Towards Robust Multimodal Physiological Foundation Models: Handling Arbitrary Missing Modalities

Wei-Bang Jiang[1,2]*, **Xi Fu**[1]*, **Yi Ding**[1]†, **Cuntai Guan**[1]†
[1]Nanyang Technological University [2]Shanghai Jiao Tong University
935963004@sjtu.edu.cn,FUXI0010@e.ntu.edu.sg,{ding.yi,ctguan}@ntu.edu.sg

Abstract

Multimodal physiological signals, such as EEG, ECG, EOG, and EMG, are crucial for healthcare and brain-computer interfaces. While existing methods rely on specialized architectures and dataset-specific fusion strategies, they struggle to learn universal representations that generalize across datasets and handle missing modalities at inference time. To address these issues, we propose PhysioOmni, a foundation model for multimodal physiological signal analysis that models both homogeneous and heterogeneous features to decouple multimodal signals and extract generic representations while maintaining compatibility with arbitrary missing modalities. PhysioOmni trains a decoupled multimodal tokenizer, enabling masked signal pre-training via modality-invariant and modality-specific objectives. To ensure adaptability to diverse and incomplete modality combinations, the pre-trained encoders undergo resilient fine-tuning with prototype alignment on downstream datasets. Extensive experiments on four downstream tasks, emotion recognition, sleep stage classification, motor prediction, and mental workload detection, demonstrate that PhysioOmni achieves state-of-the-art performance while maintaining strong robustness to missing modalities. Our code and model weights will be released.

1 Introduction

Multimodal physiological signals, such as electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), and electromyography (EMG), have garnered increasing interest in brain-computer interfaces (BCI) due to their ability to capture diverse physiological and cognitive states [29]. EEG reflects neural activity, ECG monitors cardiac rhythms, EOG tracks eye movements, and EMG records muscle activations, collectively offering a comprehensive representation of human physiological responses. These modalities have been widely applied across various domains, including cognitive load assessment [1], emotion recognition [20], motor imagery [5], and sleep stage classification [22], driving advances in both medical diagnostics and BCI. To enhance performance, numerous algorithms have been proposed to effectively integrate multiple modalities, leveraging their complementary information to improve accuracies across applications [32, 4, 9, 33, 17].

Over the past two years, several EEG foundation models, including LaBraM, have emerged, demonstrating the feasibility of using masked EEG models for pre-training on large-scale EEG datasets [18, 38, 39]. These models have notably improved performance and generalization, laying the groundwork for general unsupervised representation learning. However, a gap persists in the development of general pre-trained foundation models for multimodal physiological signals. Although some studies have explored the integration of multiple physiological signals to boost performance, they are often limited by two key factors: either using multiple modalities during training but relying on a single

*Equal contribution

†Corresponding authors

Preprint. Under review.

modality for testing, or being tailored to specific downstream tasks, which restricts their generalizability across diverse datasets. For instance, Fang *et al.* introduced multimodal foundation models designed for sleep stage classification using masked autoencoders [10], while Brant-X employed contrastive learning for multi-level alignment between EEG and EXG but relied solely on EEG for downstream tasks [44]. Developing a universal multimodal physiological foundation model capable of extracting semantic representations while handling arbitrary missing modalities presents several key challenges:

1) Decoupling homogeneous and heterogeneous features: Multimodal physiological signals encompass both shared and unique patterns. Effectively disentangling modality-invariant (homogeneous) and modality-specific (heterogeneous) features is crucial for robust multimodal learning.

2) Unified multimodal representation learning: Beyond feature decoupling, integrating physiological signals with diverse characteristics to derive effective and generalizable representations remains a fundamental challenge in multimodal representation learning.

3) Handling arbitrary missing modalities: While leveraging all available modalities during training, ensuring robust performance under incomplete modalities at inference time is highly challenging, requiring strategies that maximize adaptability while minimizing performance degradation.

With regard to above challenges, we propose PhysioOmni, a universal multimodal physiological foundation model pre-trained on diverse multimodal datasets, including EEG, ECG, EOG, and EMG signals. Our approach begins with training a decoupled multimodal tokenizer, where one shared codebook and four private codebooks disentangle multimodal embeddings into modality-invariant and modality-specific codes. These discrete codes serve as the foundation for masked signal modeling, enabling the encoders to learn universal representations across modalities. During fine-tuning, we introduce homogeneous representation mapping, which projects features from different modalities into a common space. To handle arbitrary missing modalities, we incorporate prototype alignment and modality-specific prediction, ensuring robust adaptation across different modality combinations. We comprehensively evaluate PhysioOmni on four popular BCI tasks: emotion recognition, sleep stage classification, motor prediction, and workload detection. Our approach achieves state-of-the-art (SOTA) performance across both unimodal and multimodal settings, underscoring the effectiveness of decoupled multimodal learning and resilient fine-tuning. Our key contributions are threefold:

1) Decoupled multimodal tokenizer: We design a multimodal tokenizer that disentangles modality-invariant and modality-specific features using a shared codebook for common patterns and private codebooks for unique characteristics, enhancing multimodal fusion.

2) Masked signal pre-training: We extend masked signal modeling to multimodal physiological signals, enabling the model to learn both generic and semantic representations. Leveraging the decoupled tokenizer, we design modality-invariant and modality-specific code prediction, ensuring the model captures both shared and unique characteristics of each signal type.

3) Resilient fine-tuning with prototype alignment: To handle arbitrary missing modalities, we introduce homogeneous representation mapping, prototype alignment, and modality-specific prediction, enabling the model to dynamically adapt to varying modality combinations during inference while maintaining robust performance despite incomplete input signals.

2 Related Work

2.1 Foundation Models for EEG

EEG foundation models, designed to handle arbitrary configurations while learning robust and versatile representations, have gained significant attention in recent years. BIOT [42] introduces a Biosignal Transformer that tokenizes channels into patches, enabling cross-data learning across heterogeneous biosignal formats (EEG and ECG). LaBraM [18] leverages vector-quantized neural spectrum prediction and masked EEG modeling to facilitate cross-dataset learning, significantly enhancing performance across diverse EEG tasks. EEGPT [38] pre-trains a 10-million-parameter Transformer using mask-based dual self-supervised learning and spatio-temporal representation alignment to improve EEG representation learning. CBraMod [39] employs a criss-cross Transformer to separately model spatial and temporal dependencies while incorporating an asymmetric conditional positional encoding scheme for greater adaptability.

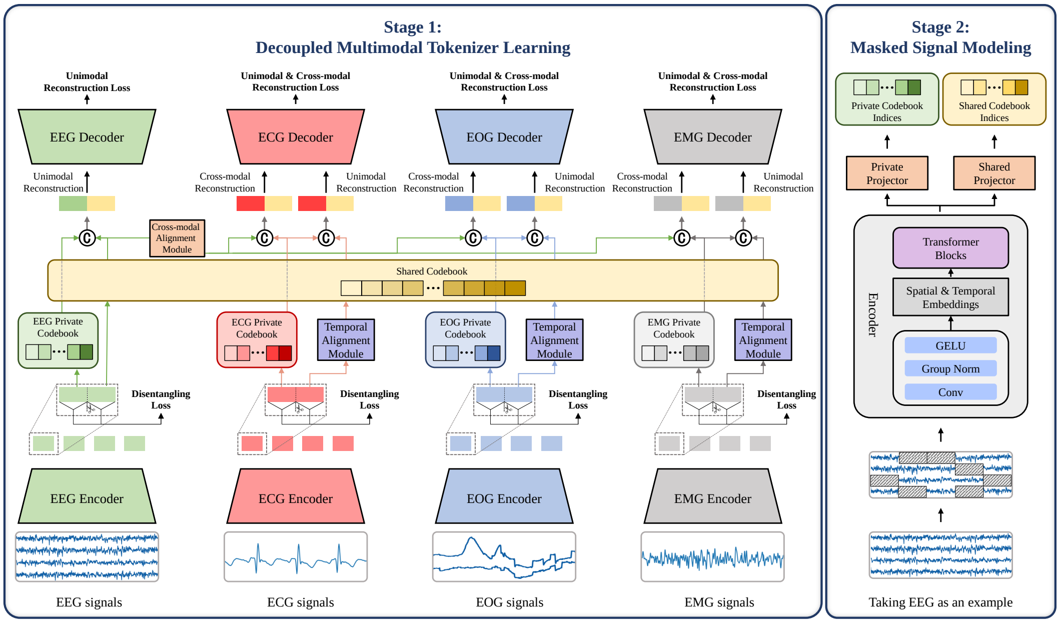

Figure 1: Overview of the decoupled multimodal tokenizer learning framework and masked signal modeling. **Left**: The tokenizer learning framework consists of a shared codebook, private codebooks, encoders, decoders, temporal alignment modules, and cross-modal alignment modules. **Right**: The masked signal modeling process includes two projectors for private and shared code prediction.

