A Notaion Table

The listed the notations used in the main text in the following Table 4.

Table 4: Notations used in BIOT

Symbols	Descriptions
$I \in \mathbb{N}^+$	the number of channels in biosignal, such as 16
$J \in \mathbb{R}^+$	the length of biosignals, such as $10s \times 200Hz$ (use a complete sample for ease of notation)
$\mathbf{S} \in \mathbb{R}^{I \times J}$	the mulit-channel biosignal
$\mathbf{S}[i] \in \mathbb{R}^J$	the i-th channel of the biosignal.
$r \in \mathbb{R}^+$	the sampling rate, such as 200Hz
$t \in \mathbb{R}^+$	the length of the biosignal token, such as $1s \times 200Hz$
$p \in \mathbb{R}^+$	the overlap length between two neighboring tokens, such as $0.5s \times 200Hz, t > p$
$k \in \mathbb{N}^+$	the k-th token in the i-th channel is denoted by $\mathbf{S}[i, (t-p)(k-1) : (t-p)(k-1)+t]$ which has length t
$K \in \mathbb{N}^+$	max number of tokens in one channel
$N \in \mathbb{N}^+$	the number of tokens in tokenized biosignal "sentence"
$l_1 \in \mathbb{N}^+$	the dimension of the token embedding
$\mathbf{X} \in \mathbb{R}^{N \times l_1}$	the tokenized biosignal "sentence"
$l_2 \in \mathbb{N}^+$	the new dimensions of tokens after self-attention module
$\mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^Q \in \mathbb{R}^{l_1 \times l_2}$	the key, value, and query matrices in self-attention module
$\mathbf{E}, \mathbf{F} \in \mathbb{R}^{d \times N}$	the low-rank projection matrices
$d \in \mathbb{N}^+$	the reduced rank for self-attention matrices, $d \ll N$
$\mathbf{H} \in \mathbb{R}^{N \times l_2}$	the output of the self-attention module
$\tilde{\mathbf{S}} \in \mathbb{R}^{I \times J}$	the perturbed biosignal sample
$\mathbf{Z}, \tilde{\mathbf{Z}}$	the real and predicted embeddings of $\mathbf{S}$
T	the temperature hyperparameter in the contrastive loss
$\mathbf{I}$	the identity matrix in the contrastive loss, has the size of (batch size, batch size)
$a \in [0, 1, 2, 3, 4, 5]$	the discrete uniformly distributed variable for the number of masked segments in Section 3.3
$b \in [0, 1, 2, 3, 4]$	the discrete uniformly distributed variable for the number of masked channels in Section 3.3

B Details of Datasets and Experimental Settings

B.1 More for Datasets and Processings

We provide more descriptions on each dataset in this section.

For EEG datasets. First, the 16 derivations (in 10-20 international system) are "FP1-F7", "F7-T7", "T7-P7", "P7-O1", "FP2-F8", "F8-T8", "T8-P8", "P8-O2", "FP1-F3", "F3-C3", "C3-P3", "P3-O1", "FP2-F4", "F4-C4", "C4-P4", "P4-O2".

- Sleep Heart Health Study (**SHHS**) (Zhang et al., 2018; Quan et al., 1997) is a multi-center cohort study from the National Heart Lung & Blood Institute assembled to study sleep-disordered breathing, which contains 5,445 recordings. The data is accessible upon request in their website [2]. Each recording has 14 Polysomnography (PSG) channels, and the recording frequency is 125.0 Hz. We use the C3/A2 and C4/A1 EEG channels. The dataset is released with sleep annotations. We use the existing codes [3] and split each recordings into 30-second samples. In this study, we use SHHS samples for unsupervised pre-training without the original labels.
- **PREST** is a private dataset recorded in hospital sleep lab, primarily for seizure and abnormal EEG detection purpose (such as spikes). The local IRB waived the requirement for informed consent for this retrospective analysis of EEG data. We follow the clinician's instructions and split each recordings into 10 seconds without labels. In the experiment, we use it for EEG model pre-training.
- The **CHB-MIT** database [4] (Shoeb, 2009) is publicly available, which is collected at the Children's Hospital Boston, consists of EEG recordings from pediatric subjects with intractable seizures. The dataset is under Open Data Commons Attribution License v1.0 [5] and is used to predict whether the EEG recordings contain seizure signals. Each recording initially contains 23 bipolar channels and

