FREQUENCY-AWARE MASKED AUTOENCODERS FOR MULTIMODAL PRETRAINING ON BIOSIGNALS

Ran Liu[1,2*], **Ellen L. Zippi**[1], **Hadi Pouransari**[1], **Chris Sandino**[1], **Jingping Nie**[1,3*]
Hanlin Goh[1], **Erdrin Azemi**[1], **Ali Moin**[1]
Apple[1], Georgia Institute of Technology[2], Columbia University[3]

ABSTRACT

Leveraging multimodal information from biosignals is vital for building a comprehensive representation of people's physical and mental states. However, multimodal biosignals often exhibit substantial distributional shifts between pretraining and inference datasets, stemming from changes in task specification or variations in modality compositions. To achieve effective pretraining in the presence of potential distributional shifts, we propose a frequency-aware masked autoencoder (`bio`FAME) that learns to parameterize the representation of biosignals in the frequency space. `bio`FAME incorporates a frequency-aware transformer, which leverages a fixed-size Fourier-based operator for global token mixing, independent of the length and sampling rate of inputs. To maintain the frequency components within each input channel, we further employ a frequency-maintain pretraining strategy that performs masked autoencoding in the latent space. The resulting architecture effectively utilizes multimodal information during pretraining, and can be seamlessly adapted to diverse tasks and modalities at test time, regardless of input size and order. We evaluated our approach on a diverse set of transfer experiments on unimodal time series, achieving an average of $\uparrow 5.5\%$ improvement in classification accuracy over the previous state-of-the-art.

1 INTRODUCTION

Physical and mental states of an individual are manifested by a variety of physiological responses or *biosignals*. For example, electroencephalography (EEG) can decode human emotions by monitoring their brain activities (Liu et al., 2010), electromyography (EMG) can detect facial expressions such as smiling by recording muscle contractions (Canento et al., 2011), and a combination of these modalities can help decode a person's affective states. The effective use of multimodal information can not only build better and more resilient representations of the human body and mental states (Bachmann et al., 2022; Smith & Gasser, 2005; De Sa & Ballard, 1998), but also help researchers understand how each biosignal contributes to each physiological state and how their information overlaps (Bird et al., 2020).

Recently, in language-vision domains, large-scale multimodal pretraining has demonstrated remarkable generalization and zero-shot capabilities (Huang et al., 2021; Bachmann et al., 2022; Radford et al., 2021), outperforming small-scale models that are trained on specific downstream tasks (Kirkpatrick et al., 2017; Radford et al., 2019). In light of these advancements, we investigate whether similar pretraining can be applied to the biosignal domain.

In this work, we propose to incorporate frequency information in time series to enable multimodal pretraining on biosignals, where we use frequency domain information to help the encoder to address the distributional shift issues. We propose a simple, yet effective, multi-head frequency filter layer with fixed-size Fourier-based operator that directly parameterizes the representation of biosignals in the frequency space. The proposed layer can be easily incorporated into the transformer, giving a *frequency-aware (FA) encoder* that is both expressive and computationally efficient. Furthermore, to extend the frequency awareness into a multimodal pretraining setting, we couple the FA encoder with a *frequency-maintain (FM) pretraining strategy*. Combining the two techniques, our proposed

*Work completed during internship at Apple. Contact: rliu361@gatech.edu, amoin@apple.com.

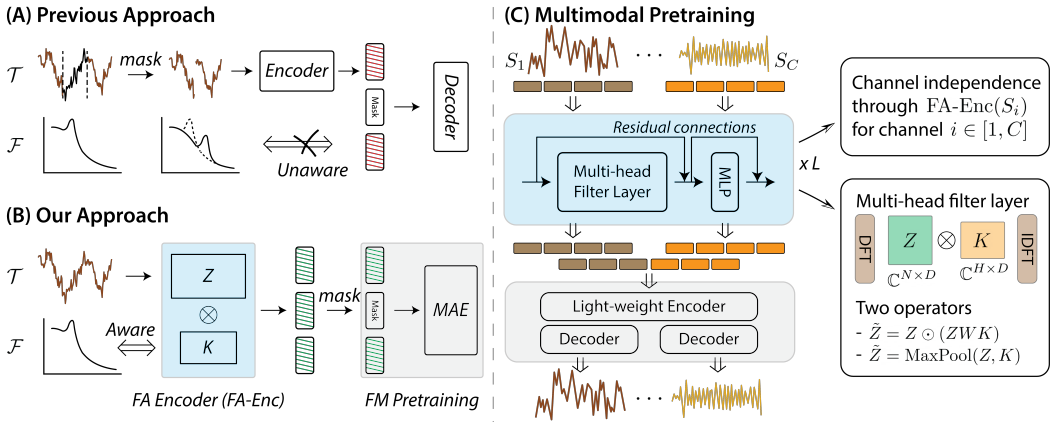

Figure 1: *Overview*. **(A)** Previous approaches perform masking in the time domain, which causes shifts in the frequency components. Also, the encoders are unaware of the frequency information in time series. **(B)** To address the issues, we propose bioFAME, which (i) builds frequency awareness by directly learning frequency filters in the representation space, and (ii) performs masked autoencoding in the latent space to maintain frequency information during pretraining. **(C)** We implement bioFAME in the multimodal pretraining scheme, where the frequency-aware encoder (FA-Enc$(\cdot)$) processes signals in a channel-independent manner, and extracts representations with multi-head filter layer with fixed-size Fourier operators. The frequency-maintain pretraining strategy further performs masked autoencoding in the latent space with separate reconstruction to guide the effective mixing of multimodal information.

