 1(c)	MR-SSL	Transformer	C×N	✓
CBraMod (ICLR [Wang *et al.*, 2025])	✓	✓			Fig. 1(c)	MR-SSL	Transformer	C×N	✓
HeartLang (ICLR [Jin *et al.*, 2025])			✓		Fig. 1(d)	MR-SSL	Transformer	C×N (C=12)	
PaPaGei (ICLR [Pillai *et al.*, 2025])				✓	Fig. 1(b)	C-SSL	ResNet	NA*	
Pulse-PPG (UBiComp [Saha *et al.*, 2025])				✓	Fig. 1(b)	C-SSL	ResNet	NA*	
CSBrain (NeurIPS [Zhou *et al.*, 2025])	✓	✓			Fig. 1(c)	MR-SSL	Transformer	C×N	✓
SPOTR (ours)	✓	✓	✓	✓	Fig. 2	SPOTR	Transformer	C+N	✓

Table 1: Overview of baselines. * NA indicates that the "Number of tokens" column is not applicable or non-Transformer architectures.

Signal modality	Dataset	Task	Label	# Channels	Duration	# Subjects
iEEG	Mayo [Nejedly *et al.*, 2020]	Intermittent epileptiform discharges (IEDs) detection	2-class	1	3s	15
	FNUSA [Nejedly *et al.*, 2020]	Intermittent epileptiform discharges (IEDs) detection	2-class	1	3s	12
EEG	ISRUC [Khalighi *et al.*, 2016]	Sleep Stage Classification	5-class	2	30s	100
	Siena [Detti *et al.*, 2020]	Seizure detection	2-class	29	4s	13
	NTUH-BIS [Liu *et al.*, 2018]	Anesthesia status detection	4-class	1	30s	23
	MDD [Wajid, 2016]	Major depressive disorder detection	2-class	19	5s	64
	Schizo-28 [Olejarczyk and Jernajczyk, 2017]	Schizophrenia detection	2-class	19	10s	28
ECG	CPSC2018 [Ng *et al.*, 2018]	Heart disease detection	9-class	12	10s	6877
	PTB-sub [Wagner *et al.*, 2020]	Heart disease detection (Fine-grained)	23-class	12	10s	18885
	PTB-super [Wagner *et al.*, 2020]	Heart disease detection (Coarse-grained)	5-class	12	10s	18885
	PTB-rhythm [Wagner *et al.*, 2020]	Heart disease detection (Rhythm)	12-class	12	10s	18885
	Europe ST-T [Taddei *et al.*, 1992]	Heartbeat Classification	5-class	2	1s	90
PPG	PPG-BP [Liang *et al.*, 2018]	Hypertension detection	2-class	3	2.1s	219
	PPG-Dalia [Reiss *et al.*, 2019]	Human activity recognition	9-class	1	8s	15
	WESAD_v [Schmidt *et al.*, 2018]	Emotion classification	2-class	1	10s	15
	WESAD_a [Schmidt *et al.*, 2018]	Emotion classification	2-class	1	10s	15
	CLBP [Kachuee *et al.*, 2015]	Hypertension detection	2-class	1	10s	12

Table 2: Overview of downstream datasets. "# Channels" and "# Subjects" denote the number of channels and subjects in the dataset.

2022; Kalyakulina *et al.*, 2020; Bousseljot *et al.*, 1995; Tihonenko *et al.*, 2008; Alday *et al.*, 2020; Ribeiro *et al.*, 2021], **3** PPG datasets [Mehrgardt *et al.*, 2022; Nemcova *et al.*, 2021; Garde *et al.*, 2014], and **one** multimodal dataset [Lee *et al.*, 2022]. In total, it provides **over 17 million signal samples** from **more than 450,000 subjects**, spanning diverse acquisition settings. More details are provided in the Appendix.

Downstream Datasets

As shown in Table 2, we conduct a comprehensive evaluation of SPOTR on **17** downstream datasets spanning multiple physiological signal modalities, including iEEG, EEG, ECG, and PPG. For all the downstream datasets, we resampled the signals to 200 Hz and applied a 50/60 Hz notch filter (depending on the acquisition region) to suppress power-line interference. More details are provided in the Appendix.

