Brant-X: A Unified Physiological Signal Alignment Framework

Daoze Zhang
Zhejiang University
zhangdz@zju.edu.cn

Zhizhang Yuan
Zhejiang University
zhizhangyuan@zju.edu.cn

Junru Chen
Zhejiang University
jrchen_cali@zju.edu.cn

Kerui Chen
Zhejiang University
chenkr@zju.edu.cn

Yang Yang*
Zhejiang University
yangya@zju.edu.cn

ABSTRACT

Physiological signals serve as indispensable clues for understanding various physiological states of human bodies. Most existing works have focused on a single type of physiological signals for a range of application scenarios. However, as the body is a holistic biological system, the inherent interconnection among various physiological data should not be neglected. In particular, given the brain's role as the control center for vital activities, *electroencephalogram* (EEG) exhibits significant correlations with other physiological signals. Therefore, the correlation between EEG and other physiological signals holds potential to improve performance in various scenarios. Nevertheless, achieving this goal is still constrained by several challenges: the scarcity of simultaneously collected physiological data, the differences in correlations between various signals, and the correlation differences between various tasks. To address these issues, we propose a unified physiological signal alignment framework, *Brant-X*, to model the correlation between EEG and other signals. Our approach (1) employs the EEG foundation model to data-efficiently transfer the rich knowledge in EEG to other physiological signals, and (2) introduces the *two-level alignment* to fully align the semantics of EEG and other signals from different semantic scales. In the experiments, *Brant-X* achieves state-of-the-art performance compared with task-agnostic and task-specific baselines on various downstream tasks in diverse scenarios, including sleep stage classification, emotion recognition, *freezing of gaits* detection, and eye movement communication. Moreover, the analysis on the arrhythmia detection task and the visualization in case study further illustrate the effectiveness of *Brant-X* in the knowledge transfer from EEG to other physiological signals. The model homepage is at https://github.com/DaozeZhang/Brant-X/.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**.

KEYWORDS

Physiological signal, Multi-channel time series, Contrastive learning, Alignment, Healthcare

ACM Reference Format:
Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. 2024. Brant-X: A Unified Physiological Signal Alignment Framework . In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671953

1 INTRODUCTION

Physiological signals, as indispensable biomarkers, characterize the underlying complexities of the human body and encapsulate a wide range of critical information about an individual's health, with great significance for health monitoring, disease diagnosis, and treatment [22, 42]. Among these, several key signals including *electroencephalogram* (EEG), *electrooculography* (EOG), *electrocardiogram* (ECG) and *electromyogram* (EMG), are especially essential in capturing primary physiological manifestations [49]. For instance, EEG signals, which record neural activity in the brain, have been utilized to study different stages of sleep and human emotions, aiding in diagnosing sleep-related disorders and emotional health issues [47]. Also, EOG signals, owing to their ability to monitor potential changes during eyeball movements, have proved instrumental in enabling communication for individuals living with *neurodegenerative disorders* [10, 59]. Moreover, ECG signals, which record the fluctuation of the heart's bio-electric activities, have been widely employed in investigations relating to cardiac health and diseases [66]. Finally, EMG signals capture the electrical activity of human muscles, helping the diagnosis and rehabilitation training of neuromuscular diseases [2]. The applications of these physiological signals allow clinicians to monitor individual health in real-time and make data-driven decisions, holding far-reaching implications for many research fields like healthcare.

Despite each physiological signal records the physiological conditions of its corresponding body part, it is worth noting that the body functions as an integrated biological system rather than some independent components [68]. Thus, there exists an inherent interconnection among different physiological signals. Among these, given the brain's role as the epicenter for controlling vital activities, EEG exhibits *significant correlations with synchronous physiological signals* from other body parts [29]. Specifically, in some scenarios, since the information of single-type signal may be insufficient or noisy, ignoring this correlation can lead to great performance losses. Taking sleep staging as an example, as shown in Fig. 1(a), although

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '24, August 25–29, 2024, Barcelona, Spain.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671953

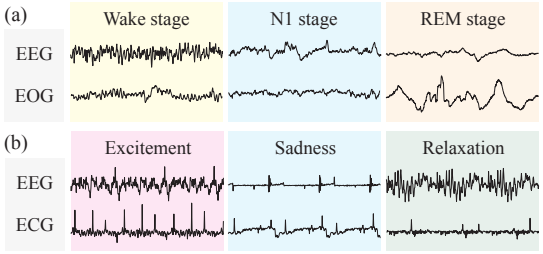

Figure 1: Illustration of inherent correlations between EEG and other physiological signals. (a) The waveform patterns in EEG and EOG vary with different sleep stages, especially with the REM stage, which is marked by rapid oscillations in EOG. (b) Excitement boosts heart beats in ECG with enhanced β waves evident in EEG. Sadness slows heart rate and increases the brain activity in low-frequency α band. During relaxation, ECG presents a stable heart rate with heightened high-frequency EEG θ waves.

EEG records different brainwaves in different stages, the rapid oscillations of EOG are particularly essential criteria for the *rapid eye movement* (REM) stage. Moreover, Sharma et al. [51] has also shown that introducing EOG signals can bring a relative improvement of 14.98% in accuracy. Besides, the correlations between EEG and other signals also exists in other scenarios: **(1) EEG&EOG**: For individuals with neurodegenerative disorders who can only express their thoughts and achieve interaction through eye movements, EEG and EOG can contribute to the development of assistive communication systems [59]. **(2) EEG&ECG**: During different emotional states in Fig. 1(b), brain signals and heartbeats consistently present different patterns, such that EEG and ECG can be utilized for emotion recognition [20]. **(3) EEG&EMG**: Since the abrupt muscle rigidity (named *freezing of gaits*, FoG) of Parkinson's disease is related to a complex interplay between motor, cognitive and affective factors, EEG and EMG can be employed in FoG detection to enhance patient safety and quality of life [74]. Hence, the correlations between EEG and other physiological signals (refered to as "EXG" in this paper, including EOG, ECG, and EMG) hold potential to improve performance in a variety of scenarios. Therefore, our work focuses on establishing an EEG-centric unified framework for modeling the correlation between EEG and EXG, which exploits the combined information of EEG and EXG to contribute to various application scenarios. However, current researches leave much to be explored in this direction, primarily due to the following challenges.

From the viewpoint of data, **simultaneously collected EEG and EXG signals face a conspicuous lack of data.** Due to the acquisition costs, ethical restrictions, and a lack of emphasis on the signal correlation in current machine learning research, the majority of physiological data records only a single type of signal, such as EEG datasets of several terabytes in size [19]. In contrast, available multi-type physiological datasets, which contain various physiological data collected simultaneously, are much smaller in scale, most being less than a few gigabytes. Therefore, the scarcity of simultaneously collected EEG and EXG data poses challenges in training a unified framework for modeling the correlation between EEG and EXG.

From a method perspective, **there exist significant inherent differences in correlations between EEG and different EXG signals.** Different types of physiological signals differ greatly in their inherent properties such as amplitude and bandwidth [58]. To satisfy the sampling theorem [50], the huge gap in bandwidth further leads to differences in sampling rates. Specifically, due to the gap in bandwidth, the sampling rates for EOG, ECG, and EMG may vary respectively within the ranges of 50-100Hz, 250-500Hz, and 1000-2000Hz. These discrepancies are also evident in other features like typical waveforms and rhythmicity [58]. The above factors result in vast inherent differences in correlations between EEG and different EXGs, posing a challenge to the unified modeling method of EEG-EXG correlation.

From the viewpoint of task, **in different scenarios, various downstream tasks depend on different correlations even between EEG and the same EXG.** Given that different application scenarios involve different physiological activities of body organs, different downstream tasks need to capture different correlations between EEG and even the same EXG. Specifically, since the physiological changes during sleep is relatively slow, in sleep staging task, the EEG-EOG correlation is required to capture on a scale up to 30sec, which is defined as a sleep stage [7]. In contrast, in eye movement communication task, eyeball movements may occur in less than 1sec, depending on different EEG-EOG correlation from sleep staging [24]. Therefore, it is challenging to capture different EEG-EXG correlations for various downstream scenarios.

To tackle the above issues, we propose a contrastive-learning-based framework named *Brant-X*, to efficiently align EEG and EXG signals from different semantic scales for the modeling of correlation between EEG and EXG. To address the scarcity of simultaneously collected EEG and EXG data, our intuitive idea is to use models trained with a large amount of EEG data to empower the representation learning on EXG signals. Inspired by large language models that are widely applied in other research fields like computer vision [39, 63], we employ the EEG foundation model Brant-2 [70, 73], which is pre-trained on 4TB brain signal data and contains 1B parameters. Based on this, we summarize existing public multi-type physiological datasets[1], to perform data-efficient knowledge transfer from EEG to EXG. Observing that the gaps between tasks primarily stem from the differences in *semantic scales* of correlation, to address the gaps among various signals and tasks, we introduce the *two-level alignment* that aligns the semantics of EEG and EXG at both patch- and sequence-level. The patch-level alignment overcomes finer inherent differences and captures EEG-EXG correlation at a smaller semantic scale, while the sequence-level one aligns coarser differences and captures the correlation at a larger scale. Moreover, we adopt the *sampling augmentation* to enhance model robustness to different sampling rates. Using the above methods, data and model resources in EEG are extended to empower the research on other physiological signals, paving a new avenue to model the correlations between various physiological signals.

