Xiao Gu[1], Yuxuan Shu[1], Jinpei Han[1], Yuxuan Liu[1], Zhangdaihong Liu[1], James Anibal[1], Veer Sangha[1], Edward Phillips[1], Bradley Segal[1], Yuxuan Liu[1], Hang Yuan[1], Fenglin Liu[1], Kim Branson[1], Patrick Schwab[1], Danielle Belgrave[1], Lei Clifton[1], Dimitris Spathis[1], Vasileios Lampos[1], A Aldo Faisal[1], and David A Clifton[1]

[1]Affiliation not available

August 24, 2025

Foundation Models for Biosignals: A Survey

Xiao Gu†, Yuxuan Shu†, Jinpei Han†, Yuxuan Liu, Zhangdaihong Liu, James Anibal, Veer Sangha, Edward Phillips, Bradley Segal, Yuxuan Liu, Hang Yuan, Fenglin Liu, Kim Branson, Patrick Schwab, Danielle Belgrave, Lei Clifton, Dimitris Spathis, Vasileios Lampos, A. Aldo Faisal, David A. Clifton

Abstract—Foundation models, which are neural networks pretrained on large-scale data, have emerged as a powerful paradigm, in languages, images, and increasingly general time series, for learning generalizable representations across diverse tasks. As this paradigm extends to biomedical domain, there is growing interest in developing foundation models for biosignals, including inertial measurement unit (IMU), electrocardiography (ECG), electroencephalogram (EEG), photoplethysmogram (PPG), and various wearable sensing signals. These biosignals form the physiological and behavioral "languages" of the human body, carrying rich diagnostic and prognostic value across a range of clinical settings. However, unlike natural language or vision, efforts to build foundation models for biosignals lack a cohesive roadmap to organize diverse modalities, design strategies, and optimize pretraining paradigms. In this survey, we focus on three converging directions shaping the landscape of biosignal foundation models: (i) *training from scratch using large biosignal datasets* (ii) *adapting general time series models to biomedical domain* (iii) *leveraging (multi-modal) large language models for biosignal analysis.* These directions either aim to build dedicated biosignal foundation models or repurpose advances in vision, language, and general time series modeling to meet biomedical needs. Together, they highlight complementary opportunities for clinically meaningful biosignal modeling. With further challenges and open research directions, this work provides a comprehensive roadmap to building foundation models that meet the unique demands of biomedical sensing.

Index Terms—Foundation models, Biosignals, Time series analysis, Digital health, Biomedical sensing, Large language models.

1 Introduction

Biomedical sensing is a fundamental source of health informatics. It captures the intrinsic physiological, physical, and behavioral "languages" of the human body. These "languages" offer valuable insights into our internal functioning, enabling both individuals and healthcare professionals to understand what is happening within the body and to characterize how we behave and interact with the world around us. These time series signals, from physiological signals, such as electrocardiography (ECG), photoplethysmography (PPG), electroencephalography (EEG), to motion and acoustic recordings, such as inertial measurement unit (IMU), span a wide spectrum of health informatics. Over the past decades, research efforts have focused on analyzing and interpreting these biomedical signals using digital tools, leading to the development of a wide range of biomedical signal processing and analysis approaches.

As digital health technologies are becoming increasingly pervasive and biomedical sensing devices more affordable and accessible [1], the volume of biomedical data has grown exponentially. Meanwhile, novel applications of these sensing informatics are explored across various specialized medical fields and levels of care. This rapid expansion highlights the urgent need for scalable, flexible, and generalizable modeling frameworks capable of addressing the complexity and diversity inherent in biosignals and healthcare.

Foundation models, which are networks pretrained on large-scale, diverse datasets typically via self-supervised learning and subsequently adaptable to a broad range of downstream tasks [2], have revolutionized the landscape of machine learning. This paradigm breaks away from the traditional one-task one-dataset one-model approach, offering a more scalable and generalizable framework. Conceptually, this aligns well with the increasing volume, diversity, and complexity of biomedical sensing data and associated tasks, making them particularly suitable for addressing the diverse analytical needs emerging in this field.

Thus far, foundation models have achieved significant progress in the realm of natural language processing (NLP) and computer vision (CV), showcasing remarkable capabilities in handling a wide range of tasks. Transformative work like GPT series [3], LLaMA series [4], [5], and SAM [6], have demonstrated superior performance in scalability and generalizability. Particularly, large language models (LLMs) and their multimodal extensions (e.g., vision-language models) have achieved significant breakthroughs in integrating and interpreting multimodal data by leveraging pretrained knowledge. These successes have recently inspired similar developments in general time series analysis, extending the foundation model paradigm to sequential data beyond text and images [7], [8].

This has catalyzed a growing body of research working on foundation models for biosignals. These efforts typically

†*X.G., Y.S, and J.H. contributed equally to this work.*
X.G., Z.L., J.A., V.S., E.P, B.S, F.L, D.A.C are from Department of Engineering Science, University of Oxford, UK.
Y.S, V.L. are from Centre for Artificial Intelligence, Department of Computer Science, University College London, UK.
J.H., Y.L., A.A.F. are from Department of Computing, Imperial College London, UK.
Y.L. is from School of BME, Shanghai Jiao Tong University, China.
Y.H. is from Big Data Institute and Nuffield Department of Population Health, University of Oxford, Oxford, UK.
K.B, P.S, D.B. are from GlaxoSmithKline, UK.
L.C is from Nuffield Department of Parimary Care Health Sciences, University of Oxford, UK.
D.S. is from Google Research, UK and the University of Cambridge, UK.

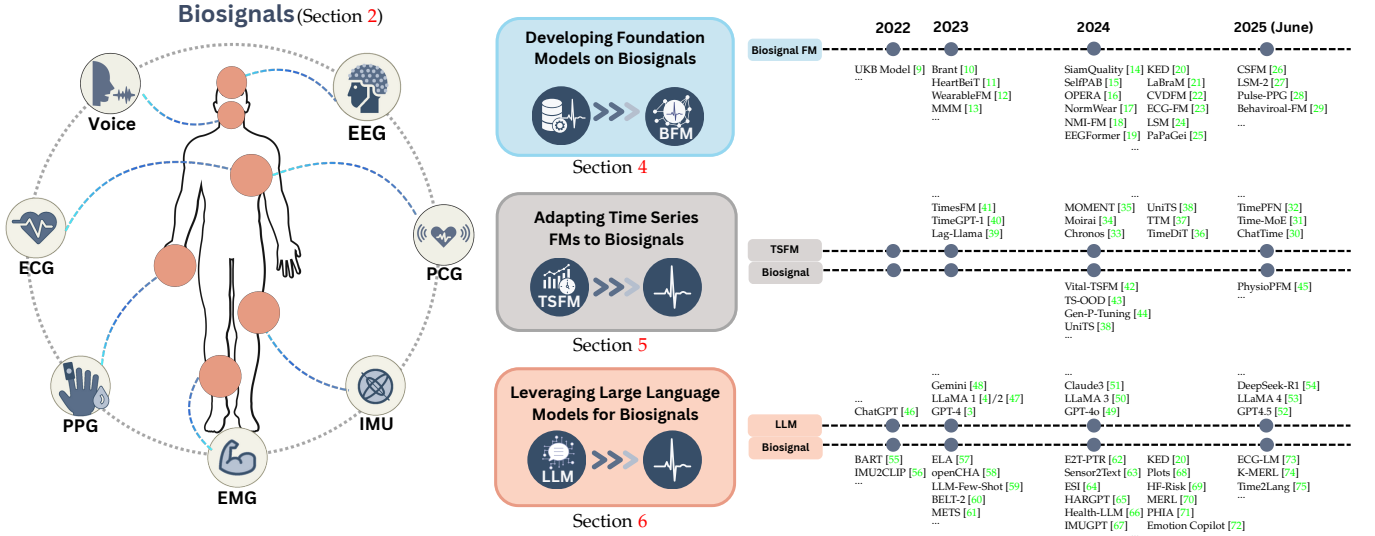

Fig. 1: Overview of three directions in foundation models for biosignals. Left: Common biosignals such as ECG, EEG, and PPG. Center: Three strategic modeling directions—(i) developing biosignal foundation models (BFMs) from scratch on large biosignal corpora, (ii) adapting general time series foundation models (TSFMs) to biosignals, and (iii) leveraging large language models (LLMs) for biosignal interpretation. Right: Timeline of representative models from 2022 to June 2025, sorted by first online dates. A curated list is available at https://github.com/guxiao0822/awesome-biosignal-foundation-model for community contributions and ongoing updates.

follow established paradigms for training foundation models on text and images, or leverage existing pretrained models from language, vision, or general time series domains and adapt them for the unique characteristics of biosignals. Despite this growing interest, there lack of a cohesive roadmap that summarizes and examines this rapidly developing field. Unlike NLP or CV domains where foundational pipelines have matured, biosignal foundation modeling still lacks standardized practices for data curation, pretraining objectives, model architecture, evaluation protocols, etc.

In light of this, there is a pressing need for a unified and structured overview that consolidates recent advances, identifies common challenges, and outlines principled strategies for building and adapting foundation models for biosignals. While recent surveys have explored related aspects of foundation models in health and medicine, none offer a dedicated or in-depth treatment of biosignal modeling as a distinct and complex domain. Specifically:

First, surveys on *general time series foundation models* (TSFMs) [7], [8] either treat biosignals as a minor subsection, or organize their concepts without considering the unique challenges associated with biosignals. While biosignals are indeed time series, they differ significantly from typical time series data (e.g., financial data) in terms of data processing, methodological design, task settings, and deployment requirements. These demand specialized strategies that are largely absent from general-purpose time series surveys.

Second, surveys on *biomedical foundation models* [76], [77] have primarily focused on modalities such as medical imaging and clinical text (e.g., radiology reports, clinical notes). While these reviews provide valuable insights into large-scale modeling in healthcare, they offer little guidance on how foundation models can be designed for biosignals.

To mitigate this gap, this paper serves as a survey that combines conceptual clarity, practical modeling strategies, and domain-specific insights, for biosignal foundation models. We organize the discussion around three major directions in the development, adaptation, and leveraging of foundation models for biosignals, as illustrated in Figure 1:

- **Developing foundation models from scratch on large biosignal corpora.** We present a comprehensive roadmap for developing foundation models from scratch, tailored specifically to biosignals. This not only serves as a practical training recipe but also highlights the unique challenges of biosignal data, with potential solutions.
- **Adapting general time series foundation models to biosignals.** The emergence of foundation models for general-purpose time series forecasting offers new opportunities for biosignal modeling. We examine how these models can be adapted to accommodate the structure, semantics, and constraints of biosignal data.
- **Leveraging emerging (multi-modal) large language models for biosignals.** The rapid advancement of LLMs and their extensions to multimodal domains, particularly vision-language models, has opened new possibilities for biosignal interpretation. In this direction, we outline the diverse functional roles that LLMs can play in biomedical signal processing and analysis.

Following the introduction, in Section 2, we review commonly used biosensing modalities, outline their distinct modeling challenges, and summarize open-access datasets that support model development. Section 3 categorizes the major downstream tasks in biosignal analysis, ranging from low-level signal denoising and imputation to high-level predictive modeling and outcome forecasting. Finally, in Sections 4, 5, and 6, we present the three primary development pathways, followed by the challenges and future directions in biosignal foundation models in Section 7.

2 Overview of Biosignal Data Types

The development of biosignal foundation models is highly dependent on the nature ofthe training data. These data types can be broadly categorized into raw waveforms, which capture direct physiological or mechanical measurements, and numeric health metrics that represent derived or integrated health indicators, as in Figure 2.

2.1 Raw Waveforms

Raw waveform data represents the direct, unprocessed, continuous recordings of biological, physiological, and physical processes captured at high temporal resolution. These can be categorized into two main non-exclusive types: *physiological signals* that measure biological electrical activity and blood flow, *biomechanical signals* that capture physical movement and force, or record biological sounds.

2.1.1 Physiological signals

Physiological signals encompass the direct measurement of the body's electrical activity or hemodynamic responses.

Electrocardiography (ECG) records the heart's electrical activity via electrodes placed on the body surface. The signal exhibits a quasi-periodic pattern with distinct waveform components (P, QRS, T) within a typical frequency range of 0.05–50Hz. Clinical ECG is commonly acquired using a 12-lead configuration, offering spatial information critical for diagnosing conduction abnormalities and regional pathologies. Ambulatory monitors (e.g., Holter) and wearable devices (e.g., smartwatches) provide reduced-lead recordings (1–3 leads), enabling long-term monitoring but with limited spatial resolution. Clinical ECG datasets typically contain 10-second, 12-lead recordings with structured diagnostic annotations [78], and they can be linked to auxiliary records such as comorbidities, textual reports, or mortality outcomes [79]. In contrast, wearable datasets provide longer recordings with variable conditions, often lacking standardized lead placement and ground-truth diagnostic labels.

Photoplethysmography (PPG) captures blood-volume changes using optical sensors and is commonly used in wearables for heart rate, respiratory, and oxygen saturation monitoring. PPG signals contain components at multiple frequency bands (e.g., cardiac: 0.5–2Hz), with desired recording duration dependent on the task. For example, shorter windows for heart rate estimation and longer for capturing slower physiological rhythms [80], [81]. PPG is sensitive to various factors (e.g. sensor placement, skin tone, device geometry), leading to inter-subject variability and noise [82]. Furthermore, the lack of standardized sensor montage across various sensor placement sites (e.g., wrist, finger, ear) introduces waveform variability.

Electroencephalography (EEG) captures the brain's electrical activity via scalp-mounted electrodes, with signal characteristics heavily influenced by both the electrode configuration and neural dynamics [83]. EEG spans multiple frequency bands: delta (∼0.5-4Hz), theta (∼4-8Hz), alpha (∼8-13Hz), beta (∼13-30Hz), and gamma (>30Hz), each linked to distinct brain states and cognitive functions. This calls for frequency-domain transformation strategies for preprocessing and feature extraction. Clinical EEG commonly follows the 10–20 system for electrode placement (channel configurations may vary across studies). EEG data are widely used in tasks such as sleep staging, emotion recognition, and neurorehabilitation.

Electromyography (EMG) measures muscle activation by recording electrical potentials generated during muscle contractions. Surface EMG is most commonly used in non-invasive settings and can be configured as either bipolar recordings or high-density electrode arrays. The high-density setting captures richer spatial patterns, but introduces greater variations in channel count and spatial configuration across studies [84]. Its frequency content usually ranges from 20 to 250Hz, or higher (500Hz [85]). This calls for higher sampling rates (∼1000Hz) and increasing storage and computational efficiency demands. In practice, EMG datasets lack a standardized montage and vary widely in electrode placement, number of channels, and muscle groups recorded. While guidelines such as SENIAM [86] offer muscle-specific recommendations, sensor placement is often adapted per participant due to anatomical variability.