[2] https://sleepdata.org/datasets/shhs

[3] https://github.com/ycq091044/ContraWR/tree/main/preprocess

[4] https://physionet.org/content/chbmit/1.0.0/

[5] https://physionet.org/content/chbmit/view-license/1.0.0/

we select the 16 standard derivations in the experiments. We utilize the existing preprocessing [6] and follow the typical practices to further split each recordings into 10-second non-overlapping samples by default. Since the dataset is highly imbalanced, we use 5 seconds as overlaps to split the seizure regions (which could potentially double the positive samples). After processing, the positive ratio in the training set is around 0.6%.

- **IIIC Seizure** is requested from Ge et al. (2021); Jing et al. (2023), and we follow the license and usage statements in Jing et al. (2023). The samples follow 16 derivations and span 10-second signals at 200Hz. This dataset is used for predicting one of the six classes: lateralized periodic discharges (LPD), generalized periodic discharges (GPD), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), Seizure types, and Other.
- TUH Abnormal EEG Corpus (**TUAB**) (Lopez et al., 2015) and TUH EEG Events (**TUEV**) (Harati et al., 2015) is accessible upon request at Temple University Electroencephalography (EEG) Resources [7]. We process both datasets to follow the 16 EEG derivations.

For ECG datasets. We use the Cardiology collection to pre-train the ECG models and apply it on downstream supervisd PTB-XL task.

- The **Cardiology** collection (Alday et al., 2020) is publicly available at physionet [8], which was used in the PhysioNet/Computing in Cardiology Challenge 2020. This collection is under Creative Commons Attribution 4.0 International Public License [9]. In this study, we use five sets from the training portion of the collection (It has in total six sets. Another one overlaps with the PTB-XL dataset, and thus we drop it for pre-training), which contains recordings from CPSC2018 (6,877 recordings), CPSC2018Extra (China 12-Lead ECG Challenge Database – unused CPSC 2018 data, 3,453 recordings), St Petersburg Incart (12-lead Arrhythmia Database, 74 recordings), PTB (Diagnostic ECG Database, 516 recordings), Georgia (12-Lead ECG Challenge Database, 10,344 recordings). For preprocessing, we extract 10-second samples from each recording with 0.5s as the overlapping window (for obtaining more unsupervised trianing corpus). All the samples are merged together as an unsupervised pre-training ECG corpus of nearly 0.5 million samples. We pre-train a Pre-trained `BIOT` (Cardiology-12) on all the channels and a Pre-trained `BIOT` (Cardiology-6) on the first 6-channels of all samples. The sample sizes are different from the below PTB-XL dataset.
- Physikalisch-Technische Bundesanstalt (**PTB-XL**) [10] (Wagner et al., 2020) is a publicly available large dataset of 12-lead ECGs from 18885 patients. It is under the Creative Commons Attribution 4.0 International Public License [11]. The raw waveform data was annotated by up to two cardiologists, who assigned potentially multiple ECG statements to each record up to 27 diagnoses: 1:1st degree AV block, 2:Atrial fibrillation, 3:Atrial flutter, 4:Bradycardia, 5:Complete right bundle branch block, 6:Incomplete right bundle branch block, 7:Left anterior fascicular block, 8:Left axis deviation, 9:Left bundle branch block, 10:Low QRS voltages, 11:Nonspecific intraventricular conduction disorder, 12:Pacing rhythm, 13:Premature atrial contraction, 14:Premature ventricular contractions, 15:Prolonged PR interval, 16:Prolonged QT interval, 17:Q wave abnormal, 18:Right axis deviation, 19:Right bundle branch block, 20:Sinus arrhythmia, 21:Sinus bradycardia, 22:Sinus rhythm, 23:Sinus tachycardia, 24:Supraventricular premature beats, 25:T wave abnormal, 26:T wave inversion, 27:Ventricular premature beats. We following clinical knowledges and further groups them into six broader categories: Arrhythmias, Bundle branch blocks and fascicular blocks, Axis deviations, Conduction delays, Wave abnormalities, Miscellaneous. Each recordings can be associated to multiple categories. In this paper, we conduct the "Arrhythmias" phenotyping prediction task. If the recordings have at least one diagnosis belonging to the Arrhythmias group, then we label them as positive, otherwise as negative.