2.2 Multimodal Models in BCI

While many studies incorporate multimodal signals to enhance BCI performance, a universal foundation model that generalizes across modalities and datasets while extracting generic representations remains lacking. VBH-GNN [26] utilizes multimodal physiological signals and variational Bayesian heterogeneous graphs for cross-subject emotion recognition. Brant-X [44] aligns physiological signals by leveraging an EEG foundation model for knowledge transfer but is limited to EEG-based downstream tasks. CIMSleepNet [31] addresses arbitrary modality missing in sleep staging through a modal imagination module and contrastive learning, alongside temporal attention for better context representation. However, these models either rely on multimodal inputs during training but evaluate on a single modality or are tailored to specific tasks, limiting their generalizability across datasets.

3 Method

In essence, PhysioOmni consists of three training stages: 1) Joint learning of a shared codebook and four modality-specific private codebooks through decoupled multimodal tokenizer training; 2) Masked signal modeling to learn generic representations by predicting private and shared codes from masked inputs; 3) Resilient fine-tuning of pre-trained encoders with self-alignment to handle arbitrary missing modalities in downstream tasks. Figure 5 shows the first two stages.

Given paired multimodal samples $X = \{(x_i^e, x_i^c, x_i^o, x_i^m)\}_{i=1}^{N}$, where $x_i^j \in \mathbb{R}^{C_j \times T_j}$, with C_j and T_j representing the number of channels and time points, we segment the signals into patches $x_i^j \in \mathbb{R}^{N_j \times P_j}$ before passing them into the networks, where $N_j = \frac{C_j T_j}{P_j}$ represents the number of patches and P_j is the patch size for modality j. We denote the sampling rate for modality j as S_j.

Model Architecture. We adopt the encoder architecture of LaBraM [18] for all modalities due to its effectiveness and simplicity. A lightweight temporal encoder, consisting of several 1-D convolutional layers followed by GroupNorm [40] and GELU [15] activation, is used to extract temporal features within each patch. To incorporate temporal and channel-specific information for each modality, learnable temporal and and spatial embeddings are added to the extracted features before feeding them into the Transformer blocks. Unlike LaBraM, we incorporate RMSNorm [43] and SwiGLU

[30] into the Transformer blocks, as done in LLaMA [36]. The same backbone is used for both multimodal tokenizer training and masked signal modeling.

3.1 Decoupled Multimodal Tokenizer Training

At this stage, we aim to extract decoupled, compact, and semantically meaningful representations from multimodal physiological signals to facilitate subsequent masked signal modeling in pre-training. To achieve this, we first define four modality-specific encoders $\mathcal{E}^e, \mathcal{E}^c, \mathcal{E}^o, \mathcal{E}^m$ for EEG, ECG, EOG, and EMG, respectively, to extract both private and shared latent embeddings from the input signals:

$$z_i^{ep}, z_i^{es} = \mathcal{E}^e(x_i^e) \quad z_i^{cp}, z_i^{cs} = \mathcal{E}^c(x_i^c) \quad z_i^{op}, z_i^{os} = \mathcal{E}^o(x_i^o) \quad z_i^{mp}, z_i^{ms} = \mathcal{E}^m(x_i^m), \tag{1}$$

where z_i^{jp} and z_i^{js} represent the private and shared embeddings for modality j, obtained by splitting the encoder outputs. Unlike prior studies [13, 25, 41] that rely on separate encoders for decoupled embeddings, we employ a single encoder per modality to reduce computational overhead.

Codebook Optimization. We design five learnable codebooks: one shared codebook ($\mathcal{V}^s \in \mathbb{R}^{K \times D}$) and four private codebooks ($\mathcal{V}^j \in \mathbb{R}^{K \times D}$ for modality j), where K is the codebook size and D is the code dimension. These codebooks extract both modality-invariant and modality-specific features from multiple modalities. Notably, the shared codebook operates at the largest temporal scale (EEG). The shared and private embeddings from the encoders retrieve the closest codebook entry by looking up their respective codebooks and replacing themselves with the nearest code:

$$\hat{z}_i^{jp} = \mathcal{V}^j[\arg\min_k \|\ell_2(z_i^{jp}) - \ell_2(v_k^j)\|_2] \quad \hat{z}_i^{js} = \mathcal{V}^s[\arg\min_k \|\ell_2(\text{TA}(z_i^{js})) - \ell_2(v_k^s)\|_2], \tag{2}$$

where $v_i \in \mathcal{V}$ is the code and $k \in [1, K]$. To enhance stability, we apply ℓ_2 normalization to the embeddings, which is equivalent to selecting codes based on cosine similarity. Since the temporal scale (patch size) differs across modalities, the Temporal Alignment module (TA) is crucial for aligning smaller-scale modalities with EEG. It is implemented by a cross-attention layer, where a query aggregates $\frac{P_j}{P_e}$ patches into a single embedding:

$$\text{TA}(z_i^{js}) = \text{CrossAttention}(q, W_j^K z_i^{js}, W_j^V z_i^{js}). \tag{3}$$

where $q \in \mathbb{R}^{1 \times D}$ is a learnable query, and W_j^K and W_j^V are the key and value projection weights.

Given the unique characteristics of each modality, EEG and EMG signals are crucial in the frequency domain [8, 24], while EOG reflects eye movements, and ECG exhibits periodic patterns. We propose reconstructing the Fourier amplitude for EEG and EMG signals, while preserving the original signals for EOG and ECG. The codes from the private and shared codebooks are first ℓ_2 normalized, concatenated, and then fed into the decoders $\mathcal{D}^e, \mathcal{D}^c, \mathcal{D}^o, \mathcal{D}^m$ to reconstruct the target signals:

$$o_i^e = \mathcal{D}^e(\hat{z}_i^{ep} \| \hat{z}_i^{es}) \quad o_i^c = \mathcal{D}^c(\hat{z}_i^{cp} \| \hat{z}_i^{cs}) \quad o_i^o = \mathcal{D}^o(\hat{z}_i^{op} \| \hat{z}_i^{os}) \quad o_i^m = \mathcal{D}^m(\hat{z}_i^{mp} \| z_i^{ms}), \tag{4}$$

where o_i denotes the reconstructed signals, and $\|$ represents concatenation operator. The training loss for codebook optimization is then formulated as:

$$\mathcal{L}_{CB} = \sum_i (\sum_j \underbrace{\|o_i^j - x_i^j\|_2^2}_{\text{reconstruction loss}} + \sum_l (\underbrace{\|\boldsymbol{sg}(\ell_2(z_i^l)) - \ell_2(\hat{z}_i^l)\|_2^2}_{\text{VQ loss}} + \underbrace{\|\ell_2(z_i^l) - \boldsymbol{sg}(\ell_2(\hat{z}_i^l))\|_2^2}_{\text{commitment loss}})), \tag{5}$$

where $j \in \{e, c, o, m\}$ and $l \in \{e, c, o, m, s\}$. The stop-gradient operation is denoted as $\boldsymbol{sg}$. In this framework, the decoders are optimized via the reconstruction loss, the codebooks are updated using the VQ loss, and the encoders are refined through the commitment loss. We apply z-score normalization to the reconstruction target in each sample to enhance the stability of convergence. Additionally, we employ an exponential moving average strategy to ensure stable codebook updates[37].