3.3 Implementation Details

All experiments were conducted on a Linux server with an NVIDIA A100 80GB GPU, using PyTorch 2.0.1 and CUDA 12.2. For pretraining, we used AdamW with $\beta = (0.9, 0.999)$, weight decay 0.1, and a peak learning rate of 2×10^{-4}. Each dataset was trained with a per-step batch size of 256 and gradient accumulation of 20, resulting in an effective batch size of 5120. We pretrained for 3 epochs with a learning-rate schedule including a 10% warm-up; bfloat16 precision was used for computational efficiency.

For downstream evaluation, we used subject-independent train/validation/test splits for all tasks and report the mean and standard deviation of test-set performance over 5 independent runs with different random seeds. We kept the optimizer as AdamW with the same β and weight decay, but used a constant learning rate of 2×10^{-4} and a batch size of 128. All downstream runs were trained for up to 200 epochs with early stopping based on validation AUC (patience = 20 epochs). Unless otherwise specified, downstream training was performed in float32 for stable optimization.

4 Results

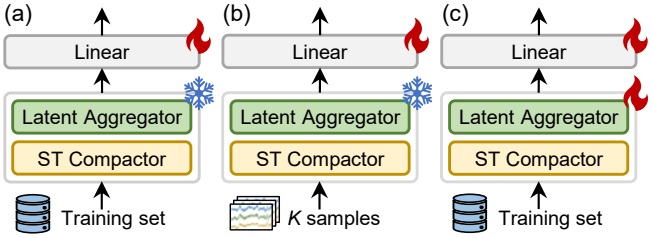

Figure 3: Experimental setup.

We evaluate three adaptation protocols: **(i) Linear probing**, where the pretrained backbone is frozen and only a single linear classification head is trained (Fig. 3(a)); **(ii) Few-shot classification**, where we construct a support set by randomly sampling $K \in 1, 2, 4, 8$ labeled examples per class from the training set, freeze the backbone, and train only the same linear head (Fig. 3(b)); and **(iii) Full fine-tuning**, where we train the backbone and the classification head under the same downstream configuration (Fig. 3(c)). In all the experiments, we use AUC as the evaluation metric.

Model	Dataset / Applicable scope	Mayo mean	Mayo std.	FNUSA mean	FNUSA std.	Ave
iEEG						
MOMENT	*Time-series*	50.46	3.29	50.02	2.83	50.24
LaBraM	*EEG/iEEG-only*	63.69	4.60	69.32	5.78	66.51
CBraMod	*EEG/iEEG-only*	46.65	17.67	62.83	4.16	54.74
CSBrain	*EEG/iEEG-only*	59.64	2.67	55.91	0.64	57.78
SPOTR	*Physiological signals*	**88.92**$_{\uparrow 25.23}$	0.06	**87.52**$_{\uparrow 18.20}$	0.16	**88.22**$_{\uparrow 21.71}$

Model	Dataset / Applicable scope	ISRUC mean	ISRUC std.	Siena mean	Siena std.	NTUH-BIS mean	NTUH-BIS std.	MDD mean	MDD std.	Schizo-28 mean	Schizo-28 std.	Ave
EEG												
MOMENT	*Time-series*	71.16	0.89	62.45	5.17	55.09	2.17	79.13	4.25	68.73	8.85	67.31
LaBraM	*EEG/iEEG-only*	81.63	2.43	66.87	9.39	63.67	4.87	81.23	4.15	53.58	5.56	69.40
CBraMod	*EEG/iEEG-only*	77.20	1.22	58.25	2.86	54.53	1.59	90.03	1.37	68.04	6.16	69.61
CSBrain	*EEG/iEEG-only*	60.57	2.27	55.68	3.38	54.75	1.26	78.37	3.26	59.47	4.61	61.77
SPOTR	*Physiological signals*	**92.83**$_{\uparrow 11.21}$	0.03	**76.87**$_{\uparrow 10.00}$	0.90	**87.69**$_{\uparrow 24.02}$	0.13	**95.31**$_{\uparrow 5.28}$	0.12	**87.79**$_{\uparrow 19.06}$	8.87	**88.10**$_{\uparrow 18.49}$