For human activity sensory data. Human activity recognition (**HAR**) dataset [12] (Anguita et al., 2013) is publicly available at UCI machine learning repository. The data is collected from smartphone accelerometer and gyroscope data with 3D coordinates to detect six actions: walking, walking

[6] https://github.com/bernia/chb-mit-scalp

[7] https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml

[8] https://physionet.org/content/challenge-2020/1.0.2/

[9] https://physionet.org/content/challenge-2020/view-license/1.0.2/

[10] https://physionet.org/content/ptb-xl/1.0.1/

[11] https://physionet.org/content/ptb-xl/view-license/1.0.1/

[12] https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones

upstairs, walking downstairs, sitting, standing, laying. The samples are already splitted and provided in the original datasets.

B.2 More for Experimental Settings

For model implementation, the SPaRCNet code is requested from the authors (Jing et al., 2023), the ContraWR code is downloaded and modified upon the github [13], CNN-Transformer is easily implemented following the Fig. 3 of the original paper (Peh et al., 2022), FFCL (Li et al., 2022) combines a CNN model and a LSTM model for learning separete representations and then merges them before the final prediction layer, the implementation of ST-Transformer refer to this repo [14]. The linear-complexity attention module is referred to this repo [15] in our `BIOT` implementation.

For all EEG tasks, we resample the datasets into 200Hz. The ECG tasks use 500Hz, and the HAR tasks use 50Hz by default. For each specific tasks, we have to adjust the baseline model architectures (e.g, number of layers, input channel sizes, etc) accordingly since the input data have various formats. While for our `BIOT`, we only adjust the fft size based on their sampling rate (200Hz for EEG, 1000Hz for ECG, 100Hz for HAR) and use 0.5s, 0.2s, and 0.1s as the hop length (i.e., overlaps) in three signal types. These model configurations are chosen by testing several combinations based on the validation performance and we select the best one. For our `BIOT` model, we use 8 as the number of head, 4 as the number of transformer layers, and $T = 2$ as the temperature in unsupervised pre-training by default. We use the Adam optimizer with learning rate 1×10^{-3} and 1×10^{-5} as the coefficient for L2 regularization by default. We use the pytorch lightning framework (with 100 as the max epoch) to handle the training, validation, and test pipeline by setting AUROC as the monitoring metirc for binary classification and Coken's Kappa as the monitoring metric for multi-class classification in the validation. More details can refer to our Supplementary codes. Below, we provide the definition of each metric used in the paper.

Balanced Accuracy is defined as the average of recall obtained on each class. It is used for both binary classification and multi-class classification.

AUC-PR is the area under the precision recall (PR) curve for binary classification task.

AUROC is the area under the ROC curve, summarizing the ROC curve into an single number that describes the performance of a model for multiple thresholds at the same time. It is used for binary classification.

Coken's Kappa is a statistic that measures inter-annotator agreement, which is usually used for imbalanced multi-class classification task. The calculation can refer to sklearn metrics [16].

Weighted F1 is used for multi-class classification in this paper, which is a weighted average of individual F1-scores from each class, with each score weighted by the number of samples in the corresponding class.

C Additional Results

This section provides additional experimental results to support claims in the main paper.

C.1 Additional Experiments on TUEV and TUAB

We have provided the supervised learning results on EEG dataset IIIC Seizure and CHB-MIT in the main text. For completeness, we provide similar comparison results on TUAB and TUEV below in Table 5 6, which show a similar trend that our `BIOT` shows better performance against baseline models, and the pre-trained `BIOT` models can bring significant improvements on two downstream tasks, especially on TUEV. For TUEV, we also append the results of all different pre-trained models (e.g., train from scratch, supervised training, unsupervised training, etc) in the end in Table 6.