Cross-modal Reconstruction. EEG signals encapsulate rich cognitive and physiological information, as they directly reflect neural activities [6]. To encourage the shared codebook to capture common patterns across modalities, we designate EEG as an anchor and leverage its shared embeddings to reconstruct the other modalities with their private codes:

$$\overline{o}_i^c = \mathcal{D}^c(\hat{z}_i^{cp} \| \text{CMA}(\hat{z}_i^{es})) \quad \overline{o}_i^o = \mathcal{D}^o(\hat{z}_i^{op} \| \text{CMA}(\hat{z}_i^{es})) \quad \overline{o}_i^m = \mathcal{D}^m(\hat{z}_i^{mp} \| \text{CMA}(\hat{z}_i^{es})), \tag{6}$$

where CMA denotes the Cross-modal Alignment module. Similar to TA, CMA employs cross-attention to handle varying temporal scales across modalities. It extends the shared

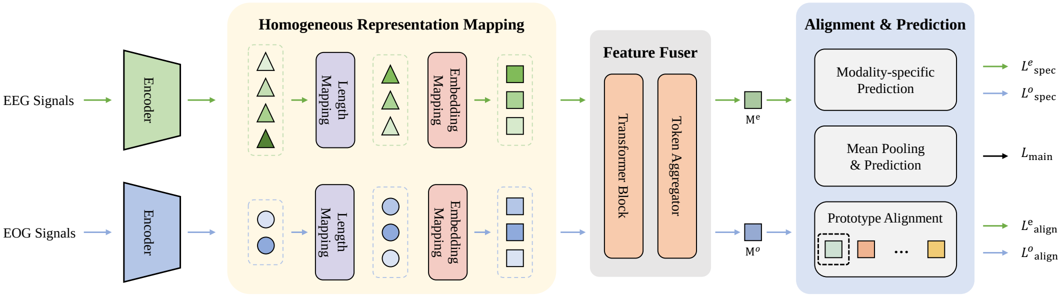

Figure 2: Schematic of the resilient fine-tuning process, illustrated with EEG and EOG as examples. Outputs from all modality encoders undergo Homogeneous Representation Mapping, followed by a Feature Fuser, which consists of a Transformer Block and a Token Aggregator for multimodal feature integration. Prototype alignment ensures robustness to missing modalities in downstream tasks, while modality-specific prediction preserves the performance of individual modalities.

EEG codes by a factor of $\frac{P_e S_j}{P_j S_e}$ to match the patch count of each modality $j \in \{c, o, m\}$: CrossAttention$(Q, W_j^K z_i^{js}, W_j^V z_i^{js})$, where $Q \in \mathbb{R}^{\frac{P_e S_j}{P_j S_e} \times D}$ is the learnable query matrix. The cross-modal reconstruction loss is then computed as:

$$\mathcal{L}_{CR} = \sum_i \sum_j \|\overline{o}_i^j - x_i^j\|_2^2. \tag{7}$$

Disentangling Loss. Beyond ensuring that the shared codebook captures modality-invariant information, it is equally crucial to distinguish shared and private embeddings, ensuring they encode distinct aspects of the input. To mitigate redundancy between embeddings, we introduce a disentangling loss that enforces soft orthogonality:

$$\mathcal{L}_D = \sum_i \sum_j \text{sim}(z_i^{jp}, z_i^{js}), \tag{8}$$

where cosine similarity is used to encourage orthogonality.

Ultimately, the total loss for decoupled multimodal tokenizer training integrates these constraints:

$$\mathcal{L}_T = \mathcal{L}_{CB} + \alpha_1 \mathcal{L}_{CR} + \alpha_2 \mathcal{L}_D, \tag{9}$$

where α_1 and α_2 are weighting factors balancing the loss terms.

3.2 Masked Signal Pre-training

Masked signal modeling, an effective representation learning paradigm across various domains [7, 14, 18], is extended here to learn generic and semantic representations for multimodal signals. For each modality j, we randomly generate a mask $\mathcal{M} = \{0, 1\}^{N_j}$ with a mask ratio r_j. All masked patches are replaced by a learnable mask token, producing a corrupted sample $\widetilde{x}_i^j$. The encoder is then trained to predict the corresponding codebook indices for each masked patch. Since the embeddings are decoupled into shared and private codebooks, we employ two projection heads to recover the masked patches. Due to varying temporal scales, shared codes may span multiple patches. To ensure alignment, we duplicate the shared codes $\frac{P_e S_j}{P_j S_e}$ times. The training objective is formulated as:

$$\mathcal{L}_M = -\sum_i \sum_{m_k=1, m_k \in \mathcal{M}} \log p(\hat{z}_i^{jp}, \hat{z}_i^{js} | \widetilde{x}_i^j). \tag{10}$$

3.3 Resilient Fine-tuning with Prototype Alignment

In this stage, we aim to maintain the representational capacity of individual modalities while addressing the challenge of missing modalities. Figure 2 provides an overview of this process.

Homogeneous Representation Mapping. Each modality-specific input x_i^j is first encoded using its respective pre-trained encoder $\mathcal{E}^j$, producing feature representations $z_i^j \in \mathbb{R}^{n^j \times d^j}$, where n^j and

d^j denote the sequence length and dimension, respectively. Since different modalities yield features of varying lengths and dimensions, we introduce Homogeneous Representation Mapping to project them into a unified feature space $\mathbb{R}^{n \times d}$ [45].

Following encoding, a modality-specific Length Mapping module ℓ_m standardizes feature lengths by transforming the extracted representations into a common-length format $f_i^j \in \mathbb{R}^{n \times d^j}$:

$$f_i^j = \ell_m(z_i^j) = \mathcal{R}(z_i^j, \widetilde{z}_i^j), \quad \widetilde{z}_i^j = \text{softmax}(\mathcal{H}^j(z_i^j)) \in \mathbb{R}^{n^j \times n}, \quad \mathcal{R}(a, b) = b^T a. \tag{11}$$

Here, the encoded representations z_i^j are first transformed by a modality-specific head $\mathcal{H}^j$, which consists of linear and normalization layers, to produce $\widetilde{z}_i^j$, a probability distribution over spatial locations obtained via a softmax operation. To facilitate cross-modality alignment, a Reorganization Function $\mathcal{R}$ restructures spatial distributions based on the encoded feature representations. Finally, a modality-specific Embedding Mapping module e_m projects the reorganized embeddings into a common feature space using a linear transformation: $\hat{f}_i^j = e_m(f_i^j) = \text{Linear}(f_i^j) \in \mathbb{R}^{n \times d}$.

Fusion & Alignment. Given the encoded feature $\hat{f}_i^j$, the Feature Fuser module integrates information from multiple tokens using a shared Transformer block, followed by a Token Aggregator, which applies a 1D convolution layer per modality to aggregate sequential information:

$$h_i^j = \text{Aggregator}(\text{Transformer}(\hat{f}_i^j)) \in \mathbb{R}^d. \tag{12}$$

To enhance robustness against missing modalities while ensuring effective modality fusion, we introduce three losses in the Alignment & Prediction module. First, to align unimodal features within a common space, we propose Prototype Alignment, where a set of learnable prototypes $\mathcal{U} = \{u_1, ..., u_{|\mathcal{U}|}\} \in \mathbb{R}^{|\mathcal{U}| \times d}$ is shared across all modalities. Specifically, for an n-class classification task, each sample is encouraged to be close to the unique prototype corresponding to its class, setting $|\mathcal{U}| = n$. For regression tasks, $|\mathcal{U}|$ is a hyperparameter, and each sample is encouraged to align with the nearest prototype. Therefore, the alignment loss is defined as:

$$\mathcal{L}_{\text{align}} = \sum_i \sum_j \left\| h_i^j - u_{h_i} \right\|_2^2, \tag{13}$$

where $u_{h_i} = \mathcal{U}[\text{label}(i)]$ (classification), $u_{h_i} = \mathcal{U}[\arg\min_k \|\ell_2(h_i^j) - \ell_2(u_i)\|_2]$ (regression). This loss ensures that the feature representations h_i^j align with their respective prototypes.

The main prediction loss $\mathcal{L}_{\text{main}}$ is computed by averaging the features h_i^j across all modalities, reinforcing the effectiveness of multimodal fusion, which is also used as the final prediction at test stage. Additionnaly, a modality-specific loss $\mathcal{L}_{\text{spec}}^j$ is introduced to preserve the independent representational capacity of each modality. For $\mathcal{L}_{\text{main}}$ and $\mathcal{L}_{\text{spec}}^j$, we employ cross-entropy loss with label smoothing for multi-class classification, binary cross-entropy (BCE) loss for binary classification, and mean squared error (MSE) loss for regression tasks. The total fine-tuning loss is a weighted sum of all components:

$$\mathcal{L}_F = \gamma_m \mathcal{L}_{\text{main}} + \sum_j \gamma_j \mathcal{L}_{\text{spec}}^j + \gamma_a \mathcal{L}_{\text{align}}, \tag{14}$$

where γ values control the trade-off between different loss terms.

4 Experiments

4.1 Downstream Datasets

We consider 4 popular downstream tasks in BCI for evaluating PhysioOmni, where EEGMAT is presented in Appendix E. The detailed information is presented in Table 1:

- **SEED-VII** [16] (emotion recognition): SEED-VII features seven emotions (happiness, sadness, neutral, fear, disgust, surprise, and anger) of 20 subjects who underwent 4 sessions on different days. Each session contains 20 trials of video clips lasting 2-5 minutes. For data partitioning, all subjects' data are combined, with the first 10 trials designated as the training set, the middle 5 trials as the validation set, and the final 5 trials as the test set.