approach bioFAME is systematically evaluated on a publicly available one-to-many transfer learning benchmark (Zhang et al., 2022), giving an average of $\uparrow 5.5\%$ improvements in classification accuracy over the previous state-of-the-art, showing consistency across datasets of different input lengths, sampling rates, and diverse sources of modalities.

2 Method

Preliminaries: Discrete Fourier Transform (DFT) for Token Mixing Consider a sequence $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times D}$ of N tokens of D-dimensions, transformers aim to learn the interactions across tokens, typically through the self attention operation. Recently, mixing tokens with frequency-based operations through DFT and IDFT is shown to be a computationally efficient alternative (Rao et al., 2021; Guibas et al., 2021), as it considers global-wise information mixing. The token mixing process is theoretically grounded by the Fourier Neural Operators (Li et al., 2020), which is often implemented in its discrete form (denote as $\mathcal{K}$) as such:

$$(\mathcal{K}(X))(\boldsymbol{x}_i) = \mathcal{F}^{-1}(R \cdot \mathcal{F}(X))(\boldsymbol{x}_i), \forall i \in [1, N] \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represents the DFT and IDFT processes, respectively. Ideally, R should be the Fourier transform of a periodic function which admits a Fourier series expansion. For the sake of simplicity, it is often implemented as learnable weights of shape $\mathbb{C}^{N \times D}$.

2.1 Frequency-aware transformer with multi-head frequency filters

Multi-head Frequency Filter Layer We propose to manipulate the frequency representation with a multi-head frequency filters $K \in \mathbb{C}^{H \times D}$, where H is the total number of heads. Given a sequence of tokens $X \in \mathbb{R}^{N \times D}$, we first perform DFT along the sequence dimension to obtain its representation in the frequency space as $Z \in \mathbb{C}^{N \times D}$. To obtain the manipulated features in frequency space $\tilde{Z} \in \mathbb{C}^{N \times D}$, we first compute queries $Q = ZW$, where $W \in \mathbb{R}^{D \times H}$ is a learnable matrix that is used to combine processed information across different filters. The resulting queries are used to re-weight the kernels to obtain $\tilde{Z}$ through the below operations:

$$\tilde{Z} = Z \odot (QK) = Z \odot (ZWK) \tag{2}$$

where $\odot$ is the Hadamard product. We show in Appendix C that the operation is equivalent to a weighted summation between each modulated frequency representation matrix, where the weights

are self-generated through the queries. In Appendix, we also show an alternative maxpooling operator $\tilde{Z} = \mathrm{MaxPool}(Z, K)$ with additional nonlinearity.

The resulting modulated frequency representation $\tilde{Z}$ is later recovered in time space through $\tilde{X} = \mathcal{F}^{-1}(\tilde{Z})$ with IDFT (see Figure 1(C)). We denote the whole process as $\text{Freq-L}(\cdot)$, which is computationally efficient, transferrable across different input lengths and sampling rates, and can be easily implemented in a few lines of code.

Add $\text{Freq-L}(\cdot)$ **into the Transformer** The transformer architecture has revolutionized many domains. Following Nie et al. (2022), we first patchify the biosignals by dividing them into chunks, compute representations for each patch, and then feed the resulting patches into a transformer. Specifically, for a signal $\mathbf{s} \in \mathbb{R}^L$ where L is the total length of the sequence, we divide them into sequences of $S = [\mathbf{s}_1, ...\mathbf{s}_N]$, where each patch $\mathbf{s}_i \in \mathbb{R}^P$ has a size of P. An initial MLP is used to compute representation $\mathbf{x}_i = \mathrm{MLP}(\mathbf{s}_i) \in \mathbb{R}^D$, and the sequence is later stacked into $X_0 \in \mathbb{R}^{N \times D}$.

We replace the multi-head self-attention with our proposed multi-head frequency filter layer $\text{Freq-L}(\cdot)$ to mix the information across the sequence of tokens, which gives the FA transformer encoder layer as below:

$$X_{\ell+1} = X_\ell + \text{Freq-L}\left(X_\ell\right) + \text{FF}\left(X_\ell + \text{Freq-L}\left(X_\ell\right)\right), \ell = \{0, \ldots, L-1\} \tag{3}$$

where the representation is passed into the proposed $\text{Freq-L}(\cdot)$ layer and projection layers $\text{FF}(\cdot)$ with residual connections, as shown in Figure 1(C).