To validate the effectiveness of *Brant-X*, extensive experiments show that *Brant-X* achieves SOTA performance on various downstream tasks across diverse scenarios involving EEG and EXG signals, including sleep stage classification, emotion recognition, freezing of gaits detection, and eye movement communication. The analysis on the arrhythmia detection task and the visualization in case

[1]For the details about the review of public multi-type physiological datasets, please refer to https://github.com/DaozeZhang/Brant-X/

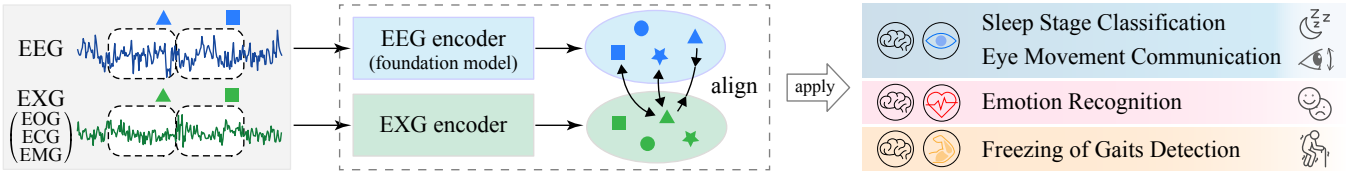

Figure 2: Overview of the physiological signal alignment framework *Brant-X*. Firstly, based on the EEG foundation model, the EXG encoder is trained by the alignment between simultaneously collected EEG and EXG data. Then, the EEG and EXG encoders, capable of learning strong representations from EEG and EXG signals, are applied to various downstream tasks in diverse scenarios.

study further demonstrate that *Brant-X* can effectively transfer the knowledge from EEG to EXG signals through alignment. Overall, our key contributions comprise:

- We are the first to design a unified EEG-centric alignment framework to model the correlations between EEG and other physiological signals, which can be applied to various scenarios.
- Based on the EEG foundation model, we adopt the two-level alignment for data-efficient knowledge transfer from EEG to EXG signals, which combines the semantics of EEG and EXG to jointly improve the performance on downstream tasks.
- We validate *Brant-X* through extensive experiments on multiple downstream tasks involving various physiological signals. Moreover, the analysis and visualization illustrate the effectiveness of *Brant-X* in knowledge transfer from EEG to EXG.

2 PROPOSED METHOD

In this section, we introduce the technical details of the proposed framework *Brant-X*. Specifically, as shown in the upper left part of Fig. 3, we first split the EEG and EXG sequences into continuous data patches. Then, considering the variance in sampling rates between physiological signals in different scenarios, we adopt the sampling augmentation (lower left corner of Fig. 3) to enhance the model's robustness to changes in sampling rates. As shown in middle part of Fig. 3, the EEG patches and EXG patches, along with the augmented patches, are fed into the EEG and EXG encoder, respectively, to acquire the representation of each data patch. Here we employ the EEG foundation model Brant-2 as the EEG encoder of our framework (details in Sec. 2.2). During the unsupervised training process, we propose the two-level alignment (right part of Fig. 3), which aligns the simultaneously collected patches and sequences at both patch- and sequence-level. After the unsupervised alignment, the representations of EEG and EXG data output by the two encoders will be aggregated via the attention mechanism for various tasks in diverse scenarios.

2.1 Problem Formulation

First, we formalize the definitions of the four downstream tasks where the experiments are conducted. The collection of physiological signals relies on signal collection pads, referred to as *electrodes*, distributed on the body part to be monitored. Multiple electrodes simultaneously record the bioelectric activity of the corresponding organs, generating a multi-*channels* time series. Formally, given S EEG signal sequences $\{\boldsymbol{x}_i\}_{i=0}^{S-1}$ that correspond to S physiological processes, each data sequence $\boldsymbol{x}_i \in \mathbb{R}^{C\times L}$ includes C channels with a length of L timestamps. The simultaneously collected EXG signals, including $\tilde{C}$ channels, are denoted as $\{\tilde{\boldsymbol{x}}_i\}_{i=0}^{S-1}$, where $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{\tilde{C}\times L}$. According to different tasks or scenarios, each multi-type sequence $\{\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i\}$ is annotated with a label $y_i \in \mathbb{R}$ by professional physicians. Based on the above, our research problems can be defined as:

Definition 2.1. Given EEG data sequences $\{\boldsymbol{x}_i\}_{i=0}^{S-1}$ and EXG data sequences $\{\tilde{\boldsymbol{x}}_i\}_{i=0}^{S-1}$, with the corresponding labels $\{y_i\}_{i=0}^{S-1}$, the aim is to classify each multi-type sequence $\{\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i\}$ to determine which class it belongs to.

2.2 Foundation Models for EEG

Due to the scarcity of simultaneously collected physiological data, it is challenging to build a unified framework for the modeling of correlations between EEG and EXG. To address this issue, we adopt the brain signal foundation model to perform data-efficient knowledge transfer from EEG to EXG. To the best of our knowledge, only the series of works named Brant currently serves as open-source foundation models on brain signals, including Brant and Brant-2. Specifically, Zhang et al. [73] provide the first off-the-shelf foundation model named Brant for intracranial EEG (iEEG)[2] signals, which contains 500M parameters pre-trained on 1.01TB iEEG data. Based on Brant, Yuan et al. [70] propose the foundation model for brain signals named Brant-2. It consists of over 1B parameters and is pre-trained on as much as nearly 4TB mixed data (with 2.3TB iEEG data from 26 subjects and 1.6TB EEG data from about 15,000 subjects). Our choice to use Brant-2 as the EEG encoder in our framework was two-fold. Firstly, it is pre-trained on a large corpus of brain signal data and can learn powerful representations from EEG signals. More importantly, it uses pre-training data with different sampling rates, resulting in a heightened level of robustness towards changes in sampling rates.

2.3 Overall Architecture

Patching. Given that physiological data are bioelectric signals, the semantic information of the physiological states can only be collectively expressed with multiple sampling points, rather than a single one. Therefore, we split a whole data sequence into several consecutive patches to aggregate semantic information within patches and reduce computation demand [43].

Formally, as shown in the upper left part of Fig. 3, given the i-th multi-channel EEG data sequence $\boldsymbol{x}_i \in \mathbb{R}^{C\times L}$ where C denotes the number of EEG channels and L denotes the number of timestamps (length of the sequence), we split $\boldsymbol{x}_i$ with length M to generate a set of non-overlapping patches $\{\boldsymbol{x}_{i,j}\}_{j=0}^{P-1}$, where $\boldsymbol{x}_{i,j} \in \mathbb{R}^{C\times M}$ and

[2]Compared to EEG signals recorded on the surface of the scalp, iEEG relies on implanted electrodes to measure deep brain activity. However, due to the required cranial surgery and ethical restrictions, the application of iEEG is not as widespread as EEG.

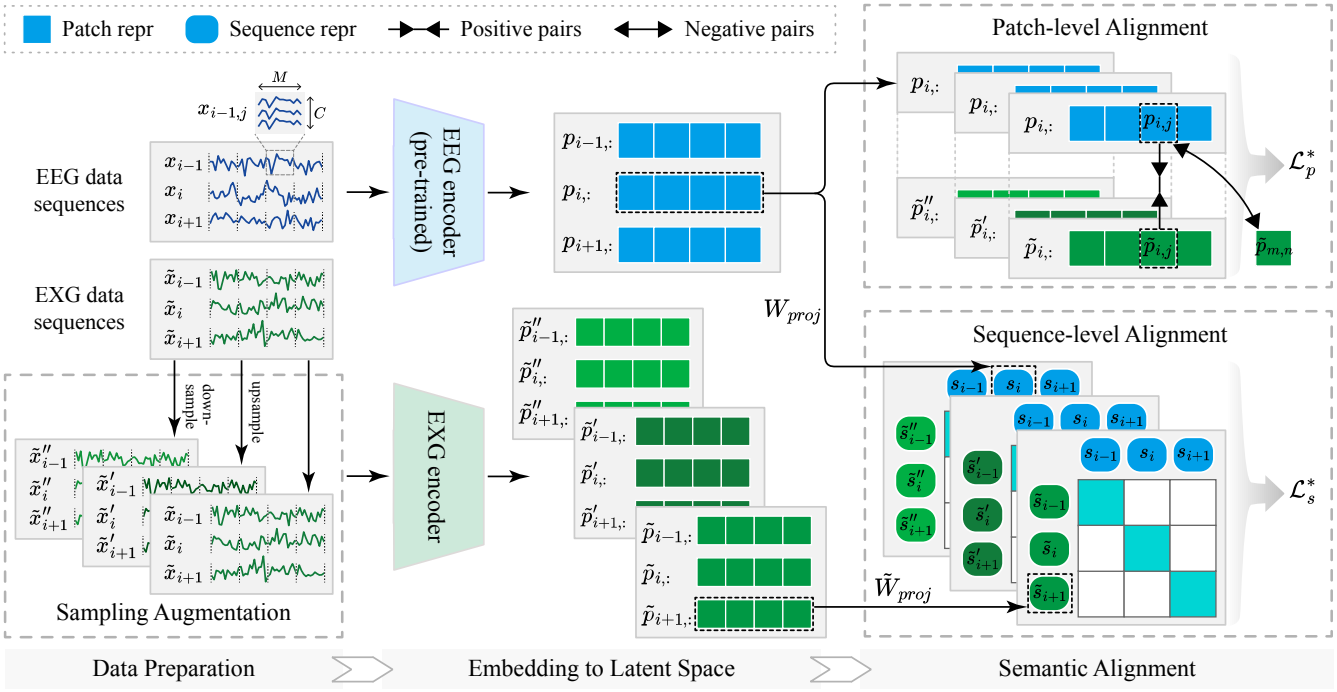

Figure 3: Architecture of *Brant-X*. In the data preparation stage, EXG data are upsampled and downsampled for data augmentation. Then, EEG and EXG data are fed into the EEG encoder and EXG encoder respectively to obtain the representations of data patches. Finally, we align the simultaneously collected EEG and EXG patches and the corresponding EEG and EXG sequences by two-level alignment.