[13] https://github.com/ycq091044/ContraWR

[14] https://github.com/eeyhsong/EEG-Transformer

[15] https://github.com/lucidrains/linear-attention-transformer

[16] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

Table 5: Additional Supervised Learning Results on TUAB

Models	TUAB (abnormal detection)		
	Balanced Acc.	AUC-PR	AUROC
SPaRCNet	0.7896 ± 0.0018	0.8414 ± 0.0018	0.8676 ± 0.0012
ContraWR	0.7746 ± 0.0041	0.8421 ± 0.0104	0.8456 ± 0.0074
CNN-Transformer	0.7777 ± 0.0022	0.8433 ± 0.0039	0.8461 ± 0.0013
FFCL	0.7848 ± 0.0038	0.8448 ± 0.0065	0.8569 ± 0.0051
ST-Transformer	0.7966 ± 0.0023	0.8521 ± 0.0026	**0.8707 ± 0.0019**
(Vanilla) BIOT	**0.7925 ± 0.0035**	**0.8707 ± 0.0087**	0.8691 ± 0.0033
Pre-trained BIOT (PREST)	0.7907 ± 0.0050	0.8752 ± 0.0051	0.8730 ± 0.0021
Pre-trained BIOT (PREST+SHHS)	$\boxed{0.8019 \pm 0.0021}$	0.8749 ± 0.0054	0.8739 ± 0.0019
Pre-trained BIOT (6 EEG datasets)	0.7959 ± 0.0057	$\boxed{0.8792 \pm 0.0023}$	$\boxed{0.8815 \pm 0.0043}$

* **Bold** for the best model (trained from scratch) and $\boxed{\text{box}}$ for the best pre-trained models.

Table 6: Additional Supervised Learning Results on TUEV (All-in-one-table comparison)

Models	TUEV (event type classification)		
	Balanced Acc.	Coken's Kappa	Weighted F1
(Training from scratch in Section 3.2			
SPaRCNet	0.4161 ± 0.0262	0.4233 ± 0.0181	0.7024 ± 0.0104
ContraWR	0.4384 ± 0.0349	0.3912 ± 0.0237	0.6893 ± 0.0136
CNN-Transformer	0.4087 ± 0.0161	0.3815 ± 0.0134	0.6854 ± 0.0293
FFCL	0.3979 ± 0.0104	0.3732 ± 0.0188	0.6783 ± 0.0120
ST-Transformer	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
(Vanilla) BIOT	0.4682 ± 0.0125	0.4482 ± 0.0285	0.7085 ± 0.0184
(Unsupervised pre-trained models in Section 3.4):			
Pre-trained BIOT (PREST)	0.5207 ± 0.0285	0.4932 ± 0.0301	0.7381 ± 0.0169
Pre-trained BIOT (PREST+SHHS)	0.5149 ± 0.0292	0.4841 ± 0.0309	0.7322 ± 0.0196
(Supervised pre-trained models in Section 3.5):			
Pre-trained BIOT (pre-trained on CHB-MIT with 8 channels and 10s)	0.4123 ± 0.0087	0.4285 ± 0.0065	0.6989 ± 0.0015
Pre-trained BIOT (pre-trained on CHB-MIT with 16 channels and 5s)	0.4218 ± 0.0117	0.4427 ± 0.0093	0.7147 ± 0.0058
Pre-trained BIOT (pre-trained on CHB-MIT with 16 channels and 10s)	0.4344 ± 0.0065	0.4719 ± 0.0231	0.7280 ± 0.0126
Pre-trained BIOT (pre-trained on IIIC seizure with 8 channels and 10s)	0.4956 ± 0.0552	0.4719 ± 0.0475	0.7214 ± 0.0220
Pre-trained BIOT (pre-trained on IIIC seizure with 16 channels and 5s)	0.4894 ± 0.0189	0.4881 ± 0.0045	0.7348 ± 0.0056
Pre-trained BIOT (pre-trained on IIIC seizure with 16 channels and 10s)	0.4935 ± 0.0288	0.5316 ± 0.0176	0.7555 ± 0.0111
Pre-trained BIOT (pre-trained on TUAB with 8 channels and 10s)	0.4980 ± 0.0384	0.4487 ± 0.0535	0.7044 ± 0.0365
Pre-trained BIOT (pre-trained on TUAB with 16 channels and 5s)	0.4954 ± 0.0305	0.5053 ± 0.0079	0.7447 ± 0.0049
Pre-trained BIOT (pre-trained on TUAB with 16 channels and 10s)	0.5256 ± 0.0348	0.5187 ± 0.0160	0.7504 ± 0.0102
(Supervised + unsupervised pre-trained model in Section 3.6):			
Pre-trained BIOT (6 EEG datasets)	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082