Model	Dataset / Applicable scope	CPSC2018 mean	CPSC2018 std.	PTB-sub mean	PTB-sub std.	PTB-super mean	PTB-super std.	PTB-rhythm mean	PTB-rhythm std.	Europe ST-T mean	Europe ST-T std.	Ave
ECG												
MOMENT	*Time-series*	58.90	0.92	62.44	1.07	61.99	1.76	63.07	2.94	68.87	1.40	63.05
ST-MEM	*ECG-only*	53.56	0.38	56.44	1.25	56.98	1.40	55.89	0.96	51.02	2.25	54.78
HeartLang	*ECG-only*	69.87	0.39	77.31	0.28	77.44	0.33	66.65	0.40	56.38	1.21	69.53
SPOTR	*Physiological signals*	**92.33**$_{\uparrow 22.46}$	0.03	**90.21**$_{\uparrow 12.90}$	0.02	**89.10**$_{\uparrow 11.66}$	0.01	**82.84**$_{\uparrow 16.19}$	0.04	**82.47**$_{\uparrow 13.60}$	0.19	**87.39**$_{\uparrow 17.86}$

Model	Dataset / Applicable scope	PPG-BP mean	PPG-BP std.	PPG-Dalia mean	PPG-Dalia std.	WESAD$_v$ mean	WESAD$_v$ std.	WESAD$_a$ mean	WESAD$_a$ std.	CLBP mean	CLBP std.	Ave
PPG												
MOMENT	*Time-series*	47.14	14.74	68.07	2.56	67.63	2.47	61.68	3.44	52.98	0.40	59.50
PaPaGei	*PPG-only*	**55.97**	14.29	51.47	0.33	53.85	2.99	47.53	3.67	49.81	0.22	51.73
Pulse-PPG	*PPG-only*	46.58	13.24	57.28	1.37	49.31	8.17	45.88	3.96	51.85	0.74	50.18
SPOTR	*Physiological signals*	51.99$_{\downarrow 3.98}$	8.26	**82.12**$_{\uparrow 14.05}$	0.05	**69.03**$_{\uparrow 1.40}$	0.49	**62.06**$_{\uparrow 0.38}$	0.19	**55.52**$_{\uparrow 2.54}$	0.25	**64.14**$_{\uparrow 4.64}$

Table 3: Linear-probing performance comparison. Bold denotes the best result and underline denotes the second best. For SPOTR, subscripts report the absolute AUC gap relative to the strongest baseline: ↑ denotes the improvement over the second-best model when SPOTR attains the top performance, whereas ↓ denotes the gap to the best-performing model otherwise.

4.1 Linear Probing

Table 3 summarizes linear-probing performance, which directly reflects the quality of pretrained representations under lightweight adaptation.

On iEEG, SPOTR achieves the best performance with an average AUC of 88.22%, surpassing the strongest iEEG/EEG SSL baseline by over 20%. On EEG, SPOTR is consistently superior across the five datasets, reaching an average AUC of 88.10% and exceeding EEG-specific SSL baselines. On ECG, SPOTR again delivers the strongest results, achieving an average AUC of 87.39% and improving over the best ECG-specific SSL baseline by 17.86%. On PPG, SPOTR attains the highest average AUC of 64.14%, outperforming PPG-specific baselines (PaPaGei and Pulse-PPG) on most datasets. Since MOMENT is evaluated on all datasets, we further summarize results over all 17 datasets in Table 3. SPOTR achieves an overall mean AUC of **80.86%**, exceeding MOMENT by **19.11%**.

Across four physiological modalities, SPOTR consistently surpasses both the general time-series baseline and modality-specific SSL models in the linear-probing setting, supporting its practicality as a universal and lightweight-transfer pretraining framework for physiological signals.

4.2 Few-shot Classification

As shown in Fig. 4, under 1/2/4/8-shot settings, SPOTR achieves the best or near-best AUC on four representative datasets, each corresponding to a different modality: ISRUC (EEG), Mayo (iEEG), CPSC2018 (ECG), and PPG-Dalia (PPG). Across all modalities, SPOTR exhibits a stable and consistent improvement as more labeled examples are provided, indicating that it can make good use of very limited labeled data. Moreover, the relatively tight performance dispersion across runs suggests robust adaptation under limited supervision. Overall, these results support SPOTR for practical few-shot scenarios, where annotation is scarce but reliable adaptation is still needed.