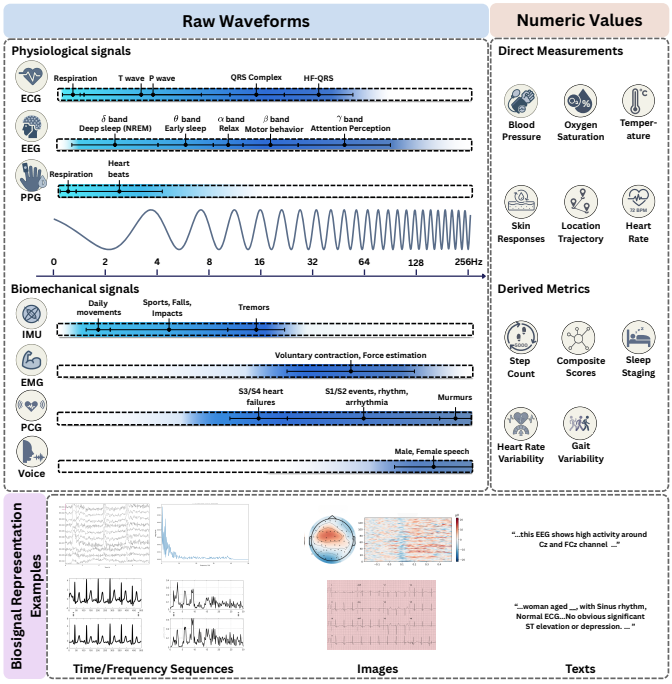

Fig. 2: Illustration of representative biosignals and corresponding input representations used in foundation models. Commonly used biosignals are categorized into raw waveforms (top left, along with their typical frequency ranges of interest) and numeric values (top right). At the bottom, we highlight how biosignals like ECG and EEG can be represented as sequences, images, or texts for model input.

2.1.2 Biomechanical signals

Mechanical signals in biomedical sensing capture physical movements, forces, and displacements produced by the body. They provide crucial information about physical activity, movement disorders, respiratory function, overall mobility patterns, etc., in both clinical and daily life settings.

Inertial measurement units (IMUs) use tri-axial accelerometers, gyroscopes, and sometimes magnetometers for motion tracking. The frequency content of IMU signals depends on the activity: from walking (1–2Hz), to fine motor tasks (up to 10–20Hz [97]). Sensor placement and configuration vary significantly across applications. Trunk-mounted IMUs capture gross body movement, while limb-mounted or wrist-worn devices record finer motion patterns. Moreover, the number of sensors varies from a single IMU in wrist-worn sensors [9] to multi-IMUs used in clinical gait analysis.

Bioacoustic signals refer to physiological sounds generated by the body, including heart sounds (often represented by Phonocardiogram (PCG)), respiratory sounds, cough sounds, and speech (typically recorded as general audio signals). These signals cover a wide frequency spectrum. For example, heart sounds typically lie between 20-150 Hz,

TABLE 1: Overview of large-scale datasets that are available for training biosignal foundation models from large biosignal corpora. Datasets are grouped by modality and ordered by scale. # Individuals indicates the number of unique subjects available and duration refers to total cumulative signals hours, where available. Studies are reported in previous usage of biosignal modeling or foundation model development.

Datasets	Data Types		Additional Information	# Individuals	# Duration (hr)	Used Studies
	Waveforms	Numerics				
UK Biobank	ECG, PPG, IMU	health metrics	linked clinical database	-	-	[9], [69], [87]
MC-MED	ECG, PPG, Resp	vital signs	linked clinical database	70K	-	[88]
MIMIC-III-WDB	ECG, PPG, ABP, Resp	vital signs	linked clinical database	30K	3M	[25]
VitalDB	ECG, PPG, ABP, Resp	vital signs	linked clinical database	6K	-	[25]
PulseDB	ECG, PPG, ABP	-	-	5K	50M	[14]
MESA	ECG, PPG, EEG	-	sleep stage label	2k	-	[25]
VTaC	ECG, PPG, ABP	-	false ICU alarm label	2K	-	[89]
CODE	ECG	-	disease and mortality label (partial)	2M	-	[90]
MIMIC-IV-ECG	ECG	-	linked clinical database, report	160K	2K	[23]
PhysioNet2020	ECG	-	disease label, report (PTB-XL)	40k	90	[91]
eICU	-	vital signs	linked clinical database	139K	-	[42]
HiRID	-	vital signs	linked clinical database	34K	-	[92]
UCSF-PPG	PPG	-	-	21K	600K	[14]
TUEG	EEG	-	seizure detection label	15K	27k	[13], [91], [93]
MOABB	EEG	-	brain-computer interface label	>1K	-	[94]
SEED Series	EEG	gaze metrics (partial)	emotion label (partial)	-	-	[13], [21]
emg2pose	EMG	-	gesture label	193	370	[95]
emg2qwerty	EMG	-	typing label	108	346	[96]
HUNT4	IMU	-	linked clinical database	35K	-	[15]
Capture24	IMU	-	human activity label (partial)	151	4K	[9]
Ego4D	IMU, Audio	gaze metrics	action labels	923	4k	[56]
COVID-19 Sounds	Audio	-	self-reported disease label	36K	552	[16]

while respiratory events and speech can extend into the kilohertz range [98]. The characteristics of these signals are highly influenced by sensor placement, surrounding acoustic conditions, and individual anatomy, leading to substantial variation across datasets.

2.2 Numeric Values

Modern biomedical sensing systems provide not only raw waveform signals but also more abstract readings. These can be broadly categorized into "first-order" **direct measurements**, which are device readings with minimal processing, and "second-order" **derived health metrics**, which require more complex analysis or integration of multiple signals.

Direct measurements include core vital signs such as heart rate (HR), blood pressure, oxygen saturation (SpO_2), respiratory rate (RR), and body temperature and are widely used in both hospital monitoring and home-based care. Consumer wearables also report activity-related metrics like movement intensities, derived from inertial or GPS sensors through lightweight firmware-level processing. Emerging sensors also allow for direct monitoring of skin conductance, hydration, and glucose levels, enabling continuous and non-invasive tracking in real-world settings.

Derived health metrics, on the other hand, require more sophisticated post-processing, often combining multiple signals or longer temporal windows. For instance, heart rate variability (HRV) and ECG morphology metrics (e.g., QRS duration, QT interval) involve statistical or signal-domain feature extraction beyond the capabilities of standard hardware outputs. In the domain of mobility, indices like step counts, gait stability, and step regularity rely on analyzing temporal patterns of movement data and are increasingly used in geriatric and neurological assessments.

2.3 Open-Sourced Biosignal Datasets

Building on the above discussions, we summarize representative open-source datasets in Table 1, with an emphasis on large-scale, population-level datasets and curated collections that are relevant for training and evaluating foundation models. Their availability may be subject to specific data agreements. For each dataset, we also indicate the associated metadata or auxiliary resources, such as linkage to clinical databases and annotations, and highlight benchmark studies or prior work that have utilized the dataset.

2.4 Biosignal Representations

On the other hand, as preliminaries, we discuss here the input format of biosignals to computational models. They are originally represented as raw time series. We define an input time series in shape $\mathbb{R}^{C\times T}$, where C denotes the number of channels, and T is the number of time steps. The specific format in which time-series data are fed into computational models can vary considerably across different applications and signal types. Below, we outline common practices to structure this data, with examples shown in Figure 2.

Time and/or frequency sequences. Most existing research uses raw time series data as model input. In contrast, some studies use specialized biomedical signal processing to transform time series into frequency-domain (e.g., via Fourier transforms) or time-frequency domain (e.g., via wavelet transforms) representations, which are particularly effective for signals with complex, nonlinear patterns and dynamics, such as EEG and EMG. For example, Yi *et al.* [13] transformed EEG into differential entropy feature vectors to capture essential frequency-based characteristics. Zhang *et al.* [99] incorporated both temporal and frequency inputs for biosignals to enforce cross-domain consistency. Regardless of whether the signal is raw or transformed, the model input remains an ordered sequence of values.

Images. Beyond sequence-based input, signals can be transformed into images for model input through various methods: (i) *Temporal-Frequency Transformation*: Many signals (e.g. EEG and voice recordings) contain rich time-frequency information, making spectrogram particularly effective [15], [16], [22]; (ii) *Topographical Transformation*: In signals like EEG or high-density EMG, spatially informed channels can be mapped into image-like "topographs". These layouts help localize spatially specific information, providing deeper insight into underlying physiological processes; (iii) *Plot*: Additionally, time-domain signals (e.g. ECG and PPG) can be directly plotted as images, facilitating the use of advanced image-based architectures to capture their underlying temporal patterns [68], [100].

An added benefit of this approach is that it enables the use of pretrained vision-based and vision-language models with minimal architectural changes by converting time series into compatible image inputs [68].

Texts. With the advent of LLMs, which operate on natural language inputs, some approaches encode time-series data or their statistical summaries as text [71], [101]. This allows LLMs to process and interpret numerical sequences in a unified manner along with conventional language data.

3 Overview of Biosignal Tasks

Building on the discussion of data types and representations, we examine the analytical tasks that underpin biosignal understanding, which we group into 3 categories: (i) signal enhancement and augmentation – improving signal quality, utility, and diversity; (ii) signal interpretation and pattern recognition – extracting patterns from complex time series; (iii) predictive modeling and outcome forecasting – linking biosignals to health/behavioral trajectories for prognosis, risk forecasting, and long-term outcome prediction.

Figure 3 illustrates these typical tasks, and highlights their relationships to representative tasks in general time series analysis (i.e., forecasting, imputation, classification, and anomaly detection [102]). It can be observed that time series modeling tasks in the biomedical domain partially overlap with general-purpose formulations. However, biosignal tasks are fundamentally grounded in physiological and clinical contexts, requiring task definitions and evaluation criteria that extend beyond those used in general ones.

3.1 Signal Enhancement and Augmentation

One foundational task for biosignal analysis is to improve data integrity through imputation, augmentation, and restoration. This is because biosignal data often suffers from poor quality, continuity, and completeness, whether from clinical monitors or wearables.

Meanwhile, there also exists data collection bias, such as the underrepresentation of rare diseases or minority populations. To address this, low-level signal modeling is used to denoise corrupted inputs, impute missing segments, or synthesize one modality from another.

Data imputation and restoration. Improving signal quality is essential in biosignal processing, as sensor artifacts, movement, or environmental interference often lead to missing or incomplete recordings. Traditional imputation methods use statistical interpolation or models trained on clean data. Contrarily, foundation models leverage pretraining objectives to perform more robust imputation [38], such as masking and reconstructing signals across datasets/modalities.

Data generation and augmentation. Biosignal datasets often lack diversity in subjects and conditions, particularly in real-world or resource-constrained settings, making data generation and augmentation crucial. Traditional augmentations like jittering or scaling may lack physiological realism. Generative models address this by synthesizing condition-specific and realistic waveforms [103], [104], enhancing dataset diversity, especially for rare conditions. This approach has been extended to multimodal contexts, enabling cross-modal reconstruction. For example, models can infer one modality from another, such as generating medical imaging from biosignal [105], and from PPG to ECG [106].

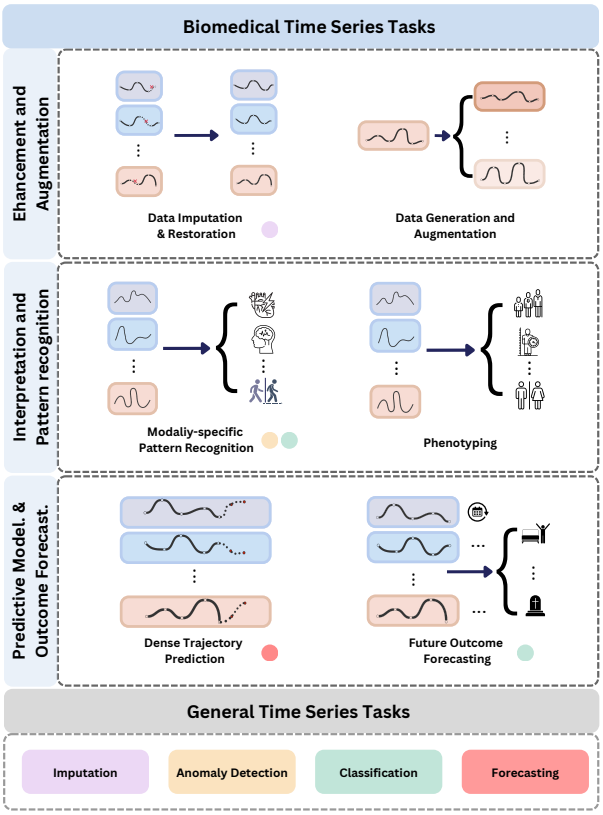

Fig. 3: Illustration of representative biomedical time series tasks and their distinctions from general time series analysis tasks. We annotate tasks that share (partial) overlap with general time series objectives using corresponding colored dots next to each biomedical time series task. For instance, ECG signal restoration conceptually overlaps with the general task of imputation (purple).

3.2 Signal Interpretation and Pattern Recognition

Pattern recognition is a core task for biosensing models, spanning modality-specific tasks and general phenotyping. Foundation models can learn to extract and interpret complex, clinically or behaviorally relevant patterns from high-dimensional, often noisy biomedical time series.

Modality-specific pattern recognition tasks. Most tasks in this category involve classification, with some regression tasks, and their specificity is closely related to the signal modality. For example in ECG, foundation models have shown competitive performance in detecting various cardiac pathologies [11], [22], [23]. Similarly, EEG-based foundation models also show promise in diagnosing neurological and mental health disorders [107], and decoding cognitive and affective states [13], [19], [21]. For wearable IMU data, activity recognition [15], [66], [68] and fall detection [68] remain common tasks supporting broad healthcare applications. In PPG and multimodal wearables, recent work focuses on estimating physiological signals (e.g. heart rate and respiratory rate [14], [108]) and more holistic health metrics like fatigue, stress, or sleep quality [66].

Phenotyping tasks. Foundation models can extract high-level representations related to physiological or behavioral traits. Phenotyping often involves inferring underlying health states or patient subtypes, including identifying digital biomarkers (e.g. age, sex) and uncovering previously unrecognized clinical subgroups, offering insights into disease

heterogeneity. Studies using large-scale wearable datasets have shown that foundation models can predict a wide array of health [109] and demographic variables (e.g., age, gender, and BMI) [108], capturing patterns that would be inaccessible to single-task or modality-constrained models.

3.3 Predictive Modeling and Outcome Forecasting

Another key task of biosignals is predictive modeling, which uses current or historical biosignal data to forecast future events or states. This includes tasks like early warning system development and survival analysis, where the output labels correspond to outcomes that occur after the input window, rather than concurrently. The relevant tasks can be broadly categorized into two types based on prediction horizons and data resolutions.

Dense trajectory prediction. The first type involves forecasting fine-grained, high-resolution future values that follow immediately after a given input window. This task is central to time series modeling [7], where continuous, precise prediction is essential [110], [111]. This has been extended to biomedical domain, vital-sign-based early warning system development [42], [112], and motion trajectory prediction [113], or even broader health trajectories [114].

Future outcome prediction. The second type of predictive modeling focuses on forecasting discrete clinical outcomes or events that may occur minutes, hours, or even months after the input signal window. This includes tasks such as predicting in-hospital mortality, sepsis onset, ICU admission, readmission risk, or long-term disease progression. Outcome prediction targets higher-level, event-based outcomes, often over extended time horizons. A key subset of this category is survival analysis, which models the time until a specific event (e.g., death or deterioration) and accounts for censoring, where the outcome is not observed for all patients within the study period. Biosignal-based models have shown promise in forecasting deterioration risk on diverse tasks [115], and foundation models could further enhance this by modeling complex temporal relationships and learning time-to-event distributions [116] across large datasets. However, the benchmarking of this task itself remains challenging due to the intrinsic difficulty of collecting high-quality, long-term labeled data, and the conditional, sparse nature of many clinical events.