2.2 Frequency-maintain pretraining with latent masking

Masked Autoencoding in the Latent Space Masked autoencoder (MAE) is a self-supervised pretraining framework, which masks out input patches and predicts the missing patches using the rest present patches. The architecture typically contains an transformer encoder that processes non-masked patches, follows by a decoder, usually a lightweight transformer, that reconstructs the original patches (He et al., 2022).

To preserve the frequency information while being able to perform pretraining based on the masked autoencoding strategy, we perform *masked autoencoding in the latent space*. Specifically, denote our frequency-aware transformer encoder as $\text{FA-Enc}(\cdot)$, full sequence of biosignals S is learnt through $\text{FA-Enc}(\cdot)$ to obtain $X_L = [\boldsymbol{x}_1^L, \boldsymbol{x}_2^L, ..., \boldsymbol{x}_N^L]$. We sample a random set of patches based on a fixed masking ratio without replacement, and then process the resulting sequence with a lightweight transformer (second) encoder. We later pad the masked patches with mask tokens, and pass the resulting sequence into a lightweight transformer decoder to reconstruct the original signal, where the i-th reconstructed patch corresponds to $\boldsymbol{s}_i$. Denote the masked autoencoder as $\mathrm{MAE}(\cdot)$, `bioFAME` aims to optimize the below objective:

$$\mathcal{L} = \frac{1}{\Omega} \sum_{i \in \Omega} l(\boldsymbol{s}_i, \mathrm{MAE}(\text{FA-Enc}(S))[i]) \tag{4}$$

where i is the token index, Ω is the set of masked tokens, and l is an error term which is set as mean squared error (MSE) in this work. We show in Section 3 that the performance is robust if we remove $\mathrm{MAE}(\cdot)$ and only keep $\text{FA-Enc}(\cdot)$ at test time. We note that this is the first work that finds using the masked autoencoding objective itselfis effective on biosignals (Zhang et al., 2022).

3 Experiments

Datasets and Experimental Details We evaluate the model's generalization ability by transferring it on a diverse set of unimodal time series downstream tasks, following Zhang et al. (2022). The transfer experiments include a set of four downstream tasks: **Epilepsy** (Andrzejak et al., 2001); **SleepEOG** (Kemp et al., 2000); **ExpEMG** (Goldberger et al., 2000); **FD-B** (Lessmeier et al., 2016). For model pretraining, we used the SleepEDF dataset (Kemp et al., 2000) as in (Eldele et al., 2021; Zhang et al., 2022), where the single-channel EEG (channel Fpz-Cz) is commonly used for unimodal pretraining. In this work, we also used an additional EEG channel (Pz-Oz) and an additional modality (EOG) from SleepEDF to perform multimodal pretraining with the same train/test split.

I. Generalization with modality or task association.

Models	Epilepsy (EEG)				SleepEOG			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
TS-TCC (Eldele et al., 2021)	92.53	94.51	81.81	86.33	69.65	61.56	61.49	61.16
TF-C (Zhang et al., 2022)	94.95	**94.56**	89.08	91.49	69.58	62.04	68.05	64.15
PatchTST (Nie et al., 2022)	95.01	91.66	**92.96**	92.27	68.00	61.20	**68.28**	63.26
bioFAME (scratch)	90.41	84.64	86.29	85.33	68.29	60.03	66.10	61.81
bioFAME (unimodal)	**95.51**	**94.02**	91.57	**92.72**	**70.03**	**63.37**	68.00	**65.05**
bioFAME (multimodal)	**95.71**	93.57	**92.82**	**93.18**	**71.55**	**64.80**	**68.70**	**66.62**
Δ(uni, multi)	↑0.20	↓0.45	↑1.25	↑0.46	↑1.52	↑1.43	↑0.70	↑1.57

II. Generalization without explicit association.

Models	ExpEMG				FD-B (Electromechanics)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
TS-TCC (Eldele et al., 2021)	78.89	58.51	63.10	59.04	54.99	52.79	63.96	54.18
TF-C (Zhang et al., 2022)	81.71	72.65	81.59	76.83	69.38	75.59	72.02	74.87
PatchTST (Nie et al., 2022)	92.68	90.87	94.51	92.07	67.03	71.96	75.57	70.09
bioFAME (scratch)	93.17	88.58	94.10	89.97	67.92	76.45	76.51	76.20
bioFAME (unimodal)	**98.05**	**97.07**	**96.63**	**96.40**	**76.58**	**83.28**	**82.85**	**82.63**
bioFAME (multimodal)	**98.54**	**96.67**	**98.95**	**97.64**	**78.18**	**84.99**	**84.01**	**83.75**
Δ(uni, multi)	↑0.49	↓0.40	↑2.32	↑1.24	↑1.60	↑1.71	↑1.16	↑1.12

Table 1: *Transfer experiments on unimodal time series*. All benchmark models are pretrained on the same single-lead EEG. All variants of our model is based on the same architecture, where bioFAME (scratch) is trained from scratch, bioFAME (unimodal) follows the same pretraining as baselines, and bioFAME (multimodal) is pretrained on the multimodal version of the data. Model standard deviation are in Appendix A.5.