$P = \lfloor L/M \rfloor$ is the number of patches in this sequence. For the EXG data, we apply the same patching process as above, and the symbols are also similar. Specifically, we use $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{\tilde{C} \times \tilde{L}}$ to denote the i-th EXG sequence, where $\tilde{C}$ is the number of EXG channels and $\tilde{L}$ is the sequence length. Also, $\{\tilde{\boldsymbol{x}}_{i,j}\}_{j=0}^{P-1}$ denotes the set of patches, where $\tilde{\boldsymbol{x}}_{i,j} \in \mathbb{R}^{\tilde{C} \times M}$.

Sampling Augmentation. Considering that different physiological signals exhibit large differences in sampling rates, models that learn representations from physiological data must be sufficiently robust to changes in sampling rates. For EEG, as presented in Sec. 2.2, Brant-2 utilize pre-training data at various sampling rates, making it fairly robust to changes in sampling rates. Hence, serving as the EEG encoder of our framework, it is capable of handling differences in the sampling rate of EEG data.

For the EXG signals, to address the issue of various sampling rates, we adopt sampling augmentation to enhance the model's robustness to changes in sampling rate. Specifically, as shown in the lower left corner of Fig. 3, we both upsample the original data to twice its original rate and downsample it to half, producing two sets of augmented data with different sampling rates. Formally, given the original EXG data patches $\{\tilde{\boldsymbol{x}}_{i,j}\}_{j=0}^{P-1}$, we upsample the data to twice its sampling rate, generating the upsampled data patches $\{\tilde{\boldsymbol{x}}'_{i,j}\}_{j=0}^{P-1}$ where $\tilde{\boldsymbol{x}}'_{i,j} \in \mathbb{R}^{\tilde{C} \times 2M}$. Similarly, the original patches are also downsampled to half the sampling rate, thus obtaining the downsampled data patches $\{\tilde{\boldsymbol{x}}''_{i,j}\}_{j=0}^{P-1}$ where $\tilde{\boldsymbol{x}}''_{i,j} \in \mathbb{R}^{\tilde{C} \times \lfloor M/2 \rfloor}$.

In subsequent representation learning and semantic alignment sections, the original data $\tilde{\boldsymbol{x}}$, along with the upsampled data $\tilde{\boldsymbol{x}}'$ and downsampled data $\tilde{\boldsymbol{x}}''$, will be fed into the EXG encoder for model learning purposes.

Embedding to Latent Space. For EEG data, as shown in the middle part of Fig. 3, we feed it directly into the pre-trained EEG encoder (details in Sec. 2.2) to obtain the EEG representation. Formally, P consecutive patches $\{\boldsymbol{x}_{i,j}\}_{j=0}^{P-1}$ from the i-th EEG data sequence $\boldsymbol{x}_i$ will be input into the EEG encoder, yielding the representations $\{\boldsymbol{p}_{i,j}\}_{j=0}^{P-1}$ of these patches, where $\boldsymbol{p}_{i,j} \in \mathbb{R}^{D_p}$ denotes the representation of the j-th patch from the i-th sequence of EEG data, and D_p denotes the dimension of patch representations.

When it comes to EXG data, it will be fed into the EXG encoder to obtain its representation. Formally, all the patches $\{\tilde{\boldsymbol{x}}_{i,j}\}_{j=0}^{P-1}$ from the i-th EXG data sequence $\tilde{\boldsymbol{x}}_i$ are input into the EXG encoder, generating their representations $\{\tilde{\boldsymbol{p}}_{i,j}\}_{j=0}^{P-1}$, where $\tilde{\boldsymbol{p}}_{i,j} \in \mathbb{R}^{D_p}$. Given that the focus of our work is the alignment framework, the specific architecture of the EXG encoder can be flexible. For the technical details of the EXG encoder used in this paper, please refer to App. A.

Similarly, the upsampled EXG patches $\{\tilde{\boldsymbol{x}}'_{i,j}\}_{j=0}^{P-1}$ and the downsampled patches $\{\tilde{\boldsymbol{x}}''_{i,j}\}_{j=0}^{P-1}$ undergo the same process, obtaining the representations $\{\tilde{\boldsymbol{p}}'_{i,j}\}_{j=0}^{P-1}$ and $\{\tilde{\boldsymbol{p}}''_{i,j}\}_{j=0}^{P-1}$ of augmented EXG data.

2.4 Two-level Alignment

We adopt two-level alignment that fully aligns the semantics of EEG and EXG signals at patch- and sequence-level, to overcome

inherent differences and capture the correlation between EEG and EXG at different semantic scales.

Patch-level Alignment. At a finer grain, we align EEG and EXG data at patch-level by placing the simultaneous EEG and EXG patches close together in the latent space, while mapping unrelated patches further apart. As shown in the upper right part of Fig. 3, since our EEG encoder is pre-trained on a large amount of data (Sec. 2.2), it is reasonable to believe it can output representative representations of EEG patches. Therefore, we set the EEG representation $\boldsymbol{p}_{i,j}$ as the anchor. The anchor $\boldsymbol{p}_{i,j}$ and the simultaneously collected EXG patch $\tilde{\boldsymbol{p}}_{i,j}$ are set as the positive sample pair. Negative samples are randomly selected from the representations $\{\tilde{\boldsymbol{p}}_m\}_{m\neq i}$ from other EXG data sequences. It is noteworthy that, contrary to the sequence-level alignment described later, we can't randomly select the representations $\{\tilde{\boldsymbol{p}}_{i,n}\}_{n\neq j}$ from the EXG sequence $\tilde{\boldsymbol{p}}_i$ as negative samples. This is because these representations $\{\tilde{\boldsymbol{p}}_{i,n}\}_{n\neq j}$ and the anchor originate from the same physiological process and may have a temporal dependency between them. Formally, for the anchor $\boldsymbol{p}_{i,j}$, the negative sample set $Z^p_{i,j}$ is randomly sampled from all the negative samples $\{\tilde{\boldsymbol{p}}_{m,n} | m \neq i, n = 0, ..., P-1\}$. The InfoNCE [44] loss is applied to retain the maximum mutual information between positive pairs:

$$\mathcal{L}_p = \frac{1}{SP}\sum_i\sum_j -\log\frac{\exp(\boldsymbol{p}_{i,j}^{\mathsf{T}}\tilde{\boldsymbol{p}}_{i,j}/t_p)}{\sum_{\tilde{\boldsymbol{p}}_{m,n}\in Z^p_{i,j}}\exp(\boldsymbol{p}_{i,j}^{\mathsf{T}}\tilde{\boldsymbol{p}}_{m,n}/t_p)}, \tag{1}$$

where t_p denotes the temperature hyperparameter to adjust scale, and $\mathcal{L}_p$ denotes the InfoNCE loss between EEG and original EXG data in patch-level alignment.

Similarly, the same alignment process would also exist between the EEG data and the two sets of augmented EXG data. These two losses are denoted as $\mathcal{L}'_p$ and $\mathcal{L}''_p$, respectively. Overall, the optimization objective of patch-level alignment is given by:

$$\mathcal{L}^*_p = \mathcal{L}_p + \mathcal{L}'_p + \mathcal{L}''_p. \tag{2}$$

Sequence-level Alignment. At a coarser granularity level, we employ sequence-level alignment to align the corresponding sequence in the latent space. To aggregate the representations of patches from a data sequence, we firstly perform a linear projection $W_{proj} \in \mathbb{R}^{D_s\times PD_p}$ on all patch representations $\boldsymbol{p}_{i,:} \in \mathbb{R}^{P\times D_p}$ from sequence $\boldsymbol{x}_i$, thus obtaining the sequence representation $\boldsymbol{s}_i \in \mathbb{R}^{D_s}$, where D_s denotes the dimension of sequence representations:

$$\boldsymbol{s}_i = W_{proj}\left(\text{Flatten}(\boldsymbol{p}_{i,:})\right). \tag{3}$$

This linear projection is applied similarly for EXG data $\tilde{\boldsymbol{p}}_{i,:}$ and the augmented data $\tilde{\boldsymbol{p}}'_{i,:}, \tilde{\boldsymbol{p}}''_{i,:}$ as well, yielding the sequence representations $\tilde{\boldsymbol{s}}_i$, $\tilde{\boldsymbol{s}}'_i$ and $\tilde{\boldsymbol{s}}''_i$ respectively.

After obtaining the sequence representations, we set the representations of simultaneously collected EEG and EXG sequences ($\boldsymbol{s}_i$ and $\tilde{\boldsymbol{s}}_i$) as positive sample pairs, while all other sequence pairs are set as negative pairs. Formally, the negative sample set Z^s_i of sequence $\boldsymbol{s}_i$ is randomly sampled from all the negative samples $\{\tilde{\boldsymbol{s}}_m | m \neq i\}$. The sequence-level InfoNCE loss $\mathcal{L}_s$ for the EEG and the original EXG data can be given as follows:

$$\mathcal{L}_s = \frac{1}{S}\sum_i -\log\frac{\exp(\boldsymbol{s}_i^{\mathsf{T}}\tilde{\boldsymbol{s}}_i/t_s)}{\sum_{\tilde{\boldsymbol{s}}_m\in Z^s_i}\exp(\boldsymbol{s}_i^{\mathsf{T}}\tilde{\boldsymbol{s}}_m/t_s)}, \tag{4}$$

where t_s denotes the temperature hyperparameter. As shown in the bottom right part of Fig. 3, following the common practice in CLIP [48], we adopt a similarity matrix to optimize this objective.

