The following publication S. Chiu, Z. Chen, J. Wu and K. C. Tan, "Multi-Modal EEG Datasets and Benchmarks for EEG-Based Neural Decoding Research," 2025 21st International Conference on Computational Intelligence and Security (CIS), Nanning, China, 2025, pp. 1-5 is available at https://doi.org/10.1109/CIS69366.2025.11433934.

Multi-modal EEG Datasets and Benchmarks for EEG-Based Neural Decoding Research

Szuchi Chiu
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong, SAR
szu-chi.chiu@connect.polyu.hk

Zhige Chen
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong, SAR
zhige.chen@connect.polyu.hk

Jibin Wu*
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong, SAR
jibin.wu@polyu.edu.hk

Kay Chen Tan
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong, SAR
kctan@polyu.edu.hk

***Abstract*—The growing availability of publicly shared EEG datasets involving audio, video, and linguistic stimuli offers great potential for advancing multi-modal neural decoding research. However, significant heterogeneity in file formats, annotation structures, and temporal alignment across datasets has hindered their direct usability and cross-dataset benchmarking. In this study, we present a unified preprocessing pipeline applied to six diverse EEG datasets that cover auditory attention, language comprehension, and visual perception. The proposed pipeline standardizes EEG representations and multimodal annotations, allowing the resulting data to be aligned in time and directly usable for large-scale modeling tasks. We summarize the resulting unified database and highlight its consistency in structure and format. To demonstrate usability, we provide benchmark classification results on label-eligible datasets using both traditional methods (rLDA) and deep learning (EEGNet). The resulting resource enables consistent multimodal EEG analyses and provides a structured basis for advancing research on cross-modal alignment and neural decoding. The accompanying benchmark experiments serve to illustrate the usability of the standardized datasets; achieving competitive performance was not the aim of these evaluations.**

***Keywords*—EEG, multimodal datasets, preprocessing, temporal alignment, neural decoding, dataset standardization**

I. INTRODUCTION

The use of electroencephalography (EEG) to study neural responses elicited by naturalistic stimuli, including speech, videos, and written language, has gained increasing attention, as it enables the examination of brain activity in settings that more closely reflect everyday cognition. EEG captures dynamic neural responses across perceptual, attentional, and linguistic domains, making it a valuable modality for multimodal neural decoding research [1], [2]. EEG has also been used in interactive and cognitively demanding settings such as BCI-VR systems [3], demonstrating its relevance beyond conventional laboratory paradigms.

Recent releases of large-scale EEG datasets, ranging from tri-modal audio–video–EEG recordings [4] and narrative-based multi-modal corpora [5], to structured visual object recognition datasets [1], [2] reflect a growing interest in developing unified models capable of handling heterogeneous sensory modalities.

Despite this increasing availability, publicly available multimodal EEG datasets exhibit substantial structural differences, including incompatible data formats, diverse annotation conventions, and variations in how sampling rates and event timings are defined. Such irregularities hinder the direct reuse of these datasets and complicate efforts to compare results across studies or to design models that generalize reliably beyond a single corpus, thereby limiting progress in multimodal EEG research [2], [6]. Recent tools such as EEGUnity [7] similarly point to the practical difficulties researchers face when working across heterogeneous EEG datasets and the need for more consistent preprocessing standards.

To address these challenges, this study proposes a unified preprocessing framework applied to six publicly accessible multi-modal EEG datasets involving auditory attention, visual perception, and sentence-level language comprehension. The pipeline standardizes EEG formats, unifies heterogeneous annotation structures, extracts audio/text/video stimulus modalities, and aligns them temporally with EEG signals. The resulting outputs follow a consistent structure that supports comparative evaluation and scalable modeling.

The main contributions of this work are summarized as follows:

- We document and resolve several practical sources of inconsistency across datasets—such as differences in annotation formats, sampling rates, and stimulus alignment—producing a unified multi-modal EEG resource.
- We organize the processed outputs into a unified database with consistent EEG and stimulus representations, enabling reproducible analyses in future multi-modal EEG

---

*Corresponding author

studies.

- Implementation of a baseline benchmarking pipeline using the standardized datasets to evaluate neural decoding performance across multiple heterogeneous EEG corpora.
- Provision of a transparent and reusable preprocessing framework that facilitates cross-dataset experimentation, reduces redundant implementation effort, and supports scalable multi-modal modeling.

II. METHODOLOGY

This section presents the methodology used to standardize and process six multi-modal EEG datasets. We first describe the characteristics of the datasets, followed by a unified pre-processing pipeline designed to align the EEG with the corresponding sensory modalities.