C.2 Additional Experiments on CHB-MIT

This section performs a similar experiment on CHB-MIT, similar to Section 3.2. We pre-train on the training set of IIIC Seizure (which has 16 channels and 10s duration), TUAB (which has 16 channels and 10s duration), TUEV (which has 16 channels and 5s duration) and fine-tunes on CHB-MIT (which has 16 channels and 10s duration). All datasets use 200Hz sampling rate. We design five sets of configurations for the pre-trained datasets: **Format (i)** uses the first 8 channels and 10s duration; **Format (ii)** uses the full 16 channels but only the first 5s recording; **Format (iii)** uses full 16 channels and full 10s recording; **Format (iv)** uses 8 channels and 5s recording, and **Format (v)** uses full 16 channels and 2.5s recording. The last two are only for the TUEV dataset. During fine-tuning, we then remove the prediction layers from these pre-trained model and add a new prediction layer to fit the CHB-MIT dataset.

The results are reported in Figure 5, which shows that the supervised pre-training on both IIIC seizure and TUEV can help improve the downstream performance on CHB-MIT task compared to training from scratch. The reason is that IIIC Seizure is on multiple seizure type classification while CHB-MIT is on binary "seizure or not" classification, and the context of both tasks are fairly related. Although TUEV is not entirely on seizure related classification, some classes in TUEV are seizure subtypes (such as GPED, PLED), and thus its supervisd pre-trained models can also bring benefits for the CHB-MIT task.

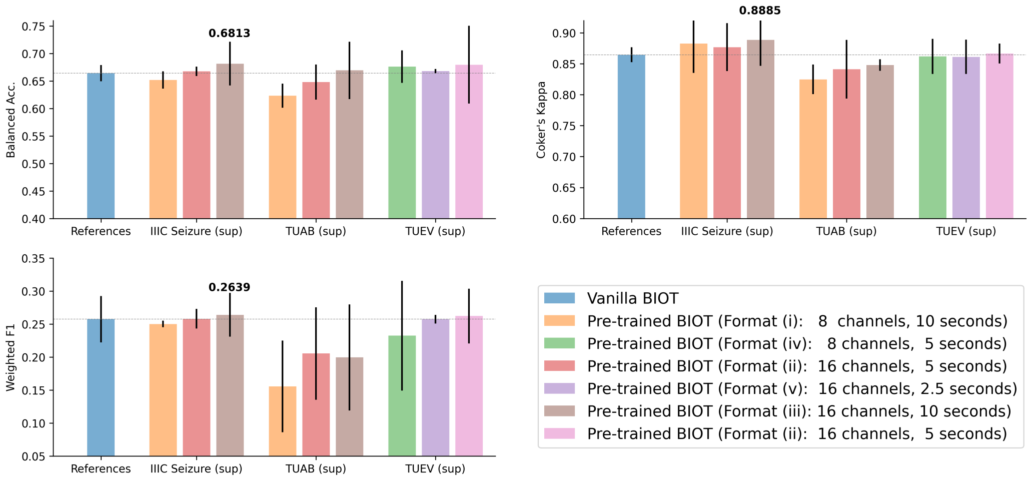

Figure 5: Fine-tuned on CHB-MIT from different supervised pre-trained models. IIIC Seizure and TUAV datasets follow Format (i)(ii)(iii), while TUEV follows the Format (iv)(v)(ii).

C.3 Ablation Studies on Hyperparameters

This section provides ablation studies on three hyperparameters in data processing: target sampling rate r, token length t, and the overlap size p between two neighboring tokens. We use two EEG datasets as example: IIIC Seizure and TUAB. The default configuration in the main paper is **(1) sampling**: $r = 200Hz$, **(2) token length**: $t = 1s \times r$, **(3) overlaps**: $p = 0.5s \times r$ as reference.

We also show an illustration for tokenization with overlap in Figure 6.