Likewise, we carry out the same alignment process between EEG and augmented EXG data, resulting in two losses $\mathcal{L}'_s$ and $\mathcal{L}''_s$ in the same form. The overall loss in sequence-level alignment is:

$$\mathcal{L}^*_s = \mathcal{L}_s + \mathcal{L}'_s + \mathcal{L}''_s. \tag{5}$$

Finally, the objective of joint optimization is obtained by adding the patch-level and sequence-level alignment losses $\mathcal{L}^*_p$ and $\mathcal{L}^*_s$.

3 EXPERIMENT

3.1 Experimental Setup

Alignment. To align the simultaneously recorded EEG and arbitrary EXG data, the training data used for unsupervised alignment is collectively assembled from three datasets: CAP [57], ISRUC [32], and HMC [4], which include EEG, EOG, ECG, and EMG signals. Overall, the alignment training data includes 359 recordings from 267 subjects. The alignment is performed on a Linux system with 2 CPUs (AMD EPYC9654 96-Core Processor) and 2 GPUs (NVIDIA Tesla A100 80G). The learning rate of EEG encoder is set as 1×10^{-5} for finetuning, while the EXG encoder is trained with a higher learning rate of 3×10^{-4}.

Downstream Tasks. Here we introduce the four downstream tasks used to validate the effectiveness of our *Brant-X*, along with the datasets, setups and and evaluation metrics.

• *Sleep Stage Classification.* In sleep health research, sleep staging refines human understanding of sleep states and patterns, which holds significance for the prevention and diagnosis of sleep-related diseases [47]. According to the American Academy of Sleep Medicine (AASM) manual [7], sleep occurs in five stages: wake, N1, N2, N3, and REM. Among these, N1 to N3 are *non-rapid eye movement* sleep, with each stage leading to progressively deeper sleep. Hence, sleep stage classification is a five-class classification problem.

As for the dataset, the Sleep-EDF datasets [31] are very popular in sleep staging researches. The Sleep-EDF-78 dataset contains 153 whole-night polysomnographic sleep recordings from sleep cassette studies, containing 100Hz EEG and EOG data from 78 subjects aged 25-101 years (37 males and 41 females). Data are segmented into 30sec epochs and manually annotated by experts. The Sleep-EDF-20 dataset, which contains 39 recordings from 20 subjects, is also used in our study to facilitate the comparison with the existing methods.

The experiment is conducted on EEG and EOG signals in a subject-independent setting. We divide the subjects into training, validation, and test sets in a 3:1:1 ratio. The experiments are repeated on all subjects to obtain overall results. The evaluation metrics include accuracy, sensitivity, specificity, macro F1 score, and Cohen's kappa κ.

• *Emotion Recognition.* Automatic emotion recognition has made a remarkable entry in the domain of biomedical, brain-computer interface, smart environment, safe driving and so on [28]. Emotions are categorized into two types: (1) discrete emotions like joy, fear and sadness; and (2) multi-dimensional emotions on three emotion dimensions: arousal, valence, and dominance dimensions.

Existing works [30, 35, 36, 54, 56] mainly focus on the recognition of multi-dimensional emotions, so the task can be regarded as three independent binary classification problems: low/high valence, low/high arousal and low/high dominance.

The DREAMER dataset [30] is used to conduct experiments on emotion recognition task. It contains EEG (128Hz) and ECG (256Hz) data of 23 subjects (14 males and 9 females) when they are watching 18 film clips. Each film clip has an average length of 199s, which is thought to be sufficient for eliciting single emotion. After watching a film clip, emotion statuses are labeled as low or high on the three emotion dimensions, serving as the labels for emotion recognition.

The experiment in this task is conducted on EEG and ECG signals in a subject-independent setting. We split subjects into training, validation, and test sets in a 3:1:1 ratio and repeat the experiments on all subjects. The evaluation metrics are mainly accuracy [35, 36, 56], with some studies [30, 54] also including the F1 score and the AUC of precision-recall curve.

• *Freezing of Gaits Detection.* FoG, which refers to the interruption of the motion caused by the brain's incompetence to deal with concurrent cognitive and motor request, affects about 50%-80% of Parkinson's disease patients as one of the severest manifestations. Thus, accurate detection of FoG can significantly improve patients' life quality and promote personalized treatment [74]. The FoG detection task is a binary classification problem, that is, determining whether FoG appears during a walking process.

The FoG dataset [74] is used in this work, which includes EEG and EMG signals (1000Hz) collected from 12 Parkinson's disease patients (6 males and 6 females) aged 57-81 years with disease durations between 1 and 20 years. The valid data lasts for 3h42min, including 2h14min of normal gait and 1h28min of freezing of gait, labeled by two qualified physicians.

The experiment in this task is conducted on EEG and EMG. The training, validation, and test data are randomly split in a 3:1:1 ratio. We also repeat the experiments to obtain the overall results. As a classification problem, the evaluation metrics used for this task are accuracy, precision, recall and F1 score.

• *Eye Movement Communication.* Due to paralysis caused by neurodegenerative disorders like *amyotrophic lateral sclerosis* (ALS), many patients lost almost all their communication abilities [24], and only have remnant oculomotor control to form words, phrases, and sentences using a speller system [59]. The speller system works on a binary principle where the patient responds to auditory questions by moving their eyes to say "yes" and not moving the eyes for "no". Therefore, the eye movement communication task is also a binary classification problem (yes or no).

The dataset published by Jaramillo-Gonzalez et al. [24] is used for the eye movement communication experiment. The dataset contains EEG and EOG data (500Hz) recorded from four patients suffering from ALS. Data are recorded during 2-10 visits, each visit consisting of an average of 3.22 days with 5.57 sessions recorded per day. Due to the inconsistency in EOG channels across different files in the dataset, we exclude files lacking specific EOG channels to conduct the experiment.

The experiment in this task is conducted using EEG and EOG. Experiments are conducted on training, validation, and test data split 8:1:1 and are repeated on all data files. The evaluation metrics are accuracy, precision, recall and F1 score.

As for data pre-processing, for the three tasks except eye movement classification, we did not perform filtering or other processing, directly using the preprocessed data of the original datasets. For the eye movement dataset, as the publisher didn't filter, we applied 45Hz low-pass filtering and z-score normalization.

Baselines. As a unified unsupervised alignment framework for physiological signal modeling, we compare *Brant-X* with the advanced self-supervised or unsupervised methods designed for general time series on all the downstream tasks, including TF-C [75] and SimMTM [15]. Also, to compare *Brant-X* with the methods that performs time series classification based on pre-trained language models, we set OneFitsAll [76] and Time-LLM [27] as a baseline. As for the supervised methods, MiniRocket [14] is selected as our baseline due to its efficiency and versatility. Furthermore, we compare *Brant-X* with the SOTA methods that are specially designed for each task, to demonstrate the effectiveness of *Brant-X* in various scenarios. These task-specific or signal-specific supervised methods includes: (1) TinySleepNet [55], XSleepNet [45], L-SeqSleepNet [46], SleepHGNN [25], SleepKD [34], and SleepDG [62] for sleep stage classification; (2) MLF-CapsNet [36], EEG-Conformer [52], Lin et al. [35] and Wang et al. [64] for emotion recognition; (3) Aly and Youssef [5], Batool and Javeed [6] and Goel et al. [17] for freezing of gaits detection; and (4) eyeSay [77], Adama and Bogdan [1] and Hossieny et al. [23] for eye movement communication. More details about these baselines are given in App. B.

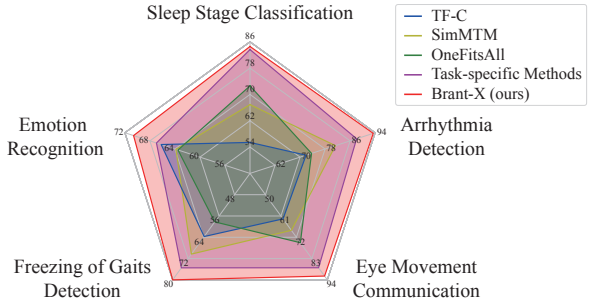

Figure 4: Overall performance comparison on various tasks.

3.2 Experimental Results

Fig. 4 summarizes the overall accuracy of *Brant-X* and other baselines on various downstream tasks (including the arrhythmia detection in Sec. 3.3). Since the task-specific methods vary across different tasks, we use "Task-specific Methods" to collectively represent their best results on each task. As shown in Fig. 4, compared with other baseline methods, *Brant-X* achieves SOTA performance on all of the five tasks, illustrating the effectiveness of our framework in various scenarios. Detailed comparisons on each task are discussed in following paragraphs, where in all the tables we mark values ranking the first (**v**), second (v) and third (v*) in each column.

The performance comparison on sleep stage classification task is given in Tab. 1. It shows that *Brant-X* achieves top rankings in almost all performance metrics, demonstrating that *Brant-X* can effectively transfer the knowledge from EEG to EOG signals,

Table 1: Average performance on the sleep stage classification task.

