BEYOND HEARING: LEARNING TASK-AGNOSTIC ExG REPRESENTATIONS FROM EARPHONES VIA PHYSIOLOGY-INFORMED TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrophysiological (ExG) signals offer valuable insights into human physiology, yet building foundation models that generalize across everyday tasks remains challenging due to two key limitations: (i) insufficient data diversity, as most ExG recordings are collected in controlled labs with bulky, expensive devices; and (ii) task-specific model designs that require tailored processing (*i.e.*, targeted frequency filters) and architectures, which limit generalization across tasks. To address these challenges, we introduce an approach for scalable, task-agnostic ExG monitoring in the wild. We collected 50 hours of unobtrusive free-living ExG data with an earphone-based hardware prototype to narrow the data diversity gap. At the core of our approach is *Physiology-informed Multi-band Tokenization (PiMT)*, which decomposes ExG signals into 12 physiology-informed tokens, followed by a reconstruction task to learn robust representations. This enables adaptive feature recognition across the full frequency spectrum while capturing task-relevant information. Experiments on our new *DailySense* dataset—the first to enable ExG-based analysis across five human senses—together with four public ExG benchmarks, demonstrate that *PiMT* consistently outperforms state-of-the-art methods across diverse tasks.

1 INTRODUCTION

Electrophysiological (ExG) signals, including electroencephalography (EEG), electromyography (EMG), electrooculography (EOG), and electrocardiography (ECG), provide critical insights into neural, muscular, ocular, and cardiovascular activities. They enable a wide range of physiological applications, from gaze tracking (Merino et al., 2010) and emotion recognition (Gkintoni et al., 2025) to sleep staging (Nguyen et al., 2016) and seizure detection (JW et al., 2016). Recent advances in deep learning have improved ExG analysis by developing data-driven training approaches (Song et al., 2022; Jiang et al., 2024) that capture complex temporal and spectral patterns for various physiological tasks. Building on this, foundation models, which have demonstrated remarkable success across domains by leveraging large-scale data to learn general-purpose representations (Narayanswamy et al., 2025), offer a promising opportunity for advancing everyday ExG analysis.

However, ExG foundation models remain underexplored due to two limitations: (i) insufficient dataset diversity and (ii) task-specific model design. First, ExG datasets are typically collected in controlled environments (Zheng & Lu, 2015; Katsigiannis & Ramzan, 2018; Wang et al., 2023) using bulky, expensive devices (*e.g.*, EEG headsets (Duvinage et al., 2013)). This setup restricts both scale and diversity across tasks, leaving free-living ExG data largely untapped. Second, existing ExG models are highly task-specific, relying on tailored processing pipelines, *i.e.*, architectures optimized for a fixed frequency band, which limits their generalization. For example, gaze tracking methods are designed to capture low-frequency bands (0.1∼15 Hz) (Merino et al., 2010), whereas emotion recognition relies on higher EEG bands (8∼30 Hz) (Gkintoni et al., 2025). As a result, a model trained for gaze tracking cannot be directly applied to emotion recognition, highlighting the lack of transferability across tasks.

To address the first challenge, we collected free-living ExG data in unobtrusive settings, constructing the *DailySense* dataset. For this, we prototyped *NeuroBuds*, an earphone-based ExG sensing device. Unlike traditional bulky systems, NeuroBuds is lightweight, low-cost, and portable while still capturing rich physiological signals: near-ear EEG, EMG from facial muscles, and EOG from eye movements. This design enables long-term data collection, overcoming the constraints of lab-based recordings. Leveraging this platform, we collected 50 hours of free-living ExG recordings from 22 participants engaged in unconstrained daily activities. Furthermore, we gathered 20 hours of targeted task-specific data spanning the five human senses (*i.e.*, sight, hearing, taste, touch, and smell), establishing the first benchmark for evaluating model performance across diverse tasks.

Moreover, we propose *Physiology-informed Multi-band Tokenization (PiMT)*, an approach designed to learn task-agnostic ExG representations. Instead of relying on a task-specific narrow band or a single wide-band input, PiMT decomposes ExG data into 12 fixed sub-band tokens, each corresponding to distinct physiological modalities. For instance, the [0.5∼4 Hz] band captures to EEG delta waves, which are informative for sleep staging (Elsaid & Labanowski, 2017), whereas the [15∼45 Hz] band reflects low-frequency EMG activities, relevant for muscle activation and motor tasks (Allison & Fujiwara, 2002). These structured tokens provide the encoder with fine-grained access to diverse spectral features, enabling the model to capture task-relevant information while remaining agnostic to any specific task. Coupled with self-supervised reconstruction objectives, we train a robust, transferable representations that generalize effectively across diverse downstream tasks.

To evaluate our approach, we benchmark PiMT against the state-of-the-art ExG training approaches. Specifically, we evaluate it on our newly introduced DailySense benchmark, which spans tasks across the five human senses, along with four widely used datasets covering diverse ExG applications, including emotion recognition, sleep staging, and brain–computer interface (BCI) tasks. Extensive experiments demonstrate that PiMT achieves robust performance and strong generalization across both DailySense and public datasets. Our key contributions are as follows:

- We identify the key limitations of existing ExG frameworks—insufficient dataset diversity and task-specific model design—that hinder generalization to real-world applications.
- We introduce NeuroBuds, an earphone-based prototype for unobtrusive, long-term ExG monitoring. Leveraging NeuroBuds, we curate DailySense, a dataset containing 50 hours of free-living recordings and 20 hours of task-specific data spanning the five human senses.
- We propose PiMT, a task-agnostic ExG training approach that incorporates a novel, physiology-informed multi-band tokenization scheme. This enables automatic extraction of task-relevant features across the entire frequency spectrum.
- Extensive experiments on DailySense spanning six distinct tasks and four public ExG benchmarks show PiMT achieves state-of-the-art performance, with an average F1 score of 87% over baseline models.

Together, these contributions establish the foundation for scalable, real-world ExG analysis, bridging wearable sensing technology and foundation models for deeper human understanding.

2 Related Work

To enable effective analysis of ExG signals and uncover valuable physiological patterns, recent approaches can be categorized into following three main groups: (i) *Conventional deep learning frameworks*, such as EEGNet (Lawhern et al., 2018) and DeepConvNet (Schirrmeister et al., 2017), leverage temporal and spatial convolutions to extract features directly from raw ExG signals. (ii) *Transformer-based models*, which capture local and long-term temporal dependencies, are well-suited for complex, high-dimensional ExG signals. Early efforts such as EEGConformer (Song et al., 2022) combine convolution and attention to jointly model local and global patterns. PatchTST (Nie et al., 2023) introduces patch-wise attention and independent channel encoding, while Medformer (Wang et al., 2024) enhances feature extraction through multi-scale patching and cross-channel attention. Most recently, Bidirectional-Mamba (Zhu et al., 2024) applies bidirectional state-space modeling for efficient long-range dynamics. (iii) *Self-supervised learning methods* aim to learn generalizable representations from unlabeled ExG signals using proxy tasks such as masked modeling or contrastive learning. BrainBERT (Wang et al., 2023) first applied BERT-style masked modeling to intracranial

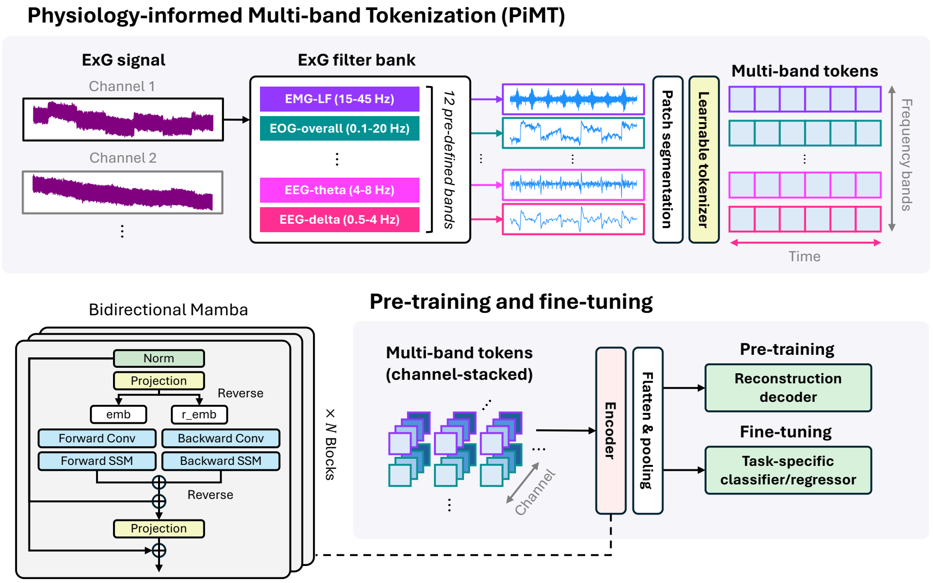

Figure 1: Overview of PiMT. ExG signals are decomposed into 12 sub-bands via Physiology-informed Multi-band Tokenization (PiFT). A Bidirectional-Mamba encoder processes the tokens, and the model is pre-trained with reconstruction tasks before fine-tuning on downstream tasks.

EEG spectrograms. BIOT (Yang et al., 2023) extends this idea to cross dataset via patch-token transformers, and BrainWave (Yuan et al., 2024) further scaled it to foundation models trained on large clinical datasets. However, despite these advances, existing approaches typically focus on specific ExG tasks and modalities, and are primarily evaluated on lab-controlled datasets. This limitation presents an opportunity to develop more generalized and robust representations from free-living ExG data. NeuroBuds addresses these gaps by introducing a unified, frequency-agnostic framework trained on real-world data, improving both robustness and practical usability.