4 Developing Foundation Models from Scratch on Large Biosignal Corpora

In this section, we discuss the methodologies for developing biomedical time series foundation models. We provide detailed instructions on every stage of the process, from data preparation, architecture designs, model training, to deployment, regarding developing biosignal foundation models from scratch. At each step, the unique challenges posed by biosignals are highlighted, along with potential strategies to address them. We also summarize recent foundation models for biomedical time series data in Table 2. Models listed in the bottom part of the table represent self-supervised approaches, which are not explicitly claimed as foundation models but are included for reference.

4.1 Input Tokenization

Tokenization in time series modeling refers to the transformation of signal into structured representations, such as patches or learnable tokens, that can be processed by transformer-based or other sequence models. Though optional, this step can be beneficial depending on the model architecture or biomedical signals' characteristics. This section outlines common tokenization strategies for converting raw time series data into model-compatible representations.

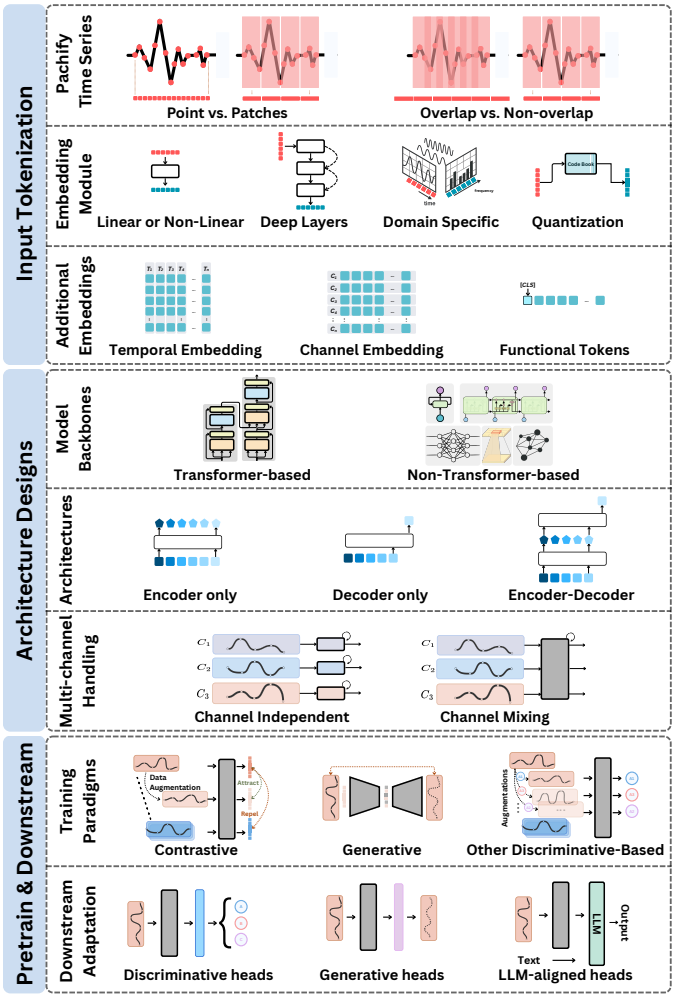

Fig. 4: Overview of training recipes across different stages of developing biosignal foundation model from scratch. It summarizes core components involved in the training, including input tokenization strategies, architectural configurations, and pretraining and downstream adaptation paradigms.

4.1.1 *Patchify Time Series*

Segmentation, often referred to as "patchify", is a key step in tokenizing time series data. This involves partitioning the time series along the time dimension, enabling models to capture local and global dependencies across patches.

Point vs. Patches. Traditional time series modeling treats each time point as a token, allowing explicit capture of temporal dependencies but resulting in high token counts and computational cost for long signals. Inspired by "patchify" operation in images [121], recent methods propose segmenting time series into fixed-length patches and consider each patch as a token [122] to reduce computational overhead.

Nevertheless, there is currently no standard protocol for determining the optimal patch size. In practice, the patch size is often chosen based on the minimum physiologically / clinically meaningful segment for a given signal, taking into account the signal's sampling frequency. Recent work [123], [124] explored variable patch sizes to capture information

TABLE 2: Overview of current foundation models developed directly on large biosignal corpora. Models are ordered by first online year (reverse chronological). For each model, we list supported modalities, representation type, tokenization strategy, model architecture, training paradigm, and available model/dataset scale. Some models may not explicitly identify as foundation models but follow similar large-scale, general-purpose training practices.

Work	Modality					Representation	Tokenization	Architecture	Training	Model Size	Dataset Size (hr)
	ECG	PPG	IMU	EEG	Others						
Pulse-PPG [28]		✓				time sequence	-	CNN	contrastive	28.5M	55k
CBraMod [117]				✓		time sequence	patch + domain-specific	Transformer	generative	4M	27k
HeartLang [118]	✓					time sequence	patch + CNN	Transformer	generative	510M	2k
NeuroLM [119]				✓		time sequence	patch + CNN	Transformer	generative	1.7B	25k
PaPaGei [25]		✓				time sequence	-	CNN	contrastive	5M-139M	57k
RelCon [120]			✓			time sequence		CNN	contrastive	-	0.7M
BrainWave [93]				✓		time sequence	patch + domain-specific	Transformer	generative	-	41k
CVDFM [22]	✓				PCG	image (Mel transform)	-	Transformer	generative	92M	-
ECG-FM [23]	✓					time sequence	patch+CNN	Transformer	contrastive+generative	-	3.6k
EEGFormer [19]				✓		time sequence	patch + quantization	Transformer	generative	-	-
KED [20]	✓					time sequence	-	CNN	contrastive	-	2.2k
LaBraM [21]				✓		time sequence	patch + CNN	Transformer	generative	5.8M	2.5k
LSM [24]					Numerics	image (plot)	-	Transformer	generative	110M	40M
NMI-FM [18]					EMG	time sequence	-	CNN + RNN	generative	60M	($\geq$ 6.5k subjects)
NormWear [17]	✓	✓	✓	✓	EMG	time sequence	patch + domain-specific	Transformer	generative	58M	15k
OPERA [16]					Respiratory	image (Mel transform)	patch + CNN	Transformer	contrastive+generative	4M,21M,31M	404
SelfPAB [15]			✓			image (TF transform)	patch + linear	Transformer	generative	76M	100k
SiamQuality [14]		✓				time sequence	-	CNN	contrastive	2.2M	$\geq$ 600k
MMM [13]				✓		frequency sequence	linear	Transformer	generative	0.3M	-
Brant [10]				✓		time sequence	patch + domain-specific	Transformer	generative	500M	2.6k
HeartBeiT [11]	✓					image (plot)	patch + quantization	Transformer	generative	86M	11.8k
wearableFM [12]	✓	✓				time sequence	-	-	contrastive	-	51k
UKB Model [9]			✓			time sequence	-	CNN	other discriminative	10M	16.8M

at different levels of temporal granularity for medical time series. Notably, this challenge is not unique to foundation models; window length has long been considered a tunable parameter in biomedical signal analysis [125].

Additionally, for specific types of biosignals, domain-specific customizations have been explored. For signals with inherent periodicity, such as ECG and PPG, researchers have proposed to segment the signal into heartbeat-aligned units rather than using arbitrary temporal slices. This ensures that each token corresponds to a physiologically meaningful cycle, thereby preserving cyclic information for better modeling performance [118], [126].

Overlap vs. Non-Overlap. In sequence segmentation, the stride size (time steps between the starting points of consecutive patches) controls the degree of overlap between patches. This modeling design was first explored in PatchTST [122], demonstrating that using non-overlapping patches improves the computational efficiency by reducing the number of segmented patches. Subsequent work [35], [41] further highlight its practicality in time series analysis. Non-overlapping segmentation has also been widely adopted in biomedical time series. This approach has proven effective across a range of signal types, including both waveforms [10], [19] and numeric values [92].

4.1.2 Embedding Module

The embedding module transforms the input, whether segmented or in its original time series form, into feature matrices that enable representation learning in attention-based and other machine learning models.

Linear vs. non-linear projections.

The embedding in time series models is often implemented using a linear projection layer [127], [128], although nonlinear projectors have also been explored [129]. Most contemporary approaches, particularly in foundation models, favor simple linear embeddings for their efficiency, scalability, and effectiveness.

Deep embedding layers. Deeper architectures have been used to extract the embeddings from biosignals segments. For instance, existing works on EEG [21], [130] and ECG [23] applied CNN blocks to convert raw segments to embeddings. These methods can bypass explicit segmentation because operations like convolutions inherently aggregate information over input patches.

Domain-specific embeddings. Embedding processes can be optimized using domain-specific transformations that leverage inherent characteristics of the data [91], [117]. For example, BioT [91] applies a fast Fourier transform (FFT) to each segment before embedding, effectively exploiting frequency-domain patterns in biomedical time series.

Quantization-based embeddings. Quantization-based embedding method, inspired by methodologies such as VQ-VAE and VQ-GAN, uses a discrete codebook to map time series patches into a predefined set of quantized representations, facilitating efficient sequence modeling. Commonly used in image generation [131], such techniques have recently been used in time series analysis [19], to enhance model interpretability and efficiency. The quantized vectors serve as discrete representations, and they need to be further projected into continuous embeddings.

4.1.3 Positional Encoding and Functional Tokens

Beyond the basic time series embeddings, auxiliary embeddings, such as temporal and channel-specific encoding, are commonly incorporated to better capture interdependencies and contextual relationships within the data.

Temporal positional embeddings. Temporal information is commonly encoded using fixed sinusoidal embeddings [127] or hard-coded timestamp features [122]. Additionally, learnable methods, such as relative positional embeddings [132], and rotary position embeddings [133], have also been explored to facilitate temporal modeling.

Channel-wise positional embeddings. Beyond temporal encodings, prior work has explored using channel-wise positional embeddings to incorporate sensor or channel-specific information. These can be achieved using learnable parameters or heuristic designs. For instance, Yi *et al.* [13] considered different EEG channels as a 2-dimensional manifold, and introduced a multi-dimensional positional encoding to encode the 2D spatial information for EEG. This facilitates cross-dataset EEG pretraining to learn unified topology-agnostic representations.

Functional tokens. In addition to these various positional embeddings, functional tokens were introduced to further improve the model's capacity. A common approach is to append a [CLS] token to incorporate the whole-sequence-level information. Recent studies have also explored additional prefix tokens to encode data-specific information,

such as signal types [134], data resolutions [37], [135], and pretraining tasks [38]. This can be particularly useful for training foundation models on data from various sources.

4.2 Architecture Designs

With the appropriate preprocessing and optional tokenization, the next step is to select the machine learning architecture for modeling biomedical time series. Below, we discuss several approaches tailored to the unique challenges.

4.2.1 Transformers or Other Sequence Modeling Models

With the rise of deep learning models and the pursuit of capturing the complex dynamics in time series data, various modeling architectures have been explored to enhance the modeling capacity of intricate temporal patterns and dependencies. These include models based on Recurrent Neural Networks (RNNs) [136], [137], Graph Neural Networks (GNNs) [138], [139], Convolutional Neural Networks [25], [120], and Transformers [110], [122].

In the general time series analysis, there is a growing shift in modeling architectures from CNN/RNN/GNNs towards Transformer-based models, particularly in the development of general time-series foundation models [34], [39], [41], [140]. One possible explanation for this is that Transformer models are good at model parallelization, enabling efficient training on large-scale data [141]. They also excel at integrating non-time series modalities [142], and capturing long-term dependencies [122], [143], which enhances their ability to capture autoregressive patterns within individual time series over time. Nevertheless, the inherent inductive biases of non-Transformer architectures (e.g., translation invariance in CNNs, adjacency structures in GNNs, memory gating in RNNs, etc.) would help reduce potential overfitting and are particularly useful with limited data. As a result, the design of current biosignal foundation models remains diverse, as noted in Table 2, balancing the advantages of Transformers with the practicality of specialized, domain-aligned approaches.

Transformers. Most time series foundation models employ Transformer-based architecture, either waveform data [12], [15], [118] or numeric health metrics [24]. This enables the use of state-of-the-art generative methods for self-supervised pretraining. Within these Transformer-centric designs, various architectural adaptations have been explored. For instance, Yang *et al.* [91] applied a linear attention mechanism, which was effective in pretraining long-term biomedical time series with light computational cost.

Others. Several studies have explored CNN-based structures in foundation models for biosignals. These include PaPaGei [25] for PPG, UKB Model [9] and RelCon [120] for Wrist-worn Accelerometer. Transformer-based models, while powerful, typically require large datasets to perform well [121]. Additionally, in many current biosignal applications where classification is the main downstream task, the generative capabilities of Transformers may not be essential. Consequently, convolutional architectures offer a more practical and efficient solution [9], [25].

4.2.2 Encoders, Decoders, or Hybrid

In time series modeling, encoders map inputs into lower-dimension representations, while decoders generate outputs by conditioning on these representations or prior tokens. Based on this, architectures can be categorized into encoder-only, decoder-only, and encoder-decoder (hybrid) frameworks, originating from Transformer-based language models [144]. As these structures are increasingly adopted in recent time series foundation models, it is essential to systematically examine their structural characteristics and implications for modeling biomedical time series.

Encoder only. Encoder-only architectures encode the time series to latent representations. These architectures have been successfully employed for pretraining models using RNN [145] and Transformer-based encoders [110], [122]. Encoder-only designs are well-suited for self-supervised pretraining with contrastive learning objectives (see Section 4.3.2) as they excel at contextual representation learning, aligning with contrastive learning's goal of extracting robust, informative representations.

Decoder only. Decoder-only models generate sequences autoregressively, predicting each token based on previous outputs. This architecture is well-suited for forecasting and generative tasks. Pioneering work in general time series, TimesFM [41] and recent studies [30], [31], [146] have explored decoder-only models for time series forecasting foundation models. However, decoder-only models remain relatively underexplored in biomedical applications, as most practical use cases involve fixed-length biosignals (i.e. 10s for ECGs) and focus on tasks such as event prediction or classification, rather than step-wise generation.

Encoder-Decoder. Encoder-decoder models [132], [147] combine the strengths of both encoders and decoders, allowing for a more flexible design to capture input dependencies. However, balancing both components can potentially make training more challenging. On the other hand, encoder-decoder architectures are well-suited for masked generative pretraining, such as masked autoencoding (MAE) [148], where the encoder encodes visible (unmasked) input parts and the decoder reconstructs masked sections. While encoder-only architectures can also perform masked pretraining, encoder-decoder models naturally facilitate reconstruction objectives due to their decoding capabilities [148], [149]. Recent work in biomedical domains has leveraged encoder-decoder frameworks for pretraining purposes for tasks such as physiological signal modeling and multimodal integration [17], [21], [24].

4.2.3 Channel Independence or Mixing

In general time series modeling, multi-channel data are handled using either channel-independent [39], [40], [41] or channel-mixing methods [34], [37]. Channel-independent methods model each channel using its own history values, which is effective when strong seasonal patterns exist [150]. In contrast, channel-mixing methods capture inter-channel dependencies to model richer temporal dynamics.