Methods \ Metrics	Sleep-EDF-20					Sleep-EDF-78				
	Acc.	Sens.	Spec.	Macro F1	Kappa	Acc.	Sens.	Spec.	Macro F1	Kappa
TF-C [75]	55.42 ±1.39	31.52 ±1.09	86.07 ±0.39	26.04 ±0.21	30.74 ±1.52	53.90 ±4.03	31.35 ±2.40	85.80 ±1.34	26.00 ±2.09	29.32 ±6.43
SimMTM [15]	66.91 ±1.89	53.47 ±1.58	90.61 ±1.63	53.21 ±1.95	53.25 ±2.02	63.06 ±2.67	59.12 ±3.88	91.21 ±1.56	57.07 ±2.13	53.07 ±3.42
OneFitsAll [76]	72.60 ±1.51	63.50 ±8.36	92.76 ±1.12	61.61 ±5.80	61.81 ±3.50	68.50 ±2.19	56.58 ±4.16	91.34 ±0.86	54.24 ±1.96	55.21 ±3.07
Time-LLM [27]	80.31 ±2.63	76.53 ±3.15	94.53 ±2.95	71.64 ±3.02	70.22 ±2.84	78.08 ±2.96	67.44 ±3.73	94.13 ±3.01	66.09 ±3.25	68.04 ±3.14
MiniRocket [14]	81.60 ±1.55	72.63 ±1.80	95.15 ±1.12	72.82 ±2.01	72.79 ±1.96	78.36 ±1.93	69.76 ±2.44	94.08 ±1.76	70.18 ±2.35	69.46 ±2.46
TinySleepNet [55]	83.64 ±2.31	**81.60**±2.60	96.05 ±2.08	77.54 ±2.55	77.63 ±2.29	**83.49**±2.24	80.25*±2.65	**96.02**±2.11	76.64 ±2.61	76.41 ±2.59
XSleepNet [45]	80.93 ±2.34	75.78 ±2.21	94.79 ±2.54	76.71*±2.59	74.31 ±2.32	81.83*±2.30	80.50 ±2.28	95.74*±2.58	75.28*±2.66	75.44*±2.37
L-SeqSleepNet [46]	82.90*±2.12	78.42 ±2.25	95.86*±2.00	74.90 ±2.22	76.47 ±2.24	80.84 ±2.18	72.75 ±2.54	95.19 ±2.34	72.67 ±2.38	74.94 ±2.51
SleepHGNN [25]	81.15 ±1.96	74.23 ±2.10	94.93 ±1.96	72.88 ±2.17	73.35 ±2.16	77.35 ±2.13	69.94 ±2.48	94.04 ±2.02	69.56 ±2.39	68.65 ±2.41
SleepKD [34]	82.44 ±2.40	78.20 ±2.54	94.78 ±2.34	74.11 ±2.72	76.87*±2.63	80.19 ±2.85	72.95 ±2.88	94.95 ±2.69	72.65 ±2.84	74.86 ±2.93
SleepDG [62]	81.92 ±2.27	79.12*±2.35	95.75 ±2.68	74.74 ±2.53	76.43 ±2.47	79.95 ±2.42	73.31 ±2.41	93.57 ±2.63	72.21 ±2.59	74.16 ±2.68
Brant-X	**84.58**±1.98	80.18 ±2.23	**96.36**±1.89	**77.63**±2.13	**79.29**±2.18	82.84 ±2.21	**81.85**±2.42	95.91 ±2.08	**77.04**±2.30	**76.67**±2.49

Table 2: Average performance on the emotion recognition task.

Methods \ Metrics	Valence			Arousal			Dominance		
	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC
TF-C [75]	66.20 ±3.76	78.09 ±5.02	69.71 ±7.34	76.45*±11.36	85.86 ±8.23	80.40 ±10.27	78.17 ±9.64	87.01 ±6.57	85.20 ±4.81
SimMTM [15]	63.84 ±5.93	75.52 ±5.90	69.73 ±4.02	76.16 ±7.97	86.21*±5.23	76.42 ±12.39	78.54 ±3.94	87.81 ±2.54	82.84 ±7.99
OneFitsAll [76]	63.51 ±6.66	76.93 ±5.03	64.71 ±11.22	73.88 ±7.28	83.84 ±6.14	76.75 ±7.33	77.41 ±4.94	86.92 ±3.31	85.59*±5.60
Time-LLM [27]	68.03 ±5.82	72.22 ±5.27	**80.83 ±6.04**	76.39 ±7.45	85.63 ±6.18	80.28 ±7.73	80.10 ±4.81	88.92 ±3.45	79.68 ±5.07
MiniRocket [14]	60.54 ±7.09	65.68 ±6.80	64.36 ±8.05	75.73 ±8.89	85.75 ±7.64	77.90 ±10.46	75.11 ±5.91	85.28 ±5.14	86.69 ±7.16
MLF-CapsNet [36]	65.67 ±2.87	77.06 ±3.87	71.05 ±4.66	74.56 ±7.49	84.98 ±5.32	79.80 ±10.92	77.13 ±2.36	86.94 ±1.35	82.61 ±8.21
EEG Conformer [52]	59.82 ±7.05	69.53 ±6.93	71.94*±11.91	73.07 ±9.67	83.21 ±7.41	75.11 ±7.65	81.82*±6.05	89.50*±4.13	83.19 ±9.87
Lin et al. [35]	66.47 ±6.85	79.50*±5.04	67.10 ±8.94	75.54 ±7.81	85.87 ±5.14	79.06 ±6.65	78.46 ±5.04	87.83 ±3.12	79.40 ±6.50
Wang et al. [64]	66.95*±8.30	79.84 ±6.11	66.20 ±10.73	76.47 ±9.29	86.44 ±6.14	80.29*±7.58	81.87 ±5.31	89.96 ±3.22	83.97 ±5.81
Brant-X	**70.61**±4.01	**80.51**±3.81	72.48 ±4.10	**78.64**±8.56	**87.59**±5.71	**82.14**±7.98	**83.54**±5.27	**90.97**±3.16	**90.19**±4.94

Table 3: Average performance on the FoG detection task.

Methods	Acc.	Prec.	Rec.	F1
TF-C [75]	63.72 ±1.83	61.28 ±2.91	76.14 ±6.02	67.77 ±2.52
SimMTM [15]	70.32 ±4.22	69.05 ±8.65	74.79*±7.98	71.88 ±0.20
OneFitsAll [76]	58.22 ±3.31	56.62 ±4.13	71.80 ±16.41	62.98 ±4.19
Time-LLM [27]	72.73 ±2.98	74.23*±4.75	69.14 ±5.54	71.62 ±4.66
MiniRocket [14]	73.42 ±2.02	72.73 ±2.07	72.11 ±0.69	72.18 ±1.38
Aly et al. [5]	72.12 ±3.31	71.09 ±3.41	73.54 ±5.10	72.24 ±3.60
Batool et al. [6]	75.49 ±2.35	75.50 ±2.56	74.62 ±3.22	75.04 ±2.59
Goel et al. [17]	74.18*±2.10	73.49 ±3.59	74.58 ±2.64	73.96*±1.98
Brant-X	**80.14**±1.33	**81.97**±1.56	**76.73**±3.80	**79.21**±1.86

Table 4: Average performance on the eye movement communication task.

Methods	Acc.	Prec.	Rec.	F1
TF-C [75]	62.50 ±4.52	61.90 ±4.18	65.00 ±9.61	63.41 ±3.94
SimMTM [15]	68.42 ±5.42	74.95 ±4.16	56.06 ±10.62	63.83 ±8.54
OneFitsAll [76]	74.81 ±3.88	75.90 ±3.72	72.69 ±8.64	74.26 ±3.61
Time-LLM [27]	81.78 ±4.37	84.71 ±4.65	77.54 ±7.36	80.96 ±6.14
MiniRocket [14]	71.43 ±5.37	63.64 ±6.40	93.31 ±8.63	75.68 ±6.07
eyeSay [77]	80.24 ±6.61	84.43 ±4.66	74.75 ±9.81	79.18 ±7.35
Adama et al. [1]	87.75 ±8.41	87.86 ±7.30	87.41*±10.71	87.56 ±8.70
Hossieny et al. [23]	83.34*±5.19	86.33*±4.74	79.02 ±7.04	82.47*±5.77
Brant-X	**92.04**±3.13	**90.99**±3.96	**93.42**±2.87	**92.17**±3.06

combining the information of both EEG and EOG signals to learn the high-level semantic information therein. The baselines on general time series did not yield good results, mainly because these models struggle to overcome the huge gap in inherent features between various physiological signals, and do not model correlations from different semantic scales like *Brant-X* does.

As shown in Tab. 2, on emotion recognition task, *Brant-X* achieves SOTA performance compared to all the baselines. Compared to the baselines designed solely for EEG, Wang et al. [64] claims the second spot, because it adopt the same strategy as *Brant-X* for combining the information of both EEG and ECG, thereby demonstrating stronger learning capabilities. However, *Brant-X* still surpasses Wang et al. [64] on all metrics, benefiting from alignment training based on contrastive learning.

The overall results on the freezing of gaits detection and eye movement communication tasks are given in Tab. 3 and Tab. 4, respectively. *Brant-X* defeats all the baseline methods on these two

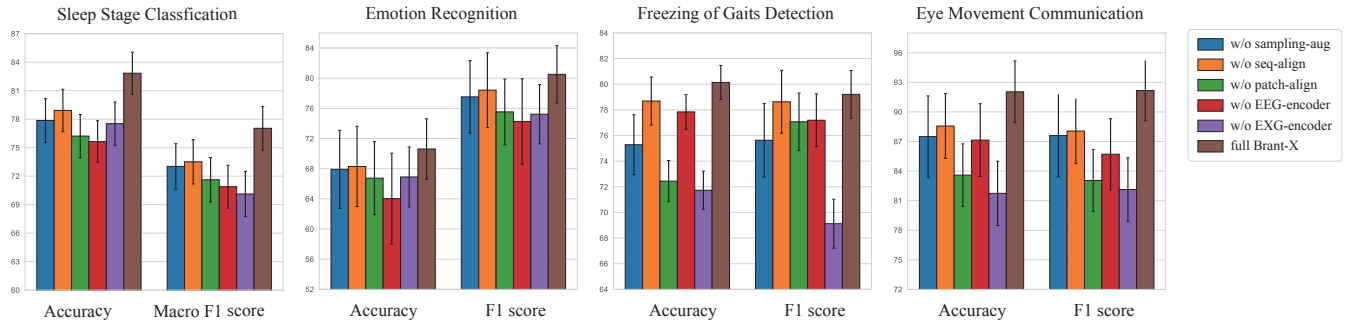

Figure 5: Results of the ablation study on all the downstream tasks.

tasks, showing its ability to learn representations from simultaneously collected EEG, EMG, and EOG data. Batool et al. [6] and Adama et al. [1] achieve the second-best performance on these two tasks, respectively, mainly because they explicitly extract the frequency domain information of physiological signals as inherent features for physiological data modeling.