3 Learning Task-agnostic ExG Representation

Motivation. Real-world ExG tasks are often associated with distinct physiological frequency bands. For example, gaze tracking with EOG signals typically relies on low-frequency components in 0.1∼15 Hz range (Merino et al., 2010), whereas EEG-based emotion recognition depends on higher-frequency bands, such as 8∼30 Hz (Gkintoni et al., 2025). Prior methods either design task-specific models (Gao et al., 2024; Altaheri et al., 2023) or apply narrow-band filters (Farhana et al., 2023; Apicella et al., 2021), both of which limit generalization across tasks. While a wide-band filter (*e.g.*, 0∼100 Hz) offers broader coverage, it suffers from loss of physiological features and poor task adaptation. We aim to develop a task-agnostic method that generalizes across tasks without relying on task-specific customization.

Overview. We propose a training framework that enables NeuroBuds to generalize effectively across diverse tasks. Figure 1 provides the overview. First, *Physiology-informed Multi-band Tokenization (PiMT)* decomposes the input into 12 physiology-informed sub-bands, producing tokens that that grant the model fine-grained access to task-relevant features across different frequency ranges. Next, a Bidirectional-Mamba encoder generates embeddings from the tokenized representations. To leverage unlabeled free-living data, we introduce a *Reconstruction-based Pre-training* to learn robust ExG representations. The pre-trained encoder is then fine-tuned on downstream tasks.

3.1 Physiology-informed Multi-band Tokenization (PiMT)

To enable the task-agnostic framework, we design a two-step tokenization pipeline that converts raw ExG signals into structured embeddings: (i) physiology-informed frequency decomposition via an *ExG Filter Bank*, and (ii) *Patch Segmentation and Tokenization* to generate input tokens.

ExG Filter Bank. Instead of relying on task-specific frequency bands, we design a fixed filter bank grounded in established physiological knowledge of ExG signals (Niedermeyer & Lopes da Silva, 2005; Nunez & Srinivasan, 2006; Task Force, 1996). Concretely, we define 12 canonical sub-band filters spanning key physiological modalities: EEG-delta (0.5$\sim$4 Hz), EEG-theta (4$\sim$8 Hz), EEG-alpha (8$\sim$13 Hz), EEG-beta (13$\sim$30 Hz), EEG-gamma (30$\sim$100 Hz), EMG-Low-Frequency (15$\sim$45 Hz), EMG-Mid-Frequency (45$\sim$95 Hz), EMG-High-Frequency (95$\sim$100 Hz), EOG-overall (0.1$\sim$20 Hz), ECG-Low-Frequency (0.03$\sim$0.12 Hz), ECG-High-Frequency (0.12$\sim$0.488 Hz), and the QRS complex (8$\sim$50 Hz).

Multi-band Filtering. ExG signals are decomposed into complementary spectral components by simultaneously applying all filters in the bank. This decomposition provides the model with fine-grained, physiologically relevant features that span multiple modalities and tasks, rather than forcing reliance on a single-band representation. Formally, given a multi-channel ExG signal $X_c \in \mathbb{R}^T$, where X_c is from channel c over T time steps, we apply the N_F filters to obtain band-specific signals $X_{f,c} \in \mathbb{R}^T$, where $f \in \{1, \ldots, N_F\}$. Each $X_{f,c}$ retains only the components within band f, serving as the foundation for subsequent tokenization.

Patch Segmentation and Tokenization. The band-specific signal $X_{f,c}$ is segmented into non-overlapping patches, *i.e.*, $\mathbf{p}_{f,c,l} \in \mathbb{R}^w$, where w denotes the patch size and l indexes its temporal position. This segmentation improves computational efficiency and facilitates modeling long-range temporal dependencies (Nie et al., 2023). Together, each patch is contextualized by three dimensions: frequency f, channel c, and time l, forming a structured 3D representation of the ExG input. Finally, each patch $\mathbf{p}_{f,c,l}$ is projected into a latent embedding $e_{f,c,l} \in \mathbb{R}^d$ through a learnable tokenizer, where d denotes the embedding dimension. Specifically, we use a single linear layer shared across all tokens to map each patch into a fixed-dimensional embedding space.

3.2 Encoder

We adopt Bidirectional-Mamba for its strong ability to capture long-range sequential dependencies (Schiff et al., 2024; Shams et al., 2024). A detailed analysis of its effectiveness compared with standard Transformers on ExG data is provided in Appendix H. Furthermore, since PiMT introduces an additional frequency dimension that increases sequence length, Mamba is especially suitable: it achieves linear-time complexity in sequential modeling, whereas Transformers suffer from quadratic complexity (Gu & Dao, 2024).

To fully leverage the rich structure of multi-channel ExG signals, we organize the input tokens along three axes, *i.e.,* frequency, channel, and time, in a fixed scanning sequence. Based on empirical validation, we adopt a frequency-first (f), channel-second (c), and time-last (l) ordering scheme. To achieve this, the embeddings $e_{f,c,l}$ are flattened into a sequence following the $(f \times c \times l)$ order and then passed into the encoder. The encoder produces contextualized representations $\mathbf{z}$, which are subsequently fed into the downstream task heads.

3.3 Pre-training from Free-living ExG Data

Most existing ExG models are trained on lab-controlled datasets using task-specific designs, limiting their ability to generalize to real-world conditions. In contrast, ExG signals collected in free-living environments provide richer diversity and broader coverage of human activities, enabling models to learn more robust and general-purpose representations. To exploit this free-living unlabeled data, we adopt a self-supervised pre-training based on reconstruction objectives. Our design choice is motivated by prior work showing that reconstruction outperforms alternatives, *i.e.*, contrastive learning, when training physiological foundation models on unlabeled data (Narayanswamy et al., 2025). To ensure robust feature extraction, we define six distinct reconstruction tasks, each paired with a dedicated decoder, which are jointly used to train the encoder E_θ.

Autoencoding: Given a sequence of patches $\mathbf{p}$ generated from a raw signal $\mathbf{x}$, the encoder E_θ maps it into a latent representation $\mathbf{z} = E_\theta(\mathbf{p})$. The decoder D_ϕ^{AE} reconstructs the original signal $\hat{\mathbf{p}}^{\text{AE}} = D_\phi^{\text{AE}}(\mathbf{z})$. This task encourages the encoder to capture temporal features while reducing noise.

Masked Reconstruction: To enforce contextual learning, we employ masked reconstruction (Devlin, 2018; He et al., 2022). The patches $\mathbf{p}$ are partially masked along time, channel, and frequency dimensions, producing a corrupted version $\mathbf{p}^{\text{mask}}$. The encoder processes the masked input to yield $\mathbf{z}^{\text{mask}} = E_\theta(\mathbf{p}^{\text{mask}})$. The decoder D_ϕ^{MR} recovers the original signal, generating $\hat{\mathbf{p}}^{\text{MR}} = D_\phi^{\text{MR}}(\mathbf{z}^{\text{mask}})$.

Frequency Domain Feature Reconstructions: To capture spectral information, we incorporate two frequency-domain reconstruction tasks. We first apply the Fast Fourier Transform (FFT) to obtain amplitude $\mathbf{p}^{\text{A}}$ and phase $\mathbf{p}^{\text{P}}$. Two decoders, D_ϕ^{A} and D_ϕ^{P}, reconstruct these features from the encoded representation $\mathbf{z}$, producing $\hat{\mathbf{p}}^{\text{A}}$ and $\hat{\mathbf{p}}^{\text{P}}$, respectively. Specifically, the two decoders aim to recover the original frequency domain signals based on: $\hat{\mathbf{p}}^{\text{A}} = D_\phi^{\text{A}}(\mathbf{z})$ and $\hat{\mathbf{p}}^{\text{P}} = D_\phi^{\text{P}}(\mathbf{z})$, respectively.

Masked Frequency Domain Reconstructions: To enhance the model's capacity to infer spectral features from incomplete inputs, we apply the same frequency reconstruction tasks to masked input signals. Two additional decoders, D_ϕ^{MA} and D_ϕ^{MP}, reconstruct the amplitude and phase, producing $\hat{\mathbf{p}}^{\text{MA}}$ and $\hat{\mathbf{p}}^{\text{MP}}$ from $\mathbf{z}^{\text{mask}}$. The new decoders aim to recover the original frequency domain signals based on: $\hat{\mathbf{p}}^{\text{A}} = D_\phi^{\text{MA}}(\mathbf{z}^{\text{mask}})$ and $\hat{\mathbf{p}}^{\text{P}} = D_\phi^{\text{MP}}(\mathbf{z}^{\text{mask}})$, respectively.

To jointly optimize the self-supervised objectives, we assign each task an independent decoder that reconstructs a specific aspect of the input signal, while sharing the encoder. Training is guided by mean absolute error (MAE) losses between the original and reconstructed signals. We combine these losses into a single objective by weighting each task-specific loss with a coefficient λ, which controls its relative contribution. The overall reconstruction loss is thus a weighted sum across all tasks. Each decoder is implemented as a lightweight MLP designed to reconstruct the target sequence. The λ values were empirically selected, and details are provided in Appendix G.

3.4 Fine-tuning

Building on the representations learned from free-living data, we fine-tune the model to diverse downstream tasks (*e.g.*, sight, hearing, taste, touch, and smell). The pre-trained encoder serves as a feature extractor, while task-specific decoders are trained on labeled data. For classification tasks, we aggregate the encoder's patch-wise outputs into a fixed-length feature vector via mean pooling. The vector is then passed through a fully connected classification decoder trained with cross-entropy loss. For continuous regression tasks, such as gaze tracking, we employ a linear decoder operating at the patch level to generate sequential outputs, which are then aggregated into the final prediction. The model is optimized using a standard regression loss.