A. Dataset Description

To provide a standardized benchmark for multi-modal EEG research, we curated and unified six publicly available datasets, as summarized in Table I. These datasets encompass a diverse range of cognitive tasks, including auditory attention, memory load, visual cueing, and language comprehension. Each dataset contains EEG, audio/video, and text modalities, thus providing a comprehensive resource to evaluate multi-modal neural decoding approaches.

TABLE I
DATASETS USED IN THE STUDY

Dataset	**Ref.**	**Paradigm**	**Modalities**
zen_1158410	[8]	Speech-in-noise N-back task	EEG, electrooculography (EOG), audio (speech envelope)
VK-KUL	[9]	Silent video perception	EEG, video
ds_004408	[10]	Sentence comprehension	EEG, audio (speech envelope)
ds_004718	[11]	Naturalistic speech listening	EEG, audio (speech envelope)
cEEGrid	[12]	Auditory attention decoding	ear-EEG, audio (speech envelope)
DTU	[13]	Auditory attention decoding	EEG, audio (speech envelope)

B. Dataset-Specific Challenges

Although all datasets contain EEG recordings paired with sensory stimuli, their raw structures and annotation formats differ substantially. These inconsistencies motivate the need for a unified preprocessing framework. The specific issues encountered across the datasets in our processing pipeline are summarized below:

- **Heterogeneous EEG formats.** Raw EEG files appear in `.mat`, `.bdf`, `.edf`, and BIDS-formatted structures across datasets.
- **Divergent annotation schemes.** Event markers appear as nested MATLAB structs (zen_1158410), TSV event logs (OpenNeuro datasets), TextGrid tiers (ds_004408 and ds_004718), or JSON files, requiring unification.
- **Differences in sampling rates.** Sampling frequencies range from 128 Hz to 512 Hz, making consistent resampling necessary for aligned processing.
- **Variable channel montages.** Some datasets use 64-channel BioSemi systems, others use 32-channel caps, and cEEGrid provides 10–12 ear-EEG channels with non-standard geometry. (e.g., the electrode distribution mismatch challenge discussed in [14]).
- **Temporal misalignment.** Audio envelopes, word boundaries, and video frames may use independent clocks, requiring explicit alignment to the EEG time axis.

In addition, each dataset exhibits distinct structural challenges:

zen_1158410: event markers embedded in MATLAB structs; varying trial durations; audio envelopes reconstructed from raw recordings.

VK-KUL: video timestamps recorded as frame indices requiring conversion to absolute time; EEG stored in dataset-specific MATLAB files.

ds_004408 and ds_004718: multi-tier TextGrid files with word/phoneme boundaries requiring parsing; continuous recordings without clear trial segmentation.

cEEGrid: non-standard ear-EEG montage; occasional audio–EEG drift.

DTU: dual-speaker audio requiring separate envelope extraction; triggers and envelopes requiring careful alignment.

These observations prompted the development of a unified preprocessing workflow aimed at handling the datasets in a consistent and comparable way.

C. Dataset Processing Approach

To facilitate consistent usage and fair comparison across datasets, we employed a unified preprocessing pipeline that jointly handles EEG signals and their corresponding sensory modalities. As illustrated in Fig. 1, the pipeline includes EEG format conversion, annotation integration, modality extraction, and temporal alignment.

All datasets were downloaded from public repositories Zenodo (https://zenodo.org) or OpenNeuro (https://openneuro.org) and decompressed prior to processing. EEG signals were parsed and converted using the MNE-Python toolbox [15], with appropriate annotations of events. In parallel, associated stimulus modalities, including raw audio waveforms or envelopes, video streams, and linguistic annotations (e.g., TextGrid files, word-level transcripts), were extracted from each dataset.

These modalities were then temporally aligned with the EEG data based on time stamps, and stored as synchronized EEG-modality sample pairs, along with their corresponding index metadata for downstream processing.

III. WITHIN-DATASET EXPERIMENTS: RESULTS AND DISCUSSION

This section presents the classification results obtained using two representative algorithms: a linear baseline and a deep learning model applied to the preprocessed multi-modal EEG

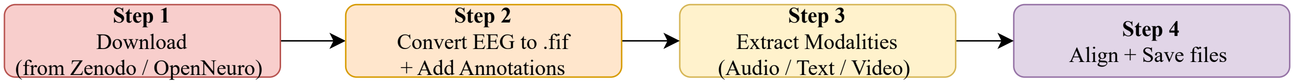

Fig. 1. Unified pipeline for EEG, audio, text, and video modality alignment.

datasets. The goal of these experiments is to demonstrate the usability and consistency of the standardized datasets rather than to pursue competitive performance benchmarks.

A. Eligible Datasets for Classification

Among the six unified datasets, only three contain discrete trial-level labels necessary for supervised classification:

- **zen_1158410**: four-class task condition labels based on the N-back paradigm.
- **VK-KUL**: four-class visual category labels associated with video segments.
- **DTU**: four-class auditory attention labels defined by the attended speaker's gender and direction.