3.3 Ablation Study

To evaluate the effectiveness of each component in *Brant-X*, we conduct ablation experiments on four model variants, including: (1) *Brant-X w/o sampling-aug*: *Brant-X* without the sampling augmentation during alignment; (2) *Brant-X w/o patch-align*: *Brant-X* without the patch-level alignment; (3) *Brant-X w/o seq-align*: *Brant-X* without the sequence-level alignment; (4) *Brant-X w/o EEG-encoder*: *Brant-X* without the EEG encoder during downstream evaluation after alignment; (5) *Brant-X w/o EXG-encoder*: *Brant-X* without the EXG encoder during downstream evaluation after alignment.

The comparison results of the ablation experiments on the four downstream tasks are presented in Fig. 5. It demonstrates that *Brant-X* outperforms other variants on all metrics of all the tasks, evidencing the contribution of each component in our framework. Compared to the full *Brant-X*, the performance of *Brant-X w/o sampling-aug* decreases, showing the boost of model robustness against variable sampling rates provided by the sampling augmentation. Also, *Brant-X w/o patch-align* and *Brant-X w/o seq-align* show a decrease in performance, suggesting that the two-level alignment can align EEG and EXG signals from different semantic scales to learn informative representations from physiological data. For sleep staging and emotion recognition, *Brant-X w/o EEG-encoder* drops greatly in performance, as EEG signals play an important role in these scenarios. This corroborates the significance of the brain as a central control in vital activities, as we emphasized in Sec. 1.

EXG Encoder Analysis. As a supplement to the *Brant-X w/o EEG-encoder* in the ablation experiments, we extend our assessment to more tasks using the standalone EXG encoder, to validate whether the EXG encoder can learn useful representations from EXG data during the alignment training. Specifically, we conduct experiments with the aligned EXG encoder on ECG data (without incorporating the EEG encoder on EEG data) on the arrhythmia detection task. More details about this task and the results are given in App. C. As shown in Tab. 5, *Brant-X* achieves SOTA performance on the arrhythmia detection task, showing that the alignment training indeed enables the EXG encoder to learn the representations from ECG signals, and then effectively classify cardiac rhythms.

3.4 Case Study

Fig. 6 displays four similarity matrices between patch representations of two multi-type physiological data sequences, $\{\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i\}$ and $\{\boldsymbol{x}_j, \tilde{\boldsymbol{x}}_j\}$. The vertical axis represents the patch representations of two EEG sequences, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and the horizontal axis represents the patch representations of two EXG sequences, $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$. Thus, four similarity matrices are given in Fig. 6. The darker the colour of each small square, the higher the normalised similarity between the representations of two corresponding patches.

Among these, matrix (a) (or (d)) indicates the similarity of patch representations of simultaneously collected EEG sequence $\boldsymbol{x}_i$ (or $\boldsymbol{x}_j$) and EXG sequence $\tilde{\boldsymbol{x}}_i$ (or $\tilde{\boldsymbol{x}}_j$). It presents an overall darker colour, demonstrating the correlations between patches from the simultaneously collected EEG and EXG sequence. Moreover, the diagonal of matrix (a) (or (d)) is particularly dark, indicating that the simultaneous EEG and EXG patches are well-aligned. However, as for matrices (b) and (c), they have an overall lighter colour, suggesting little to no correlation between patch representations of non-simultaneously collected EEG and EXG data. These four similarity matrices in this case illustrate well that the two-level alignment can bring the representations of simultaneous EEG and EXG data closer, while distancing irrelevant sequences, such that *Brant-X* can perform knowledge transfer from EEG to EXG.

4 RELATED WORK

Physiological Signal Modeling. With the maturation of physiological recording technology and the advancement of machine learning methods, physiological signal modeling has captivated many researchers. Initially, researchers mainly focus on model learning on a single type of signal. A large body of works propose to use time series [18, 36, 37, 45, 52, 55, 71] or graph [8, 12, 35] data structures with supervised [18, 35–37, 45, 52, 55] or self-supervised [8, 12, 71] learning paradigms for various tasks on EEG signals. Recently, some large EEG models [26, 70, 73] also emerged, which break through the limitations of different tasks on EEG. Also, methods based on feature engineering or supervised learning are proposed to learn representations from EOG [1, 23, 77], ECG [3, 40], and EMG [41, 72] signals. Additionally, to fully mine the potential semantics of physiological data, research attention has been drawn to the modeling of

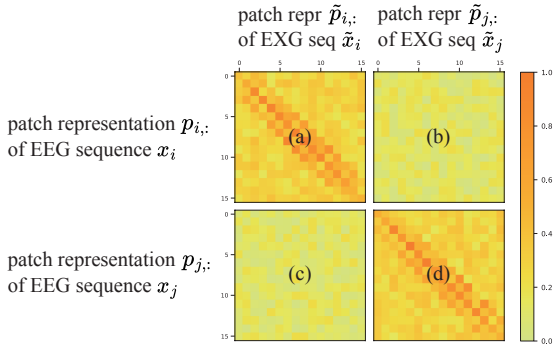

Figure 6: Similarity matrices of patch representations. The vertical axis represents the patch representations of two EEG sequences, x_i and x_j, and the horizontal axis represents the patch representations of two EXG sequences, $\tilde{x}_i$ and $\tilde{x}_j$.

multi-type signals. Jia et al. [25] consider the interactivity of EEG, EOG and EMG signals, imporving the SOTA performance on sleep staging task. Wang et al. [64] fuse the features from single-lead EEG and ECG data for emotion recognition. Aly and Youssef [5] propose an approach that integrates EEG with EMG signals, boosting the performance of hand and wrist motion control. However, most works on physiological signals are task-specific or signal-specific. They neither leverage foundation models (thus are hindered by the data scarcity), nor possess a unified framework for various physiological signals across a range of tasks.

Multimodal Alignment. To capitalize on the information consistency in multimodal data, contrastive-learning-based alignment strategy has achieved impressive results in many fields like image-text [48, 67, 69]. For physiological research, methods [9, 11] conduct alignment to the features of images for medical images segmentation. Wang et al. [61] aligns the paired medical image and radiology reports (text) for image classification and object detection, etc. Fan et al. [16] propose a domain adaptation approach to bridge the gap between the EEG data distribution of source and target domains for sleep staging. Lv et al. [38] reinforce features by aligning the visual and acoustic modality within video clips for emotion recognition. However, these studies primarily explore the consistency among text, image, or audio modalities, none of which explicitly align the simultaneously collected physiological signals.

Time seires Modeling. Time series (TS) analysis has been utilized in many real-world applications, including finance, meteorology, healthcare, and so on, attracting more and more researchers. Wu et al. [65] propose TimesNet as a task-general backbone to discover the multi-periodicity adaptively for TS analysis. Dong et al. [15] propose to recover masked time points by the weighted aggregation of multiple neighbors outside the manifold for TS modeling. Zhou et al. [76] propose a unified model that leverages language or vision models for TS analysis. Jin et al. [27] present a reprogramming framework named Time-LLM to repurpose large language models for general TS forecasting. However, most TS works cannot adapt well to high-frequency physiological signals and ignore the correlation between physiological signals.

5 CONCLUSION

In this work, we are the first to propose a unified physiological signal alignment framework, *Brant-X*. Based on the EEG foundation model, we summarize available multi-type physiological datasets, to transfer the rich knowledge from the EEG foundation model to EXG signals. We adopt the two-level alignment that aligns the semantics of EEG and EXG data at both patch- and sequence-level, to adapt to various downstream scenarios. In this way, EEG is viewed as a bridge between the EEG foundation model and EXG data, allowing the data and model resources in the EEG field to empower the research on other physiological signals, paving a new avenue to model the correlations between various physiological signals. In the future, motivated by the positive results of *Brant-X*, it would be intriguing to explore further studies along this research line on more physiological signals.

Acknowledgment. This work is supported by National Natural Science Foundation of China (No. 62322606, No. 62441605) and SMP-IDATA Open Youth Fund.

REFERENCES

[1] Sophie Adama and Martin Bogdan. 2021. Yes/No Classification of EEG data from CLIS patients. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 5727–5732.
[2] Muhammad Al-Ayyad, Hamza Abu Owida, Roberto De Fazio, Bassam Al-Naami, and Paolo Visconti. 2023. Electromyography Monitoring Systems in Rehabilitation: A Review of Clinical Applications, Wearable Devices and Signal Acquisition Methodologies. *Electronics* 12, 7 (2023), 1520.
[3] Negin Alamatsaz, Leyla Tabatabaei, Mohammadreza Yazdchi, Hamidreza Payan, Nima Alamatsaz, and Fahimeh Nasimi. 2024. A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection. *Biomedical Signal Processing and Control* 90 (2024), 105884.
[4] Diego Alvarez-Estevez and Roselyne M Rijsman. 2021. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one* 16, 8 (2021), e0256111.
[5] Heba Aly and Sherin M Youssef. 2023. Bio-signal based motion control system using deep learning models: A deep learning approach for motion classification using EEG and EMG signal fusion. *Journal of Ambient Intelligence and Humanized Computing* 14, 2 (2023), 991–1002.
[6] Mouazma Batool and Madiha Javeed. 2022. Movement Disorders Detection in Parkinson's Patients using Hybrid Classifier. In *International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, 213–218.
[7] Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. 2017. AASM scoring manual updates for 2017 (version 2.4). , 665–666 pages.
[8] Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. 2023. MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 130–141.
[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems* 33 (2020), 12546–12558.
[10] Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. 2016. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology* 12, 9 (2016), 513–525.
[11] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. 2020. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging* 39, 7 (2020), 2494–2505.
[12] Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. 2022. Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2741–2751.
[13] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. 2017. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In *Computing in Cardiology (CinC)*. IEEE, 1–4.
[14] Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. 2021. Minirocket: A very fast (almost) deterministic transform for time series classification. In

Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 248–257.