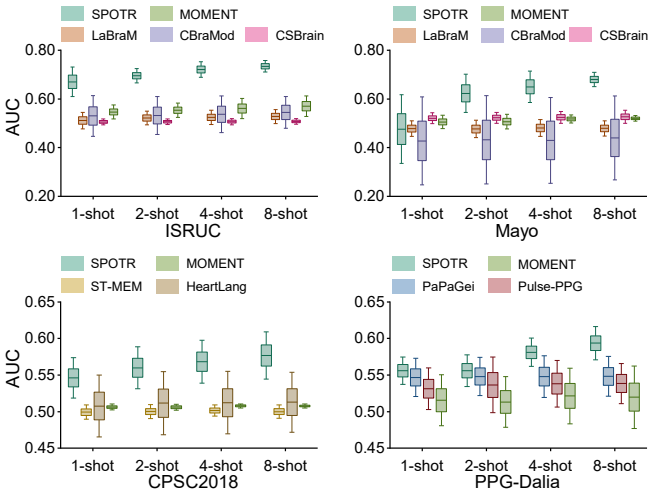

Figure 4: Few-shot classification results. Boxes show the 25–75th percentiles with the median line; whiskers indicate the spread across repeated runs under each 1/2/4/8-shot setting.

4.3 Full Fine-tuning

As shown in Table 4, under the full-parameter fine-tuning setting, SPOTR remains competitive across all four modalities. It achieves the best average performance on iEEG (90.25%) and EEG (89.61%), and shows the clearest advantage on ECG with the highest mean score (92.70%). On PPG, SPOTR is overall comparable to the strongest PPG-specific baseline, with only a marginal difference in average performance. Overall, even when end-to-end fine-tuning reduces the gaps among pretraining paradigms, SPOTR still provides robust transfer across modalities and datasets.

iEEG					
Dataset / Model	**Mayo** mean	std.	**FNUSA** mean	std.	Ave
MOMENT	**89.69**	0.03	89.11	0.01	89.40
LaBraM	86.20	1.23	87.71	1.58	86.96
CBraMod	88.17	1.79	87.08	2.23	87.63
CSBrain	87.63	1.51	87.22	9.56	87.43
SPOTR	89.28$_{\downarrow 0.41}$	1.32	**91.21**$_{\uparrow 2.10}$	0.81	**90.25**$_{\uparrow 0.85}$
EEG					
Dataset / Model	**ISRUC** mean	std.	**Siena** mean	std.	Ave
MOMENT	87.91	0.01	83.29	1.13	85.60
LaBraM	93.52	0.27	77.95	1.20	85.74
CBraMod	**96.98**	0.23	74.04	3.83	85.51
CSBrain	93.46	0.20	57.82	4.86	75.64
SPOTR	95.46$_{\downarrow 1.54}$	0.16	**83.76**$_{\uparrow 0.47}$	3.01	**89.61**$_{\uparrow 3.87}$
ECG					
Dataset / Model	**CPSC2018** mean	std.	**Europe ST-T** mean	std.	Ave
MOMENT	79.48	1.47	85.79	0.63	82.64
ST-MEM	74.64	1.08	82.34	2.18	78.49
HeartLang	94.46	0.19	69.60	2.77	82.03
SPOTR	**95.45**$_{\uparrow 0.99}$	0.39	**89.94**$_{\uparrow 4.15}$	2.07	**92.70**$_{\uparrow 10.06}$
PPG					
Dataset / Model	**PPG-Dalia** mean	std.	**WESAD$_v$** mean	std.	Ave
MOMENT	80.20	0.03	68.22	0.13	74.21
PaPaGei	82.41	0.23	66.04	2.81	74.23
Pulse-PPG	**84.14**	0.48	67.43	3.10	**75.79**
SPOTR	82.30$_{\downarrow 1.84}$	0.81	**68.87**$_{\uparrow 0.65}$	1.97	75.59$_{\downarrow 0.20}$

Table 4: Full fine-tuning performance comparison.