The remaining datasets—ds004408, ds004718, and cEEGrid—do not contain explicit class labels.

- **ds004408** and **ds004718** are naturalistic speech listening datasets designed primarily for temporal and semantic regression analyses.
- **cEEGrid** contains ear-EEG recordings with extremely limited samples and no discrete trial categories, making classification infeasible.

Therefore, classification experiments are conducted exclusively on zen_1158410, VK-KUL, and DTU to ensure methodological consistency and avoid conflating classification with regression style tasks.

B. Experiment Settings

To evaluate the utility and effectiveness of the proposed multi-modal EEG database, we conducted experiments using two different algorithms:

(1) Regularized Linear Discriminant Analysis (rLDA) [16]: A classical linear classifier that maximizes inter-class separability while applying regularization to avoid overfitting. EEG recordings were first segmented into epochs based on stimulus annotations. Then each trial was reshaped from (C, T) into a one-dimensional characteristic vector of length $C \cdot T$.

(2) EEGNet [17]: A compact convolutional neural network designed for EEG-based decoding tasks. The input EEG epochs were cropped to a fixed duration of sample time points and reshaped into 4D tensors of shape $(N, 1, C, samples)$, where N is the number of trials and C is the number of channels. The network comprises the following layers:

- **Temporal Convolution:** 8 filters with kernel size $(1, 64)$ and padding $(0, 32)$
- **Depthwise Convolution:** 16 filters with kernel size $(C, 1)$, grouped by 8
- **Separable Convolution:** 16 filters with kernel size $(1, 16)$ and padding $(0, 8)$
- **Pooling and Dropout:** Average pooling layers of size $(1, 4)$ and $(1, 8)$ are applied after each block, followed by dropout layers with a probability of 0.25
- **Classifier:** A fully connected layer maps the extracted features to the output

Table II summarizes the architectural components and training configurations of the two classifiers used in our study. The rLDA serves as a traditional linear baseline, while EEGNet represents a compact deep neural network optimized for EEG decoding tasks. Key differences in input formatting, network depth, and learning mechanisms are described.

TABLE II
MODEL ARCHITECTURE SUMMARY

Model	Architecture and Details
rLDA	*Regularized Linear Discriminant Analysis* • Input: Flattened EEG epochs (n_trials × n_channels × n_times) • scikit-learn LDA implementation • 5-fold cross-validation • Output: Class probabilities No iterative optimization or backpropagation is required
EEGNet	*Compact Convolutional Neural Network* • Input shape: (n_trials × 1 × channels × samples) • First Conv2D: kernel=(1, 64), filters=8, padding=(0, 32), `BatchNorm2D` • DepthwiseConv2D: kernel=(chans, 1), filters=16, groups=8, `BatchNorm2D`, `ELU`, AvgPool2D=(1,4), Dropout=0.25 • SeparableConv2D: kernel=(1,16), filters=16, padding=(0,8), `BatchNorm2D`, `ELU`, AvgPool2D=(1,8), Dropout=0.25 • Output: Predicted label $\in \{0, 1, ..., C-1\}$, where C = number of classes • 5-fold Stratified Cross-Validation Trained for 50 epochs using Adam optimizer (learning rate = 0.001)

C. Benchmark Results

We conducted benchmark experiments on preprocessed datasets to evaluate their utility for classification-based EEG decoding. Table III summarizes the output characteristics of each dataset, including the number of converted EEG files, approximate total size, number of samples used in classification (i.e., trials or segments) and the number of target classes where applicable.

Consequently, classification experiments were performed only on the three label-eligible datasets: `zen_1158410`, `VK-KUL`, and `DTU`. As shown in Table IV, we applied two representative methods: rLDA and EEGNet.

The results, averaged over 5-fold stratified cross-validation, demonstrate that EEGNet consistently outperforms rLDA across all evaluated datasets, highlighting its superior ability

TABLE III
OUTPUT SUMMARY OF EACH DATASET

Dataset	Output Files	Size (GB)	Samples	Targets
zen_1158410	22	1.63	40	4
VK-KUL	28	38.5	13	4
ds004408	19	18.5	20	0
ds004718	20	4.5	47	0
cEEGrid	20	2.23	3	0
DTU	18	6.75	30	4

TABLE IV
CLASSIFICATION ACCURACY USING RLDA AND EEGNET

Dataset	rLDA Accuracy	EEGNet Accuracy
Zenodo 1158410	0.250 ± 0.137	0.300 ± 0.061
VK-KUL	0.321 ± 0.179	0.400 ± 0.081
DTU	0.463 ± 0.031	0.5963 ± 0.025

to capture complex spatiotemporal EEG patterns. These accuracies serve as **baseline evaluations** that confirm the usability and consistency of the standardized datasets for classification tasks, rather than establishing new state of the art performance.