For bioFAME, we used a 4-layer encoder, 8-head filter with 64 dimensions. The model was trained using an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a learning rate of 0.001. We repeated experiments with five random seeds for major results, and three random seeds for ablations.

Unimodal Pretraining Achieves SOTA Following previous works Zhang et al. (2022), we first performed pretraining on a single-channel EEG from the SleepEDF dataset, and then fine-tuning on a small amount of data from the downstream tasks. The performance of our proposed architecture is shown in Table 1. We show that with the same unimodal pretraining setup on single-channel EEG, our model consistently outperforms state-of-the-art benchmarks in most experiments, giving $\uparrow 4.2\%$ improvments in accuracy. These results demonstrate that bioFAME is effective in terms of transfer on different tasks, with robustness to domain shifts across tasks, subjects, sampling rate, and sensors. Surprisingly, our architecture, without any pretraining (scratch), also provides robust performance on many datasets, different from previously reported results (Zhang et al., 2022).

Multimodal Pretraining Further Improve Performance While the Fpz-Cz EEG channel is shown to be the most informative channel for the pretraining task and typically provides robust prediction performance on its own (Supratak et al., 2017), in this work, we explore whether using additional multimodal information from the same task can further boost the pretraining performance. As shown in Table 1, for bioFAME, including multimodal information during pretraining provides better results than unimodal pretraining in general, consistently outperforming unimodal pretraining. Training on multimodal data also improves the model's stability by giving a lower standard deviation, as shown in Appendix B.4. Note that in previous work (Zhang et al., 2022), including multimodal information hurt performance rather than helped. This suggests that bioFAME can effectively utilize and combine information across modalities, resulting in better performance on downstream tasks. We hypothesize that pretraining on multiple modalities exposes the model to a more diverse range of frequency components, improving the model's few-shot generalization.

Robustness for Modality Mismatch Scenarios We consider two modality mismatch scenarios as shown in Figure 2(A): (i) Modality substitution, where one modality is replaced by another modality; and (ii) Modality dropout, where only a subset of modalities is present at test time. We show the model's performance with modality substitution in Figure 2(B), where the model is pretrained with { EEG Fpz-Cz; EOG; EMG }. Each of the pretraining modality is replaced with another channel to examine the performance degradation (more details in Appendix B.3). Our model gives

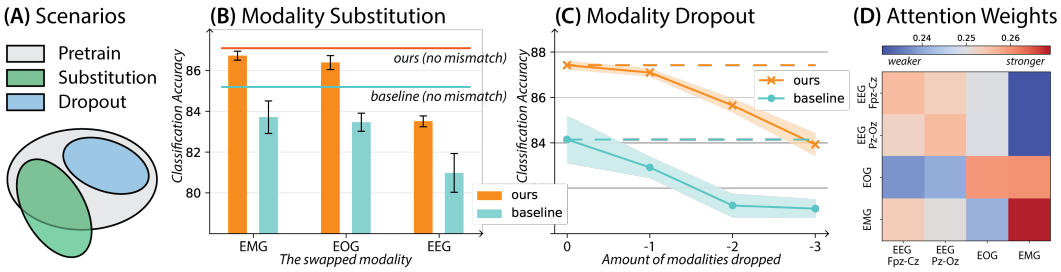

Figure 2: *Multimodal evaluation results*. **(A)** Two modality mismatch scenarios are considered: Modality substitution and modality dropout. **(B)** When a modality is swapped with another available one, or **(C)** when modalities are dropped out at test time, our model gives lower performance degradation when comparing to a robust baseline. **(D)** By visualizing the attention weights across modalities, we can understand how modalities are associated with each other.

better performance than the robust baseline PatchTST (Nie et al., 2022), exhibiting less performance degradation. In terms of modality dropout, we pretrained the model with { EEG Fpz-Cz; EEG Pz-Oz; EOG; EMG }, and we dropped an increasing amount of modalities till there is only one modality left (see Figure 2(C)). We see that `bio`FAME is more resistant to unexpected modalities dropout in comparison to the baseline. Unlike many other baselines that contain spatial layers, `bio`FAME can be applied at test time even when there are unexpected amount of channels while exhibiting resilience towards modality mismatch scenarios. This study further demonstrated that `bio`FAME presents a robust model when used in real-world scenarios.

Visualizing the Connections Across Modalities To understand how the information across different channels affects each other, we visualized the averaged attention matrix to examine the relationship across modalities. As shown in Figure 2(D), for each channel (row), the intensity of its attention or connection to the other channels can be visualized by the color (red means stronger connections). Interestingly, we notice that while each channel would rely on its own information the most, they tend to focus on the stronger modalities, which is the EEG Fpz-Cz channel in our case. Moreover, interesting asymmetry is observed for EOG-EMG, as EOG correlates more to the EMG while the opposite does not hold. We hypothesize that this is because facial movement would produce moving artifacts for EOG on the temple, while the opposite connection does not hold. This observation demonstrates that `bio`FAME can be used by researchers to further understand the information overlap across modalities (Bird et al., 2020).