[15] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. 2023. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. In *Advances in Neural Information Processing Systems*.

[16] Jiahao Fan, Hangyu Zhu, Xinyu Jiang, Long Meng, Chen Chen, Cong Fu, Huan Yu, Chenyun Dai, and Wei Chen. 2022. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), 205–216.

[17] Kshitij Goel, Neetu Sood, and Indu Saini. 2023. Ensemble Technique based Parkison's Disease Detection from FOG and EEG Signals. In *World Conference on Communication & Computing (WCONF)*. IEEE, 1–5.

[18] Peiliang Gong, Ziyu Jia, Pengpai Wang, Yueying Zhou, and Daoqiang Zhang. 2023. ASTDF-Net: Attention-Based Spatial-Temporal Dual-Stream Fusion Network for EEG-Based Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 883–892.

[19] A Harati, S Lopez, I Obeid, J Picone, MP Jacobson, and S Tobochnik. 2014. The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In *2014 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE, 1–5.

[20] Muhammad Anas Hasnul, Nor Azlina Ab Aziz, Salem Alelyani, Mohamed Mohana, and Azlan Abd Aziz. 2021. Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review. *Sensors* 21, 15 (2021), 5015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Rongyan He, Hao Liu, Yan Niu, Huiqing Zhang, Guy M Genin, and Feng Xu. 2022. Flexible miniaturized sensor technologies for long-term physiological monitoring. *npj Flexible Electronics* 6, 1 (2022), 20.

[23] Radwa Hossieny, Manal Tantawi, Mohamed Fahmy Tolba, et al. 2022. Developing a Method for Classifying Electro-Oculography (EOG) Signals Using Deep Learning. *International Journal of Intelligent Computing and Information Sciences* 22, 3 (2022), 1–13.

[24] Andres Jaramillo-Gonzalez, Shizhe Wu, Alessandro Tonin, Aygul Rana, Majid Khalili Ardali, Niels Birbaumer, and Ujwal Chaudhary. 2021. A dataset of EEG and EOG from an auditory EOG-based communication system for patients in locked-in state. *Scientific data* 8, 1 (2021), 8.

[25] Ziyu Jia, Youfang Lin, Yuhan Zhou, Xiyang Cai, Peng Zheng, Qiang Li, and Jing Wang. 2023. Exploiting Interactivity and Heterogeneity for Sleep Stage Classification Via Heterogeneous Graph Neural Network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[26] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. *arXiv preprint arXiv:2405.18765* (2024).

[27] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[28] Kranti Kamble and Joydeep Sengupta. 2023. A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals. *Multimedia Tools and Applications* (2023), 1–36.

[29] Jessleen K Kanwal, Emma Coddington, Rachel Frazer, Daniela Limbania, Grace Turner, Karla J Davila, Michael A Givens, Valarie Williams, Sandeep Robert Datta, and Sara Wasserman. 2021. Internal state: dynamic, interconnected communication loops distributed across body, brain, and time. *Integrative and Comparative Biology* 61, 3 (2021), 867–886.

[30] Stamos Katsigiannis and Naeem Ramzan. 2017. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 98–107.

[31] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* 47, 9 (2000), 1185–1194.

[32] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. 2016. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine* 124 (2016), 180–192.

[33] Ana Karina Kirby, Sidharth Pancholi, Zada Anderson, Caroline Chesler, Thomas H Everett IV, and Bradley S Duerstock. 2023. Time and frequency domain analysis of physiological features during autonomic dysreflexia after spinal cord injury. *Frontiers in Neuroscience* 17 (2023).

[34] Heng Liang, Yucheng Liu, Haichao Wang, Ziyu Jia, and Brainnetome Center. 2023. Teacher assistant-based knowledge distillation extracting multi-level features on single channel sleep EEG. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*. 3948–3956.

[35] Xuefen Lin, Jielin Chen, Weifeng Ma, Wei Tang, and Yuchen Wang. 2023. EEG emotion recognition using improved graph neural network with channel selection. *Computer Methods and Programs in Biomedicine* 231 (2023), 107380.

[36] Yu Liu, Yufeng Ding, Chang Li, Juan Cheng, Rencheng Song, Feng Wan, and Xun Chen. 2020. Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Computers in Biology and Medicine* 123 (2020), 103927.

[37] Yuchen Liu and Ziyu Jia. 2022. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*.

[38] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.

[39] Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183* (2022).

[40] Wissal Midani, Wael Ouarda, and Mounir Ben Ayed. 2023. DeepArr: An investigative tool for arrhythmia detection using a contextual deep neural network from electrocardiograms (ECG) signals. *Biomedical Signal Processing and Control* 85 (2023), 104954.

[41] Mansooreh Montazerin, Soheil Zabihi, Elahe Rahimian, Arash Mohammadi, and Farnoosh Naderkhani. 2022. ViT-HGR: Vision transformer-based hand gesture recognition from high density surface EMG signals. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 5115–5119.

[42] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H Falk. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion* 76 (2021), 355–375.

[43] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations* (2023).

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

[45] Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2021. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 5903–5915.

[46] Huy Phan, Kristian P Lorenzen, Elisabeth Heremans, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B Mikkelsen, and Maarten De Vos. 2023. L-SeqSleepNet: Whole-cycle long sequence modelling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics* (2023).

[47] Huy Phan and Kaare Mikkelsen. 2022. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement* 43, 4 (2022), 04TR01.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

[49] Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. 2020. Deep learning in physiological signal data: A survey. *Sensors* 20, 4 (2020), 969.

[50] Claude E Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE* 37, 1 (1949), 10–21.

[51] Manish Sharma, Anuj Yadav, Jainendra Tiwari, Murat Karabatak, Ozal Yildirim, and U Rajendra Acharya. 2022. An automated wavelet-based sleep scoring model using eeg, emg, and eog signals with more than 8000 subjects. *International Journal of Environmental Research and Public Health* 19, 12 (2022), 7176.

[52] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), 710–719.

[53] Petre Stoica, Randolph L Moses, et al. 2005. *Spectral analysis of signals*. Vol. 452. Pearson Prentice Hall Upper Saddle River, NJ.

[54] Mingyi Sun, Weigang Cui, Shuyue Yu, Hongbin Han, Bin Hu, and Yang Li. 2022. A Dual-Branch Dynamic Graph Convolution Based Adaptive TransFormer Feature Fusion Network for EEG Emotion Recognition. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2218–2228.

[55] Akara Supratak and Yike Guo. 2020. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 641–644.

[56] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. 2020. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing* (2020).

[57] Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, et al. 2001. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep medicine* 2, 6 (2001), 537–554.

[58] Nitish V Thakor. 2017. Biopotentials and electrophysiology measurement. *Measurement, Instrumentation, and Sensors Handbook* (2017), 64–1.
[59] Alessandro Tonin, Andres Jaramillo-Gonzalez, Aygul Rana, Majid Khalili-Ardali, Niels Birbaumer, and Ujwal Chaudhary. 2020. Auditory electrooculogram-based communication system for ALS patients in transition from locked-in to complete locked-in state. *Scientific reports* 10, 1 (2020), 8452.
[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[61] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 33536–33549.
[62] Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. 2024. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 265–273.
[63] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Thirty-seventh Conference on Neural Information Processing Systems* (2023).
[64] Xiaoman Wang, Jianwen Zhang, Chunhua He, Heng Wu, and Lianglun Cheng. 2023. A Novel Emotion Recognition Method Based on the Feature Fusion of Single-Lead EEG and ECG Signals. *IEEE Internet of Things Journal* (2023).
[65] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The eleventh international conference on learning representations*.
[66] Qiao Xiao, Khuan Lee, Siti Aisah Mokhtar, Iskasymar Ismail, Ahmad Luqman bin Md Pauzi, Qiuxia Zhang, and Poh Ying Lim. 2023. Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review. *Applied Sciences* 13, 8 (2023), 4964.
[67] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19163–19173.
[68] Ning Yi, Haitao Cui, Lijie Grace Zhang, and Huanyu Cheng. 2019. Integration of biological systems with electronic-mechanical assemblies. *Acta biomaterialia* 95 (2019), 91–111.
[69] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).
[70] Zhizhang Yuan, Daoze Zhang, Junru Chen, Gefei Gu, and Yang Yang. 2024. Brant-2: Foundation Model for Brain Signals. *arXiv preprint* (2024).
[71] Zhizhang Yuan, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. 2023. PPi: Pretraining Brain Signal Model for Patient-independent Seizure Detection. In *Thirty-seventh Conference on Neural Information Processing Systems*.
[72] Soheil Zabihi, Elahe Rahimian, Amir Asif, and Arash Mohammadi. 2023. Trahgr: Transformer for hand gesture recognition via electromyography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).
[73] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. 2023. Brant: Foundation Model for Intracranial Neural Signal. In *Thirty-seventh Conference on Neural Information Processing Systems*.
[74] Wei Zhang, Zhuokun Yang, Hantao Li, Debin Huang, Lipeng Wang, Yanzhao Wei, Lei Zhang, Lin Ma, Huanhuan Feng, Jing Pan, et al. 2022. Multimodal data for the detection of freezing of gait in Parkinson's disease. *Scientific data* 9, 1 (2022), 606.
[75] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems* 35 (2022), 3988–4003.
[76] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*.
[77] Jiadao Zou and Qingxue Zhang. 2021. eyeSay: Eye electrooculography decoding with deep learning. In *IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–3.