4 DailySense: Free-living ExG Data across Five Human Senses

We built *DailySense*, an ExG dataset collected through earphones, designed to enhance the ExG dataset diversity beyond traditional lab-controlled settings and to enable benchmarking across a broad spectrum of daily life tasks. DailySense includes data from 22 participants, including 50 hours of unlabeled free-living recordings and 20 hours of labeled task-specific data spanning the five fundamental human senses.[1] Compared with existing lab-based ExG benchmarks, which often involve a similar number of participants but shorter recording durations (*e.g.*, DREAMER includes 23 participants with approximately 20 hours of data), DailySense provides a more diverse and comprehensive dataset, laying a stronger foundation for generalizable ExG representation learning.

Data Collection Platform. To collect free-living ExG data, we developed *NeuroBuds*, an earphone-integrated ExG sensing prototype. Unlike traditional head-mounted ExG devices that are bulky and expensive ($10,000–$50,000), NeuroBuds employs an earhook-style form factor that is low-cost and compact, and well-suitable for scalable, long-term daily use. The device integrates amplification, digitization, onboard storage, and wireless transmission into a lightweight PCB (4.2 cm × 2.2 cm, 20 g, $80). During data collection, participants wore earphones with integrated electrodes and carried the PCB as shown in Figure 2. The around-the-ear electrodes can then capture ExG signals, including EEG (sites T7–T10, FT7–FT10, TP7–TP10), auricular electrodes for EMG, and lateral electrodes for EOG, providing cognitive, muscular, and ocular coverage. We detail hardware design in Appendix A, and the physiological rationale behind the design and signal quality for each modality in Appendix B.

[1] Our data collection was approved by the Institutional Review Board (IRB). We are also currently working with our legal team to determine the possibility of publicly or conditionally sharing the dataset.

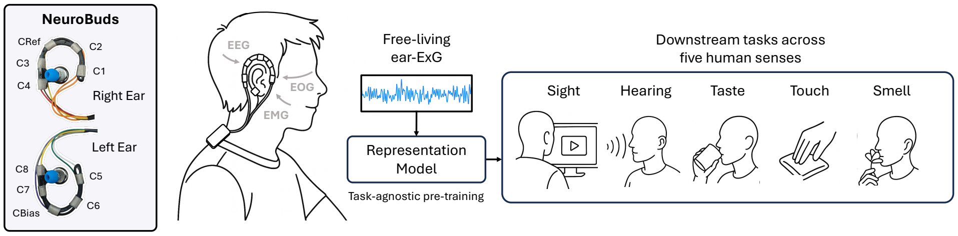

Figure 2: Overview of *DailySense* dataset. Using our earphone-based ExG analysis device, NeuroBuds, we collect free-living unlabeled data for task-agnostic pre-training and labeled data spanning five human senses, serving as benchmarks for downstream tasks.

Data Collection Protocol. DailySense contains (i) unlabeled data of daily life and (ii) labeled data spanning the five human senses. A total of 22 participants (ages 23∼62, 16 men, 6 women) wore NeuroBuds during daily routines without restrictions, performing natural activities such as walking, eating, talking, and facial movements. This produced 50 hours of free-living ear-ExG recordings. Furthermore, we curated six benchmark tasks covering the five human senses: (1) gaze tracking, (2) interest inference while watching videos (sight), (3) interest inference while listening to audio (hearing), (4) surface texture classification (touch: rough vs. smooth), (5) taste classification (sweet vs. sour), and (6) smell classification (floral vs. sour). Data were collected in a task-controlled environment with up to seven participants per task, producing 20 hours of labeled recordings. All experimental protocols followed prior brain-computer interface studies (Amini et al., 2022; Iravani et al., 2019; Namazi & Kulish, 2016; Vo et al., 2023; Xia et al., 2023). Further experimental details are provided in Appendix C.

Data Processing. Following established protocols (Jiang et al., 2024), we applied minimal preprocessing steps, including notch filtering (50/60 Hz), resampling to 200 Hz, and normalization. The signals were segmented into non-overlapping 4-second windows. To improve the robustness of the model, we augmented the training data with small additive noise. Appendix D provides visualizations of the collected signals.

5 Experiments

5.1 Experimental Setup

Baselines. We benchmarked PiMT against baselines including a traditional machine learning model (SVM), ExG-specific neural architectures (DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), and EEGConformer (Song et al., 2022)), and general-purpose time-series models (Time-Series Transformer (TST) (Zerveas et al., 2021), PatchTST (Nie et al., 2023), and Bidirectional-Mamba (Zhu et al., 2024)). Among them, we emphasize PatchTST as a strong baseline—an advanced masked reconstruction model built on a Transformer backbone that independently models each ExG channel and excels at capturing long-range temporal modeling.

Benchmark Datasets. We evaluated PiMT on DailySense along with four widely used ExG benchmarks: DREAMER (Katsigiannis & Ramzan, 2018) and SEED (Zheng & Lu, 2015) for emotion recognition, Sleep-EDF (Kemp et al., 2000) for sleep stage classification, and BCI Competition IV 2b (Leeb et al., 2008) for motor imagery. Dataset details are provided in Appendix E.

Implementation Details and Metrics. Our implementation consists of two primary stages: (i) pre-training on free-living data and (ii) task-specific fine-tuning. We first pre-train PiMT on 50 hours of free-living data using a masked reconstruction objective, then fine-tune it on each downstream dataset using an 8:2 train-test split for each participant. For evaluation, we report the mean squared error (MSE) in angular (°) units for gaze tracking and macro-averaged F1-scores for all classification tasks. All tasks are repeated three times with different random seeds, and we report the corresponding standard deviation. Additional training details, resource specifications, and hyperparameter tuning are provided in Appendix F and Appendix G.

Table 1: Performance of PiMT and baselines on DailySense. Classification results are in F1-score, and gaze tracking performance is in angular error. Best results are highlighted in **bold**.

	Classification (↑)						Regression (↓)
Method	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
Without pre-training							
SVM	0.665 ± 0.078	0.610 ± 0.126	0.556 ± 0.114	0.554 ± 0.107	0.510 ± 0.084	0.579	6.60° ± 1.27°
EEGNet	0.753 ± 0.137	0.712 ± 0.149	0.709 ± 0.088	0.643 ± 0.097	0.669 ± 0.063	0.697	6.52° ± 1.24°
DeepConvNet	0.680 ± 0.174	0.706 ± 0.129	0.633 ± 0.074	0.638 ± 0.075	0.636 ± 0.062	0.659	7.04° ± 1.31°
TST	0.773 ± 0.125	0.705 ± 0.104	0.731 ± 0.068	0.669 ± 0.116	0.667 ± 0.096	0.709	6.54° ± 1.30°
PatchTST	0.771 ± 0.146	0.749 ± 0.113	0.731 ± 0.092	0.686 ± 0.119	0.681 ± 0.049	0.724	6.47° ± 1.28°
EEGConformer	0.738 ± 0.127	0.752 ± 0.141	0.688 ± 0.062	0.678 ± 0.102	0.670 ± 0.047	0.705	6.53° ± 1.28°
Bidirectional-Mamba	0.820 ± 0.102	0.858 ± 0.113	0.733 ± 0.060	0.762 ± 0.101	0.722 ± 0.067	0.779	6.53° ± 1.16°
PiMT (Ours)	**0.858** ± **0.084**	**0.885** ± **0.125**	**0.790** ± **0.077**	**0.807** ± **0.113**	**0.753** ± **0.069**	**0.819**	**6.11°** ± 1.20°
With pre-training							
PatchTST	0.807 ± 0.146	0.786 ± 0.146	0.697 ± 0.099	0.700 ± 0.131	0.670 ± 0.082	0.732	6.42° ± 1.33°
PiMT (Ours)	**0.964** ± **0.028**	**0.961** ± **0.038**	**0.801** ± **0.064**	**0.860** ± **0.118**	**0.793** ± **0.069**	**0.876**	**6.00°** ± **1.13°**

Table 2: F1-score on four public ExG benchmarks across various tasks.

Baselines	DREAMER	SEED	Sleep-EDF	BCI Competition IV 2b
PatchTST	0.889 ± 0.085	0.756 ± 0.093	0.810 ± 0.005	0.657 ± 0.008
Bidirectional-Mamba	0.875 ± 0.090	0.750 ± 0.107	0.796 ± 0.002	0.646 ± 0.015
PiMT (Ours)	**0.910** ± **0.074**	**0.820** ± **0.121**	**0.822** ± **0.006**	**0.693** ± **0.004**

5.2 Evaluation on DailySense

Table 1 shows the F1-scores of PiMT compared with the baselines on DailySense. Overall, the results demonstrate that ear-ExG combined with PiMT can effectively capture five human senses, achieving up to 81.9% F1-score and as low as 6.11°gaze error, even without pre-training. Notably, PiMT outperformed all baselines, achieving a 4% improvement in F1-score and a 0.41°reduction in gaze error. We attribute this generalizability to PiMT's ability to interpret task-relevant frequency bands, a capability essential for handling diverse ExG-based tasks characterized by heterogeneous frequency-band features. We also observed that the Mamba-based backbone contributed significantly to performance gains; detailed comparisons against Transformer-based variants are reported in Appendix H.