4 Conclusion

In this work, we proposed a frequency-aware masked autoencoder that performs pretraining on multimodal biosignals. Our proposed method leverages a frequency-aware encoder with fixed-size Fourier-based operator to extract representation on biosignals, and uses a frequency-maintain pretraining module to perform pretraining. Our empirical experiments show that our model achieves state-of-the-art performance on a set of transfer experiments, where the models, both pretrained on unimodality and multimodality, can be adapted to effectively classify time series with varying input lengths, sensors, and sampling rates.

References

Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pp. 348–367. Springer, 2022.

Jordan J Bird, Jhonatan Kobylarz, Diego R Faria, Anikó Ekárt, and Eduardo P Ribeiro. Cross-domain mlp and cnn transfer learning for biological signal processing: Eeg and emg. *IEEE Access*, 8:54789–54801, 2020.

Filipe Canento, Ana Fred, Hugo Silva, Hugo Gamboa, and André Lourenço. Multimodal biosignal sensor data handling for emotion recognition. In *SENSORS, 2011 IEEE*, pp. 647–650. IEEE, 2011.

Virginia R De Sa and Dana H Ballard. Category learning through multimodality sensing. *Neural Computation*, 10(5):1097–1117, 1998.

Thomas Donoghue, Matar Haller, Erik J Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H Lara, Joni D Wallis, Robert T Knight, et al. Parameterizing neural power spectra into periodic and aperiodic components. *Nature neuroscience*, 23(12):1655–1665, 2020.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

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

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

FN Hooge, TGM Kleinpenning, and Lode KJ Vandamme. Experimental studies on 1/f noise. *Reports on progress in Physics*, 44(5):479, 1981.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM Society European Conference*, volume 3, 2016.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based human emotion recognition and visualization. In *2010 international conference on cyberworlds*, pp. 262–269. IEEE, 2010.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *arXiv preprint arXiv:2206.08496*, 2022.

APPENDIX

A ADDITIONAL RESULTS

A.1 MULTI-MODAL EVALUATIONS AND VISUALIZATIONS

Datasets and Experimental Details We investigated how well the model performs when applied to real-world cases in which multimodal information is available at test time. To understand this, we systematically studied different combinations of the EEG Fpz-Cz, EEG Pz-Oz, EOG, EMG, and the respiration channels of the SleepEDF dataset (Kemp et al., 2000), which are simultaneously recorded. We followed the same train/val/test split as in Eldele et al. (2021) while attaching the multimodal information instead of using only the unimodal information. We utilized the same model setup, aside from that we follow Section 2.2 to expand the training and testing under multimodal designs with weight sharing and channel independence. We also implemented two variants of multimodal latent expansion methods as in Appendix C.

A.2 ABLATIONS EXPERIMENTS ON TRANSFERABILITY

We performed a set of ablation experiments to understand what makes `bio`FAME robust under the transfer experiments setting. In Table 2, we first studied the effect of the frequency-aware (FA) and frequency-maintain (FM) modules by either replacing the FA module with a self-attention transformer; or by replacing the FM module with a normal masking procedure. We found both approaches, when applied independently, improve the performance of a baseline variant by a significant margin ($\approx 3\%$). Combining both modules gives the best performance, further boosting the effect of each individual component ($\approx 5\%$). We also tested whether it is possible to discard the second encoder at test time, which would indicate whether or not the FA encoder plays a major role in learning. Interestingly, we show that discarding the second encoder at test time gives almost identical performance in the unimodal setting. However, when multimodal information is used for pretraining, discarding the second encoder would give a performance that is lower than the unimodal result, while keeping the second encoder increases the unimodal performance by $\approx 1\%$ instead (see Table 3). We hypothesize that it is beneficial to retain the second encoder at test time under the multimodal setting because it is responsible for merging the information present across the multimodal data. Finally, in Table 4, we investigate how different patch sizes and masking ratios affect the performance of our model. We show that `bio`FAME gives stable performance when the patch size is relatively small, giving robust performance under a range of masking ratios.

FA	FM	Acc.
✗	✗	80.68
✓	✗	84.09
✗	✓	83.53
✓	✓	**85.04**

Table 2: Average accuracy without FA/FM modules.

Enc-2	Modality	Acc.
✗	Uni	85.04
	Multi	83.92
✓	Uni	85.05
	Multi	**85.99**

Table 3: The effect of keeping the 2nd encoder for multimodal pretraining.

		Masking ratio		
		0.3	0.5	0.7
Patch	10	83.86	84.05	82.70
	20	84.11	**85.04**	83.86
	50	80.88	80.84	80.64

Table 4: The effect of different masking ratios and patch sizes.