- **HMC** [2] (sleep stage classification): HMC was developed for automatic sleep scoring, focusing on five sleep stages: wake, NREM-1, NREM-2, NREM-3, and REM. The dataset consists of randomly selected patient recordings from a diverse population undergoing polysomnographic (PSG) examinations for various sleep disorders. It includes full-night polysomnographic recordings from 151 subjects, with the first 100 used for training, the next 25 for validation, and the remaining 26 for testing.
- **FBM** [3] (motor prediction): This dataset includes full-body motion capture (66 targets) from approximately 10 walking trials conducted by 10 able-bodied individuals during gait tasks on level ground, ramps, and stairs, with the advantage of unconstrained movement using a motion capture system with wireless IMUs. We use a data stride of 50 ms. For model evaluation, data from all subjects are combined, with the last trial used for testing, the second-to-last trial for validation, and the remaining trials for training.

Table 1: Information of datasets used for downstream evaluation.

Dataset	Modality (#Channel)	Sampling Rate	Duration	#Sample	Task
SEED-VII	EEG (62), EOG, ECG	1000 Hz	1 second	281,679	7-class classification
HMC	EEG (4), EOG, EMG	256 Hz	30 seconds	137,243	5-class classification
FBM	EEG (60), EOG, EMG (12)	1000 Hz	2 seconds	166,028	Regression
EEGMAT	EEG (19), ECG	500 Hz	4 seconds	2,088	Binary classification

4.2 Experimental Setup

Preprocessing & Basic Settings. All modalities undergo a consistent preprocessing pipeline comprising a bandpass filter, a notch filter, and resampling. The specific parameters are tailored to each modality's characteristics: EEG (bandpass: 0.1–75 Hz, notch: 50 or 60 Hz, resampling: 200 Hz), EOG (bandpass: 0.1–75 Hz, notch: 50 or 60 Hz, resampling: 200 Hz), ECG (bandpass: 0.5–60 Hz, notch: 50 or 60 Hz, resampling: 500 Hz), and EMG (bandpass: 5–200 Hz, notch: 50, 100, 150 Hz or 60, 120, 180 Hz, resampling: 500 Hz). Signals are scaled to μV units, and all values are divided by 100 for normalization. The patch size is set to 200 for EEG (1 second) and 100 for ECG (0.2 second), EOG (0.5 second), and EMG (0.2 second). This preprocessing largely follows existing methods [18, 19], which are widely adopted. Additional details on hyperparameter settings are provided in Appendix A.

Training Data & Environment Settings. To enhance the model's generalization ability, we extensively gather diverse multimodal physiological data from multiple datasets, mostly containing three or more modalities, as detailed in Appendix B. Our experiments are conducted on four NVIDIA A100-80G GPUs with Python 3.12.8 and PyTorch 2.5.1 + CUDA 12.4. The best models are selected based on their performance on the validation set and subsequently evaluated on the test set with full modalities. To ensure reliability and comparability, we report the average and standard deviation across three random seeds.

Table 2: The results of different methods on SEED-VII.

Method	Training Modality	Test Modality	Balanced Accuracy	Cohen's Kappa	Weighted F1
EEG-Conformer [34]	EEG	EEG	0.1775±0.0014	0.0428±0.0015	0.1482±0.0030
BIOT [42]	EEG	EEG	0.3151±0.0047	0.2044±0.0054	0.3210±0.0055
	EOG	EOG	0.1888±0.0015	0.0529±0.0015	0.1656±0.0042
	ECG	ECG	0.2186±0.0015	0.0925±0.0021	0.2212±0.0033
LaBraM-Base [18]	EEG	EEG	0.3456±0.0028	**0.2391**±0.0025	**0.3511**±0.0027
CBraMod [39]	EEG	EEG	0.3237±0.0026	0.2126±0.0028	0.3270±0.0026
Fu *et al.* [11]	EEG+EOG	EOG	0.1449±0.0017	0.0025±0.0021	0.0428±0.0069
	EEG+ECG	ECG	0.1703±0.0054	0.0337±0.0066	0.1148±0.0129
FeatFusion	EEG+EOG+ECG	EEG+EOG+ECG	0.3475±0.0047	0.2301±0.0045	0.3437±0.0042
PhysioOmni	EEG+EOG+ECG	EEG	**0.3479**±0.0054	0.2316±0.0045	0.3442±0.0025
		EOG	**0.2079**±0.0052	**0.0737**±0.0051	**0.2104**±0.0084
		ECG	**0.2302**±0.0049	**0.1062**±0.0065	**0.2425**±0.0058
		EEG+EOG	0.3521±0.0048	0.2375±0.0038	0.3494±0.0019
		EEG+ECG	0.3558±0.0075	0.2431±0.0052	0.3550±0.0031
		EOG+ECG	0.2587±0.0053	0.1417±0.0074	0.2744±0.0071
		EEG+EOG+ECG	**0.3642**±0.0065	**0.2539**±0.0041	**0.3647**±0.0025

Table 3: The results of different methods on HMC.

Method	Training Modality	Test Modality	Balanced Accuracy	Cohen's Kappa	Weighted F1
EEG-Conformer [34]	EEG	EEG	0.6767±0.0200	0.5886±0.0397	0.6550±0.0463
BIOT [42]	EEG	EEG	0.6862±0.0041	0.6295±0.0113	0.7091±0.0147
	EOG	EOG	**0.6192**±0.0049	0.5553±0.0020	**0.6595**±0.0038
	EMG	EMG	0.3705±0.0117	0.2072±0.0050	0.3946±0.0065
LaBraM-Base [18]	EEG	EEG	0.7286±0.0101	0.6812±0.0073	0.7554±0.0024
CBraMod [39]	EEG	EEG	0.7177±0.0072	0.6653±0.0057	0.7388±0.0052
Fu *et al.* [11]	EEG+EOG	EOG	0.4885±0.0704	0.3333±0.1251	0.3754±0.1670
	EEG+EMG	EMG	**0.3987**±0.0057	0.1937±0.0052	0.228±0.0019
SleepMG [27]	EEG+EOG+EMG	EEG+EOG+EMG	0.6924±0.0091	0.6328±0.0102	0.7168±0.0073
FeatFusion	EEG+EOG+EMG	EEG+EOG+EMG	**0.7478**±0.0038	0.6981±0.0004	0.7728±0.0010
PhysioOmni	EEG+EOG+EMG	EEG	**0.7289**±0.0010	**0.6880**±0.0097	**0.7635**±0.0053
		EOG	0.6066±0.0073	**0.5554**±0.0023	0.6533±0.0026
		EMG	0.3914±0.0113	**0.2454**±0.0095	**0.4104**±0.0108
		EEG+EOG	0.7404±0.0018	0.7063±0.0105	0.7755±0.0058
		EEG+EMG	0.7300±0.0062	0.6958±0.0070	0.7680±0.0028
		EOG+EMG	0.6026±0.0038	0.5717±0.0108	0.6602±0.0082
		EEG+EOG+EMG	0.7377±0.0056	**0.7120**±0.0085	**0.7779**±0.0031

Table 4: The results of different methods on FBM.