In biomedical signal analysis, capturing inter-channel dependencies is often crucial for clinical interpretation, as physiological signals frequently exhibit cross-channel correlations and stochastic abnormalities. However, foundation models face a key challenge in handling channel heterogeneity across datasets [37] (see also Section 4.4.1). To facilitate inter-channel dependency modeling, some methods keep a fixed number of signals [20], [100] or randomly sample a fixed subset of channels [13]. Others flatten the multiple features into a univariate sequence [151], but this often

leads to inefficiency and sparsity, especially with an extensive number of channels and long sequences. Alternatively, TTM [37] adopts a two-stage approach where they first pretrain on univariate series and then conduct fine-tuning on an additional module with inter-channel modeling.

4.3 Training Paradigms

The self-supervised training objectives can be generally categorized into generative and discriminative approaches [1]. In this section, we discuss existing state-of-the-art self-supervised strategies that have been successfully applied to biomedical time series foundation models, as well as novel methods from the broader time-series domains that hold potential for future foundation model development.

4.3.1 Generative Modeling

Self-generation. This strategy typically involves predicting missing/masked tokens within input sequences. Common approaches include next-token prediction and MAE-based frameworks, which reconstruct masked parts of the input from their unmasked context. Although such adoptions on biosignals [23] have shown good performance, these masking strategies are often domain-specific, different from next-token prediction or uniformly random masking.

Narayanswamy *et al.* [24] considered the different types of generation tasks in wearable sensor data, as random imputation, temporal interpolation, sensor imputation, and temporal extrapolation, with the aim to generalize the model capabilities across different real-world scenarios. These masking strategies are domain-inspired, designed to mimic practical signal degradation or missingness. There are several domain-inspired strategies to explore innovative masking strategies during generative modeling, such as LaBraM [21], MMM [13], and BrainBert [152].

Cross-representation generation. In certain scenarios, especially with signals exhibiting significant noise or non-stationarity in the temporal domain (e.g., EEG), it can be advantageous to generate alternative representations instead of directly reconstructing the original signal. For example, an EEG foundation model proposed by Jiang *et al.* [21] adopts an MAE framework to reconstruct the spectral representations of masked segments, thereby bypassing modeling noisy temporal signals directly.

Cross-modality generation. This typically reconstructs signals across different modalities or signal types. For instance, in [105], an ECG-MRI cross-modality reconstruction framework is proposed to learn shared representations. In [22], the MAE framework is extended to multi-modality with ECG and PCG, to enable cardiovascular disease diagnosis.

4.3.2 Contrastive Modeling

Unlike generative pretraining, discriminative pretraining focuses on proxy discriminative tasks, with self-supervised contrastive learning being a dominant approach. This method learns representations by maximizing the similarity between positive pairs while pushing apart negative pairs, enabling the model to capture meaningful distinctions in the data. Various contrastive learning frameworks, such as SimCLR, BYOL, and SWAV, have been developed, and in this section, we discuss how to effectively define positive and negative pairs in the context of biomedical time series.

Segment level. This line of contrastive strategy is based on applying different augmentations on the same sequence. These can include temporal, spatial, and spectral augmentations or transformations. For instance, Kiyasseh *et al.* [153] and Zhang *et al.* [99] have utilized such augmentations to enhance model robustness to variations in biomedical signals. Biot [91] explored a channel-wise masked strategy, ensuring that signals collected with different channel settings can be effectively aligned and learned in a unified manner.

Time level. Considering the long-term nature of biomedical time series, especially for continuous monitoring, leveraging the temporal relationship between different signals can be a good option. For instance, Lan *et al.* [154] introduced a stationarity test module to detect abrupt changes in neighboring frames, ensuring that segments with sudden variations are treated distinctly during analysis. This enhances the model's sensitivity to significant temporal fluctuations. Lee *et al.* [155] proposed a soft contrastive strategy, which assigns varying degrees of contrast based on the temporal distance between signal pairs, allowing the model to capture nuanced temporal dependencies. Differently, Raghu *et al.* [156] grouped sequences along a patient's trajectory and applied diverse augmentations at each timestamp, aiming to enforce consistency across different temporal scales.

Individual level. At a broader scale, some contrastive learning strategies aim to preserve individual-specific characteristics by ensuring that records from the same individual remain similar in representation space. Sangha *et al.* [157] and Abbaspourazad *et al.* [12] have both leveraged such contrast for the self-supervised learning of biosignals, and highlight its superiority to the simplistic segment-level contrast [158].

Domain knowledge level. Rather than strictly adhering to segment, time, or individual-level contrastive learning strategies [107], [159], recent studies have explored constructing contrastive pairs based on commonly used domain-specific features. These handcrafted features can help better capture clinically meaningful similarities by leveraging bespoke characteristics intrinsic to biosignals. For instance, PaPaGei [25] utilizes morphological features extracted from PPGs to construct similarity pairs across participants, promoting feature-level consistency during training. Similar strategies have been introduced in earlier works [160], [161] on other biosignal types, as well.

Cross-modality contrast. Beyond single-modality contrastive learning, recent studies have also explored constructing positive pairs across modalities. This approach is particularly effective in scenarios where multiple biosignals are collected simultaneously [22], [162] or where auxiliary data sources, such as clinical reports, are available to provide richer supervision [62], [70]. In the latter case, representations of texts are typically derived using pretrained LLMs to align with biosignal representations. A more detailed discussion on leveraging LLMs for semantic alignment between biosignal and text can be found in Section 6.3.

4.3.3 Other Discriminative Based

Apart from the aforementioned generative tasks, other approaches, such as augmentation prediction [9], have been explored. However, these methods may be less adaptable across different biomedical applications. In contrast, generative and contrastive learning frameworks provide more versatile and effective strategies for self-supervised training.

4.4 Key Challenges and Solutions

The aforementioned pretraining paradigms outline state-of-the-art methods for self-supervised learning on biosignals. Building on this foundation, another critical aspect of training large-scale models is scaling up the volume of data. However, this process introduces significant challenges in harmonizing datasets from diverse sources, particularly due to variations in channel configurations and sampling frequencies. In the following section, we review current strategies that can be applied to address these challenges in the context of biosignal data.

4.4.1 Addressing Channel Heterogeneity in Biosignals

One significant challenge in sensor format unification is the varying channel configurations. This refers to cases where the number of channels is not fixed across samples or training/deployments. This often arises when biosignals are aggregated from heterogeneous sources (e.g., clinical monitoring systems across hospitals or consumer wearables from different manufacturers), leading to scenarios where models must handle missing, optional, or varying input channels. Addressing this scenario requires models that are not only robust to channel variability but also generalizable across diverse input configurations. We provide an overview of existing strategies addressing this problem.

Data alignment. This line of research generally follows two main strategies: (i) treating channels independently, and (ii) synthesizing or reconstructing missing channels. A straightforward approach is to consider each channel as an independent univariate signal and aggregate the outputs, which is commonly adopted in general multivariate time series analysis, as discussed in Section 4.2.3. However, in biosignals, where channels often carry inherent spatial, anatomical, or topological relationships (e.g., EEG electrode positions), treating channels as fully independent can overlook crucial inter-channel dependencies, and may not be ideal for addressing channel heterogeneity.

To better preserve these dependencies, more advanced methods have explored synthesizing missing channels using either classical spatial interpolation techniques [163] or pretrained generative models [164], [165]. When certain channels are missing or fewer channels are available than expected, interpolation methods aim to fill in the missing data or upsample the signals to a standardized set of channels. This harmonizes the input across heterogeneous datasets while maintaining domain-informed spatial structure.

Architecture designs that focus on (i) channel-aware modules and (ii) channel-mix modules are also used to address the channel heterogeneity problem.

The inherent sequential flattening of transformer architecture with channel-wise encoding can directly support biosignals [13], [91], [117], which is particularly useful for biosignal settings where different subsets of channels would be collected given a fixed number of channels. For instance, different ECG lead configurations in 12-lead ECGs, and EEG cap settings under standard 10-20. Different from Transformer, some other methods like graph [139] or channel-dependent module ensembles are considered [162].

On the other hand, channel-mixing modules aim to fuse information across channels and ideally support flexible or multiple input configurations, namely, any-variate. Luo *et al.* [17] proposed such a design, allowing the model to integrate features from varying channel wearables. These approaches are particularly useful when inter-channel relationships are crucial for downstream performance, yet the set of available channels varies across samples or datasets.

These architectures are normally combined with additional training strategies below to yield robust performance.

Training strategies. This line of work typically focuses on two main approaches: (i) enforcing cross-channel consistency and (ii) enabling cross-channel generation.

A dominant strategy involves deliberately masking out specific subsets of channels during training to improve the model's robustness to missing or variable inputs [21], [91]. This setup encourages the model to maintain consistent representations across different channel combinations (cross-channel consistency [91]), or to reconstruct masked channels from the remaining ones (cross-channel generation [21]).

Other strategies include knowledge distillation, where a model trained on the full set of channels teaches another model to work with fewer inputs [166], and curriculum learning, which gradually introduces more challenging, lower-channel scenarios during training [167]. However, these methods may become less effective when channel configurations vary drastically across samples.

4.4.2 Addressing Resolution Heterogeneity in Biosignals

Another common challenge in biosignal modeling is the heterogeneity of sampling frequencies. As discussed in Section 2, biosignals collected from different devices or sources often exhibit substantial disparities in temporal resolution. Moreover, for practical purposes, the same type of signal may be acquired or processed at different sampling rates depending on the application context, device constraints, or preprocessing protocols. While standard strategies such as resampling can align signals with moderate differences, they become ineffective when the resolution mismatch is extreme (e.g., from second to daily), thus failing to preserve important high-frequency features or adequately enhance coarse signals. Similar to the last section, we categorize existing approaches into three groups. In practice, waveform datasets often rely on data alignment techniques such as resampling, whereas numeric values (such as vital signs) may more likely benefit from the rest two.

Data alignment. One common strategy to address this issue is to perform unification on the data part, which includes (i) resampling and (ii) frequency-invariant representations.

Resampling aligns signals to a shared frequency that covers the semantic meaning of specific biosignals [21], [117], as illustrated in Section 2. For instance, Wang *et al.* [21] resampled multiple EEG datasets to 200Hz, and PaPaGei [25] to 125Hz of PPG signals. This is particularly common in the preprocessing of waveform datasets, where unifying sampling rates facilitates batching, input formatting, and downstream validation. This is straightforward, although prior studies have shown that some clinically meaningful signal components, may reside in higher frequency bands that exceed typical resampling targets [168]. Moreover, this approach may not generalize well to non-waveform numerical values, for instance, heart rate, where large heterogeneity may exist between recordings.

Another strategy is to seek frequency-invariant representations. For instance, during tokenization, rather than treat-

ing signals as original temporal values, BIOT [91] converts signals into spectral representations (i.e., FFT) as input.

Architectural design. The majority of biosignal foundation models address the challenge of aligning series with different frequencies by applying resampling as a preprocessing step, as mentioned above. While this approach is effective, resampling inherently distorts the data, and it may not be applicable to numeric health metrics collected at vastly different temporal resolutions.

Existing progress in general time series models offers alternatives in modeling structures. Some studies adopt adaptive feature extraction methods that capture temporal patterns across multiple scales dynamically: Lag-LLama [39] captures multi-frequency patterns through lagged features from specific temporal hierarchies (e.g., hour-of-day), while TTM [37] employs learnable resolution tokens to dynamically adapt to varying timescales. DAM [169], on the other hand, takes a different approach by combining predefined multi-scale patterns (from minutes to years) with adaptive weighting, providing a more scalable solution without explicit frequency alignment. Additionally, other studies adopt hierarchical structures to process time series data at different frequency levels, assigning each layer to capture patterns at a specific temporal resolution. For example, MOIRAI [34] proposed to have multiple projection layers, each tailored to handle data corresponding to a specific frequency. These methods demonstrate that resampling is not the only viable strategy, shedding light on the modeling options in handling mixed-frequency time series data.

Training strategy. Enriching the training data can also be effective when it comes to improving the model's ability to handle diverse frequencies. TimesFM [41] incorporates synthetic data generated from different frequencies. This explicitly tailored enrichment of training data helps in improving the generalizability of foundational models towards the underrepresented granularities.

4.5 Downstream adaptation

The primary value of biosignal foundation models lies in their adaptability across diverse downstream tasks. Depending on the task formulations, different types of heads can be applied for adaptation. Broadly, these can be grouped into three categories: discriminative, generative, and language-aligned. We hereby provide an overview of the use cases in prior studies.

Discriminative heads. Discriminative heads are designed for tasks framed as regression or classification, as discussed in Section 3, including domain-specific pattern recognition [22], [93], [170], future outcome forecasting [171], and phenotyping [12], [25]. These heads are typically implemented using linear probing [25], [120], multilayer perceptrons (MLPs) [28], [153], or other classification models that map latent embeddings to target labels or values. Because they are lightweight and task-specific, they are well-suited for scenarios where labels are available for fine-tuning on downstream applications.

Generative heads. In addition to discriminative tasks, several biosignal tasks are framed as generative, i.e., dense trajectory prediction, data restoration/augmentation/generation. Generative heads are employed in these settings to model signal dynamics. These heads typically employ decoder architectures such as a transformer decoder or MLPs to generate continuous outputs conditioned on learned representations. For instance, AnyECG [172] formulates corrupted lead reconstruction as a downstream task, utilizing a dedicated lead decoder to generate missing signal segments.

LLM-aligned heads. In addition to conventional discriminative and generative tasks, language-aligned heads are emerging to bridge the gap between biosignal representations and natural language interpretation. These heads enable either sensor-to-text generation or sensor–text fusion, facilitating multimodal understanding. For instance, LLaSA [173] and NeuroLM [119] leverage pretrained biosignal models as feature extractors, forwarding the resulting embeddings to LLMs to perform sensor-related query tasks or generate natural language descriptions. Further examples are presented in Section 6.4.

5 Adapting General Time Series Foundation Models to Biosignals

In addition to developing biosignal foundation models from scratch, there is a growing trend in adapting general time series foundation models for biosignals. This has been supported by advances in self-supervised training and the availability of large-scale time series data from diverse domains. Prior work in time series foundation models has demonstrated their capacity in capturing generalizable dependencies [33], [37], [41]. Building on this, it is important to explore the existing time series foundation models for general applications and how they can be adapted for biomedical time series, with the potential to improve their performance in health-related tasks.

5.1 Background and Preliminaries

Current time series foundation models mainly focus on improving forecasting capabilities, with a variety of architectural choices being explored. Transformer-based models remain the most widely adopted, often categorized into encoder-only and decoder-only variants. For example, Lag-llama [39] and TimesFM [41] use a transformer decoder to generate predictions autoregressively. Moment [35], on the other hand, employs an encoder-only architecture and leverages temporal patches for time series analysis. Beyond transformers, alternative architectures such as MLPs [37], [179] and state-space models like Mamba [180] have also gained attention for their efficiency and scalability in sequence modeling. An overview of these models and the range of time series tasks they support is provided in Table 3, including their architectural structure, input–output design, modeled variable dependencies, dataset scale, model size, and whether they support probabilistic outputs. We also summarize which tasks each model supports, which are forecasting (F), classification (C), anomaly detection (D), and data imputation (I). These general time series tasks align closely with the core biosignal tasks outlined in Section 3.