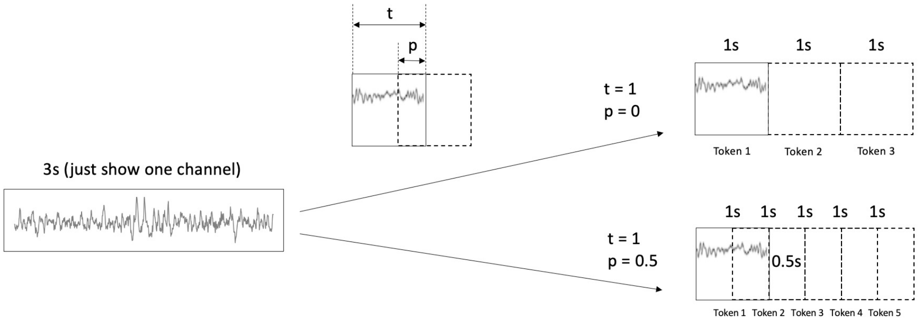

Figure 6: Illustration on tokenization with token length t and overlap p. The example shows one channel with 3 seconds. The configuration of $(t = r, p = 0)$ gives 3 tokens while the configuration of $(t = r, p = 0.5r)$ gives 5 tokens. Different configurations lead to different lengths.

C.3.1 Ablation Study on Target Sampling Rate r

In this experiment, we fix the coefficient "1" and "0.5" in (2)(3) and conduct ablation study on the target sampling rate r. The original IIIC Seizure data is at 200Hz and the TUAB data is at 256Hz. For IIIC Seizure, we vary the sampling rate to 26Hz, 50Hz, 100Hz, 150Hz, and 200Hz. For TUAB, we vary the sampling rate to 50Hz, 100Hz, 150Hz, 200Hz, 250Hz, and 300Hz. We use the tool $scipy.signal.resample$. The evaluations are conducted under three different random seeds and the mean and standard deviation values are reported.

For IIIC Seizure, we can observe that a higher sampling rate could give slightly better performance, especially on balanced acc. and coken's kappa. The reason is that higher sampling rate can preserve more detailed (high-frequency) biosignal information. The results on TUAB shows that the performances increase and then decrease slightly during increasing the sampling rate. We guess that with the increasing of r, initially the performance improves due to obtaining more information. Later, higher sampling rate does not bring more benefits but unnecessary frequency bands might incur some noise. We also conjecture that different tasks might have diverse sensitivity to the the frequency bands. For example, the task on IIIC seizure is to classify different seizure types, which may need to capture minor clues from high-frequency waves (such as Gamma waves (50-100Hz)), while the TUAB dataset is for abnormal detection, and using brain waves under 50Hz might be enough for the task. In sum, the target sampling rate r should be selected based on the predicting targets.

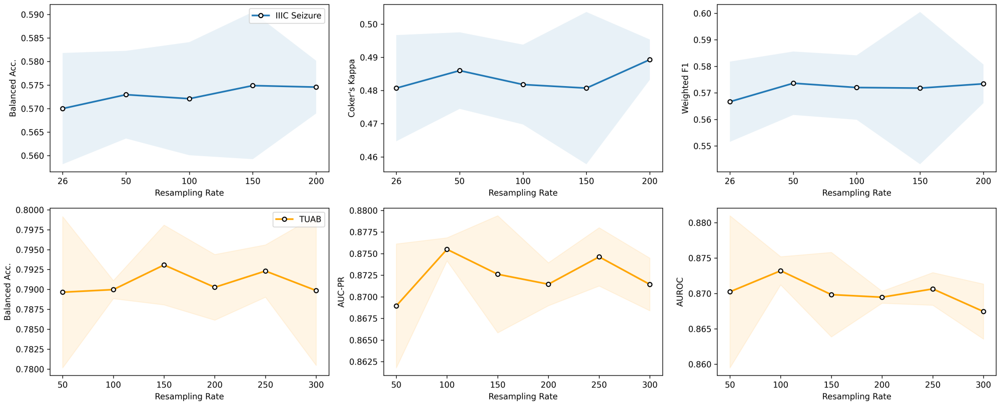

Figure 7: Ablation Study on Target Sampling Rate r

C.3.2 Ablation Study on Token Lengths t

In this experiments, we fix (1)(3) and conduct ablation study on the coefficient in token length t. Both datasets have 10s as the entire sample length and 0.5s as the overlap lengths. For both of the datasets, we vary the token length coefficient to 0.75s, 1s, 1.5s, 2s, 2.5s, 5s. The evaluations are conducted under three different random seeds and the mean and standard deviation values are reported.