Method	Training Modality	Test Modality	RMSE↓	Pearson Correlation	R^2 Score
EEG-Conformer [34]	EEG	EEG	5.7243±0.4059	0.4996±0.0824	0.2053±0.1120
BIOT [42]	EEG	EEG	6.0831±0.0606	0.4614±0.0056	-0.1714±0.0771
	EOG	EOG	6.4821±0.0357	0.3469±0.0079	-0.1506±0.0370
	EMG	EMG	6.1194±0.3700	0.3950±0.1642	-0.2671±0.0753
LaBraM-Base [18]	EEG	EEG	5.0466±0.0112	**0.6567**±0.0020	**0.3668**±0.0043
CBraMod [39]	EEG	EEG	5.1216±0.0174	0.6345±0.0035	0.3490±0.0033
Fu *et al.* [11]	EEG+EMG	EEG	5.4682±0.0457	0.5391±0.0148	0.2721±0.0128
	EOG+EMG	EOG	6.2401±0.0722	0.3278±0.0323	**0.0888**±0.0227
FeatFusion	EEG+EOG+EMG	EEG+EOG+EMG	5.0511±0.1564	0.6443±0.0203	0.2674±0.0456
PhysioOmni	EEG+EOG+EMG	EEG	**4.9650**±0.0089	0.6313±0.0075	0.3122±0.0068
		EOG	**6.0321**±0.0217	**0.3838**±0.0042	0.0173±0.0156
		EMG	**5.5950**±0.1048	**0.5934**±0.0122	**0.0337**±0.0226
		EEG+EOG	4.9002±0.0312	0.6126±0.0059	0.3472±0.0079
		EEG+EMG	4.7133±0.0354	**0.6637**±0.0023	0.3613±0.0111
		EOG+EMG	5.0091±0.0184	0.5979±0.0060	0.2812±0.0038
		EEG+EOG+EMG	**4.6191**±0.0122	0.6580±0.0008	**0.3995**±0.0074

4.3 Experimental Results

The results are presented in Table 2, 3, and 4, with baseline descriptions provided in Appendix C. Overall, PhysioOmni achieves competitive performance compared to a wide range of unimodal, multimodal, and cross-modal methods. Notably, there is a trade-off in selecting the best model for PhysioOmni, as different modality combinations reach optimal performance at different training epochs, placing our approach at a disadvantage. Despite this, PhysioOmni matches or surpasses the performance of the leading baselines in most cases. In addition to achieving best results under the single EEG condition, PhysioOmni also outperforms all baselines on other unimodal (EOG, ECG, EMG) and multimodal settings. Specifically, it delivers superior performance for EOG, ECG, and multimodal scenarios, and remains competitive with LaBraM. On HMC, PhysioOmni consistently achieves the best results across all modalities except EOG. While it does not outperform the top baselines in terms of correlation and R^2, it achieves the lowest RMSE on the FBM dataset.

4.4 Ablation Study

To evaluate the contribution of key components, we conduct ablation studies on modality-specific prediction, disentangling loss, prototype alignment, cross-modal reconstruction, and the shared codebook. The results are presented in Figure 3, where lower RMSE values on FBM indicate better performance. Overall, the removal of any component generally leads to performance degradation. Notably, prototype alignment proves especially effective on SEED-VII and HMC, while modality-specific prediction is crucial across all datasets. Removing cross-modal reconstruction, disentangling loss, or the shared codebook consistently results in performance drops, underscoring their importance.

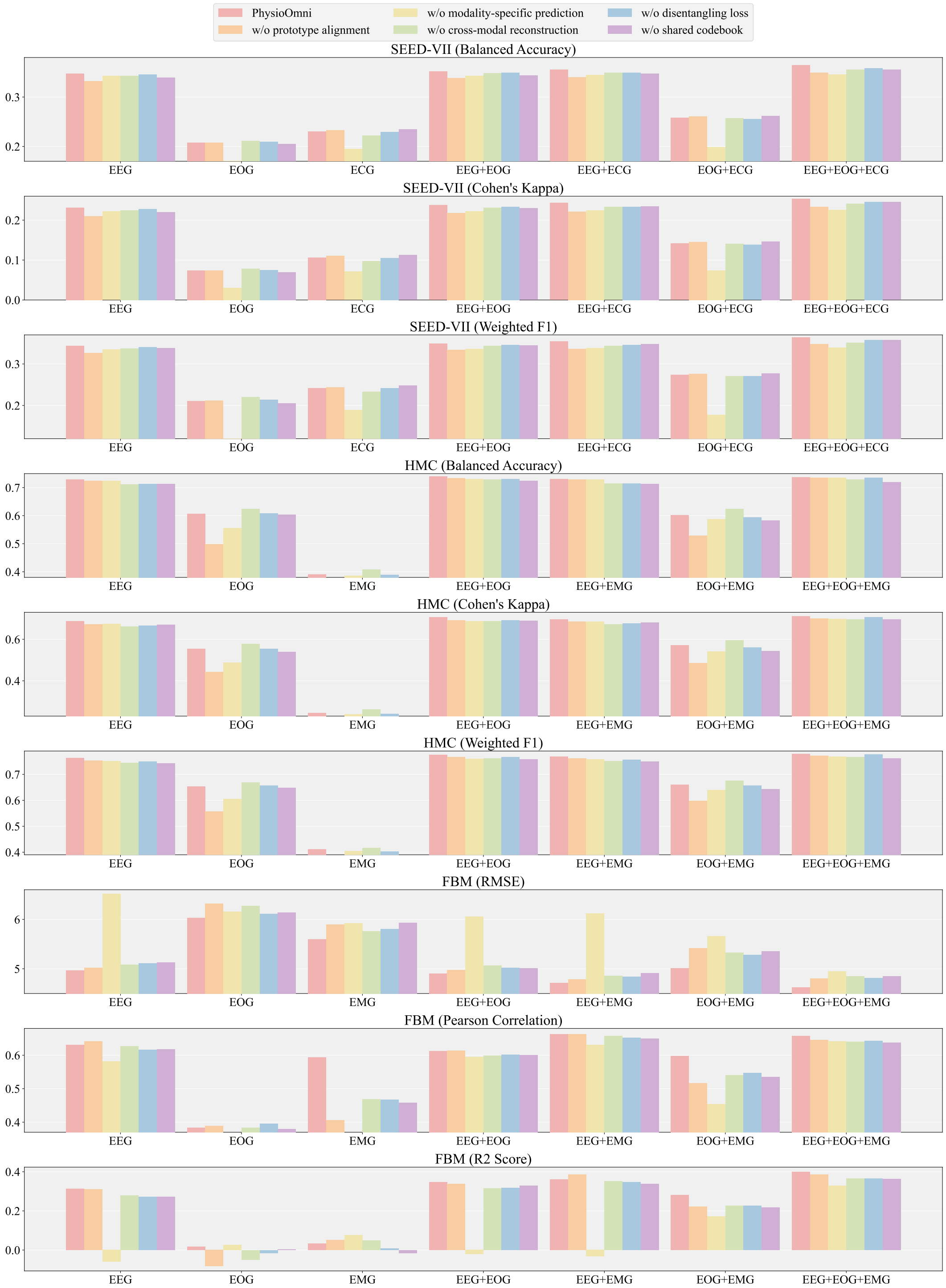

Figure 3: Ablation study of pivotal components on downstream datasets.

5 Conclusion

We propose PhysioOmni, a universal multimodal physiological foundation model that learns robust representations through decoupled tokenization and masked signal modeling, and effectively handles arbitrary missing modalities via resilient fine-tuning. Through comprehensive experiments on four BCI tasks, PhysioOmni consistently achieves SOTA across both unimodal and multimodal settings.

References

[1] Isabela Albuquerque, Abhishek Tiwari, Mark Parent, Raymundo Cassani, Jean-François Gagnon, Daniel Lafond, Sébastien Tremblay, and Tiago H Falk. Wauc: a multi-modal database for mental workload assessment under physical activity. *Frontiers in Neuroscience*, 14:549524, 2020.

[2] Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):1–27, 08 2021.

[3] Justin A Brantley, Trieu Phat Luu, Sho Nakagome, Fangshi Zhu, and Jose L Contreras-Vidal. Full body mobile brain-body imaging data during unconstrained locomotion on stairs, ramps, and level ground. *Scientific Data*, 5(1):1–10, 2018.

[4] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018.

[5] Jeong-Hyun Cho, Ji-Hoon Jeong, and Seong-Whan Lee. NeuroGrasp: Real-Time EEG Classification of High-Level Motor Imagery Tasks Using a Dual-Stage Deep Learning Framework. *IEEE Transactions on Cybernetics*, 52(12):13279–13292, 2022.

[6] Fernando Lopes da Silva. Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and clinical neurophysiology*, 79(2):81–93, 1991.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[8] O Dressler, G Schneider, G Stockmanns, and EF Kochs. Awareness and the EEG power spectrum: analysis of frequencies. *British journal of anaesthesia*, 93(6):806–809, 2004.

[9] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of EEG waves for EEG to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

[10] Ching Fang, Christopher Michael Sandino, Behrooz Mahasseni, Juri Minxha, Hadi Pouransari, Erdrin Azemi, Ali Moin, and Ellen L. Zippi. Promoting cross-modal representations to improve multimodal foundation models for physiological signals. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024.

[11] Xi Fu and Cuntai Guan. Gait Pattern Recognition Based on Supervised Contrastive Learning Between EEG and EMG. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, 2023.

[12] Nigel Gebodh, Zeinab Esmaeilpour, Abhishek Datta, and Marom Bikson. Dataset of concurrent EEG, ECG, and behavior with multiple doses of transcranial electrical stimulation. *Scientific Data*, 8(1):274, 2021.