Effect of Pre-training on Free-living Data. A key advantage of NeuroBuds is its ability to facilitate effortless collection of ExG signals, enabling large-scale pre-training. We evaluated the performance of PiMT when pre-trained on free-living data and compared it with PatchTST, which is the only baseline with a tailored pre-training strategy. As shown in Table 1, pre-training improved the average F1-score of PiMT from 81.9% to 87.6%. Similarly, PatchTST improved from 72.4% to 73.2%, whereas PiMT demonstrated a substantially larger gain. These results highlight the effectiveness of both our hardware-enabled free-living data collection and our reconstruction-based pre-training framework. Further analysis of our reconstruction-based pre-training is provided in Appendix I.

5.3 Evaluation on Public Benchmarks

To validate the generalizability of PiMT beyond the *DailySense* dataset, we evaluated it on four widely used public benchmarks covering diverse ExG tasks. We compared against two strongest baselines, PatchTST (Nie et al., 2023) and Bidirectional-Mamba (Zhu et al., 2024). As shown in Table 2, PiMT consistently outperformed all baselines across all datasets. Overall, these results demonstrate that our training strategy learns general-purpose ExG representations through physiology-informed multiband tokenization, leading to robust performance across diverse benchmarks. This confirms generalization beyond the self-collected DailySense dataset to real-world benchmarks, which underscores the potential of NeuroBuds as a unified framework for generalizable ExG representation learning.

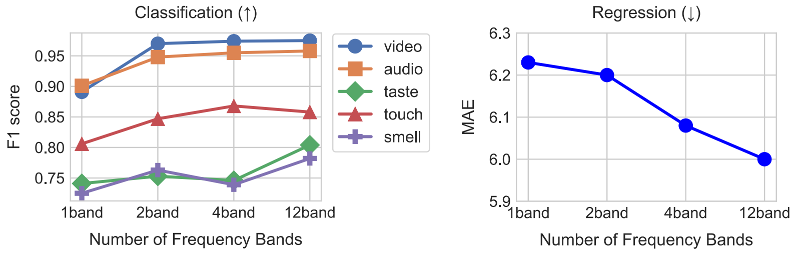

Figure 3: Comparison of different ExG tokenization strategies: 1-band (0.1∼75 Hz), 2-band (0.1∼15 Hz and 15∼75 Hz), 4-band (0.1∼5 Hz, 5∼15 Hz, 15∼35 Hz, and 35∼75 Hz), and our 12-band filter bank (described in Section 3.1).

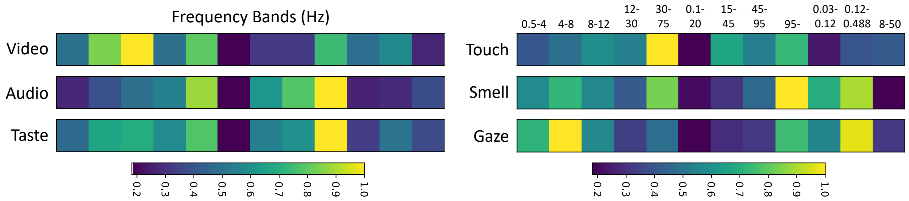

Figure 4: Saliency analysis demonstrating how the model dynamically captures task-relevant frequency bands via multi-band tokenization.

5.4 Effect of Multi-band Tokenization

To understand the impact of multi-band tokenization, we compared the model performance under different frequency-band tokenization strategies on DailySense. As shown in Figure 3, performance consistently improves as the number of bands increases. Our 12-band filter bank approach outperforms the 1-, 2-, and 4-band variants, achieving an average 4.6% F1-score gain on classification tasks and the lowest gaze-tracking error. These results suggest that fine-grained decomposition allows NeuroBuds to exploit subtle but physiologically meaningful spectral cues.

Saliency Analysis. To further understand the impact on the downstream task, we conducted a visual analysis of how different frequency bands contribute to each task. Figure 4 depicts saliency maps that highlight the contribution of each frequency-band token during inference. Importantly, we observed clear task-relevant activation patterns: (i) gaze and video tasks, which are closely linked to eye movements, exhibited strong activation in low-frequency bands (Plöchl et al., 2012), and (ii) touch, taste, smell, and auditory interest classification emphasized high-frequency components, consistent with somatosensory beta–low-gamma activity involved in processing external stimuli (Bauer et al., 2006) and peri-auricular EMG spectra reflecting near-ear muscle movements (Goncharova et al., 2003). These findings demonstrate that PiMT enables the model to dynamically focus on task-relevant frequency components without explicit supervision. We stress that this property is essential for enabling generalizable ExG-based applications in daily-life scenarios using NeuroBuds.

5.5 Impact of Pre-training Data Scale

We examine how the scale of pre-training data influences representation quality and downstream task performance. To this end, we randomly split the pre-training corpus into 80% training and 20% held-out test data, and subsampled varying proportions of the training set. Figure 5 shows the test loss across epochs under varying training data scales. As expected, larger pre-training sets consistently produced lower losses, indicating that PiMT benefits from additional data and scales effectively.

Figure 6 presents downstream results on DailySense. For classification tasks, average performance saturates around 30% of the pre-training data, suggesting diminishing returns beyond this point. In contrast, gaze regression continues to improve up to 50%, highlighting task-dependent benefits of larger pre-training scales. Overall, these findings suggest that while some tasks quickly reach saturation, others continue to benefit from larger-scale pre-training. Importantly, the consistent loss reductions in Figure 5 confirm that PiMT can effectively exploit additional data, underscoring its promise as a general-purpose ExG representation model.

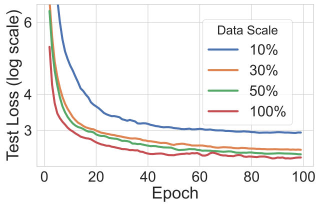

Figure 5: Test loss across different pre-training data scales.

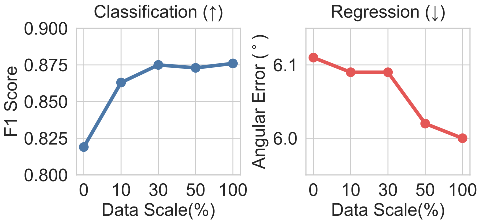

Figure 6: Downstream performance of PiMT with varying amounts of pre-training data.

5.6 Further Analysis and Ablation study

On-device Deployment and Real-time Analysis (Appendix K). We evaluated the runtime overhead of PiMT (with transformer as a backbone) on a commercial smartphone (Samsung Galaxy S24), which serves as a representative companion device for earphones. The model achieved efficient runtime performance with an average inference latency of 25 ms, memory usage of 266 MB (3.1%), and CPU utilization of 20.3%.

Leave-One-Subject-Out evaluation. To further assess cross-participant generalization, we conducted leave-one-subject-out (LOSO) experiments on DailySense. In DailySense, some participants contributed to both free-living pretraining data and task-specific data. Therefore, in LOSO, for each target user, their data was excluded from pretraining and used only for testing. Despite this constraint, PiMT achieved performance comparable to full pre-training (Figure 7), confirming its ability to generalize effectively to unseen users.

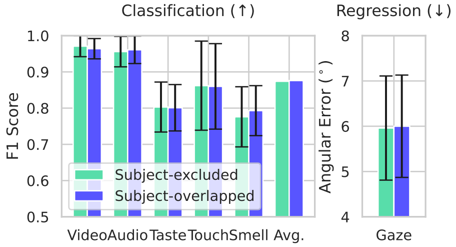

Figure 7: LOSO Performance when the target subject's data is either included or excluded.

Ablation Study (Appendix I, J). We performed an ablation study to investigate the contribution of each pre-training component to representation learning. As shown in Table 4, the complete model achieved the highest overall performance. We observed consistent performance improvements as additional components were incorporated, indicating the complementary benefits of each module. Furthermore, Table 5 shows that a patch size of 0.5 seconds yielded the best downstream performance compared to alternative configurations.

5.7 Discussion

We acknowledge several limitations of our current approach. Like many existing ExG frameworks, our methods are constrained by the limited number of subjects and challenges in personalized generalization. While the number of subjects is comparable to prior benchmarks, DailySense provides over 70 hours of high-resolution (1000 Hz) recordings across both free-living and task-specific conditions. Our LOSO experiments show promising cross-subject generalization, but performance drops when training and testing on different users highlight the persistent challenge of personalized modeling. Nonetheless, by demonstrating the effectiveness of earphone-derived free-living ExG data for representation learning, NeuroBuds provides a scalable path toward broader population-level data collection and establishes the foundation for a more generalizable framework in future work.

6 Conclusion

We tackle two long-standing barriers in ExG analysis: (i) the lack of diverse, real-world data and (ii) the reliance on task-specific model designs. To address data diversity, we developed NeuroBuds, an earphone-based sensing prototype, and curated DailySense, the first dataset with 50 hours of free-living recordings and 20 hours of task-specific ExG data spanning all five human senses. To overcome task-specificity, we propose Physiology-informed Multi-band Tokenization (PiMT), which decomposes ExG signals into structured tokens across 12 canonical sub-bands aligned with distinct physiological modalities. Combined with reconstruction-based pre-training on free-living data, PiMT learns robust, task-agnostic representations that generalize across tasks. Evaluations on DailySense and four public benchmarks demonstrate that PiMT consistently outperforms state-of-the-art baselines. Together, these contributions push ExG research beyond narrow, lab-constrained applications toward generalizable and real-world physiological sensing. Looking ahead, this work opens new opportunities in personalized health monitoring, cognitive interfaces, and scalable everyday sensing powered by ExG platforms.