A DETAILS OF THE EXG ENCODER

Because the focus of this paper is to introduce our proposed alignment framework, the specific encoder architecture can be flexible. The model architecture of the EXG encoder used in this paper are introduced here.

Since physiological signals are bioelectric signals, the time domain provides information about the amplitude and duration, while the frequency domain can reveal the oscillation patterns and underlying biological rhythms [33]. Therefore, to combine the information from both time and frequency domains, in our EXG encoder, we first calculate the power spectral density (PSD) [53], which describes the distribution of the signal's total average power over frequency, as the information in frequency domain. Then, a convolutional neural network (CNN) performs on the PSD to extract features in the frequency domain of the EXG signal. The extracted features in the frequency domain will be concatenated with the convolution-derived features in the time domain, serving as the features within a single patch. Due to the fact that physiological signals are time series, each patch has a temporal dependency with its contextual patches from the same sequence. With this in mind, the features of consecutive patches from a sequence will be fed into the Transformer [60] to obtain a more comprehensive representation that considers temporal dependencies.

Formally, all the patches $\{\tilde{\boldsymbol{x}}_{i,j}\}_{j=0}^{P-1}$ from the i-th EXG data sequence $\tilde{\boldsymbol{x}}_i$ are input into the EXG encoder, generating their representations $\{\tilde{\boldsymbol{p}}_{i,j}\}_{j=0}^{P-1}$:

$$\tilde{\boldsymbol{p}}_{i,j} = \text{Transformer}\Big(\text{CNN}_T\left(\tilde{\boldsymbol{x}}_{i,j}\right) \parallel \text{CNN}_F\left(\text{PSD}(\tilde{\boldsymbol{x}}_{i,j})\right)\Big), \quad j = 0, 1, 2, ..., P-1, \tag{6}$$

where $\tilde{\boldsymbol{p}}_{i,j} \in \mathbb{R}^{D_p}$ denotes the representation of the j-th patch from the i-th EXG data sequence $\tilde{\boldsymbol{x}}_i$, and D_p denotes the dimension of patch representations.

B DETAILS OF BASELINES

Firstly, we compare *Brant-X* with the existing self-supervised or unsupervised works on genral time series. The Details of these baseline models are given here:

- TF-C [75]: A decomposable pre-training model for general time series modeling, where the self-supervised signal is provided by the distance between time and frequency components.
- SimMTM [15]: A pre-training framework on time series to recover masked time points by the weighted aggregation of multiple neighbors outside the manifold.

Also, we compare *Brant-X* with the methods that performs time series classification based on pre-trained language models. Hence, we set OneFitsAll [76] as a baseline:

- OneFitsAll [76]: A unified model that leverages language or vision models for time series analysis, leading to a comparable or SOTA performance in all main time series analysis tasks.

Furthermore, to illustrate the effectiveness of *Brant-X* in various application scenarios, we compare our framework with the SOTA methods those are specially designed for each of the four downstream tasks. These supervised methods includes:

(1) For the sleep stage classification task:

- TinySleepNet [55]: An end-to-end model based on CNN and LSTM for automatic sleep stage scoring on raw single-channel EEG with a less number of trainable parameters.
- XSleepNet [45]: A sequence-to-sequence sleep staging model that is capable of learning a joint representation from both raw signals and time-frequency images.

Table 5: Average performance on the arrhythmia detection task.

Methods \ Metrics	Overall	N rhythm			A rhythm			O rhythm		
	Acc.	Sens.	Spec.	Prec.	Sens.	Spec.	Prec.	Sens.	Spec.	Prec.
TF-C [75]	71.91 ±2.25	81.44 ±10.07	25.37 ±10.91	64.61 ±0.67	3.51 ±2.15	96.88 ±0.99	8.89 ±2.92	22.95 ±9.85	84.08 ±9.42	39.33 ±6.10
SimMTM [15]	81.30 ±2.57	83.60 ±10.91	65.43 ±15.21	81.27 ±4.97	59.13*±10.14	95.01 ±3.28	56.27 ±12.18	49.58 ±14.21	85.00 ±8.18	58.13 ±8.02
OneFitsAll [76]	73.67 ±1.92	83.88 ±11.38	25.40 ±15.49	66.49 ±1.57	9.90 ±6.58	97.51 ±2.28	34.41 ±12.32	22.53 ±14.74	86.00 ±10.62	43.35 ±10.46
DeepArr [40]	86.94*±1.67	94.03*±4.51	66.29*±15.48	83.89 ±5.55	52.09 ±19.37	98.20*±1.72	77.34*±11.17	57.08*±19.78	91.89*±6.49	76.81*±12.13
Alamatsaz et al. [3]	88.08 ±1.45	94.97 ±2.29	67.66 ±4.92	83.59*±1.92	60.28 ±11.09	98.66 ±0.76	81.59 ±8.83	59.81 ±4.26	93.33 ±1.93	77.94 ±5.18
Brant-X	**93.40**±1.63	**96.46**±3.14	**83.28**±3.59	**90.96**±1.73	**79.83**±7.29	**99.62**±0.23	**95.19**±2.55	**78.80**±4.87	**95.21**±3.44	**87.14**±7.42

- L-SeqSleepNet [46]: A method for efficient long sequence modelling that considers whole-cycle sleep information for sleep staging, showing robustness in alleviating classification errors.
- SleepHGNN [25]: A novel sleep heterogeneous graph neural network designed to capture interactivity and heterogeneity of physiological signals for accurate sleep stage classification.

(2) For the emotion recognition task:

- MLF-CapsNet [36]: A multi-level features guided capsule network for multi-channel EEG-based emotion recognition, which can simultaneously extract features from the raw EEG signals and determine the emotional states.
- EEG-Conformer [52]: A compact convolutional Transformer to encapsulate local and global features in a unified EEG classification framework for motor imagery and emotion recognition.
- Lin et al. [35]: A graph convolution model with dynamic channel selection for emotion classification, which combines the advantages of 1D convolution and graph convolution to capture the intra- and inter-channel EEG features.
- Wang et al. [64]: An emotion recognition method based on the feature fusion of single-lead EEG and ECG signals, using various time-domain, frequency-domain, and nonlinear features.

(3) For the freezing of gaits detection task:

- Aly and Youssef [5]: A model based on CNN and LSTM that integrates EEG with EMG signals to investigate the efficiency of deep learning in hybrid systems with signal fusion for motion classification.
- Batool and Javeed [6]: A feature engineering method that uses time-frequency feature extraction strategy and CNN-BiLSTM to detect walking disorder in Parkinson's disease patients.
- Goel et al. [17]: An ensemble techniques that combines the prediction of multiple methods to improve the model performance for freezing of gaits detection on EEG signals

(4) For the eye movement communication task:

- eyeSay [77]: A multi-stage convolutional neural network to decode eye dynamics using electrooculography, towards voice-free communication for patients with amyotrophic lateral sclerosis.
- Adama and Bogdan [1]: A feature engineering method that employs features like relative power, spectral edge frequencies and symbolic mutual information for eye movement classification.
- Hossieny et al. [23]: A model based on ResNet[21] using horizontal and vertical EOG signals to determine six eye movement directions.

For some baselines that are not open source, we re-implemented them for experiments. In order to make a fair comparison, for baselines designed for only one type of physiological signal (EEG or EXG), we take their best results on the following three settings as their final results: only on EEG, only on EXG, and aggregation the representation from EEG and EXG.

C ANALYSIS ON THE ARRHYTHMIA DETECTION TASK

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, occurring in about 2% of the general population and is associated with significant mortality and morbidity through association of risk of death, stroke, heart failure and coronary artery disease [13]. Therefore, accurate rhythm classification and arrhythmia detection are vital to the prevention and treatment of heart disease. Depending on the different classifications of cardiac states, the task can be viewed as a multi-classification problem.

The AFDB dataset [13] comprises 12,186 single lead ECG recordings of 30 and 60sec long, gathered from subjects undergoing long-haul mobile ECG checking. Data are collected at 300Hz, and each sample may belong to one of four classes: (1) normal sinus rhythm, (2) AF, (3) other rhythm, or (4) too noisy to classify. In our experiment, we remove the noisy samples so that this task is a three-class classification problem.

As a supplement to *Brant-X w/o EEG-encoder* in the ablation study, we conduct this experiment with the aligned EXG encoder on ECG data (without incorporating the EEG encoder on EEG data). The experiment is conducted on training, validation and test data in a 3:1:1 ratio and repeated to obtain the overall results. For each of the above three classes, we use sensitivity, specificity and precision as metrics to evaluate the performance of our aligned EXG encoder and other baselines. We also report the overall accuracy as an overall assessment of model performance. In line with the main experiments on the four main tasks, besides TF-C [75], SimMTM [15] and OneFitsAll [76], we also compare our EXG encoder with the SOTA methods in the field of arrhythmia detection to demonstrate the effectiveness of our model. These methods includes DeepArr [40] and Alamatsaz et al. [3].

As shown in Tab. 5, our *Brant-X* beats all of the baselines on the arrhythmia detection task. This demonstrates that the phase of alignment training empowers the EXG encoder to effectively learn semantic representations from ECG and classify cardiac rhythms.