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[14] 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*, pages 16000–16009, 2022.

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[16] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A Multimodal Dataset of Six Basic Emotions with Continuous Labels for Emotion Recognition. *IEEE Transactions on Affective Computing*, pages 1–16, 2024.

[17] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Stamos Katsigiannis and Naeem Ramzan. 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):98–107, 2017.

[21] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. 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.

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

[23] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[24] Paavo V Komi and Per Tesch. EMG frequency spectrum, muscle structure, and fatigue during dynamic contractions in man. *European journal of applied physiology and occupational physiology*, 42:41–50, 1979.

[25] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6631–6640, June 2023.

[26] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. VBH-GNN: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] Shuo Ma, Yingwei Zhang, Qiqi Zhang, Yiqiang Chen, Haoran Wang, and Ziyu Jia. SleepMG: Multimodal Generalizable Sleep Staging with Inter-modal Balance of Classification and Domain Discrimination. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4004–4013, New York, NY, USA, 2024. Association for Computing Machinery.

[28] Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in neuroscience*, 10:195498, 2016.

[29] R. Sharma, V.I. Pavlovic, and T.S. Huang. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, 1998.

[30] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[31] Qi Shen, Junchang Xin, Bing Dai, Shudi Zhang, and Zhiqiong Wang. Robust sleep staging over incomplete multimodal physiological signals via contrastive imagination. *Advances in Neural Information Processing Systems*, 37:112025–112049, 2024.

[32] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2011.

[33] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

[35] Mario Giovanni Terzano and Liborio Parrino. The cyclic alternating pattern (CAP) in human sleep. In *Handbook of Clinical Neurophysiology*, volume 6, pages 79–93. Elsevier, 2005.

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[38] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

[39] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBramod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[41] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36:63529–63541, 2023.

[42] Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78240–78260. Curran Associates, Inc., 2023.

[43] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. Brant-X: A Unified Physiological Signal Alignment Framework. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4155–4166, 2024.

[45] Yunhua Zhang, Hazel Doughty, and Cees Snoek. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36:54716–54726, 2023.

[46] Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Hyperparameter Settings

Table 5: Hyperparameters for decoupled multimodal tokenizer.

Hyperparameters		EEG Encoder	Other Encoders
Temporal Encoder	Iput channels	{1,8,8}	
	Output channels	{16,16,16}	
	Kernel size	{15,3,3}	
	Stride	{8,1,1}	
	Padding	{7,1,1}	
Transformer encoder layers		12	
Transformer decoder layers		3	
Hidden size		200	100
MLP size		800	400
Attention head number		10	
Shared Codebook size		8192×64	
Private Codebook size		8192×64	
Batch size		512	
Peak learning rate		1e-4	
Minimal learning rate		1e-5	
Learning rate scheduler		Cosine	
Optimizer		AdamW	
Adam β		(0.9,0.99)	
Weight decay		1e-4	
α_1, α_2		(1,0.1)	
Total epochs		100	
Warmup epochs		10	
Data overlap		None	
Gradient clipping		None	

Table 6: Hyperparameters for masked signal pre-training.

Hyperparameters		EEG Encoder	Other Encoders
Temporal Encoder	Iput channels	{1,8,8}	
	Output channels	{16,16,16}	
	Kernel size	{15,3,3}	
	Stride	{8,1,1}	
	Padding	{7,1,1}	
Transformer encoder layers		12	
Hidden size		200	100
MLP size		800	400
Attention head number		10	
Mask ratio		0.5 (EEG, EMG), 0.7 (EOG, ECG)	
Batch size		512	
Peak learning rate		5e-4	
Minimal learning rate		1e-5	
Learning rate scheduler		Cosine	
Optimizer		AdamW	
Adam β		(0.9,0.98)	
Weight decay		0.05	
Total epochs		50	
Warmup epochs		5	
Data overlap		None	
Gradient clipping		3	

Table 7: Hyperparameters for fine-tuning.

Hyperparameters	Values
Projected feature size	128×128
MoE layers	1
Expert number	4
Hidden size	128
MLP size	512
Attention head number	10
Number of prototypes	7 (SEED-VII), 5 (HMC), 256 (FBM), 2 (EEGMAT)
Batch size	128 (EEGMAT), 512 (SEED-VII, HMC, FBM)
Loss ratio γ	SEED-VII: 1 (align), 0.5 (main), 0.5 (EEG), 0.5 (EOG), 0.5 (ECG) HMC: 1 (align), 0.1 (main), 0.01 (EEG), 4 (EOG), 0.5 (EMG) FBM: 0.01 (align), 0.1 (main), 2 (EEG), 0.5 (EOG), 1 (EMG) EEGMAT: 1 (align), 0.5 (main), 0.5 (EEG), 0.5 (EOG), 0.5 (EMG)
Peak learning rate	1e-3 (SEED-VII), 5e-4 (HMC, FBM), 1e-4 (EEGMAT)
Minimal learning rate	1e-4 (SEED-VII), 5e-5 (HMC, FBM), 1e-5 (EEGMAT)
Learning rate scheduler	Cosine
Optimizer	AdamW
Adam β	(0.9,0.999)
Weight decay	0.05
Total epochs	50
Warmup epochs	5
Gradient clipping	None
Label smoothing	0.1 (multi-class classification)

B Pre-training Datasets

We utilize a diverse collection of multimodal physiological datasets for various tasks:

- **TUEG** [28]: The TUH EEG Corpus (TUEG) is an extensive archive comprising 26,846 clinical EEG recordings collected at Temple University Hospital. Some recordings include EOG (horizontal), ECG, and EMG signals alongside EEG (21–23 channels), with sampling frequencies ranging from 250 to 1024 Hz. For pre-training, we select recordings that include at least three modalities.
- **DEAP** [23]: DEAP provides a multimodal dataset for analyzing human affective states, recorded from 32 participants as they watched 40 one-minute music video excerpts. EEG (32 channels), EOG (horizontal and vertical), and EMG signals were captured at a sampling rate of 512 Hz.
- **Sleep-EDF** [21]: This dataset contains 197 whole-night sleep recordings featuring EEG (2 channels), EOG (horizontal), and chin EMG signals, sampled at 100 Hz. Hypnograms, representing sleep patterns, were manually scored by trained technicians following the Rechtschaffen and Kales manual.
- **CAP** [35]: The CAP Sleep Database includes 108 polysomnographic recordings sampled at 128 Hz, featuring EEG (3 or more channels), EOG (horizontal and vertical), and EMG signals, with annotations for sleep stages and Cyclic Alternating Pattern (CAP).
- **GX** [12]: This dataset combines high-density EEG (30 channels) with ECG and EOG (horizontal) during transcranial electrical stimulation (tES) across 783 trials and 62 sessions, sampled at 2000 Hz. It includes nine HD-tES types targeting three cortical regions with different waveforms, with participants performing vigilance tasks, completing wellness questionnaires, and undergoing repeated sessions to assess within-participant reliability.
- **Private data**: This dataset comprises 54 recordings from 19 subjects while watching videos, with each recording lasting approximately one hour. EEG (62 channels), EOG (horizontal and vertical), and ECG signals were captured using the ESI NeuroScan System at a sampling rate of 1000 Hz.