Ethics Statement

Our data collection was approved by the Institutional Review Board (IRB), ensuring the safety of both the participants and the device prototype used in the study. For the other experiments, we used publicly available datasets, which were used in accordance with their intended purposes. There is no ethical issue with this paper.

Reproducibility Statement

Our Physiology-informed Multi-band Tokenization approach can be reproduced using the filter bank described in Section 3.1. Comprehensive experimental and implementation details are provided in Section 5, Appendix F, and Appendix G.

Usage of Large Language Models

Large Language Models (LLMs) were used solely for polishing the writing of this paper.

References

Tobii 4c eye tracker. `https://help.tobii.com/hc/en-us/articles/213414285-Specifications-for-the-Tobii-Eye-Tracker-4C`, 2016.

Garry T. Allison and Takayuki Fujiwara. The relationship between emg median frequency and low frequency band amplitude changes at different levels of muscle capacity. *Clinical biomechanics*, 17 6:464–9, 2002. URL `https://api.semanticscholar.org/CorpusID:10242245`.

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.

Ali Amini, Karim Faez, and Mahmood Amiri. Surface roughness classification in dynamic touch using eeg signals. In *2022 30th International Conference on Electrical Engineering (ICEE)*, pp. 205–209, 2022. doi: 10.1109/ICEE55646.2022.9827406.

Andrea Apicella, Pasquale Arpaia, Giovanna Mastrati, and Nicola Moccaldi. Eeg-based detection of emotional valence towards a reproducible measurement of emotions. *Scientific Reports*, 11(1): 21615, 2021.

Markus Bauer, Robert Oostenveld, Maarten Peeters, and Pascal Fries. Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas. *Journal of Neuroscience*, 26(2):490–501, 2006.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Matthieu Duvinage, Thierry Castermans, Mathieu Petieau, Thomas Hoellinger, Guy Cheron, and Thierry Dutoit. Performance of the emotiv epoc headset for p300-based applications. *Biomedical engineering online*, 12:1–15, 2013.

Ola Elsaid and Michael Labanowski. Physiology, sleep stages. 2017. URL `https://api.semanticscholar.org/CorpusID:79564518`.

Cunhang Fan, Jinqin Wang, Wei Huang, Xiaoke Yang, Guangxiong Pei, Taihao Li, and Zhao Lv. Light-weight residual convolution-based capsule network for eeg emotion recognition. *Adv. Eng. Inform.*, 61(C), August 2024. ISSN 1474-0346. doi: 10.1016/j.aei.2024.102522. URL `https://doi.org/10.1016/j.aei.2024.102522`.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. Neuript: Foundation model for neural interfaces. *arXiv preprint arXiv:2510.16548*, 2025.

Iffat Farhana, Jungpil Shin, Shabbir Mahmood, Md Rabiul Islam, and Md Khademul Islam Molla. Emotion recognition using narrowband spatial features of electroencephalography. *IEEE Access*, 11:44019–44033, 2023.

Pengxuan Gao, Tianyu Liu, Jia-Wen Liu, Bao-Liang Lu, and Wei-Long Zheng. Multimodal multi-view spectral-spatial-temporal masked autoencoder for self-supervised emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1926–1930. IEEE, 2024.

Evgenia Gkintoni, Anthimos Aroutzidis, Hera Antonopoulou, and Constantinos Halkiopoulos. From neural networks to emotional networks: a systematic review of eeg-based emotion recognition in cognitive neuroscience and real-world applications. *Brain Sciences*, 15(3):220, 2025.

Irina I Goncharova, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Emg contamination of eeg: spectral and topographical characteristics. *Clinical neurophysiology*, 114(9): 1580–1593, 2003.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Ahmed Hamid, Katherine Gagliano, Safwanur Rahman, Nikita Tulin, Vincent Tchiong, Iyad Obeid, and Joseph Picone. The temple university artifact corpus: An annotated corpus of eeg artifacts. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2020.

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.

Behzad Iravani, Artin Arshamian, Kathrin Ohla, Donald A. Wilson, and Johan N. Lundström. Non-invasive recording from the human olfactory bulb. *Nature Communications*, 11, 2019. URL `https://api.semanticscholar.org/CorpusID:195391468`.

Weibang Jiang, Liming 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. URL `https://openreview.net/forum?id=QzTpTRVtrP`.

Britton JW, Frey LC, and Hopp JLet al. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. Chicago: American Epilepsy Society, 2016.

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, 2018. doi: 10.1109/JBHI.2017.2688239.

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. doi: 10.1109/10.867928.

Daniel R Kramer, Krista Lamorie-Foote, Michael Barbaro, Morgan B Lee, Terrance Peng, Angad Gogia, George Nune, Charles Y Liu, Spencer S Kellis, and Brian Lee. Utility and lower limits of frequency detection in surface electrode stimulation for somatosensory brain-computer interface in humans. *Neurosurgical focus*, 48(2):E2, 2020.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, jul 2018. doi: 10.1088/1741-2552/aace8c. URL `https://dx.doi.org/10.1088/1741-2552/aace8c`.

Robert Leeb, Clemens Brunner, G Müller-Putz, A Schlögl, and GJGUOT Pfurtscheller. Bci competition 2008–graz data set b. *Graz University of Technology, Austria*, 16:1–6, 2008.

Louise R Manfredi, Hannes P Saal, Kyler J Brown, Mark C Zielinski, John F Dammann III, Vicky S Polashock, and Sliman J Bensmaia. Natural scenes in tactile texture. *Journal of neurophysiology*, 111(9):1792–1802, 2014.

Manuel Merino, Octavio Rivera, Isabel Gómez, Alberto Molina, and Enrique Dorronzoro. A method of eog signal processing to detect the direction of eye movements. In *2010 First International Conference on Sensor Device Technologies and Applications*, pp. 100–105. IEEE, 2010.

Hamidreza Namazi and Vladimir V. Kulish. Fractal based analysis of the influence of odorants on heart activity. *Scientific Reports*, 6, 2016. URL `https://api.semanticscholar.org/CorpusID:14622292`.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, shun liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=yb4QE6b22f`.

Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-kashani, Ann C. Halbower, and Tam Vu. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, SenSys '16, pp. 230–244, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342636. doi: 10.1145/2994551.2994562. URL `https://doi.org/10.1145/2994551.2994562`.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Jbdc0vTOcol`.

Ernst Niedermeyer and Fernando Lopes da Silva (eds.). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 5 edition, 2005. ISBN 0781751268.

Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2 edition, 2006. ISBN 019505038X.

Michael Plöchl, José P Ossandón, and Peter König. Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in human neuroscience*, 6:278, 2012.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: bi-directional equivariant long-range dna sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel, Christoph Baumgartner, Bin He, and Sándor Beniczky. The standardized eeg electrode array of the ifcn. *Clinical neurophysiology*, 128(10):2070–2077, 2017.

Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

Siavash Shams, Sukru Samet Dindar, Xilin Jiang, and Nima Mesgarani. Ssamba: Self-supervised audio representation learning with mamba state space model. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1053–1059. IEEE, 2024.

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, 2022.

ESC/NASPE Task Force. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996. doi: 10.1161/01.CIR.93.5.1043.

Hoang-Thuy-Tien Vo, Thi-Nhu-Quynh Nguyen, Do Duc Cuong, and Tuan Van Huynh. Classification taste-eeg signals using base neural network. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 107–111, 2023. doi: 10.1109/RIVF60135.2023.10471803.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36314–36341. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/3fe2a777282299ecb4f9e7ebb531f0ab-Paper-Conference.pdf`.

Xiuxin Xia, Yuchao Yang, Yan Shi, Wenbo Zheng, and Hong Men. Decoding taste information in human brain: A temporal and spatial reconstruction data augmentation method coupled with taste eeg, 2023. URL `https://arxiv.org/abs/2307.05365`.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2024.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A NeuroBuds Hardware Design

To enable large-scale, in-the-wild ExG data collection, we built an earphone-based sensing platform consisting of two main components:

Earphone-Shaped Sensing Array: To adopt a earhook-style form factor, We use a commercial earphone (Powerbeats PB123) as the backbone, and wrap conductive tape around the frame to form electrodes. Each side includes five electrodes: the top ones on the left and right act as bias and reference, while the remaining eight serve as signal channels.

Lightweight Processing Board: We design a custom printed circuit board (PCB) integrating signal amplification, digitization, wireless transmission, and onboard storage:

- **Amplification:** An bio-amplifier chip (ADS1299) and the front-end circuit support 8-channel ExG signal conditioning.
- **Digitization and Control:** An ESP32 microcontroller handles A/D conversion, peripheral control, and real-time streaming.
- **Wireless Streaming:** Microcontroller's built-in Wi-Fi/BLE enables direct transmission to phones or PCs for data collection or real-time on-device inference.
- **Storage:** A microSD slot supports continuous onboard logging.

To minimize size and weight without compromising signal integrity, we adopted highly integrated chips (ADS1299, ESP32), and designed a compact 6-layer PCB with dense layout of components on both sides to further reduce footprint. The resulting design measures just 4.2cm × 2.2cm and weighs only 20g, which is significantly smaller than existing COTS systems like OpenBCI (6.1cm × 6.1cm, 80g) or OpenEarable (5.7cm × 3cm, only support 2 ExG channel).

During use, the board is enclosed in a 3D-printed case and connected to the sensing array via Dupont wires. Users can wear the platform unobtrusively during daily activities, with the board placed in a pocket or wore on the body, enabling free-living data collection.