5.1.1 General vs. Biomedical Time Series Analysis

As noted in Table 3, most of the foundation models are primarily focused on forecasting tasks, and they are benchmarked on the general time series domain. In this sense, they may not be directly applicable to biomedical applications due to differences in domain-specific characteristics, data channels and resolutions, as well as distinct splitting strategies, evaluation settings, and task types.

TABLE 3: Overview of general-purpose foundation models for time series modeling. Note that intra-/inter- refers to intra-/inter-variable dependency. Task types (corresponding to Section 3) are abbreviated as follows: F = Forecasting, C = Classification, D = Anomaly Detection, I = Data Imputation. A dash (-) indicates that the corresponding information is not explicitly mentioned in the original paper. Dataset scale indicates the total number of time points across all datasets used; model size denotes the number of parameters in the model. Models are ordered by first online year (reverse chronological).

Model	Structure	Input Channels	Dependency	Dataset Scale	Model Size	Tasks				Used studies
						F	C	D	I	
Time-MoE [31]	decoder-only	multivariate	intra-	309B	2.4B	✓				-
TimePFN [32]	encoder-only	multivariate	intra- & inter-	1.5M	-	✓				-
Timer-XL [174]	decoder-only	multivariate	intra- & inter-	-	-	✓				-
Chronos [33]	encoder-decoder	univariate	intra-	84B	20/46/200/710M	✓				[25], [42], [43]
DAM [169]	encoder-only	univariate	intra-	-	-	✓			✓	-
Moirai [34]	encoder-only	multivariate	intra- & inter-	27B	14/91/311M	✓				[42], [43]
MOMENT [35]	encoder-only	univariate	intra-	1B	385M	✓	✓	✓	✓	[25], [42], [44]
TimeDiT [36]	encoder-only	multivariate	intra- & inter-	5B	33/120/460/680M	✓		✓		-
TimesFM [41]	decoder-only	univariate	intra-	100B	200M	✓				[42], [175]
Time-FFM [176]	encoder-only	multivariate	intra-	-	-	✓				-
ViTime [177]	decoder-only	multivariate	intra-	-	74/95M	✓				-
Lag-Llama [39]	decoder-only	univariate	intra-	360M	200M	✓				[42], [178]
TimeGPT-1 [40]	encoder-decoder	multivariate	intra- & inter-	100B	-	✓		✓		-
UniTS [38]	encoder-only	multivariate	intra- & inter-	-	8M	✓		✓	✓	[38]
TTM [37]	encoder-decoder	multivariate	intra-	1B	5M	✓				-

Domain-specific characteristics. One major challenge lies in the difference in signal characteristics between general time series and biomedical time series. General TSFMs are often trained on heterogeneous datasets from various domains such as finance, weather, traffic, and epidemiological surveillance. In contrast, biomedical time series data, e.g. ECG, EEG, and PPG, exhibit domain-specific challenges, including physiological constraints, noise artifacts, and variations across patients. Even within biomedical domains, data characteristics can significantly vary across modalities, collection settings, and clinical applications, requiring specialized adaptation strategies.

Data channels and resolutions. Biomedical time series data often involves multiple channels representing physiological signals from different sensor placements, each with varying sampling rates and resolutions. In contrast, general TSFMs are mostly designed to handle uni-channel or uniformly sampled time series, making it challenging to process multi-channel, multi-resolution biomedical signals without explicit adaptation mechanisms.

Data splitting and evaluation settings. Unlike general time series tasks, which often use temporal splits for training and testing, biomedical time series require more clinically meaningful evaluation strategies to ensure real-world applicability. This includes subject-wise splitting (ensures that the model does not see data from the same individuals during training and testing), and cross-institution generalization (how well models trained in one hospital/general population generalize to different hospitals/patient cohorts), and cross-scenario generalization (how well models generalize across varying levels of data granularity and baseline patient characteristics, e.g., general ward vs. ICU settings.)

Task types. In general time series modeling, including TSFMs, forecasting is the primary focus [33], [41], [140]. Although prior work has explored TSFMs for classification [35], [181], they are limited to generic benchmarks and do not address clinically relevant tasks. In contrast, biomedical time series analysis often involves a broader range of tasks beyond forecasting, as discussed in Section 3. As such, directly adapting general TSFMs to biomedical signals is often intractable without suitable modifications.

5.2 Adaptation Strategies

Given these disparities between general time series data and biosignals, directly applying pretrained models or employing fine-tuning methods, such as PEFT techniques, cannot fully address the challenges of adapting general TSFMs to biomedical applications. These necessitate more comprehensive adaptation strategies beyond conventional fine-tuning.

Nevertheless, there have been initial efforts to adapt general TSFMs to biosignal analysis, exploring various adaptation techniques to bridge the gap between general time series modeling and biomedical-specific applications.

5.2.1 *Feature Extraction*

For some encoder-only or encoder-decoder-based general TSFMs, such as Chronos [33] and Moment [35], pretrained encoders can be used to extract latent representations that are useful for non-generative downstream tasks, such as classification and regression. These representations capture general temporal patterns and can be leveraged without full model fine-tuning. PaPaGei [25] evaluated their proposed biomedical foundation model by comparing its performance against Chronos and Moment. To assess the adaptability of these models to biomedical tasks, they add a shallow, task-specific layer on top of the learned representations for downstream validation. While PaPaGei outperformed the general models, the results show that general TSFMs still offer transferable features that can serve as a strong starting point for biomedical adaptation.

5.2.2 *Model Adaptation for Biomedical Tasks*

Repurposing forecasting. Most time series foundation models (TSFMs) are primarily designed for forecasting. One practical way to adapt them for biomedical applications, without modifying their architectures, is to introduce a post-processing module that analyzes the forecasted values for specific medical prediction tasks. This has been applied in prior work [42], [178]. These are particularly useful for highly risky patients with continuous monitoring needs. Recently, FORMED [175] was proposed as a novel method for repurposing forecasting models for classification tasks, leveraging the pretrained backbone with minimal modifications to the task head. This method has been validated on biosignals across various medical tasks, demonstrating its effectiveness in transferring forecasting-based TSFMs to classification-based biomedical applications.

Plug-in adapters. Most general TSFM are designed for univariate analysis. To address this, Liu *et al.* [44] proposed the Generalized Prompt Tuning (Gen-P-Tuning) module, which adapts existing univariate time series foundation models (with fixed model parameters) to handle multivariate clinical time series. Instead of considering the information of

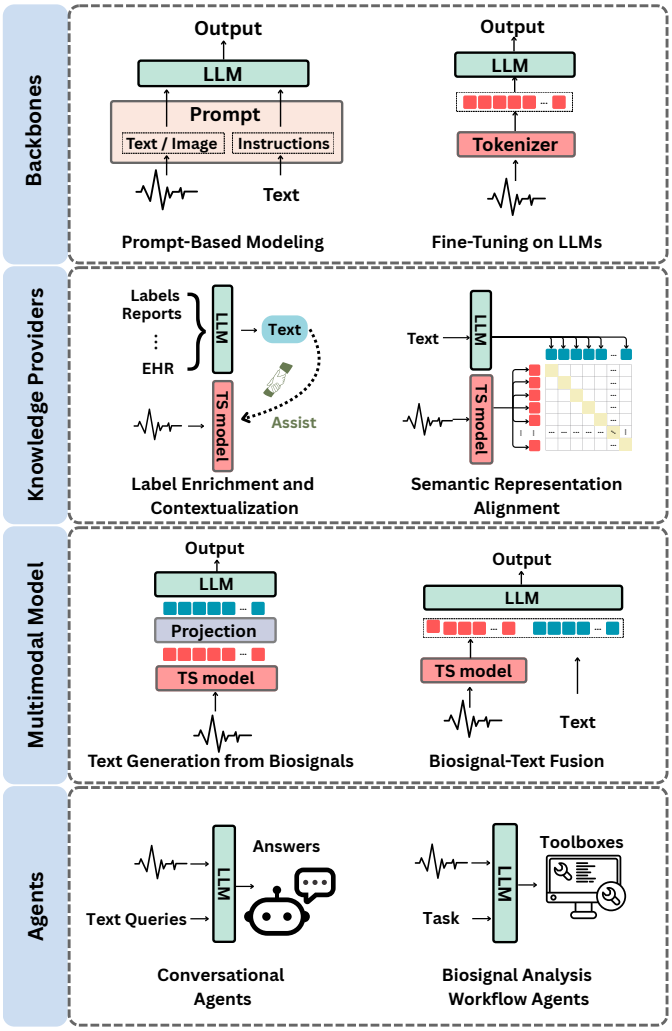

Fig. 5: Overview of the different roles of large language models (LLMs) in biomedical signal analysis. They include as backbones for biosignal modeling, as knowledge providers for biosignal model training, as interfaces for joint biosignal-text modeling, and as interactive agents for biosignal interpretation.

each channel to be independent of one another, Gen-P-Tuning concatenates the embeddings of multiple channels, enabling the model to implicitly incorporate cross-channel dependencies without requiring extensive retraining.

With the increasing availability and accessibility of general TSFMs, we anticipate more systematic validation of these adaptation methods across various types of biosignals. An important direction for future research may involve benchmarking these models on diverse biomedical datasets, evaluating their generalization capabilities, and refining adaptation techniques to improve model performance, interpretability, and clinical applicability.

6 Leveraging (Multi-Modal) Large Language Models for Biosignals

Beyond developing biosignal foundation models and adapting time series foundation models, recent research increasingly explores the integration of LLMs into biomedical time series analysis. LLMs have exhibited exceptional strengths in semantic reasoning, contextual understanding, and generative tasks. Recent developments have expanded these capabilities into multi-modal contexts, enabling LLMs to effectively handle diverse data types such as images, audio, and, increasingly, time series signals [185].

This section explores how the advances of LLMs can be leveraged for biosignal analysis. We first present fundamental concepts and methodologies for integrating existing LLM architectures into biomedical signal analysis (Section 6.1). Subseuently, we categorize multimodal LLMs applications into four distinct but interrelated roles: (i) direct modeling backbones (Section 6.2); (ii) knowledge-rich supervisory resources during training (Section 6.3); (iii) components in multimodal biosignal-text frameworks (Section 6.4); and (iv) interactive agents for dialogue-based interpretation and decision support (Section 6.5). A list of representative methods annotated with the roles of LLMs is presented in Table 4.

6.1 Background and Preliminaries

Integrating biosignals with LLMs is non-trivial due to the modality mismatch between continuous high-dimensional biosignals and token-based language representations. This section outlines the existing fundamental elements for enabling such practice. We begin by revisiting strategies for converting biosignals into formats compatible with LLMs, building on the representation techniques introduced in Section 2.4. Subsequently, we examine mechanisms for aligning biosignals with LLM-compatible inputs, and describe training paradigms and model adaptation techniques that enable LLMs to operate on biosignal tasks, with LLM-specific adaptation strategies (prompt tuning; parameter-efficient tuning, PEFT; instruction tuning) discussed in the Appendix.

6.1.1 From Biosignals to LLM-Compatible Inputs

To apply LLMs to biosignal analysis, raw signals must be transformed into a representation compatible with the token-based architecture of language models. Following the biosignal representations outlined in Section 2.4, we categorize these representations into three types based on how they interact with LLMs: *embedding-based encodings*, *token-based sequences*, and *textual numeric representations*.

Encoder-Based Embeddings. In this approach, biosignals are passed through a pretrained encoder (see Section 4.2.2) to produce embeddings and projected into the LLM's token space, often via learnable linear transformations. Common in multimodal fusion, this method considers signal embeddings as prefix tokens, similar to images or audio in recent LLM frameworks [68], [186]. Its effectiveness mostly depends on the efficacy of the pretrained biosignal encoder. It mostly comes with additional bridge modules to ensure semantic and dimensional alignment with the LLM input space, as further discussed in Section 6.1.2.

Token-Based Sequences. This strategy segments biosignals into token-like sequences [187], enabling direct reuse of pretrained LLMs as backbone architectures with minimal architectural modifications. Meanwhile, a lightweight embedding projector is typically required to ensure representational compatibility between signal-derived tokens and the LLM's input space [119], [188], as in Section 6.1.2.

Textual Numeric Representations. A third strategy encodes biosignal biomarkers or derived quantitative values as natural language phrases or numerical strings. These textual representations require no additional projection and can be directly integrated into prompts. Examples include summaries such as "heart rate variability: low" or "mean systolic

TABLE 4: Functional roles of existing approaches that leverage large language models for biosignal analysis. Listed methods are ordered by publication year (reverse chronological). We summarize each method's modality, representation strategy, and its functional use, ranging from label enrichment to multimodal fusion and agent-based interaction. Abbreviations: Enrich.- Enrichment, Context.- Contextualization, Align.-Alignment, Gen.-Generation.

Work	Modality	Representation	Backbones		Knowledge Providers			Multimodal Model		Agents	
			Prompt -Based	Fine Tuning	Label Enrich.	Medical Context.	Repre. Align.	Text Gen.	Fusion	Convers. Agents	Workflow Agents
ECG-LM [73]	ECG	text + time sequence		✓		✓		✓		✓	
Emotion-Copilot [72]	EEG	text		✓				✓	✓		
HF-Risk [69]	ECG	time sequence		✓			✓				
BELT-2 [60]	EEG	time sequence					✓	✓			
E2T-PTR [62]	EEG	feature sequence					✓	✓	✓		
ECG-Chat [182]	ECG	text + time sequence					✓	✓			
ESI [64]	ECG	time sequence			✓		✓				
Plots [68]	IMU	image	✓								
HARGPT [65]	IMU	time sequence	✓								
Health-LLM [66]	activity measures	text	✓								
IMUGPT [67]	IMU	time sequence			✓	✓					
KED [20]	ECG	time sequence		✓	✓		✓	✓			
LLMTrack [183]	IMU	text	✓			✓		✓			
MEIT [184]	ECG	time sequence			✓						
MERL [70]	ECG	time sequence			✓		✓				
PHIA [71]	activity measures	text	✓							✓	✓
PhysioLLM [101]	activity measures	text	✓							✓	
Sensor2Text [63]	activity measures	time sequence						✓			
METS [61]	ECG	time sequence			✓						
ELA [57]	EEG	time sequence		✓			✓		✓		
openCHA [58]	activity measures	text								✓	✓
BART [55]	EEG	time sequence		✓				✓			

pressure: 140 mmHg". They are particularly useful for tasks that combine high-level reasoning with interpretable signal features, like personalized health agents [71], [101].

6.1.2 Modality Bridging Strategies

While transforming biosignals into LLM-compatible formats enables basic input compatibility, a deeper challenge lies in aligning the heterogeneous representations of biosignals and natural language to support seamless integration and reasoning. This alignment is particularly crucial in settings where LLMs operate over multimodal inputs, or when the model must jointly interpret and reason about signal-derived patterns and textual clinical knowledge. We group alignment strategies into two broad categories: bridging modules and joint training with cross-modal objectives.