For each configuration, we also set the fft size to match the token length, which means that 5s token duration will bring more frequency information. However, we find that by increasing the token lengths, the model performance starts to decrease. Model performances on IIIC Seizure starts to decrease after 1s while the performance on TUAB decreases after 2s. The reason could be that (i) longer token length (i.e., frequency bands) do not provide extra benefits for learning the tasks; (ii) given the increasing token lengths t, the total biosignal "sentence" length, which is $\frac{J-t}{t-p}+1=\frac{10-t}{t-0.5}+1$, will decrease (here, $J=10r$ is the channel duration, t is the token length, $p=0.5r$ is the overlapping length). For example, with $t=5r$ as the token lengths, the number of tokens becomes 2 per channel while the number of tokens per channel is 19 in the default configuration with $t=r$. The performance drops is due to transformer models will be less beneficial in shorter "sentence"s.

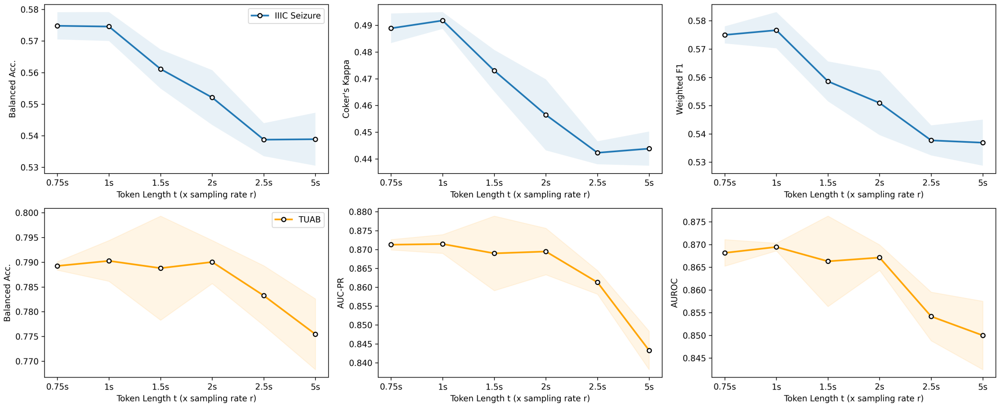

Figure 8: Ablation Study on Token Lengths t

C.3.3 Ablation Study on Overlapping Lengths p

In this experiments, we fix (1)(2) and conduct ablation study on the coefficient in the overlap length p. Both datasets have 10s as the entire sample length and 1s as the token lengths. For both of the datasets, we vary the overlap length coefficient to 0.875s, 0.75s, 0.5s, 0.25s, 0s. The evaluations are conducted under three different random seeds and the mean and standard deviation values are reported.

Based on the "sentence" length formula $\frac{J-t}{t-p}+1$, smaller overlap length p will decrease the "sentence" length. On both datasets, we find that larger overlap can brings better results due to that the biosignal "sentence" becomes longer, and Transformer models can be more beneficial in the cases. Another reason is that with larger overlaps, neighboring tokens can capture more transitioning information and help the transformer model to better capture the temporal dynamics.