A.3 PARAMETER EFFICIENCY AND ADDITIONAL ABLATIONS

Parameter efficiency To understand the parameter efficiency and the throughput of our approach, we compute the parameters and FLOPs between baselines and our approach in Table 5.

	TS2vec	TFC	TS-TCC	PatchTST	**Ours**
Params	632K	1.18M	140K	612K	243K
FLOPs	0.69B	1.38B	1.95B	35.0B	9.42B

Table 5: Comparison of parameters and FLOPs between baselines and our approach. The FLOPs are computed over a batch of SleepEDF data with batch size 64.

We can see that, bioFAME is very parameter-efficient due to its fix-size frequency filter design. With the same depth (4), heads (8), and dimensionality (64), bioFAME contains only $\approx 40\%$ parameters of the transformer baseline PatchTST. The parameter size of bioFAME also stands competitive with many CNN-based architectures. The FLOPs of bioFAME are significantly lower than the transformer baseline PatchTST ($<30\%$); yet greater than CNN-based architectures.

Additional ablations To understand the models' sensitivity towards different hyperparameters and understand if bioFAME can provide better performance with increased complexity, we conducted additional ablation experiments in Table 6 and Table 7.

dim	32	64	128	256
ExpEMG	91.1	98.05	96.48	97.78
FD-B	76.74	76.58	78.14	80.87
Avg.	83.92	87.32	87.31	89.33

Table 6: Performance of our approach with different latent dimensionality.

depth	3	4	5	6
ExpEMG	77.54	76.58	76.79	78.99
FD-B	97.78	98.05	95.55	92.59
Avg.	87.66	87.32	86.17	85.79

Table 7: Performance of our approach with different encoder depth.

We observed that increasing the latent dimensionality could further improve the performance of our approach; while increasing the network depth gives no performance gains.

A.4 Data efficiency and operator selection

Data efficiency To understand the behavior of bioFAME within the context of limited data availability, we conducted experiments aimed at gauging the architecture's efficacy when exposed to a reduced amount of labeled data during the finetuning phase. We show the performance of bioFAME in Figure 3(A), both with and without pretraining, where the performance of bioFAME is plotted when the amount of labeled data for downstream training varies from 5% to 100%. Notably, in contrast to previous work (Eldele et al., 2021), wherein architecture performance substantially deteriorated with decreased labeled data, bioFAME achieves stable results with relatively low decay of performance even without pretraining. Furthermore, the pretrained version of bioFAME gives consistently robust performance across the spectrum of labeled data proportions. We hypothesize that modeling biosignals using the Fourier function group with frequency operators improves the data efficiency of models.

Ablations on the two operators To validate the effectiveness of the Maxpool operator and the Query operator as described in Section 2.1, we examine the model's performance by varying the number of filters. We find that the Maxpool operator gives more stable results, while the Query operator seems to scale better to larger amount of filters.

A.5 Model variation

For the transfer experiments result as shown in Table 1, we provide the standard variation across five different random seeds in Table 8. Note that the entire training process, both the pretraining and the finetuning stages, is repeated to obtain the standard variation for fair evaluation. We notice that multimodal pretraining typically gives a lower standard deviation than that of unimodal pretraining, demonstrating that multimodal pretraining might help with the stability of the model, as it is exposed to a variety of frequency components.

While we believe that our diverse experiments across many datasets demonstrate the robustness of our approach under randomness, we believe that another important source of randomness comes from the data split, which is fixed in this work.

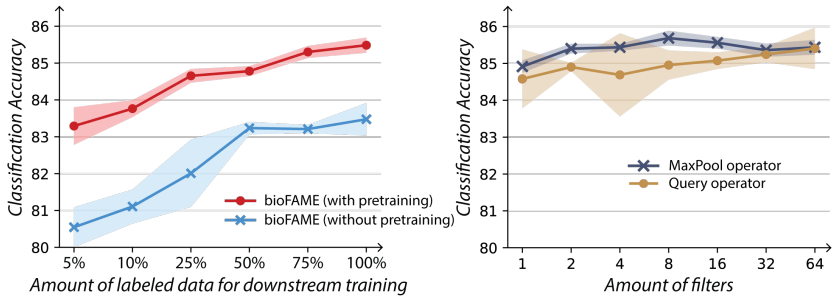

Figure 3: **(A)** We examine the performance of bioFAME under low-data regime with and without pretraining. **(B)** We examine how the MaxPool operator and Query operator would perform with different amounts of filters.