Cross-Modal Alignment Objectives. When paired signal–text datasets are available (e.g., ECG waveforms with diagnostic summaries), joint training with cross-modal objectives helps align semantic representations across modalities. These strategies are typically used during pretraining to obtain generalizable encoders that can later support diverse downstream tasks. Common approaches include contrastive learning [60], which aligns corresponding signal-text pairs, and multimodal masked modeling [62], which reconstructs missing elements in either modality. These objectives are especially effective with embedding-based or textual numeric representations, enabling improved retrieval, zero-shot generalization, and conditional generation.

Cross-Modal Bridging Modules. Another strategy introduces modules that transform biosignal embeddings into LLM-compatible token representations, using components like linear projectors, cross-attention layers, or temporal transformers (e.g., Q-formers [189]). For example, features extracted from an ECG encoder can be mapped into additional tokens [73] and integrated with LLM inputs for joint inference. This strategy is especially effective in multimodal pipelines, enabling contextual integration of signal and text, and ensuring compatibility between biosignal representations and token-based language models.

6.2 LLMs as Backbones for Biosignal Modeling

This section categorizes LLM-as-backbone approaches into two main groups: *Prompt-based* and *Fine-Tuning*. The former focuses on converting biosignals into modalities inherently understood by LLMs (e.g., text or images), while the latter leverages LLMs as backbones for sequential modeling.

Prompt-Based Modeling. A prominent line of work repurposes LLMs for biosignal interpretation, building on token-based transformations (Section 6.1.1). Two dominant strategies have emerged: text-based and image-based representations. In the text-based approach, sensor-derived signals [65], [183] or their semantically enriched summaries [66], [101], [190] are converted into structured textual formats, enabling direct interaction with standard LLMs. For example, HARGPT [65] encodes human activity sensor data into prompt templates that guide LLM-based inference for activity recognition. Similarly, PHIA [190] integrates wearable health metrics into prompts, allowing users to query physiological states via language prompts. Alternatively, some studies leverage the capabilities of multi-modal LLMs, particularly those pre-trained on vision-language tasks, by transforming biosignals into visual representations [68], [191]. This includes spectrograms, waveform plots, or topographical maps that can be directly embedded into visual prompts. ECG-Chat [182] and Plots [68] rendered biosignals as plots and fed into vision-language models, enabling biosignal interpretation without handcrafted text or domain-specific time series encoders.

Fine-Tuning on LLMs. Instead of relying on handcrafted prompts or considering biosignals as LLM-compatible inputs, another line of research explores modeling biosignals directly with LLMs. Time-LLM [187] refers to this as "reprogramming", transforming time series into tokenized sequences compatible with LLM input. These approaches typically employ lightweight adaptation modules (e.g., embedding projectors or low-rank adaptation) for efficient LLM adaptations, as seen in recent biosignal studies [119], [188]. For example, Jiang *et al.* [119] align EEG representations with text embeddings, allowing LLMs to model EEG dynamics and generalize across downstream tasks.

6.3 LLMs as Knowledge Providers for Biosignal Model Training

LLMs can also provide supervisory signals during model development by providing enriched annotations or auxiliary tasks. In this setting, LLMs act as domain-informed teachers, improving biosignal model training with richer contextual information drawn from textual knowledge, diagnostic labels, or related health informatics. In this section, we focus on three key and complementary strategies: *label enrichment*, *medical information contextualization*, and *semantic alignment*, which collectively improve supervision quality, especially when labeled data are limited or incomplete.

Label Enrichment. Biomedical time series datasets often come with coarse labels that lack clinical detail, despite being labeled by experts using established guidelines. As such, there exists rich domain knowledge that links waveform patterns to specific diagnostic interpretations. By leveraging LLMs, it becomes possible to generate or retrieve richer textual annotations, such as detailed symptom descriptions, diagnoses, or clinical observations, associated with these labels. Incorporating such textual information as additional inputs or as regularization during training helps models better capture clinically meaningful patterns and improve downstream performance [64], [67], [69]. This strategy is particularly useful in zero-shot learning scenarios, where the enriched contextual information enables better generalization to unseen tasks. For instance, recent studies [64], [70] leveraged LLMs to derive potential ECG characteristics based on given labels, either directly [70] or retrieved from clinical textbooks via Retrieval-Augmented Generation [64]. The resulting textual descriptions were then used to facilitate ECG model training via ECG-text contrastive learning. Similarly, IMUGPT [67] generated motion descriptions using LLMs to guide motion synthesis models in generating new motion sequences, which help augment training data for human activity recognition.

Medical Information Contextualization. Besides enriching labels, LLMs can also summarize and abstract essential medical information from additional modalities commonly accompanying biomedical time series (e.g., electronic health records and self-reported outcomes). These accompanying modalities are often high-dimensional and contain complex information that is not directly usable for modeling. LLMs can extract concise summaries or meaningful embeddings from such data, making them more accessible to the modeling process of these biosignals. Representative scenarios include brain–computer interface tasks [55], [192], where EEG signals are typically recorded during text-reading activities, and clinical patient databases [193], [194], where biosignal waveforms are accompanied by rich textual data such as diagnostic notes and patient histories. Zhang *et al.* [192] leveraged LLMs to extract semantically relevant words in an EEG-based reading task, which were then used to guide EEG model training. Similarly, Chan *et al.* [193] proposed an LLM-based agentic workflow to accelerate the labeling process for medical time series, addressing the challenge of limited expert-annotated data in clinical settings.

Semantic Representation Alignment. Furthermore, LLMs can serve as a bridge for learning concept-level associations between signal patterns and their semantic interpretations [20], [56], [57], [60], [195]. In this setup, LLM acts as a text encoder for derive semantic representations from biosignals associated texts, followed by alignment between biosignal and text. This approach is motivated by the observation that textual labels often capture rich inter-class relationships, such as hierarchical structures and synonymy, which can help ground and structure noisy signal representations. Existing work like BELT-2 [60] and MERL [195] use contrastive learning objectives to align paired signal-text samples, enabling generalization to unseen conditions. NormWear [17] adapts a pretrained biosignal encoder and aligns its signal embeddings with text embeddings derived from associated textual annotations by pretrained LLMs. This facilitates question answering with the biosignal encoder for clinically relevant tasks, such as interpreting physiological patterns and identifying abnormal events.

6.4 Joint Biosignal-Text Modeling with LLMs

LLMs can further be used to model biosignal and language information in tandem, either by generating one modality from the other or by integrating both during inference for joint decision-making. In many clinical and sensing applications, biosignals are inherently linked to textual information and are rarely analyzed in isolation. LLMs provide a powerful framework for capturing this multimodal interplay, enabling semantically aligned representation learning across modalities. This section reviews two principal strategies: (i) text generation from biosignals, and (ii) biosignal–text fusion for joint inference.

Text Generation from Biosignals. In this setting, the LLM serves as a conditional generator that translates biosignal-related inputs, such as raw signals, latent embeddings, or classification results, into textual outputs. This leverages the LLM's strength in language generation, enabling the creation of interpretable descriptions and associated health outcomes. Such cross-modal generation is increasingly common in various biosignal scenarios, including generating cardiologist-style reports from ECG data [20], [184], [196], or textual outputs from EEG signals [197]. Instruction-tuning based models [184], [196] and architectures that combine a pretrained biosignal model with LLMs [20] exemplify this approach by conditioning on biosignal-derived representations to generate clinically meaningful text.

Biosignal-Text Fusion. Beyond cross-modal generation, LLMs can also serve as reasoning engines that integrate biosignal features with textual context for joint decision-making [71], [72], [182]. This modality fusion is particularly important in tasks where biosignals alone lack sufficient contextual information. A pioneering work in general time series analysis [187] showed that combining contextual information with tokenized time series as input to LLMs can significantly improve performance on forecasting tasks. This conceptual initiative has also been extended to biosignal analysis, for instance, Emotion-Copilot [72] combines EEG features and demographic information together as input to fine-tune an LLM for personalized emotion recognition.

6.5 LLMs as Agents for Biosignal Interpretation

Beyond static predictions, LLMs also support interactive interfaces capable of reasoning over biosignal data. Depending on their functional roles, existing works can be broadly categorized into two groups: (i) *conversational agents*

that engage directly with users, and (ii) *workflow agents* that coordinate multi-step biosignal analysis.

Conversational Agents. Several studies have presented preliminary explorations of embedding LLMs into clinician- or patient-facing chatbots, enabling natural language access to biosignal interpretation. These models translate physiological signals into actionable advice, often grounded in personal health history or established medical guidelines. For instance, PhysioLLM [101] and Sensor2Text [63] analyze wearable sensor data and respond to user queries, making signal-level information more accessible and interpretable. Other notable systems include PH-LLM [71], openCHA [58], Emotion-Copilot [72], and LLaSA [173], each contributing unique interaction paradigms tailored to specific biosignal types and healthcare scenarios.

Biosignal Analysis Workflow Agents. Beyond dialogue, LLMs can also serve as central controllers in biosignal analysis workflows, interpreting input signals, querying external databases, calling established biosignal processing toolboxes, and integrating outputs from time series foundation models (see Sections 3.2 and 3.3). This modular design supports flexible use of diverse sensing modalities without having a separate foundation model for each. Moreover, recent studies also show that lightweight or traditional methods can achieve comparable performance to large LLMs or time series foundation models in specific tasks [42], [178]. This highlights that *when* and *how* these methods are deployed may be more critical than training increasingly large models. By leveraging LLMs as reasoning engines to selectively invoke the most appropriate tools or models for a given context, systems can achieve efficiency without compromising performance. Systems such as PHIA [190] and openCHA exemplify this capability, while models like ECG-Agent [198] and TeachMe [196] further integrate biosignal representations with clinical decision-making logic.

7 Challenges and Opportunities

While large foundation models have demonstrated promise across multiple physiological modalities, deploying these models in real-world healthcare settings requires addressing a series of practical, technical, and ethical barriers. Below, we articulate key areas that merit deeper attention from both a research and translational standpoint.

7.1 Standardization and Harmonization

Scaling foundation models depends on aggregating biosignal data from diverse sources. However, inconsistencies in signal formats, sampling rates, device configurations, and label definitions pose significant barriers. Standardization involves unifying signal-level properties (e.g., resampling, normalization, channel ordering), while harmonization addresses semantic alignment (e.g., resolving label mismatches, integrating partial modalities). Despite ongoing efforts as discussed in Section 4.4, and existing device-specific standards (e.g., IEEE 11073), the absence of universally adopted metadata schemas and acquisition protocols remains a significant obstacle. Developing interoperable ontologies, structured biosignal metadata standards, and robust multi-source harmonization pipelines is therefore essential for scalable and generalizable pretraining and deployment of biosignal foundation models.

7.2 Security, Safety and Sovereignty

Given the sensitive nature of biosignal data, the development and deployment of foundation models in healthcare must prioritize data security and safety as well as technological sovereignty. Traditional centralized training introduces significant privacy risks, especially as foundation models may inadvertently memorize or leak patient-identifiable information. To address these challenges, approaches such as federated learning and differential privacy are increasingly adopted, limiting raw data exposure and reducing the risk of data breaches. In parallel, model robustness is critical: subtle adversarial perturbations can fool biosignal models into generating incorrect predictions while appearing clinically plausible. Defensive strategies such as adversarial training and out-of-distribution detection can help to mitigate these vulnerabilities. Another pressing concern is hallucinations, which arise especially with the integration of biosignal models with LLMs. Emerging solutions in LLMs are needed to tackle this challenge in high-stakes clinical settings. In parallel, the issue of sovereignty (i.e., ensuring local control over data, models, and deployment), has become increasingly important. The use of foreign AI infrastructure raises critical concerns, particularly for biosignal-based health-related foundation models [199].

7.3 Efficiency and Deployability

Many foundation models, especially those based on LLMs and TSFMs are large and computationally expensive, limiting their deployment on wearable devices or edge platforms. Real-time inference introduces further constraints, where even modest latency can disrupt clinical workflows. Moreover, reliance on cloud-based computation may raise concerns around data privacy, especially in low-resource settings. To overcome these issues, a range of techniques has been developed, including model compression (e.g., pruning, quantization, low-rank factorization), knowledge distillation from large models to smaller students, and architectural innovations such as efficient Transformers with linear or sparse attention mechanisms. Dynamic inference strategies, such as early-exit mechanisms, further help reduce computation when predictions are confident, balancing speed and accuracy. Ultimately, the goal is to make foundation models usable in practical, embedded clinical systems without sacrificing performance.

7.4 Interpretability and Clinical Trust

Interpretability is crucial for clinical adoption. Unlike image or text, where visual or linguistic cues may offer intuitive justifications, biosignal interpretations often rely on subtle physiological patterns that are not easily discerned. Foundation models typically act as black boxes, offering little explanation for their outputs. For example, an ECG model might flag arrhythmias without indicating which waveform features triggered the decision. This opacity raises concerns about spurious correlations and hidden biases. Efforts to improve interpretability include saliency-based visualization, attribution methods, attention heatmaps, and uncertainty estimation. Making model predictions explainable in physiologically meaningful terms, as well as communicating uncertainty in high-stakes scenarios, is essential for building trust among clinicians and ensuring appropriate use.

7.5 Benchmarking and Evaluation

Current benchmarking practices in biosignal foundation modeling remain fragmented. Most studies evaluate models on custom datasets with varying metrics, experiment settings, and preprocessing strategies, making direct comparisons difficult. Unlike NLP and CV, which have well-established shared tasks and datasets (e.g., ImageNet), the biosignal domain lacks standardized benchmarks. Metrics like classification accuracy may not reflect real-world performance, where false alarms, calibration, robustness to population shifts, and generalization to new settings matter more. Efforts such as PhysioNet provide valuable open data resources, but expansion to valid biosignal benchmarks is needed. Additionally, sharing pretrained foundation models could reduce redundant training and enable fair, reproducible head-to-head comparisons. As highlighted in a recent review [200], better alignment between evaluation protocols and clinical value is urgently needed.

8 Conclusion

The recent surge of interest in foundation models has reshaped the landscape of machine learning, with transformative impacts already evident in language, vision, and increasingly, time series domains. Biomedical sensing, as a rich source of physiological and behavioral time series data, stands to benefit greatly from this paradigm shift. However, progress in this area remains fragmented, with no cohesive or unified guidelines to address the domain-specific challenges intrinsic to biosignals.

In this survey, we offered a comprehensive examination of foundation model development and application for biosignals. We structured our discussion around three major directions: developing foundation models from scratch tailored to biosignals; adapting general-purpose time series models to biomedical domains; and leveraging large (multimodal) language models for biosignal analysis. Along the way, we highlighted key challenges such as channel/resolution heterogeneity, multimodal integration, and clinical task complexity, and surveyed strategies to address them.

Despite encouraging progress, this is only the beginning. Standardized benchmarks, more robust evaluation protocols, clinically meaningful downstream tasks, and better integration of domain knowledge are all necessary steps for translating foundation models into real-world impact. As the community continues to grow, we believe a more unified and principled approach, grounded in both machine learning and clinical needs, will be essential for unlocking the full potential of biosignal foundation models. We envision this survey serves not only as a guide to the current landscape but also as a catalyst for future research at the intersection of machine learning, biosignal modeling, and healthcare.

References

[1] X. Gu *et al.*, "Beyond supervised learning for pervasive healthcare," *IEEE Rev. Biomed. Eng.*, 2023.