B Quality of EEG, EOG, and EMG Signal

Our electrode placement around the ear was carefully designed to capture EEG, EOG, and EMG signals while maintaining a compact and unobtrusive form factor. Below, we outline the physiological rationale and supporting evidence for each modality.

EEG: The electrodes align with standard around-the-ear EEG channels, *i.e.*, T7–T10, FT7–FT10, and TP7–TP10 in the 10-10 EEG system (Seeck et al., 2017). The strong classification performance on cognitive tasks demonstrates that our recordings contain reliable EEG activity.

EMG: Electrodes positioned on auricular muscles capture EMG signals linked to facial expressions. We further validated this by recording deliberate facial movements (*e.g.*, blinking, biting), which produced distinct EMG-specific patterns.

EOG: Electrodes placed on both sides of the head enable strong horizontal EOG capture, with partial vertical EOG sensitivity due to vertical displacement. Eye movement patterns (0.1–5 Hz) are clearly observed in Appendix D, and our gaze tracking accuracy (within 6.15 degrees as shown in Appendix I) further supports the presence of robust EOG signals.

This electrode configuration enables simultaneous acquisition of EEG, EMG, and EOG signals, providing a rich multimodal ExG dataset while preserving wearability for daily use.

C DailySense Data Collection Protocol

In this section, we describe the detailed protocol of our six task-specific sensory experiments. The experimental tasks included:

- **Gaze Tracking:** Participants were seated 60 cm from a 13.5-inch laptop (model: Surface Book 2, display size: 3000 x 2000 pixels, vertical refresh rate: 59 Hz). This task evaluated whether EXG signals could accurately track gaze positions. The error was quantified as the angular difference between the ExG-based gaze estimation and the ground truth obtained from a Tobii eye tracker (tob, 2016) (model: Tobii 4C Eye Tracker).
- **Auditory and Video Interest Inference:** Inspired by SEED and DREAMER datasets (Zheng & Lu, 2015; Katsigiannis & Ramzan, 2018), this experiment explored the correlation between ExG signals and engagement with visual/auditory stimuli. Participants were asked to watch or listen to video clips. After each session, they rated their interest level. Each participant watched/listened to six stimulus clips, each lasting six minutes. The goal is to classify the participant's emotional state every four seconds based on ExG responses.
- **Surface Texture Classification (Touch Perception):** Participants interacted with different textured surfaces to analyze ExG responses to tactile stimuli (Amini et al., 2022). Each participant rubbed either a rough or smooth surface for 60 continuous seconds, repeating this process 10 times for each texture. The goal is to classify the participant's touch perception every four seconds.
- **Taste Classification:** This experiment assessed ExG responses to different taste profiles (sweet vs. sour). Participants sipped a liquid and held it in their mouth for 20 seconds (Vo et al., 2023; Xia et al., 2023). To prevent cross-contamination, a 30-second rest period was enforced between different taste samples, allowing participants to rinse their mouths before proceeding to the next task. The task aims to classify the participant's taste perception every four seconds.
- **Smell Classification:** This task examined ExG signal responses to olfactory stimuli (Iravani et al., 2019; Namazi & Kulish, 2016). Participants inhaled pleasant and unpleasant odors, and the model was evaluated on its ability to distinguish between different scent categories.

Table 3 provides a comprehensive summary of the classification labels, stimulus materials, trial durations, number of sessions, and trial structures per participant for each task.

Table 3: Experimental Task Details

Task	Labels	Materials	Trial duration time	Total sessions	Total time	Rest time
Taste	Sweet	Chocolate milk	20 seconds	15	300 seconds	30 seconds
	Sour	Vinegar	20 seconds	15	300 seconds	30 seconds
Touch	Rough	Scent paper	1 minute	10	10 minutes	20 seconds
	Smooth	Silk	1 minute	10	10 minutes	20 seconds
Smell	Lavender	Lavender scent bag	20 seconds	15	300 seconds	30 seconds
	Sour	Vinegar	20 seconds	15	300 seconds	30 seconds
Video	Interesting	Comedy Clips	5 minutes	6	30 minutes	30 seconds
	Not-interesting	Lectures/Documentary	5 minutes	6	30 minutes	30 seconds
Audio	Interesting	Comedy podcast	5 minutes	6	30 minutes	30 seconds
	Not-interesting	Lectures	5 minutes	6	30 minutes	30 seconds

D VISUALIZATION OF ExG SIGNALS

We visualized the raw ExG signals measured using NeuroBuds and illustrated how they are transformed into multi-band tokens through bandpass filtering across different frequency ranges. Figure 8 shows the decomposition of the raw signal into twelve frequency bands, each of which is subsequently tokenized as part of our multi-band sequence. For gaze tracking, low-frequency bands (*e.g.*, EOG-overall and ECG-HF bands) exhibit clearer temporal patterns that align with EOG signals (Merino et al., 2010). In contrast, for tasks such as touch, low-frequency activity is less prominent, while informative features emerge in higher-frequency bands (Manfredi et al., 2014; Kramer et al., 2020).

In addition to evaluating ExG quality implicitly through downstream task performance, we also conducted a direct quantitative comparison between our earphone-based NeuroBuds prototype and a research-grade OpenBCI device. Specifically, we ran an eye-movement tracking experiment with two participants (approximately one hour of synchronized data) and computed Pearson correlations between NeuroBuds and OpenBCI channels. The average cross-system correlation reached 0.71 (statistically significant, $p < 0.001$), demonstrating that NeuroBuds capture ExG/EOG activity with high consistency relative to a laboratory-grade system. Beyond quantitative metrics, we also performed visualization analysis to assess overall signal similarity. Visual comparison of synchronized raw ExG signals shows that NeuroBuds and OpenBCI exhibit closely aligned temporal patterns, with similar waveform shapes, amplitudes, and drift trends throughout the recording as shown in Figure 9. These qualitative observations, together with the correlation analysis, further confirm that NeuroBuds produces ExG signals that closely match those from research-grade devices.

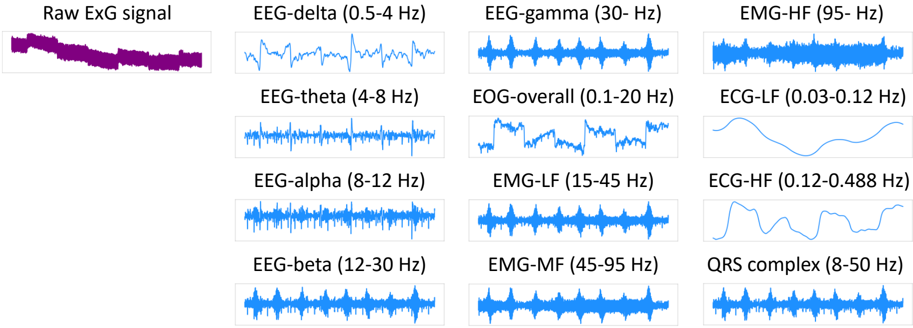

Figure 8: Raw ExG signals from DailySense dataset and their decomposition into twelve physiology-informed frequency bands for Multi-band Tokenization.

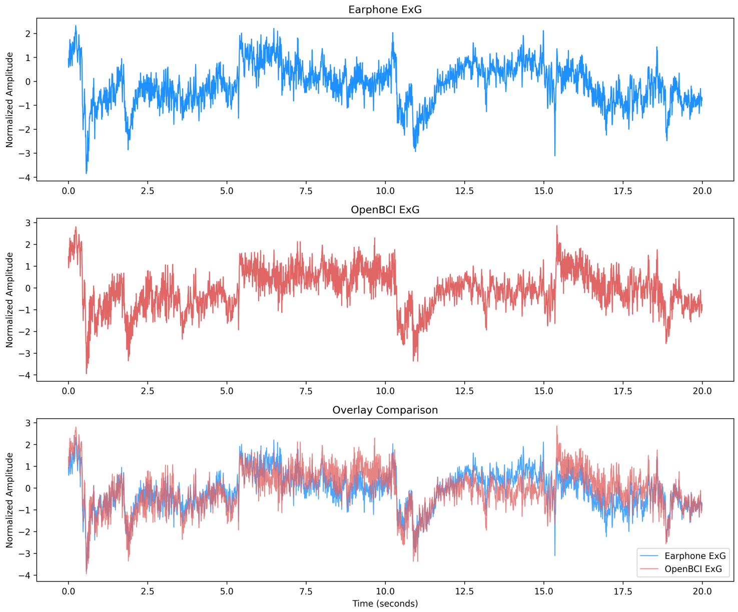

Figure 9: Visualization of synchronized ExG signals collected from NeuroBuds and a research-grade OpenBCI device with the Pearson correlation of 0.7586, demonstrating strong cross-system signal similarity.

E Benchmark datasets

We used four public benchmark datasets to further validate effectiveness of PiMT. For all datasets, we performed a random 80/20 split, assigning 80% of the data to training and the remaining 20% to testing. We followed established protocols Jiang et al. (2024) to preprocess the ExG signals.

DREAMER (Katsigiannis & Ramzan, 2018) is an EEG-based emotion recognition dataset collected from 23 participants while they watched 18 film clips designed to elicit different affective states. The dataset provides signals from electroencephalogram (14 channels at 128 Hz) and electrocardiogram (2 channels at 256 Hz). Each trial is annotated with self-reported valence, arousal, and dominance scores on a 5-point scale. We used the EEG recordings for classification of emotional states, formulating the task as binary classification based on dominance levels, where trials with dominance ≥ 3 were labeled as high and those with dominance < 3 as low.