Models	Epilepsy (EEG) Accuracy	Precision	Recall	F1	SleepEOG Accuracy	Precision	Recall	F1
bioFAME (scratch)	1.17	2.42	0.72	1.26	0.77	0.67	0.50	0.76
bioFAME (unimodal)	0.35	0.37	1.17	0.65	1.39	1.23	0.91	0.61
bioFAME (multimodal)	0.17	0.51	0.21	0.24	0.90	0.79	0.89	0.88
Δ(uni, multi)	↓0.18	↑0.14	↓0.96	↓0.41	↓0.49	↓0.44	↓0.02	↑0.27

Models	ExpEMG Accuracy	Precision	Recall	F1	FD-B (Electromechanics) Accuracy	Precision	Recall	F1
bioFAME (scratch)	2.67	3.13	2.25	3.15	1.63	1.33	1.20	1.09
bioFAME (unimodal)	2.04	2.80	5.64	4.15	2.74	1.75	2.01	2.14
bioFAME (multimodal)	1.34	3.04	0.96	2.15	1.94	1.53	1.44	1.66
Δ(uni, multi)	↓0.70	↑0.24	↓4.68	↓2.00	↓0.80	↓0.22	↓0.57	↓0.48

Table 8: The standard deviation of bioFAME for each transfer experiment.

A.6 Ablation results breakdown

In Table 9, we report the breakdown details for the average accuracy presented in Table 2 and Table 3. Our model provides robust performance across different downstream tasks consistently.

	Ablations	Epilepsy	SleepEOG	ExpEMG	FD-B
Table 2	FA✗ FM✗	95.01	68.00	92.68	67.03
	FA✓ FM✗	95.03	69.73	98.37	73.23
	FA✗ FM✓	94.81	68.41	95.94	74.97
	FA✓ FM✓	95.51	70.03	98.05	76.58
Table 3	Uni, Enc-2✗	95.91	70.17	95.94	78.16
	Multi, Enc-2✗	95.26	71.04	96.10	73.28
	Uni, Enc-2✓	95.51	70.03	98.05	76.58
	Multi, Enc-2✓	95.71	71.55	98.54	78.18

Table 9: Breakdown of model performance on different downstream tasks.

B Experimental details

B.1 Datasets details

We provide additional details about the datasets we used as follows.

SleepEDF The entire SleepEDF dataset contains 197 recordings of whole-night sleep, where the dataset contains 2-lead EEG, EOG, chin EMG, respiration rates, body temperature, and event markers. We selected a subset of the dataset from the Cassette Study following Eldele et al. (2021), where the dataset is used to study the age effects on sleep in healthy Caucasians. We further followed the

same train/validate/test split, and removed data with incomplete modalities. The recordings are segmented into 30 seconds of sleep for training, where each sample is associated with one of the five sleeping patterns/stages: Wake (W), Non-rapid eye movement (N1, N2, N3), and Rapid Eye Movement (REM).

Epilepsy The Epilepsy dataset contains single-lead EEG measurements from 500 subjects, where the brain activities are recorded for subjects with seizure. The classification task is based on if the subject is having a seizure episode during the recording session.

SleepEOG The SleepEOG dataset is a subset of the SleepEDF dataset under the Telemetry Study, where subjects are reported to have mild difficulty falling asleep, and thus intake either temazepam or placebo before sleep. The EOG channel is used for classification.

ExpEMG The ExpEMG dataset consists of single-channel EMG recordings from the tibialis anterior muscle of three healthy volunteers, where they (1) do not have history of neuromuscular disease; (2) suffer from chronic low back pain and neuropathy; and (3) suffer from myopathy due to longstanding history of polymyositis. The classification task aims to classify different conditions (subjects).

FD-B The FD-B dataset is an electromechanical dataset, where the motor currents and vibration signals of healthy or damaged motors are recorded. The classification task aims to detect different faulty conditions of the motors based on their behavior. We found that the motor movement follows a similar frequency assumption as biosignals (Hooge et al., 1981), and thus used this electromechanical dataset to provide additional validation of the transfer performance of our model.

Datasets	Train	Validate	Test	Sampling rate	Length
Epilepsy	60	20	11420	174	178
SleepEOG	1000	1000	37244	100	3000
ExpEMG	122	41	41	4000	1500
FD-B	60	21	13559	64000	5120

Table 10: Dataset split details for different downstream tasks.

We performed the transfer experiments based on the same settings as in Zhang et al. (2022), where we used the train/validate/test spilt as shown in Table 10 for downstream fine-tuning to demonstrate the few-shot generalization ability of the model across different signals.

B.2 Model training and transfer experiments details

For all experiments, we pretrain `bioFAME` for 200 epochs on the SleepEDF dataset using a batch size of 128 to obtain the weights of the model. During fine-tuning, we remove the lightweight second encoder that mixes information across modalities, and use the average token of the frequency-aware transformer encoder to perform the prediction for downsteam tasks. We fine-tune `bioFAME` for 80 epochs with a batch size of 64, using an Adam optimizer with a learning rate of 0.001 on all datasets to obtain the final results. We perform all transfer experiments under the same training setup for all downstream tasks without additional adjustment for each dataset. Note that we perform full-scale model finetuning instead of linear probing when performing the transfer experiments, because the former approach is shown to be more effective for transformers in previous works (He et al., 2022).

B.3 Multimodal setup details

The multimodal experiments are designed to tackle the challenge presented by modality mismatch scenarios, where discrepancies in biosignal recording setups between training and testing phases lead to distributional shifts. Due to the scarcity of comprehensive multimodal datasets encompassing simultaneous recording of diverse modalities of biosignals, we exclusively used the SleepEDF dataset due to its modality coverage.