[2] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] A. Grattafiori *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[6] A. Kirillov *et al.*, "Segment anything," in *Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[7] Y. Liang *et al.*, "Foundation models for time series analysis: A tutorial and survey," in *Proc. ACM SIGKDD conf. knowl. disc. and data min.*, 2024, pp. 6555–6565.

[8] J. Ye *et al.*, "A survey of time series foundation models: Generalizing time series representation with large language mode," *arXiv preprint arXiv:2405.02358*, 2024.

[9] H. Yuan *et al.*, "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data," *npj Digit. Med.*, vol. 7, no. 1, p. 91, 2024.

[10] D. Zhang *et al.*, "Brant: Foundation model for intracranial neural signal," in *Adv. Neural Inform. Process. Syst.*, 2023.

[11] A. Vaid *et al.*, "A foundational vision transformer improves diagnostic performance for electrocardiograms," *npj Digit. Med.*, vol. 6, no. 1, p. 108, 2023.

[12] S. Abbaspourazad *et al.*, "Large-scale training of foundation models for wearable biosignals," *arXiv preprint arXiv:2312.05409*, 2023.

[13] K. Yi *et al.*, "Learning topology-agnostic eeg representations with geometry-aware modeling," in *Adv. Neural Inform. Process. Syst.*, 2023.

[14] C. Ding *et al.*, "Siamquality: a convnet-based foundation model for photoplethysmography signals," *Physiol. Meas.*, vol. 45, no. 8, p. 085004, 2024.

[15] A. Logacjov *et al.*, "Selfpab: large-scale pre-training on accelerometer data for human activity recognition," *Appl. Intell.*, vol. 54, no. 6, pp. 4545–4563, 2024.

[16] Y. Zhang *et al.*, "Towards open respiratory acoustic foundation models: Pretraining and benchmarking," *arXiv preprint arXiv:2406.16148*, 2024.

[17] Y. Luo *et al.*, "Toward foundation model for multivariate wearable sensing of physiological signals," *arXiv preprint arXiv:2412.09758*, 2024.

[18] C. labs at Reality Labs *et al.*, "A generic noninvasive neuromotor interface for human-computer interaction," *bioRxiv*, pp. 2024–02, 2024.

[19] Y. Chen *et al.*, "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," *arXiv preprint arXiv:2401.10278*, 2024.

[20] Y. Tian *et al.*, "Foundation model of ecg diagnosis: Diagnostics and explanations of any form and rhythm on ecg," *Cell Rep. Med.*, vol. 5, no. 12, 2024.

[21] W.-B. Jiang *et al.*, "Large brain model for learning generic representations with tremendous eeg data in bci," *arXiv preprint arXiv:2405.18765*, 2024.

[22] G. Mathew *et al.*, "Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes," *npj Cardiovasc. Health*, vol. 1, no. 1, p. 25, 2024.

[23] J. Song *et al.*, "Foundation models for ecg: Leveraging hybrid self-supervised learning for advanced cardiac diagnostics," *arXiv preprint arXiv:2407.07110*, 2024.

[24] G. Narayanswamy *et al.*, "Scaling wearable foundation models," *arXiv preprint arXiv:2410.13638*, 2024.

[25] A. Pillai *et al.*, "Papagei: Open foundation models for optical physiological signals," in *Int. Conf. Learn. Represent.*, 2025.

[26] X. Gu *et al.*, "Sensing cardiac health across scenarios and devices: A multi-modal foundation model pretrained on heterogeneous data from 1.7 million individuals," *arXiv preprint arXiv:2507.01045*, 2025.

[27] M. A. Xu *et al.*, "Lsm-2: Learning from incomplete wearable sensor data," *arXiv e-prints*, pp. arXiv–2506, 2025.

[28] M. Saha *et al.*, "Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings," *arXiv preprint arXiv:2502.01108*, 2025.

[29] E. Erturk *et al.*, "Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions," *arXiv preprint arXiv:2507.00191*, 2025.

[30] C. Wang *et al.*, "Chattime: A unified multimodal time series foundation model bridging numerical and textual data," in *AAAI Conf. Artif. Intell.*, 2025.

[31] X. Shi *et al.*, "Time-moe: Billion-scale time series foundation models with mixture of experts," in *Int. Conf. Learn. Represent.*, 2025.

[32] E. O. Taga, M. E. Ildiz, and S. Oymak, "Timepfn: Effective multivariate time series forecasting with synthetic data," in *AAAI Conf. Artif. Intell.*, vol. 39, no. 19, 2025, pp. 20 761–20 769.

[33] A. F. Ansari *et al.*, "Chronos: Learning the language of time series," *Trans. Mach. Learn. Research*, 2024.

[34] G. Woo *et al.*, "Unified training of universal time series forecasting transformers," in *Int. Conf. Mach. Learn.*, 2024.

[35] M. Goswami *et al.*, "Moment: A family of open time-series foundation models," in *Int. Conf. Mach. Learn.*, 2024.

[36] D. Cao, W. Ye, and Y. Liu, "Timedit: General-purpose diffusion transformers for time series foundation model," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[37] V. Ekambaram *et al.*, "Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series," in *Adv. Neural Inform. Process. Syst.*, 2024.

[38] S. Gao *et al.*, "Units: A unified multi-task time series model," in *Adv. Neural Inform. Process. Syst.*, 2024.

[39] K. Rasul *et al.*, "Lag-llama: Towards foundation models for time series forecasting," in *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[40] A. Garza *et al.*, "Timegpt-1," *arXiv preprint arXiv:2310.03589*, 2023.

[41] A. Das *et al.*, "A decoder-only foundation model for time-series forecasting," in *Int. Conf. Mach. Learn.*, 2024.

[42] X. Gu *et al.*, "Are time series foundation models ready for vital sign forecasting in healthcare?" in *Machine Learning for Health (ML4H) Symposium 2024*, 2024.

[43] D. Gupta *et al.*, "Low-rank adaptation of time series foundational models for out-of-domain modality forecasting," in *Proceedings of the 26th International Conference on Multimodal Interaction*, 2024, pp. 382–386.

[44] M. Liu *et al.*, "Generalized prompt tuning: Adapting frozen univariate time series foundation models for multivariate healthcare time series," *arXiv preprint arXiv:2411.12824*, 2024.

[45] C. Wu *et al.*, "Efficient personalized adaptation for physiological signal foundation model," in *Forty-second International Conference on Machine Learning*.

[46] "ChatGPT," openai.com/index/chatgpt/, 2023.

[47] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[48] "Gemini," deepmind.google/models/gemini/, 2023.

[49] "GPT-4o," openai.com/index/gpt-4o-system-card/, 2024.

[50] "LLaMA3," ai.meta.com/blog/meta-llama-3/, 2024.

[51] "Claude3," www.anthropic.com/news/claude-3-family, 2024.

[52] "GPT4.5," openai.com/index/introducing-gpt-4-5/, 2025.

[53] "Llama 4," ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

[54] D. Guo *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[55] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 5, 2022, pp. 5350–5358.

[56] S. Moon *et al.*, "Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text," *arXiv preprint arXiv:2210.14395*, 2022.

[57] J. Qiu *et al.*, "Can brain signals reveal inner alignment with human languages?" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1789–1804.

[58] M. Abbasian *et al.*, "Conversational health agents: A personalized llm-powered agent framework," *arXiv preprint arXiv:2310.02374*, 2023.

[59] X. Liu *et al.*, "Large language models are few-shot health learners," *arXiv preprint arXiv:2305.15525*, 2023.

[60] J. Zhou *et al.*, "Belt-2: Bootstrapping eeg-to-language representation alignment for multi-task brain decoding," *arXiv preprint arXiv:2409.00121*, 2024.

[61] J. Li *et al.*, "Frozen language model helps ecg zero-shot learning," in *Medical Imaging Deep Learn.*, 2024, pp. 402–415.

[62] J. Wang *et al.*, "Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 7278–7292.

[63] W. Chen *et al.*, "Sensor2text: Enabling natural language interactions for daily activity tracking using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–26, 2024.

[64] H. Yu, P. Guo, and A. Sano, "Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text," *Trans. Mach. Learn. Research*, 2024.

[65] S. Ji *et al.*, "Hargpt: Are llms zero-shot human activity recognizers?" in *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems and Internet of Things*, 2024, pp. 38–43.

[66] Y. Kim *et al.*, "Health-llm: Large language models for health prediction via wearable sensor data," in *Proceedings of the fifth Conference on Health, Inference, and Learning*, ser. Proc. Mach. Learn. Research, vol. 248, 2024, pp. 522–539.

[67] Z. Leng *et al.*, "Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, pp. 1–32, 2024.

[68] M. Daswani *et al.*, "Plots unlock time-series understanding in multimodal models," *arXiv preprint arXiv:2410.02637*, 2024.

[69] C. Chen *et al.*, "Large language model-informed ecg dual attention network for heart failure risk prediction," *IEEE Trans. Big Data*, 2025.

[70] C. Liu *et al.*, "Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement," *arXiv preprint arXiv:2403.06659*, 2024.

[71] J. Cosentino *et al.*, "Towards a personal health large language model," in *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.

[72] H. Chen *et al.*, "Eeg emotion copilot: Optimizing lightweight llms for emotional eeg interpretation with assisted medical record generation," p. 107848, 2025.

[73] K. Yang *et al.*, "Ecg-lm: Understanding electrocardiogram with a large language model," *Health Data Science*, vol. 5, p. 0221, 2025.

[74] C. Liu *et al.*, "Knowledge-enhanced multimodal ecg representation learning with arbitrary-lead inputs," *arXiv preprint arXiv:2502.17900*, 2025.

[75] A. Pillai *et al.*, "Beyond prompting: Time2lang–bridging time-series foundation models and large language models for health sensing," *arXiv preprint arXiv:2502.07608*, 2025.

[76] M. Awais *et al.*, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[77] N. Madan *et al.*, "Foundation models for video understanding: A survey," *arXiv preprint arXiv:2405.03770*, 2024.

[78] P. Wagner *et al.*, "Ptb-xl, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, pp. 1–15, 2020.

[79] N. Strodthoff *et al.*, "Mimic-iv-ecg-ext-icd: Diagnostic labels for mimic-iv-ecg (version 1.0.1)," 2024.

[80] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, p. R1, 2007.

[81] Y. Liang *et al.*, "An optimal filter for short photoplethysmogram signals," *Sci. Data*, vol. 5, no. 1, pp. 1–12, 2018.

[82] A. Shcherbina *et al.*, "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," *J. Pers. Med.*, vol. 7, no. 2, p. 3, 2017.

[83] R. P. Rao, *Brain-computer interfacing: an introduction*, 2013.

[84] B. K. Hodossy *et al.*, "Leveraging high-density emg to investigate bipolar electrode placement for gait prediction models," *IEEE T. Hum.-Mach. Syst.*, 2024.

[85] S. Muceli and R. Merletti, "Tutorial. frequency analysis of the surface emg signal: best practices," *J. Electromyogr. Kinesiol.*, vol. 79, p. 102937, 2024.

[86] H. J. Hermens *et al.*, "Development of recommendations for semg sensors and sensor placement procedures," *J. Electromyogr. Kinesiol.*, vol. 10, no. 5, pp. 361–374, 2000.

[87] W.-H. Weng *et al.*, "Predicting cardiovascular disease risk using photoplethysmography and deep learning," *PLOS Glob. Public Health*, vol. 4, no. 6, p. e0003204, 2024.

[88] E. Chen *et al.*, "Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 45 794–45 811, 2023.

[89] L.-w. Lehman *et al.*, "Vtac: A benchmark dataset of ventricular tachycardia alarms from icu monitors," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 38 827–38 843, 2023.

[90] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nat. Commun.*, vol. 11, no. 1, p. 1760, 2020.

[91] C. Yang *et al.*, "Biot: Biosignal transformer for cross-data learning in the wild," *Adv. Neural Inform. Process. Syst.*, vol. 36, 2024.

[92] M. Burger *et al.*, "Towards foundation models for critical care time series," *arXiv preprint arXiv:2411.16346*, 2024.

[93] Z. Yuan *et al.*, "Brainwave: A brain signal foundation model for clinical applications," *arXiv preprint arXiv:2402.10251*, 2024.

[94] S. Chevallier *et al.*, "The largest eeg-based bci reproducibility study for open science: the moabb benchmark," *arXiv preprint arXiv:2404.15319*, 2024.

[95] S. Salter *et al.*, "emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation," in *Adv. Neural Inform. Process. Syst. Datasets and Benchmarks Track*.

[96] V. Sivakumar *et al.*, "emg2qwerty: A large dataset with baselines for touch typing using surface electromyography," in *Adv. Neural Inform. Process. Syst.*, vol. 37, 2024, pp. 91 373–91 389.

[97] C. Bhavana *et al.*, "Techniques of measurement for parkinson's tremor highlighting advantages of embedded imu over emg," in *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, 2016, pp. 1–5.

[98] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS one*, vol. 12, no. 5, p. e0177926, 2017.

[99] X. Zhang *et al.*, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 3988–4003, 2022.

[100] G. Narayanswamy *et al.*, "Scaling wearable foundation models," in *Int. Conf. Learn. Represent.*, 2025.

[101] C. M. Fang *et al.*, "Physiollm: Supporting personalized health insights with wearables and large language models," in *IEEE-EMBS Int. Conf. Biomed. Health. Inf.*

[102] Y. Wang *et al.*, "Deep time series models: A comprehensive survey and benchmark," *arXiv preprint arXiv:2407.13278*, 2024.

[103] B. Xiong *et al.*, "Patchemg: Few-shot emg signal generation with diffusion models for data augmentation to improve classification performance," *IEEE Trans. Instrum. Meas.*, 2024.

[104] B. van Breugel *et al.*, "Synthetic data in biomedicine via generative artificial intelligence," *Nat. Rev. Bioeng.*, vol. 2, no. 12, pp. 991–1004, 2024.

[105] A. Radhakrishnan *et al.*, "Cross-modal autoencoder framework learns holistic representations of cardiovascular state," *Nat. Commun.*, vol. 14, no. 1, p. 2436, 2023.

[106] X. Tian *et al.*, "Cross-domain joint dictionary learning for ecg reconstruction from ppg," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 936–940.

[107] Y. Wang *et al.*, "Contrast everything: A hierarchical contrastive framework for medical time-series," *Adv. Neural Inform. Process. Syst.*, vol. 36, 2024.

[108] S. Abbaspourazad *et al.*, "Wearable accelerometer foundation models for health via knowledge distillation," *arXiv preprint arXiv:2412.11276*, 2024.

[109] A. Doryab *et al.*, "Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data," *JMIR mHealth uHealth*, vol. 7, no. 7, p. e13209, 2019.

[110] Y. Liu *et al.*, "itransformer: Inverted transformers are effective for time series forecasting," in *Int. Conf. Learn. Represent.*, 2024.

[111] H. Xue and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," *IEEE Trans. Knowl. Data Eng.*, 2023.

[112] N. Hinrichs *et al.*, "Short-term vital parameter forecasting in the intensive care unit: A benchmark study leveraging data from patients after cardiothoracic surgery," *PLOS Digit. Health*, vol. 3, no. 9, p. e0000598, 2024.