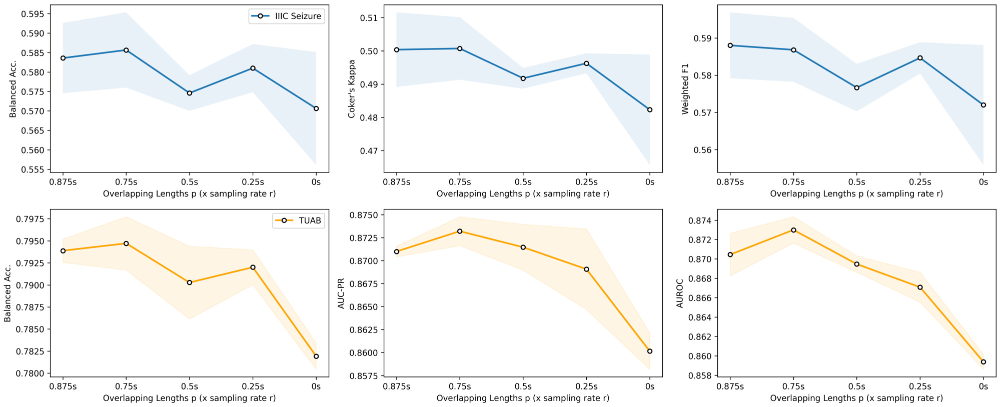

Figure 9: Ablation Study on Overlapping Lengths p Between Tokens

C.4 Running Time Comparison

This section compares the running time of all models on all supervised learning tasks (CHB-MIT, TUAB, IIIC Seizure, TUEV, PTB-XL, HAR). When recording the running time, we duplicated the environment mentioned in Section 3.2, stopped other programs, and ran all the models one by one on one GPU. We record the first 10 epochs of all models and report the per epoch mean and standard deviation of the time cost in Table 7. In the setting, we use 512 as the batch size for CHB-MIT, TUAB, TUEV, and PTB-XL, 256 as batch size for IIIC Seizure, and 64 as batch size for HAR.

The results show that our model has a similar running time profile as CNN-Transformer and the ST-Transformer models. These two baselines and our `BIOT` are all transformer based models and have the same number of heads and layers. `BIOT` uses linear self-attention module which should be faster than CNN-Transformer and the ST-Transformer, however, our sequence structure are also generally longer due to the flattening transformation. We also find that the CNN models (SPaRCNet and ContraWR) are faster than the transformer based models in our experiments.

In all, the running time of all models are in the same magnitude, and the actual running time on certain applications can vary due to the signal length, number of channels, selected number of CNN or transformer layers or attention heads. The reported running time comparisons here work as references using our selected model architectures. We think the running time of `BIOT` is acceptable given its decent performance.

Table 7: Running time comparison (seconds per epoch)

Model	CHB-MIT	TUAB	IIIC Seizure	TUEV	HAR	PTB-XL
SPaRCNet	32.4417 ± 0.9952	17.4635 ± 1.0861	24.8728 ± 2.9339	7.1237 ± 1.8801	1.9485 ± 0.0926	7.3850 ± 0.0599
ContraWR	24.7308 ± 0.7905	13.4650 ± 0.4554	14.9449 ± 0.9323	5.3337 ± 0.2003	2.3235 ± 0.1308	5.1978 ± 0.0366
CNN-Transforemr	53.3355 ± 1.8880	25.8983 ± 0.7969	25.7742 ± 1.2990	9.0415 ± 0.2501	3.0207 ± 0.0463	7.9851 ± 0.0400
FFCL	43.6103 ± 0.7927	21.6868 ± 0.5668	23.7682 ± 1.0757	7.0863 ± 0.1333	2.6649 ± 0.0973	6.5135 ± 0.0228
ST-Transformer	50.4725 ± 0.9978	24.5641 ± 0.4770	26.3251 ± 3.2320	7.9027 ± 0.0758	2.7954 ± 0.0698	11.1495 ± 0.0200
(Vanilla) BIOT	55.5780 ± 0.4229	25.2788 ± 0.1200	25.4500 ± 4.0835	8.1812 ± 0.1228	2.9560 ± 0.0563	12.1791 ± 0.0452

C.5 Discussion on handling long recordings and multiple channels

Currently, the BIOT model already have two designs to handle long recordings and more channels: (i) the current model uses linear complexity transformer, so that the complexity scales linearly with sample lengths and channel sizes; (ii) we can remove the token overlaps and enlarge the token sizes to reduce the token numbers (i.e., "sentence" length). With minor adjustments, the BIOT model can better handle long recordings and multiple channels. For long recordings, we could segment the recordings into 10-30s samples and then apply our BIOT on each sample and finally use a top level LSTM or Transformer to learn sequence embedding. For multiple channels (more than 256 channels), we can group neighboring channels or symmetric channels and tokenize them together, which could greatly shrink the final "sentence" length. However, additional adjustments are not needed in this paper and further discussion is beyond the scope of this paper.