SEED (Zheng & Lu, 2015) is a emotion recognition dataset with EEG recordings from 15 subjects. Participants watched 15 film clips (five positive, five neutral, and five negative) across three sessions. EEG was recorded from 62 channels using the ESI NeuroScan system at 1000 Hz. We used the downsampled signal at 200 Hz. The dataset provides trial-level emotion labels (positive, neutral, negative).

Sleep-EDF (Kemp et al., 2000) is a dataset used for sleep stage classification. It contains 197 whole-night polysomnographic recordings from both healthy subjects and patients with mild sleep difficulties. The EEG signals were recorded from two channels (Fpz–Cz and Pz–Oz) at 100 Hz, and the EOG signals were also sampled at 100 Hz. We used five-class scoring (W, N1, N2, N3, REM) for classification only using EEG signals.

BCI Competition IV 2b (Leeb et al., 2008) is a motor imagery dataset consisting of EEG recordings from 9 subjects across 5 sessions. Subjects were asked to perform left-hand and right-hand motor imagery tasks. Each session included multiple runs of motor imagery trials, with EEG recorded from three bipolar channels (C3, Cz, C4) at 250 Hz. The dataset defines two classes corresponding to left-hand and right-hand motor imagery.

F Implementation Details

For the Bidirectional-Mamba model, we used 8 layers with a hidden state size of 16 and an embedding dimension of 64. The decoder consisted of two fully connected layers, each with a hidden dimension of 64. For the baseline implementations, we tuned SVM using a grid search over $C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{0.01, 0.001, 0.0001\}$, and kernel types (*rbf*, *linear*, *poly*). For the other baselines, we followed their official implementations and performed grid searches to tune key hyperparameters, such as learning rate and batch size.

For the train/test split, we first segmented long sequences into 4-second windows and randomly shuffled them. We then applied a standard 80/20 division to construct the training and test sets.

G Hyper-parameter Tuning

Our implementation consists of two primary stages: representation learning and task-specific fine-tuning. During the representation learning stage, we pre-trained the model using mask and reconstruction objectives to learn robust representations transferable across various downstream tasks. The representation model is trained on the entire 40 hours of free-living data, after which it is fine-tuned on each specific task before final evaluation on the corresponding test set.

The weighting coefficients (λ) for the pretraining objectives were selected heuristically based on empirical observations. We initialized all λ values to 1 and monitored the convergence dynamics of individual loss terms. We found that the autoencoding loss ($\mathcal{L}_{\mathrm{AE}}$) and masked reconstruction loss ($\mathcal{L}_{\mathrm{MR}}$) converged more slowly than others; their weights were therefore increased to 2 to encourage balanced training. While we did not conduct a full hyperparameter sweep, this adjustment yielded more stable convergence without introducing instability.

To further optimize performance, we performed a grid search over key hyperparameters. Throughout the experiment, we used AdamW optimizer with 0.01 weight decay. During the pre-training stage, the batch size was fixed at 256, and the learning rate was scheduled from 0.01 to 0.001 using a cosine decay scheduler. For the backbone architecture, we adopted a bi-directionanl mamba model with 16 layers and a hidden dimension of 16. The masking ratio for the pretraining was fixed at 50%. During the fine-tuning stage, the batch size was fixed for all tasks, 10 for Gaze and 8 for the remaining tasks. The learning rate followed a cosine decay schedule from 0.001 to 0.00001. All experiments were repeated 3 times and the results are reported as the mean and standard deviation. All models were implemented using PyTorch, and the experimental evaluations were conducted on NVIDIA A100-SXM-80GB GPUs.

H Backbone Comparison: Mamba vs. Transformer

We adopt Bidirectional-Mamba (Zhu et al., 2024) as our backbone architecture, which has demonstrated state-of-the-art performance across various time-series tasks (Zerveas et al., 2021; Song et al., 2022). To evaluate its effectiveness on ExG signals, we compare it against Transformer-based architecture, PatchTST (Nie et al., 2023), which showed strong performance in our main evaluation (Section 5.2). For a fair comparison, we applied our Multi-band Tokenization to both Mamba- and Transformer-based models, with and without pre-training on the free-living dataset.

As shown in Figure 10, the Mamba-based model consistently outperformed the Transformer-based model under all settings, achieving a 6.4% improvement without pre-training and an 8.5% gain with pre-training. These results confirm that Mamba is a strong architectural choice for ExG signal modeling.

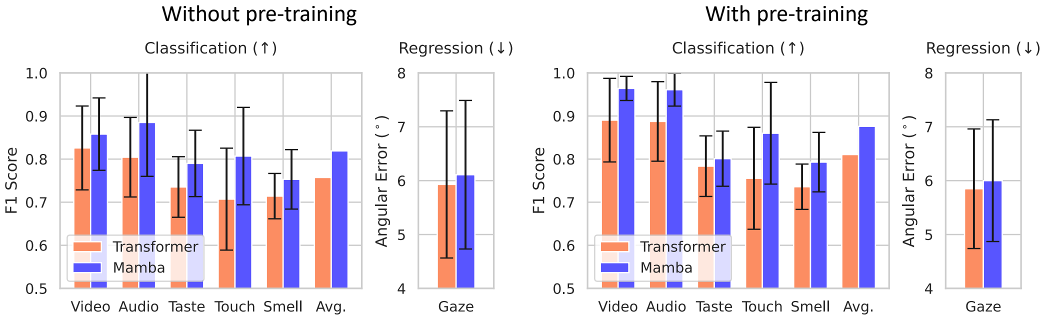

Figure 10: Comparison of Mamba and Transformer backbones.

I Effect of Pre-Training Components

Our pre-training framework on free-living data comprises six reconstruction-based tasks designed for unlabeled ExG signals: Autoencoding (AE), Masked Reconstruction (MR), (frequency) Amplitude Reconstruction (A), (frequency) Phase Reconstruction (P), Masked Amplitude Reconstruction (MA), and Masked Phase Reconstruction (MP). We assessed the contribution of each component by performing ablation experiments

Table 4 depicts the results. Although most ablation settings still achieve relatively strong performance, highlighting the overall effectiveness of pre-training, all are consistently lower than the full combination (0.876), confirming the benefit of jointly using all reconstruction tasks. The performance drops in individual ablations are modest, as complementary tasks help maintain efficacy. However, we observed task-specific sensitivities: for instance, MR, A, MA, and MP are particularly important for gaze tracking, where excluding them led to a notable increase in error. Meanwhile, AE, MR, and phase-related reconstructions strongly influence taste classification, where their removal caused meaningful performance degradation. These findings suggest that temporal- versus frequency-focused reconstruction tasks contribute differently depending on the task, reflecting the distinct feature requirements of each modality.

Table 4: Pre-Training ablation. Classification results are in F1-score, and gaze tracking performance is in angular error.

	Classification (↑)						Regression (↓)
Method	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
PiMT (Ours)	0.964 ± 0.028	0.961 ± 0.038	0.801 ± 0.064	0.860 ± 0.118	0.793 ± 0.069	0.876	6.00° ± 1.13°
w/o AE	0.970 ± 0.026	0.959 ± 0.042	0.806 ± 0.066	0.852 ± 0.125	0.778 ± 0.085	0.873	6.00° ± 1.09°
w/o MR	0.962 ± 0.030	0.956 ± 0.043	0.798 ± 0.058	0.859 ± 0.122	0.783 ± 0.082	0.872	6.10° ± 1.09°
w/o A	0.967 ± 0.028	0.955 ± 0.037	0.816 ± 0.063	0.850 ± 0.119	0.768 ± 0.079	0.871	6.19° ± 1.23°
w/o MA	0.965 ± 0.032	0.960 ± 0.040	0.818 ± 0.062	0.849 ± 0.124	0.774 ± 0.089	0.873	6.12° ± 1.20°
w/o P	0.979 ± 0.020	0.961 ± 0.041	0.803 ± 0.061	0.856 ± 0.127	0.764 ± 0.086	0.873	5.98° ± 1.15°
w/o MP	0.971 ± 0.030	0.960 ± 0.040	0.798 ± 0.062	0.857 ± 0.120	0.777 ± 0.082	0.873	6.15° ± 1.26°

Table 5: Ablation study on the impact of patch size. We report classification F1 scores (↑) and gaze regression error in degrees (↓). Our method (0.5 sec patch size) achieves the best overall balance across tasks.

	Classification (↑)						Regression (↓)
Patchsize	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
0.25 sec	0.977 ± 0.020	0.967 ± 0.036	0.776 ± 0.065	0.849 ± 0.124	0.741 ± 0.098	0.862	6.23° ± 1.26°
0.5 sec (Ours)	**0.964 ± 0.028**	**0.961 ± 0.038**	**0.801 ± 0.064**	**0.860 ± 0.118**	**0.793 ± 0.069**	**0.876**	**6.00° ± 1.13°**
1.0 sec	0.962 ± 0.034	0.945 ± 0.054	0.821 ± 0.063	0.836 ± 0.130	0.784 ± 0.088	0.870	6.10° ± 1.09°
2.0 sec	0.947 ± 0.039	0.932 ± 0.074	0.776 ± 0.101	0.823 ± 0.126	0.769 ± 0.105	0.850	6.24° ± 1.05°

J Impact of Patch Size

We discuss the impact of different patch sizes. Specifically, we selected the patch size empirically based on performance trends across tasks. A sensitivity study illustrating the effect of different patch sizes is presented in Table 5. A smaller patch size provides less contextual information for each classification window, which may limit performance. However, it benefits gaze regression, as the participant's gaze is more likely to remain fixed within a shorter temporal window. In contrast, larger patch sizes offer more temporal context for classification tasks but increase the likelihood of gaze shifts or overlapping signals from multiple classes, potentially degrading both classification and gaze estimation performance. We observed that a patch size of 0.5 seconds provides the best trade-off, yielding strong performance across both classification and regression tasks.