[113] Y. Wang *et al.*, "Motion intention prediction and joint trajectories generation toward lower limb prostheses using emg and imu signals," *IEEE Sens. J.*, vol. 22, no. 11, pp. 10 719–10 729, 2022.

[114] P. Renc *et al.*, "Zero shot health trajectory prediction using transformer," *npj Digit. Med.*, vol. 7, no. 1, p. 256, 2024.

[115] E. M. Lima *et al.*, "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nat. Commun.*, vol. 12, no. 1, p. 5117, 2021.

[116] E. Steinberg *et al.*, "Motor: A time-to-event foundation model for structured medical records," in *Int. Conf. Learn. Represent.*, 2024.

[117] J. Wang *et al.*, "Cbramod: A criss-cross brain foundation model for eeg decoding," in *Int. Conf. Learn. Represent.*, 2025.

[118] J. Jin *et al.*, "Reading your heart: Learning ecg words and sentences via pre-training ecg language model," in *Int. Conf. Learn. Represent.*, 2025.

[119] W. Jiang *et al.*, "Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals," in *Int. Conf. Learn. Represent.*, 2025.

[120] M. A. Xu *et al.*, "Relcon: Relative contrastive learning for a motion foundation model for wearable data," in *Int. Conf. Learn. Represent.*, 2025.

[121] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[122] Y. Nie *et al.*, "A time series is worth 64 words: Long-term forecasting with transformers," in *Int. Conf. Learn. Represent.*, 2023.

[123] Y. Wang *et al.*, "Medformer: A multi-granularity patching transformer for medical time-series classification," in *Adv. Neural Inform. Process. Syst.*

[124] Y. Zhang *et al.*, "Multi-resolution time-series transformer for long-term forecasting," in *International conference on artificial intelligence and statistics*, 2024, pp. 4222–4230.

[125] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bull. Am. Meteorol. Soc.*, vol. 79, no. 1, pp. 61–78, 1998.

[126] X. Li *et al.*, "Bat: Beat-aligned transformer for electrocardiogram classification," in *IEEE Int. Conf. Data Min.*, 2021, pp. 320–329.

[127] H. Zhou *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *AAAI Conf. Artif. Intell.*, 2021, pp. 11 106–11 115.

[128] H. Wu *et al.*, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Adv. Neural Inform. Process. Syst.*, vol. 34, 2021, pp. 22 419–22 430.

[129] J. Y. Zhou *et al.*, "Nonlinear time-series embedding by monotone variational inequality," *arXiv preprint arXiv:2406.06894*, 2024.

[130] Z. Wan *et al.*, "Eegformer: A transformer–based brain activity classification method using eeg signal," *Front. Neurosci.*, vol. 17, p. 1148855, 2023.

[131] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 873–12 883.

[132] Y. Shu and V. Lampos, "Deformtime: Capturing variable dependencies with deformable attention for time series forecasting," *Trans. Mach. Learn. Research*, 2025.

[133] J. Zhang *et al.*, "Elastst: Towards robust varied-horizon forecasting with elastic time-series transformer," in *Adv. Neural Inform. Process. Syst.*, 2024.

[134] X. Liu *et al.*, "Unitime: A language-empowered unified model for cross-domain time series forecasting," in *Proc. ACM Web Conf.*, 2024, pp. 4095–4106.

[135] S. Dooley *et al.*, "Forecastpfn: Synthetically-trained zero-shot forecasting," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 2403–2426, 2023.

[136] S. Lin *et al.*, "Segrnn: Segment recurrent neural network for long-term time series forecasting," *arXiv preprint arXiv:2308.11200*, 2023.

[137] Y. Jia *et al.*, "Witran: Water-wave information transmission and recurrent acceleration network for long-range time series forecasting," in *Adv. Neural Inform. Process. Syst.*, 2023.

[138] K. Yi *et al.*, "Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective," in *Adv. Neural Inform. Process. Syst.*, 2023.

[139] J. Han *et al.*, "Eeg decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks," *J. Neural Eng.*, vol. 20, no. 6, p. 066027, 2023.

[140] X. Liu *et al.*, "Moirai-moe: Empowering time series foundation models with sparse mixture of experts," *arXiv preprint arXiv:2410.10469*, 2024.

[141] A. Borzunov *et al.*, "Training transformers together," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, ser. Proc. Mach. Learn. Research, vol. 176, 2022, pp. 335–342.

[142] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[143] Z. Dai *et al.*, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[144] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.

[145] G. Lai *et al.*, "Modeling long- and short-term temporal patterns with deep neural networks," in *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 95–104.

[146] Y. Liu *et al.*, "Timer: Generative pre-trained transformers are large time series models," in *Int. Conf. Mach. Learn.*, 2024.

[147] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *Int. Conf. Learn. Represent.*, 2023.

[148] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 000–16 009.

[149] M. Lewis *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[150] A. Zeng *et al.*, "Are transformers effective for time series forecasting?" in *AAAI Conf. Artif. Intell.*, 2023, pp. 11 121–11 128.

[151] Y. Luo *et al.*, "Toward foundation model for multivariate wearable sensing of physiological signals," *arXiv preprint arXiv:2412.09758*, 2024.

[152] C. Wang *et al.*, "Brainbert: Self-supervised representation learning for intracranial recordings," *arXiv preprint arXiv:2302.14367*, 2023.

[153] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in *Int. Conf. Mach. Learn.*, 2021, pp. 5606–5615.

[154] X. Lan *et al.*, "Intra-inter subject self-supervised learning for multivariate cardiac signals," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 4, 2022, pp. 4532–4540.

[155] S. Lee, T. Park, and K. Lee, "Soft contrastive learning for time series," in *Int. Conf. Learn. Represent.*

[156] A. Raghu *et al.*, "Sequential multi-dimensional self-supervised learning for clinical time series," in *Int. Conf. Mach. Learn.*, 2023.

[157] V. Sangha *et al.*, "Biometric contrastive learning for data-efficient deep learning from electrocardiographic images," *J. Am. Med. Inf. Assoc.*, vol. 31, no. 4, pp. 855–865, 2024.

[158] H. Yèche *et al.*, "Neighborhood contrastive learning applied to online patient monitoring," in *Int. Conf. Mach. Learn.*, 2021.

[159] X. Gu *et al.*, "Transforming label-efficient decoding of healthcare wearables and "embedded" medical domain expertise," *Nat. Commun. Eng.*, 2025.

[160] H. Zhang *et al.*, "Sleeppriorcl: Contrastive representation learning with prior knowledge-based positive mining and adaptive temperature for sleep staging," *arXiv preprint arXiv:2110.09966*, 2021.

[161] M. T. Nonnenmacher *et al.*, "Utilizing expert features for contrastive learning of time-series representations," in *Int. Conf. Mach. Learn.*, 2022.

[162] R. Thapa *et al.*, "Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[163] L. Dong *et al.*, "Reference electrode standardization interpolation technique (resit): a novel interpolation method for scalp eeg," *Brain Topogr.*, vol. 34, no. 4, pp. 403–414, 2021.

[164] M. Svantesson *et al.*, "Virtual eeg-electrodes: Convolutional neural networks as a method for upsampling or restoring channels," *J. Neurosci. Methods*, vol. 355, p. 109126, 2021.

[165] J. Lai *et al.*, "Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset," *Nat. Commun.*, vol. 14, no. 1, p. 3741, 2023.

[166] X. Gu *et al.*, "Generalizable movement intention recognition with multiple heterogeneous eeg datasets," in *Int. Conf. Robot. Autom.*, 2023, pp. 9858–9864.

[167] D. Kiyasseh *et al.*, "A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions," *Nat. Commun.*, vol. 12, no. 1, p. 4221, 2021.

[168] K. Saleh *et al.*, "Ultra-high-frequency ecg assessment of qrs fragmentation predicts sudden cardiac death risk in inherited arrhythmia syndromes," *Eur. Heart J.*, vol. 43, no. Supplement_2, pp. ehac544–678, 2022.

[169] L. N. Darlow *et al.*, "Dam: Towards a foundation model for forecasting," in *Int. Conf. Learn. Represent.*, 2024.

[170] J. Li *et al.*, "An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains," *arXiv preprint arXiv:2410.04133*, 2024.

[171] E. Coppola *et al.*, "Hubert-ecg: a self-supervised foundation model for broad and scalable cardiac applications," *medRxiv*, pp. 2024–11, 2024.

[172] Y. Wang *et al.*, "Anyecg: Foundational models for electrocardiogram analysis," *arXiv preprint arXiv:2411.17711*, 2024.

[173] S. A. Imran *et al.*, "Llasa: A multimodal llm for human activity analysis through wearable and smartphone sensors," *arXiv preprint arXiv:2406.14498*, 2024.

[174] Y. Liu *et al.*, "Timer-xl: Long-context transformers for unified time series forecasting," in *Int. Conf. Learn. Represent.*, 2025.

[175] N. Huang *et al.*, "Repurposing foundation model for generalizable medical time series classification," *arXiv preprint arXiv:2410.03794*, 2024.

[176] Q. Liu *et al.*, "Time-ffm: Towards lm-empowered federated foundation model for time series forecasting," in *Adv. Neural Inform. Process. Syst.*, 2024.

[177] L. Yang *et al.*, "Vitime: A visual intelligence-based foundation model for time series forecasting," *arXiv preprint arXiv:2407.07311*, 2024.

[178] D. Gupta, A. Bhatti, and S. Parmar, "Beyond lora: Exploring efficient fine-tuning techniques for time series foundational models," *arXiv preprint arXiv:2409.11302*, 2024.

[179] H. Zhang *et al.*, "Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting," *arXiv preprint arXiv:2412.20810*, 2024.

[180] H. Ma *et al.*, "A mamba foundation model for time series forecasting," *arXiv preprint arXiv:2411.02941*, 2024.

[181] T. Zhou *et al.*, "One fits all: Power general time series analysis by pretrained lm," *Adv. Neural Inform. Process. Syst.*, vol. 36, 2023.

[182] Y. Zhao *et al.*, "Ecg-chat: A large ecg-language model for cardiac disease diagnosis," *arXiv preprint arXiv:2408.08849*, 2024.

[183] H. Yang *et al.*, "Are you being tracked? discover the power of zero-shot trajectory tracing with llms!" in *2024 IEEE Coupling of Sensing & Computing in AIoT Systems (CSCAIoT)*, 2024, pp. 13–18.

[184] Z. Wan *et al.*, "Meit: Multi-modal electrocardiogram instruction tuning on large language models for report generation," *arXiv preprint arXiv:2403.04945*, 2024.

[185] X. Zhang *et al.*, "Large language models for time series: a survey," in *Int. Joint Conf. Artif. Intell.*, 2024, pp. 8335–8343.

[186] J. Han *et al.*, "Onellm: One framework to align all modalities with language," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 26 584–26 595.

[187] M. Jin *et al.*, "Time-llm: Time series forecasting by reprogramming large language models," *arXiv preprint arXiv:2310.01728*, 2023.

[188] W. Cui *et al.*, "Neuro-gpt: Towards a foundation model for eeg," in *Proc. IEEE Int. Symp. Biomed. Imaging*, 2024, pp. 1–5.

[189] Q. Zhang *et al.*, "Vision transformer with quadrangle attention," *arXiv preprint arXiv:2303.15105*, 2023.

[190] M. A. Merrill *et al.*, "Transforming wearable data into health insights using large language model agents," *arXiv preprint arXiv:2406.06464*, 2024.

[191] T. Seki *et al.*, "Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ecg image interpretation," *Front. Cardiovasc. Med.*, vol. 12, p. 1458289, 2025.

[192] Y. Zhang *et al.*, "From word embedding to reading embedding using large language model, eeg and eye-tracking," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[193] N. Chan *et al.*, "Medtsllm: Leveraging llms for multimodal medical time series analysis," *arXiv preprint arXiv:2408.07773*, 2024.

[194] J. Huang *et al.*, "Large language models enabled multiagent ensemble method for efficient ehr data labeling," *arXiv preprint arXiv:2410.16543*, 2024.

[195] H. Yu, P. Guo, and A. Sano, "Zero-shot ecg diagnosis with large language models and retrieval-augmented generation," in *Machine learning for health (ML4H)*, 2023, pp. 650–663.

[196] R. Liu *et al.*, "Teach multimodal llms to comprehend electrocardiographic images," *arXiv preprint arXiv:2410.19008*, 2024.

[197] J. Lévy *et al.*, "Brain-to-text decoding: A non-invasive approach via typing," *arXiv preprint arXiv:2502.17480*, 2025.

[198] J. Oh *et al.*, "Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 66 277–66 288, 2023.

[199] D. L. Shrier *et al.*, "Considerations regarding sovereign ai and national ai policy," Sovereign-AI.org, Tech. Rep., 2025.

[200] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *npj Digit. Med.*, vol. 6, no. 1, p. 135, 2023.

APPENDIX

LLM Adaptation Strategies

Despite their reasoning and generation capabilities in language and vision tasks, LLMs are not inherently suitable for high-dimensional, temporally structured data like biosignals or for complex medical tasks that require domain knowledge. Bridging this gap requires targeted adaptation strategies that align LLM capabilities with biomedical signal characteristics. We thereby outline three key paradigms for adapting and tuning LLMs as below:

Prompt Tuning. Prompt tuning guides LLM behavior at inference time without modifying model parameters, using structured prompts to leverage pretrained capabilities. Common strategies include: (a) *Zero- / Few-Shot Prompting*, where the model is prompted with task instructions and, optionally, a few labeled examples (e.g., signal–label pairs) to generalize to unseen cases; (b) *Chain-of-Thought (CoT) Prompting*, which encourages multi-step reasoning by prompting the model to articulate intermediate steps, particularly valuable for clinical interpretation tasks; and (c) *Retrieval-Augmented Generation (RAG)*, which dynamically retrieves external documents (e.g., patient records, clinical guidelines) to enrich the model's contextual understanding.

Parameter-Efficient Fine-Tuning. When inference-only strategies are insufficient and full finetuning is computationally prohibitive, parameter-efficient fine-tuning (PEFT) offers a practical compromise. PEFT methods adapt LLMs by updating a small set of trainable parameters, while keeping the majority of the model fixed. Popular PEFT techniques include: (a) *LoRA* (Low-Rank Adaptation), which introduces trainable low-rank matrices within attention modules; (b) *Prefix Tuning*, which prepends learned prefix tokens to the input to guide model behavior; and (c) *Adapter Modules*, which add lightweight layers between transformer blocks for task-specific adaptation. PEFT is particularly well suited for repurposing pretrained LLMs as backbones in biosignal tasks such as classification and prediction (see Section 6.2).

Instruction Tuning. Instruction tuning involves supervised finetuning of LLMs on task-specific input–output pairs, typically in the form of natural language instructions and their corresponding expert responses. In biomedical sensing, input often includes signal and signal-derived prompts (e.g.,"Interpret the ECG waveform"), and supervising it with expert annotations (e.g., diagnostic labels, textual interpretations, or clinical decisions). Instruction tuning helps align model outputs with clinical expectations and enhances performance in structured generation tasks, such as automated reporting or event annotation. It is often used for applications like open-ended reasoning, question answering, or conversational agents in clinical support systems. For scalability, it is often combined with PEFT methods.