We first empirically assessed the representation quality of each individual channel. Similar to the findings in Supratak et al. (2017), we found that the representation capacity of different channels

follows EEG Fpz-Cz > EEG Pz-Oz > EOG > EMG > resp. Building upon these insights, we performed the modality substitution and modality dropout experiments following the below pretraining and finetuning setup.

Training modalities	Testing modalities
EEG Fpz-Cz; EOG; EMG	EEG Fpz-Cz; EOG; resp EEG Fpz-Cz; EEG Pz-Oz; EMG EEG Pz-Oz; EOG; EMG

Table 11: Modality setup for modality substitution experiments.

Training modalities	Testing modalities
EEG Fpz-Cz; EEG Pz-Oz; EOG; EMG	EEG Fpz-Cz; EEG Pz-Oz; EOG EEG Fpz-Cz; EEG Pz-Oz EEG Fpz-Cz

Table 12: Modality setup for modality dropout experiments.

B.4 Hyperparameter searching details

For transfer experiments, we performed hyperparameter searching based on results on the Epilepsy dataset, and used the same parameter setting across all transfer experiments. Specifically, we performed a grid search of learning rate of [0.0001, 0.001, 0.01], transformer depth of [2, 3, 4, 5, 6], latent dimensionality of [16, 32, 64, 128], dropout rate of [0.2, 0.3, 0.4], operator type, and filter amount correspondingly. We followed the convention for transformers and selected the MLP dimension of 128 and head dimension of 16 for `bioFAME` and the baseline transformer. We selected the optimal patch size and masking ratio based on results in Table 4. We did not search for the optimal batch size, or investigate the effect of using different activation functions or normalization techniques. For multimodal experiments, we evaluate the model's performance on the pretraining dataset, and performed the evaluation on the finetuning modalities using the best model used in pretraining. For the multimodal experiments, we performed a smaller scale grid search for the latent dimensionality and transformer depth.

C Methodology details

C.1 Additional explanation of motivation

Biosignals are often analyzed in their frequency space, where they are either studied through predefined frequency regions or through aperiodic components which typically form a 1/f-like distribution (Donoghue et al., 2020). The significance of frequency information is well-documented due to its intricate interrelation with various facets of learning, aging, as well as diseases such as ADHD or seizures. Correspondingly, modeling approaches that rely on the manual extraction and preprocessing of spectrogram features have demonstrated robust empirical performance (Supratak et al., 2017). Building upon these insights, we hypothesize that modeling biosignals employing function groups within the frequency domain could benefit the learning process by enhancing model adaptability and data efficiency. We note that this hypothesis might be violated if the frequency components carry limited information in other formats of time series datasets.

C.2 Intuition for the multi-head frequency filter layer

We provide additional intuition for the design of our multi-head frequency filter layer by breaking down the computation for each individual filter. For each k-th filter $K[k]$ inside $K \in \mathbb{C}^{H \times D}$, given latent representation $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_N]^T \in \mathbb{C}^{N \times D}$, we compute $Z^{(k)} = [\boldsymbol{z}_1 \odot K[k], \boldsymbol{z}_2 \odot K[k], ..., \boldsymbol{z}_N \odot K[k]]^T$, where $\odot$ represents the Hadamard product between each representation and the learnable filter weights. In order to learn the combination between different filters, we define weights $\boldsymbol{w}$ that compute $\tilde{Z} = \sum_{k=1}^{H} w_k Z^{(k)}$.

To increase the expressiveness of the filtering operation, instead of learning a linear combination of different filters, we borrow intuition from the computation of self-attention to compute the queries

for the kernel weights $\boldsymbol{w}$ through $\boldsymbol{w} = \boldsymbol{z}W$, where $W \in \mathbb{C}^{D \times H}$. Thus, we have:

$$\begin{aligned} \tilde{Z}[i,j] &= \sum_{k=1}^{H}(\sum_{j=1}^{D} Z[i,j]W[j,k])Z[i,j]K[k,j] \\ &= Z[i,j]\sum_{k=1}^{H}(\sum_{j=1}^{D} Z[i,j]W[j,k])K[k,j] \end{aligned} \tag{5}$$

which gives $\tilde{Z} = Z \odot (ZWK)$. In our implementation, we use the real values of latents to learn the weights of the combiner though $\boldsymbol{w} = \boldsymbol{z}_{\text{real}}W$. Similarly, based on the same intuition of combining filtered matrices, we have the max pooling operation.

C.3 Model variants for combining multimodal representations

In transfer experiments, we use the average of tokens to extract the final representations for the downstream classification. However, when having multimodal information, fixing the dimensionality of the latent representation when many modalities are present might narrow down the information from each modality, which might cause information loss. Thus, in multimodal experiments, we first average the representations from each individual modality, and then concat the representations across modalities before performing the downstream classification.