4.4 Efficiency

In the efficiency evaluation, we compare SPOTR against the general time-series foundation model MOMENT across 17 downstream datasets, reporting the mean results over four metrics: Latency (p50/p95), Throughput, and Peak GPU Memory. As shown in Table 5, SPOTR achieves substantially better inference efficiency. Specifically, the average p50 and p95 latency are reduced from 32.61 ms and 35.48 ms to 7.15 ms and 7.69 ms, corresponding to 78.1% and 78.3% lower latency, respectively. Meanwhile, throughput increases from 45.66 to 139.84 samples/s, yielding a 206.3% improvement. SPOTR also reduces peak GPU memory consumption from 532.09 MB to 256.39 MB (51.8% lower). These results demonstrate that SPOTR delivers markedly faster and more memory-efficient inference, highlighting its practicality for resource-constrained deployment. More details are provided in the Appendix.

	p50 Lat.↓	p95 Lat.↓	Thrpt.↑	Mem.(MB)↓
MOMENT	32.61	35.48	45.66	532.09
SPOTR	**7.15**	**7.69**	**139.84**	**256.39**
Δ(%)	-78.1%	-78.3%	+206.3%	-51.8%

Table 5: Efficiency comparison.

4.5 Scaling Model Size

As shown in Fig. 5, increasing the model size from SPOTR-Small (3.69M parameters) to SPOTR-Medium (13.93M) and SPOTR-Base (62.47M) leads to a clear scaling trend in pretraining. Larger models drop the loss faster early on and settle at a lower value, suggesting stronger capacity and easier optimization. The same pattern appears in linear probing. SPOTR-Base gives the best AUC on iEEG, EEG, ECG, and PPG, followed by SPOTR-Medium and then SPOTR-Small. Overall, scaling up the model improves both reconstruction training and representation transfer across signal types, which supports further scaling of model and data.

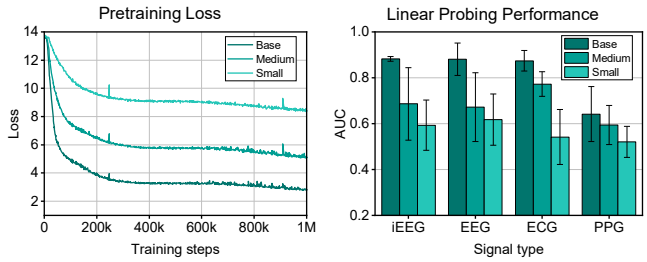

Figure 5: Scaling behavior of SPOTR with model size. Left: pretraining loss curves for different model sizes. Right: linear-probing AUC across four modalities improves consistently with model size.

5 Conclusion

In this paper, we presented SPOTR, a universal self-supervised learning framework for physiological signals that learns generalizable representations across modalities via a compress–reconstruct pretraining scheme. By compressing each input into one representation and reconstructing the signal only from this bottleneck, SPOTR discourages shortcut learning and promotes compact global representations. SPOTR further improves efficiency by shortening the token sequence before modeling. Extensive experiments demonstrate consistent gains in cross-modality generalization, with particularly strong improvements under linear probing, supporting lightweight and practical adaptation in real-world clinical scenarios, while reducing inference latency and peak GPU memory compared with a representative general-purpose time-series foundation model.

Acknowledgements

This work has been supported by Guangdong S&T Program (2025B0101140001, 2026B0101070006), Guang-

dong Provincial Key Laboratory of In-Memory Computing Chips (2024B1212020002), Shenzhen Science and Technology Program (ZDCY20250901103401002, JCYJ20241202125907011), and Beijing Natural Science Foundation (L234026, L257010). This work has been supported by the New Cornerstone Science Foundation and Financial Support for Outstanding scientific and technological innovation Talents Training Fund in Shenzhen.

References

[Alday *et al.*, 2020] Erick A Perez Alday, Annie Gu, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

[Bousseljot *et al.*, 1995] Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Type: dataset*, 1995.

[Burrello *et al.*, 2019] Alessio Burrello, Kaspar Schindler, et al. Hyperdimensional computing with local binary patterns: One-shot learning of seizure onset and identification of ictogenic brain regions using short-time ieeg recordings. *IEEE Transactions on Biomedical Engineering*, 67(2):601–613, 2019.