C Baselines

To comprehensively evaluate PhysioOmni, we compare it with the following classical and state-of-the-art baselines:

- **EEG-Conformer** [34]: A compact Convolutional Transformer that integrates local feature extraction via convolution and global feature modeling via self-attention for EEG classification. By combining temporal and spatial convolutions with self-attention, EEG-Conformer effectively captures both short-term and long-term dependencies in EEG signals.
- **BIOT** [42]: A Biosignal Transformer designed for flexible biosignal encoding, BIOT facilitates cross-data learning across diverse signal formats such as EEG and ECG. It tokenizes each channel into fixed-length segments while preserving spatio-temporal features through channel and positional embeddings, demonstrating strong generalizability via joint pre-training and fine-tuning on multiple biosignal tasks.
- **LaBraM** [18]: A unified foundation model for EEG, LaBraM enables cross-dataset learning by segmenting EEG signals into channel patches and encoding them with a vector-quantized neural tokenizer. It employs masked neural code prediction for unsupervised pre-training, leveraging approximately 2,500 hours of EEG data from 20 datasets. We fine-tune its publicly available pre-trained checkpoints on each downstream dataset.
- **CBraMod** [39]: An EEG foundation model designed to address the heterogeneity of spatial and temporal dependencies in EEG signals, CBraMod utilizes a criss-cross Transformer with separate attention mechanisms. It incorporates an asymmetric conditional positional encoding scheme for enhanced adaptability across diverse EEG formats. We fine-tune its publicly available pre-trained checkpoints on each downstream dataset.
- **Fu *et al.*** [11]: This method employs a multimodal training strategy using supervised contrastive learning, leveraging EMG signals during training to enhance EEG-based gait classification and regression. The model learns gait patterns from EEG with EMG guidance but relies solely on EEG during inference. We use multimodal signals to train and single modality to test for this method.
- **SleepMG** [27]: A multimodal generalizable sleep staging method, SleepMG balances inter-modal differences in PSG by assessing classification and domain discrimination performances across modalities. It defines modal performance metrics based on variance from the average performance and adaptively adjusts gradient updates to emphasize poorly balanced modalities.
- **FeatFusion**: We leverage our pre-trained encoders from PhysioOmni and fuse the features extracted from these encoders by the Homogeneous Representation Mapping and Feature Fuser. We concatenate the features from all modalities after Homogeneous Representation Mapping and feed them into the Transformer layer. This approach utilizes the generic representations of each modality and employs proposed fusion strategy.

D Evaluation Metrics

We consider the following metrics to comprehensively evaluate all methods:

- **Balanced Accuracy**: This metric calculates the average of recall (true positive rate) for each class, ensuring that each class contributes equally to the final accuracy score. It is particularly useful when dealing with imbalanced datasets.
- **AUC-PR**: AUC-PR is the area under the precision-recall curve, which plots precision against recall at various thresholds. It is especially useful for imbalanced datasets where the negative class is overwhelming.
- **AUROC**: AUROC is the area under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various thresholds. It provides a comprehensive view of a model's performance across all classification thresholds.
- **Cohen's Kappa**: Cohen's Kappa measures the agreement between two raters, correcting for chance agreement. A Kappa value of 1 indicates perfect agreement, while a value of 0

indicates no agreement beyond chance, with negative values indicating worse than random agreement.

- **Weighted F1**: The F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. The weighted F1 takes into account the support (the number of true instances) of each class, so it gives more importance to classes with more data.
- **RMSE**: RMSE is the square root of the average squared differences between predicted and actual values, providing a measure of the model's prediction error. Larger errors are penalized more due to the squaring of differences while a lower RMSE indicates a better fit between the model and the data.
- **R^2 Score**: R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A value of 1 indicates perfect prediction, 0 indicates that the model does not improve on simply predicting the mean of the data, and negative values suggest that the model is performing worse than a simple mean predictor.
- **Pearson Correlation**: Pearson correlation quantifies the linear relationship between two variables, with values ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value of 0 indicates no linear relationship.

For the monitor metric, AUROC is used for binary classification, Cohen's Kappa for multi-class classification, and R^2 Score for regression.

E More Experiments on EEGMAT

This dataset [46] supports the investigation of EEG characteristics during mental tasks, including Fourier power spectral, coherence, and detrended fluctuation analysis. It contains EEG recordings from 36 healthy volunteers performing a mental subtraction task, categorized into good and bad counters, and is available through the Physiobank platform for neuroscience research on cognitive workload. Subjects from number 0 to 25 are used for training, 26 to 30 for validation, and 31 to 35 for testing.

The results are presented in Table 8. Across all evaluation metrics and modality configurations, PhysioOmni consistently outperforms all competing baselines, demonstrating its strong generalization and adaptability. When trained and tested on both EEG and ECG modalities, PhysioOmni achieves the highest performance. Even in unimodal scenarios, PhysioOmni maintains its superiority, outperforming dedicated EEG and ECG baselines such as BIOT, LaBraM, and CBraMod. These results suggest that EEG generally provides stronger predictive signals than ECG for this task, yet combining both modalities leads to further performance improvements. This highlights the effectiveness of PhysioOmni in leveraging complementary information from multiple physiological signals, both in single-modality and multimodal settings.

Table 8: The results of different methods on EEGMAT.

Method	**Training Modality**	**Test Modality**	**Balanced Accuracy**	**AUC-PR**	**AUROC**
EEG-Conformer [34]	EEG	EEG	0.5141±0.0060	0.6417±0.0386	0.6455±0.0227
BIOT [42]	EEG	EEG	0.6655±0.0665	0.7189±0.0722	0.7342±0.0536
	ECG	ECG	0.5391±0.0043	0.5974±0.0293	0.5656±0.0263
LaBraM-Base [18]	EEG	EEG	0.6609±0.0204	0.7174±0.0234	0.7272±0.0165
CBraMod [39]	EEG	EEG	0.6310±0.0129	0.7073±0.0322	0.7303±0.0225
Fu *et al.* [11]	EEG+ECG	EEG	0.5241 ± 0.0734	0.5513 ± 0.0784	0.5737 ± 0.0621
	EEG+ECG	ECG	0.5586± 0.0049	0.6042± 0.0249	0.5767± 0.0548
FeatFusion	EEG+ECG	EEG+ECG	0.7219±0.0809	**0.9164**±0.0131	0.8873±0.0322
PhysioOmni	EEG+ECG	EEG	**0.7637**±0.0282	**0.8410**±0.0135	**0.8036**±0.0216
		ECG	**0.7471**±0.0496	**0.7718**±0.0798	**0.8173**±0.0569
		EEG+ECG	**0.7870**±0.0270	0.9001±0.0261	**0.8912**±0.0257

F Ablation on Frozen Encoders

We further investigate the impact of freezing the pre-trained encoders during fine-tuning, with results presented in Figure 4. As expected, freezing the encoders leads to a significant drop in downstream performance across all tasks and modality combinations. This highlights the importance of full end-to-end fine-tuning, which allows the model to adapt pre-trained representations to task-specific nuances and optimize the fusion of multimodal features. The results emphasize that while pre-training provides a strong initialization, fine-tuning remains crucial for maximizing performance in real-world downstream applications.

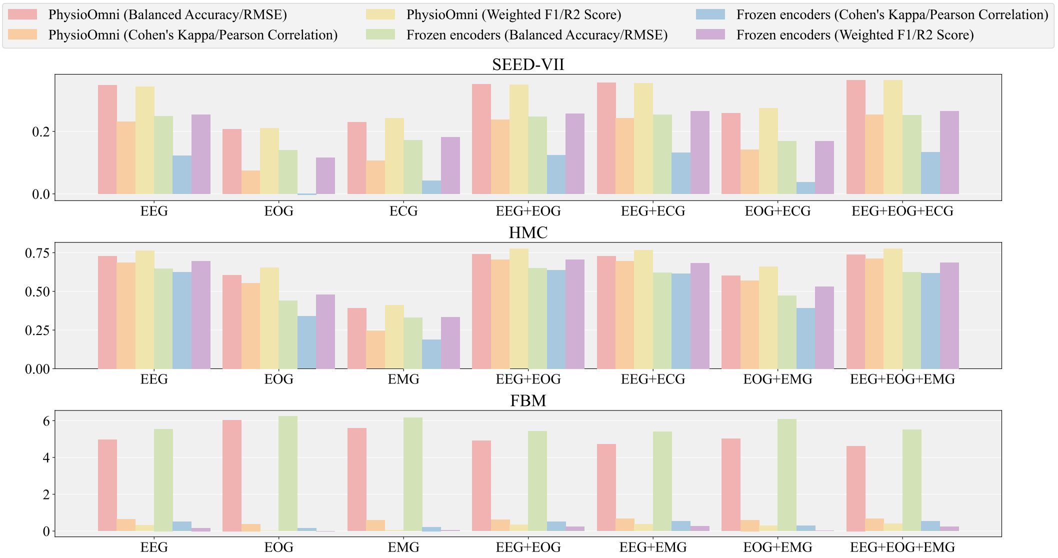

Figure 4: Ablation study on frozen encoders.