K Efficiency Analysis

Table 6: Runtime performance of PiMT on smartphone (Samsung Galaxy S24).

Metric	Value
Inference Latency	25 ms
Memory Usage	266 MB (3.6%)
CPU Usage	20.3%
Model Size (ONNX)	2.0 MB

We evaluated the runtime overhead of our method on a commercial smartphone (Samsung Galaxy S24), which serves as a practical companion device for earphone-based systems. Since NeuroBuds supports real-time data streaming via BLE, we consider a deployment scenario where inference is offloaded to the smartphone.

Because the Mamba architecture is not yet supported on Android and lacks corresponding hardware acceleration, we substituted Mamba with a Transformer of *equivalent architecture and parameter size* (e.g., number of layers, d_{model}). Prior work has shown that Transformers generally incur higher inference costs under comparable hardware acceleration (Gu & Dao, 2024). To preserve the core

Table 7: Performance of PiMT compared to PatchTST on the DailySense dataset under three data split settings: within-session, cross-session, and cross-subject.

	Classification (↑)						Regression (↓)
Method	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
Within-session							
PatchTST	0.807 ± 0.146	0.786 ± 0.146	0.697 ± 0.099	0.700 ± 0.131	0.670 ± 0.082	0.732	6.42° ± 1.33°
PiMT (Ours)	0.964 ± 0.028	0.961 ± 0.038	0.801 ± 0.064	0.860 ± 0.118	0.793 ± 0.069	0.876	6.00°± 1.13°
Cross-subject							
PatchTST	0.654 ± 0.047	0.595 ± 0.064	0.561 ± 0.047	0.553 ± 0.064	0.539 ± 0.052	0.580	7.07° ± 1.25°
PiMT (Ours)	0.612 ± 0.088	0.578 ± 0.082	0.593 ± 0.038	0.577 ± 0.092	0.571 ± 0.035	0.586	7.78° ± 0.95°
Cross-session							
PatchTST	0.658 ± 0.202	0.656 ± 0.157	0.695 ± 0.101	0.639 ± 0.097	0.611 ± 0.068	0.652	7.56° ± 1.33°
PiMT (Ours)	0.697 ± 0.249	0.698 ± 0.188	0.704 ± 0.106	0.763 ± 0.156	0.639 ± 0.146	0.700	6.98° ± 1.51°

algorithmic behavior of PiMT, we retained both the multi-band tokenization and the 3D positional embeddings.

The resulting models were exported to ONNX and evaluated using 4-second input sliding windows (200 Hz sampling, with the same preprocessing as in the main experiments). The measured runtime performance is summarized in Table 6.

Overall, the results indicate that inference can be executed in real time at up to 40 Hz with minimal resource consumption. Preprocessing operations such as filtering and windowing can be performed directly on the NeuroBuds board. In addition, given that Mamba has been reported to offer 5× higher throughput than Transformers, we anticipate supporting on-device inference with even lower overhead.

L Generalization to Unseen Sessions and Users

We also tested generalization to unseen conditions during testing in *cross-subject* and *cross-session* scenarios. Cross-subject involves training and testing on different users, while cross-session assumes the model is tested on a different session from the same user, introducing a temporal shift. These domain shifts are open challenges for ExG-based tasks, with prior work (Fan et al., 2024) reporting over a 30% drop in accuracy. As shown in Table 7, our method also experienced a performance drop under the cross-subject setting (58.6%). In the cross-session setting, NeuroBuds showed stronger robustness, achieving 70.0% compared to PatchTST's 65.2%. Generalization to unseen conditions remains a open challenge and is a focus of our future research. Nevertheless, we believe that large-scale data collection enabled by the daily usability of NeuroBuds can play a key role in improving robustness in real-world applications.

M Statistical Significance

To assess whether our method provides statistically significant improvements over the baselines, we conduct paired Wilcoxon signed-rank tests on DailySense. Each task contains 6–9 participants, and for every participant we train each model with three random seeds. For a given seed, all models share the exact same train–test split; because the split strongly influences performance, constructing pairs at the seed level ensures a fair and properly controlled comparison. For each model pair, we therefore form paired samples based on participant–seed combinations (e.g., 7 participants × 3 seeds = 21 paired samples), and the Wilcoxon test is applied to this aggregated set of paired differences. For the classification tasks, we test whether the performance differences are consistently positive (higher F1 is better), and for gaze estimation we test whether the differences are consistently negative (lower angular error is better). For the "Avg." column, we pool paired differences across all five classification tasks before applying the test. As shown in Table 8, PiMT achieves statistically significant improvements over nearly all baselines and modalities, often with extremely small p-values, demonstrating that the gains are consistent across participants and robust to seed-level variation.

Table 8: Paired Wilcoxon p-values when comparing PiMT to each baseline on DailySense. Lower values indicate stronger evidence that PiMT outperforms the baseline.

	Classification (↑)						Regression (↓)
Method	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
Without pre-training							
SVM	4.77 e-07 ***	4.77 e-07 ***	7.63 e-06 ***	7.45 e-09 ***	4.77 e-07 ***	9.61 e-20 ***	4.44 e-05 ***
EEGNet	1.64 e-04 ***	1.21 e-04 ***	1.23 e-02 *	1.49 e-08 ***	1.59 e-03 **	8.25 e-15 ***	1.01 e-03 **
DeepConvNet	9.82 e-05 ***	4.77 e-07 ***	1.64 e-04 ***	1.42 e-07 ***	4.77 e-06 ***	4.48 e-18 ***	1.03 e-07 ***
TST	1.17 e-04 ***	1.19 e-05 ***	1.92 e-02 *	4.10 e-07 ***	6.53 e-05 ***	1.22 e-15 ***	8.04 e-04 ***
PatchTST	4.32 e-04 ***	1.09 e-04 ***	3.00 e-02 *	1.49 e-08 ***	1.65 e-03 **	2.65 e-14 ***	2.14 e-03 **
EEGConformer	9.82 e-05 ***	5.25 e-05 ***	5.23 e-04 ***	1.42 e-07 ***	1.24 e-03 **	3.28 e-16 ***	1.38 e-03 **
Bidirectional-Mamba	6.14 e-03 **	1.83 e-02 *	1.56 e-02 *	1.12 e-03 **	2.30 e-02 *	1.01 e-07 ***	1.15 e-07 ***
PiMT (Ours)	–	–	–	–	–	–	–
With pre-training							
PatchTST	1.17 e-04 ***	9.54 e-07 ***	1.26 e-04 ***	7.45 e-08 ***	9.54 e-07 ***	1.91 e-18 ***	3.29 e-03 **
PiMT (Ours)	–	–	–	–	–	–	–

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. Entries marked “–” correspond to self-comparisons.

Table 9: Performance on DailySense using different pretraining sources, showing our 50-hr free-living dataset outperforms larger controlled datasets.

	Classification (↑)						Regression (↓)
Pretraining Dataset	Video	Audio	Taste	Touch	Smell	Avg.	Gaze
No PT	0.858 ± 0.084	0.885 ± 0.125	0.790 ± 0.077	0.807 ± 0.113	0.753 ± 0.069	0.819	6.11° ± 1.20°
TUAR (98.6 hrs)	0.964 ± 0.035	0.950 ± 0.049	0.791 ± 0.065	0.824 ± 0.124	0.736 ± 0.092	0.853	5.95° ± 1.17°
TUAR + TUSZ (498.6 hrs)	0.964 ± 0.024	0.950 ± 0.044	0.803 ± 0.076	0.833 ± 0.109	0.741 ± 0.094	0.858	6.03° ± 1.14°
DailySense (Ours, 50 hrs)	**0.964 ± 0.028**	**0.961 ± 0.038**	**0.801 ± 0.064**	**0.860 ± 0.118**	**0.793 ± 0.069**	**0.876**	**6.00° ± 1.13°**

N Pretraining using Public Controlled ExG Datasets

We further examine how our 50-hour free-living DailySense dataset compares to pretraining on larger publicly available ExG corpora collected in controlled settings. Specifically, we evaluate models pretrained on TUAR (Hamid et al., 2020) and TUSZ (Shah et al., 2018), two widely used pretraining datasets in recent EEG foundation models (Jiang et al., 2024; Cui et al., 2024; Fang et al., 2025). TUAR, a curated subset of TUEG, contains annotations for five artifact types—including eye movements, chewing, and muscle activity—making it relevant to our downstream tasks such as gaze tracking. TUSZ provides extensive seizure annotations and is among the largest publicly available EEG corpora. Because these datasets use electrode montages that differ from ours, we select electrodes with the closest spatial correspondence—F7, F8, T3, T4, T5, T6, O1, and O2 in the 10–20 system—for pretraining. We consider two pretraining configurations: (1) TUAR alone (98.6 hours) and (2) TUAR combined with a subset of TUSZ for a total of 498.6 hours, representing moderate- and large-scale controlled EEG datasets, respectively. As shown in Table 9, pretraining on DailySense achieves the strongest average transfer performance across all five classification tasks and yields competitive gaze estimation accuracy, despite its substantially smaller size. This highlights the power of learning more generalizable and robust representations from free-living data, which better capture the natural variability present in real-world human behavior than controlled laboratory recordings. Considering the relative ease and scalability of free-living data collection, we expect that DailySense can be expanded far more rapidly than controlled laboratory datasets, which would further amplify these performance advantages.