[Chen *et al.*, 2026] Mingzhi Chen, Yiyu Gui, et al. Ssddb: A semantic-structural dual-drive pretraining framework for brain signals. In *ICASSP*, pages 6496–6500. IEEE, 2026.

[Detti *et al.*, 2020] Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 2020.

[El-Dahshan *et al.*, 2024] El-Sayed A. El-Dahshan, Mahmoud M. Bassiouni, et al. Exhyptnet: An explainable diagnosis of hypertension using efficientnet with ppg signals. *Expert Systems with Applications*, 239:122388, 2024.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Zhenghua Chen, et al. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[Fan *et al.*, 2025] Wei Fan, Jingru Fei, et al. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *WWW*, page 5054–5064, 2025.

[Garde *et al.*, 2014] Ainara Garde, Parastoo Dehkordi, et al. Development of a screening tool for sleep disordered breathing in children using the phone oximeter™. *PLoS ONE*, 9, 2014.

[Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, et al. MOMENT: A family of open time-series foundation models. In *ICML*, pages 16115–16152, 2024.

[Gow *et al.*, 2023] Brian Gow, Tom Pollard, et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6:13–14, 2023.

[Guhdar *et al.*, 2025] Mohammed Guhdar, Abdulhakeem Mohammed, and Ramadhan J. Mstafa. Advanced deep learning framework for ecg arrhythmia classification using 1d-cnn with attention mechanism. *Knowl. Based Syst.*, 315:113301, 2025.

[He *et al.*, 2021] Kaiming He, Xinlei Chen, et al. Masked autoencoders are scalable vision learners. *CVPR*, pages 15979–15988, 2021.

[Hogan *et al.*, 2025] Robert Hogan, Sean R. Mathieson, et al. Scaling convolutional neural networks achieves expert level seizure detection in neonatal eeg. *NPJ Digital Medicine*, 8, 2025.

[Hu *et al.*, 2025] Shuaicong Hu, Yanan Wang, et al. Xsleepfusion: A dual-stage information bottleneck fusion framework for interpretable multimodal sleep analysis. *Information Fusion*, 123:103275, 2025.

[Jiang *et al.*, 2024] Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *ICLR*, pages 16405–16426, 2024.

[Jin *et al.*, 2025] Jiarui Jin, Haoyu Wang, et al. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *ICLR*, pages 8207–8227, 2025.

[Jing *et al.*, 2023] Jin Jing, Wendong Ge, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

[Kachuee *et al.*, 2015] Mohamad Kachuee, Mohammad Mahdi Kiani, et al. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. *ISCAS*, pages 1006–1009, 2015.

[Kalyakulina *et al.*, 2020] Alena Kalyakulina, Igor Yusipov, et al. Lobachevsky university electrocardiography database. *Type: Dataset.*, 2020.

[Kemp *et al.*, 2000] Bob Kemp, Aeilko H Zwinderman, et al. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[Khalighi *et al.*, 2016] Sirvan Khalighi, Teresa Sousa, et al. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–92, 2016.

[Khan *et al.*, 2022] Hassan Aqeel Khan, Rahat Ul Ain, et al. The nmt scalp eeg dataset: An open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

[Lee *et al.*, 2022] Hyung-Chul Lee, Yoonsang Park, et al. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

[Li *et al.*, 2021] Adam Li, Chester Huynh, et al. Neural fragility as an eeg marker of the seizure onset zone. *Nature neuroscience*, 24(10):1465–1474, 2021.

[Liang *et al.*, 2018] Yongbo Liang, Zhencheng Chen, et al. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific Data*, 5, 2018.

[Lin *et al.*, 2025] Hui Lin, Jiyang Li, et al. Longitudinal wrist ppg analysis for reliable hypertension risk screening using deep learning. In *ICASSP*, pages 1–5, 2025.

[Liu *et al.*, 2018] Quan Liu, Li Ma, et al. Sample entropy analysis for the estimating depth of anaesthesia through human eeg signal at different levels of unconsciousness during surgeries. *PeerJ*, 6, 2018.

[Liu *et al.*, 2022] Hui Liu, Dan Chen, et al. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific data*, 9(1):272, 2022.