G Ablation on Pre-training

To assess the effectiveness of our proposed pre-training paradigm, we conduct an ablation study by randomly initializing all encoder parameters. The results, shown in Figure 5, demonstrate a consistent performance drop across all datasets and modality combinations when pre-training is removed. This highlights the critical role of pre-training in enabling robust and generalizable representations. Without pre-training, the model struggles to extract meaningful features from the raw physiological signals, leading to degraded performance even after fine-tuning. In contrast, the pre-trained encoders

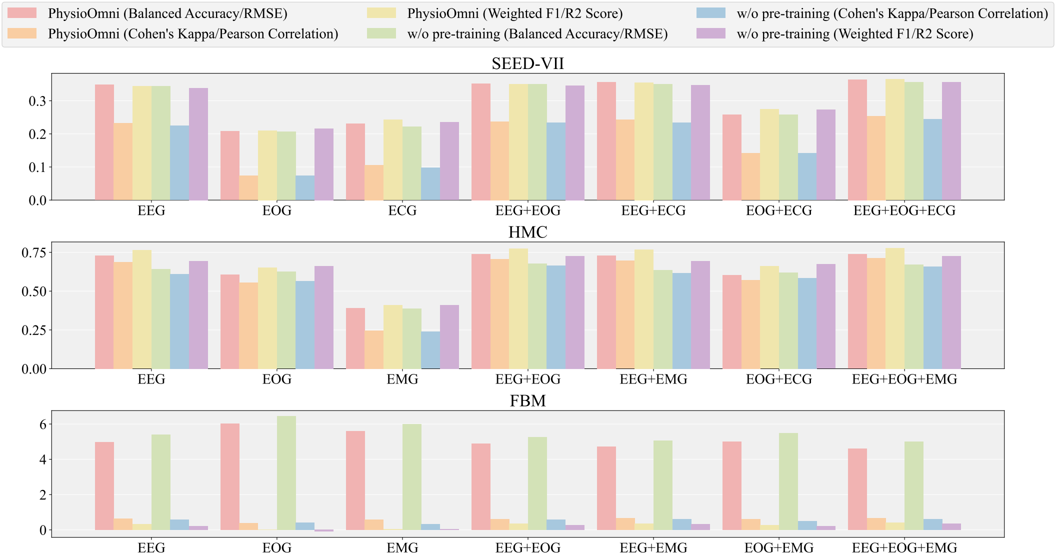

Figure 5: Ablation study on pre-training.

provide a strong initialization, capturing essential modality-specific and modality-invariant patterns that transfer well across downstream tasks. These findings confirm that our masked signal modeling and codebook-based design are effective in learning transferable representations that significantly enhance overall model performance.

H Ablation on Homogeneous Representation Mapping & Feature Fuser

Homogeneous Representation Mapping (HRM) is designed to project representations from different modalities into a common space with uniform embedding sizes, ensuring better compatibility across modalities. Meanwhile, the Feature Fuser (FF) integrates these projected representations to generate comprehensive multimodal features, enhancing the model's ability to leverage complementary information across different signals.

To assess the contribution of HRM and FF, we conduct an ablation study and present the results in Table 9. The findings demonstrate that removing the two modules generally lead to a degradation in performance on SEED-VII and HMC, underscoring their crucial role in effective modality fusion. However, on FBM, while eliminating HRM and FF leads to significant performance degration for EEG, it results in a performance boost in certain cases, particularly for EEG in combination with another modality. A possible explanation is that FBM exhibits a different modality interaction pattern compared to SEED-VII and HMC, where the added complexity of fusion modules may lead to overfitting. These results suggest that for FBM, a simpler fusion mechanism could be more advantageous than enforcing strict homogeneous representation constraints. This highlights the importance of dataset-specific tuning when designing multimodal fusion strategies.

Table 9: The ablation study on Homogeneous Representation Mapping & Feature Fuser (PhysioOmni / w/o HRM & FF).

Test Modality	**SEED-VII**		
	Balanced Accuracy	**Cohen's Kappa**	**Weighted F1**
EEG	0.3479±0.0054 / 0.3444±0.0008	0.2316±0.0045 / 0.2295±0.0011	0.3442±0.0025 / 0.3444±0.0006
EOG	0.2079±0.0052 / 0.1841±0.0243	0.0737±0.0051 / 0.0476±0.0263	0.2104±0.0084 / 0.1684±0.0459
ECG	0.2302±0.0049 / 0.2279±0.0020	0.1062±0.0065 / 0.1068±0.0043	0.2425±0.0058 / 0.2435±0.0047
EEG+EOG	0.3521±0.0048 / 0.3461±0.0023	0.2375±0.0038 / 0.2326±0.0033	0.3494±0.0019 / 0.3475±0.0020
EEG+ECG	0.3558±0.0075 / 0.3502±0.0014	0.2431±0.0052 / 0.2395±0.0005	0.3550±0.0031 / 0.3538±0.0013
EOG+ECG	0.2587±0.0053 / 0.2405±0.0091	0.1417±0.0074 / 0.1235±0.0093	0.2744±0.0071 / 0.2564±0.0098
EEG+EOG+ECG	0.3642±0.0065 / 0.3540±0.0033	0.2539±0.0041 / 0.2456±0.0038	0.3647±0.0025 / 0.3594±0.0025
Test Modality	**HMC**		
	Balanced Accuracy	**Cohen's Kappa**	**Weighted F1**
EEG	0.7289±0.0010 / 0.7115±0.0064	0.6880±0.0097 / 0.6687±0.0006	0.7635±0.0053 / 0.7460±0.0028
EOG	0.6066±0.0073 / 0.6247±0.0096	0.5554±0.0023 / 0.5669±0.0125	0.6533±0.0026 / 0.6683±0.0067
EMG	0.3914±0.0113 / 0.3934±0.0082	0.2454±0.0095 / 0.2425±0.0046	0.4104±0.0108 / 0.4133±0.0142
EEG+EOG	0.7404±0.0018 / 0.7252±0.0066	0.7063±0.0105 / 0.6886±0.0031	0.7755±0.0058 / 0.7606±0.0035
EEG+EMG	0.7300±0.0062 / 0.7132±0.0066	0.6958±0.0070 / 0.6769±0.0034	0.7680±0.0028 / 0.7513±0.0024
EOG+EMG	0.6026±0.0038 / 0.6170±0.0145	0.5717±0.0108 / 0.5800±0.0135	0.6602±0.0082 / 0.6737±0.0109
EEG+EOG+EMG	0.7377±0.0056 / 0.7252±0.0069	0.7120±0.0085 / 0.6961±0.0033	0.7779±0.0031 / 0.7651±0.0033
Test Modality	**FBM**		
	RMSE↓	**Pearson Correlation**	**R^2 Score**
EEG	4.9650±0.0089 / 5.1240±0.0075	0.6313±0.0075 / 0.6411±0.0023	0.3122±0.0068 / 0.2683±0.0065
EOG	6.0321±0.0217 / 6.0402±0.1256	0.3838±0.0042 / 0.3946±0.0077	0.0173±0.0156 / 0.0266±0.0465
EMG	5.5950±0.1048 / 5.5086±0.5068	0.5934±0.0122 / 0.4499±0.1828	0.0337±0.0226 / 0.1791±0.1108
EEG+EOG	4.9002±0.0312 / 4.8483±0.0576	0.6126±0.0059 / 0.6311±0.0069	0.3472±0.0079 / 0.3712±0.0139
EEG+EMG	4.7133±0.0354 / 4.6800±0.1101	0.6637±0.0023 / 0.6676±0.0092	0.3613±0.0111 / 0.4144±0.0162
EOG+EMG	5.0091±0.0184 / 5.3304±0.3294	0.5979±0.0060 / 0.5268±0.0759	0.2812±0.0038 / 0.2438±0.0069
EEG+EOG+EMG	4.6191±0.0122 / 4.7337±0.1369	0.6580±0.0008 / 0.6593±0.0064	0.3995±0.0074 / 0.4049±0.0228

I Limitations

While PhysioOmni demonstrates strong performance and generalizability across diverse physiological signals and downstream tasks, several limitations remain: 1) Although PhysioOmni handles arbitrary missing modalities during inference, it still relies on a fixed set of known modalities during training. Adapting to entirely new modalities not seen during pre-training remains an open challenge. 2)

Our method pre-trains individual encoders for each modality. While effective, this increases model complexity and limits parameter sharing. Developing a unified encoder architecture that can handle all physiological signals jointly is a promising direction for future work.

J Broader Impacts

PhysioOmni has the potential to advance multimodal physiological computing and BCI systems by providing a scalable and adaptable foundation model. Its ability to generalize across datasets and tasks while remaining resilient to missing modalities makes it well-suited for real-world scenarios where data completeness and consistency cannot be guaranteed. Applications include assistive healthcare systems, emotion-aware computing, cognitive load monitoring, and sleep analysis. However, care must be taken to address potential privacy concerns when deploying models trained on sensitive physiological data. Future efforts should also explore fairness, energy efficiency, and interpretability to ensure the responsible use of such foundation models in clinical and everyday applications.