[Madan *et al.*, 2023] Neelu Madan, Nicolae-Cătălin Ristea, et al. Cl-mae: Curriculum-learned masked autoencoders. *WACV*, pages 2480–2490, 2023.

[Mehrgardt *et al.*, 2022] Philip Mehrgardt, Matloob Khushi, et al. Pulse transit time ppg dataset. *PhysioNet*, 10:e215–e220, 2022.

[Miltiadous *et al.*, 2023] Andreas Miltiadous, Katerina D Tzimourta, et al. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

[Na *et al.*, 2024] Yeongyeon Na, Minje Park, et al. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *ICLR*, pages 15012–15035, 2024.

[Nejedly *et al.*, 2020] Petr Nejedly, Václav Kremen, et al. Multicenter intracranial eeg dataset for classification of graphoelements and artifactual signals. *Scientific Data*, 7, 2020.

[Nemcova *et al.*, 2021] Andrea Nemcova, Radovan Smisek, et al. Brno university of technology smartphone ppg database (but ppg). *PhysioNet*, 101:e215–e220, 2021.

[Ng *et al.*, 2018] Eddie Y. K. Ng, Feifei Liu, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 2018.

[Olejarczyk and Jernajczyk, 2017] Elzbieta Olejarczyk and Wojciech Jernajczyk. Graph-based analysis of brain connectivity in schizophrenia. *PLoS ONE*, 12, 2017.

[Pillai *et al.*, 2025] Arvind Pillai, Dimitris Spathis, et al. Papagei: Open foundation models for optical physiological signals. In *ICLR*, pages 48230–48261, 2025.

[Quan *et al.*, 1997] Stuart F Quan, Barbara V Howard, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[Rahmani *et al.*, 2025] Sana Rahmani, Reetam Chatterjee, et al. Dynamic prototype rehearsal for continual ecg arrhythmia detection. *ICASSP*, pages 1–5, 2025.

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

[Ribeiro *et al.*, 2021] Antônio H. Ribeiro, Gabriela M. M. Paixão, et al. Code-15%: a large scale annotated dataset of 12-lead ecgs. *Zenodo, Jun*, 2021.

[Saha *et al.*, 2025] Mithun Saha, Maxwell A Xu, et al. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–35, 2025.

[Schmidt *et al.*, 2018] Philip Schmidt, Attila Reiss, et al. Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.

[Shoeb, 2009] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[Taddei *et al.*, 1992] Alessandro Taddei, Giovanni Distante, et al. The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography. *European heart journal*, pages 1164–72, 1992.

[Thapa *et al.*, 2024] Rahul Thapa, Bryan He, et al. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. In *ICML*, pages 48019–48037, 2024.

[Tihonenko *et al.*, 2008] V Tihonenko, A Khaustov, et al. St petersburg incart 12-lead arrhythmia database. *PhysioBank PhysioToolkit and PhysioNet*, 2008.

[Van Dijk *et al.*, 2022] Hanneke Van Dijk, Guido Van Wingen, et al. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific data*, 9(1):333, 2022.

[Wagner *et al.*, 2020] Patrick Wagner, Nils Strodthoff, et al. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7, 2020.

[Wajid, 2016] Mumtaz Wajid. Mdd patients and healthy controls eeg data (new). *figshare, Dataset*, 2016.

[Wang *et al.*, 2024] Yihe Wang, Nan Huang, et al. Medformer: A multi-granularity patching transformer for medical time-series classification. In *NeurIPS*, pages 36314–36341, 2024.

[Wang *et al.*, 2025] Jiquan Wang, Sha Zhao, et al. Cbramod: A criss-cross brain foundation model for eeg decoding. In *ICLR*, pages 75310–75346, 2025.

[Wei *et al.*, 2023] Chen Wei, Karttikeya Mangalam, et al. Diffusion models as masked autoencoders. *ICCV*, pages 16238–16248, 2023.

[Yang *et al.*, 2023] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *NeurIPS*, pages 78240–78260, 2023.

[Zheng *et al.*, 2022] Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet*, 23:7, 2022.

[Zhou *et al.*, 2025] Yuchen Zhou, Jiamin Wu, et al. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *arXiv preprint arXiv:2506.23075*, 2025.