IV. GLOBAL-DATASET EXPERIMENTS

Before presenting the following global analyses, it is important to note that the datasets used in this section differ fundamentally from those in Section 3. Specifically, zen_1158410, ds_004408, ds_004718, and cEEGrid do not provide discrete categorical labels suitable for supervised classification. Instead, these datasets contain continuous naturalistic speech, long form narratives, or sustained attention paradigms, making them appropriate for regression based or descriptive analyses. This distinction ensures that regression and classification tasks are not conflated and that each dataset is evaluated under methods aligned with its experimental design.

A. EEG-to-Stimulus Regression

Predicting stimulus-related features directly from EEG offers a complementary perspective on how the brain encodes incoming sensory information. This approach, often operationalized using the multivariate temporal response function (mTRF) framework [18], enables the reconstruction of acoustic features such as the speech envelope. It has also been applied to study higher level processes such as semantic tracking in naturalistic speech [10].

Several datasets included in this study such as zen_1158410, ds_004408, cEEGrid, and DTU provide both EEG and time aligned speech envelopes, allowing for this type of regression analysis. The unified preprocessing pipeline plays a critical role here: by standardizing filtering, resampling, annotation formats, and temporal alignment, it ensures that stimulus envelopes and neural recordings share a consistent time base. This consistency is essential for regression based methods, which are highly sensitive to alignment errors.

Although not the primary focus of the current work, this line of analysis may complement classification-based decoding by offering a continuous measure of attention tracking. Future investigations may explore this direction to assess neural tracking consistency across datasets or to assess condition-specific differences in stimulus encoding.

Fig. 2 illustrates the general pipeline for this type of regression analysis.

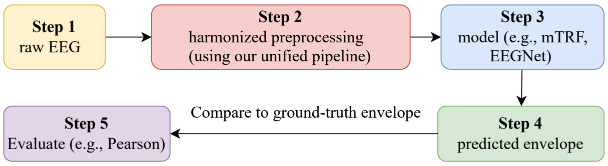

Fig. 2. Schematic pipeline for EEG-to-stimulus regression.

B. Stimulus-Type Grouping Analysis

To better understand the experimental diversity across datasets, we summarize and categorize the stimulus types and associated task attributes in Table V. This grouping highlights key distinctions such as speech-in-noise paradigms, visual versus auditory attention tasks, and naturalistic versus controlled stimuli. These differences are critical when evaluating cross-dataset generalization and designing unified modeling strategies.

TABLE V
STIMULUS TYPES AND TASK ATTRIBUTES ACROSS DATASETS

Dataset	Stimulus Type	Task Attribute
`zen_1158410`	Speech in noise	(Noise × N-back) task condition
`VK-KUL`	Silent video	Visual perception
`ds_004408`	Natural sentences	Passive listening
`ds_004718`	Cantonese + word list	Semantic processing
`cEEGrid`	Dual-speaker stories	Auditory attention (L/R)
`DTU`	Dual-talker (M/F)	Attention (gender + direction)

C. Subject Attention Dynamics

Understanding how subject attention varies across trials and tasks is an important aspect of auditory and cognitive neuroscience. Several datasets in our corpus, such as DTU, cEEGrid, and VK-KUL, provide explicit annotations of attention targets (e.g., left/right, male/female), making it feasible to investigate attention dynamics at the trial level. In contrast, datasets such as zen_1158410 or ds_004408 involve sustained or passive attention, which may manifest more subtly in neural patterns.

This opens up the possibility to explore how attention stability, shift, or fatigue differ across task types, and whether such dynamics correlate with external factors like stimulus duration, complexity, or trial order. Although we did not perform this analysis in the current work, the availability of structured labels and continuous EEG signals suggests that future studies could examine trial-by-trial variability in attention or compare intersubject differences under varying attentional demands.

Such an analysis could contribute to a more nuanced understanding of attention-related neural modulation and may support the development of more context-aware decoding strategies in real-world applications.

D. Summary and Insights

The cross-dataset analyses presented above highlight the feasibility and relevance of performing unified modeling across heterogeneous EEG corpora. First, EEG-to-stimulus regression is enabled by the availability of time-aligned speech envelopes in multiple datasets and by the consistent alignment produced by our preprocessing pipeline. Second, categorizing stimulus types reveals substantial variability in perceptual modality and task structure, which is essential for assessing cross-dataset generalization. Third, attention dynamics whether externally guided or passively sustained—differ across datasets and offer opportunities to study intra- and inter-subject variability.

Taken together, these findings motivate the development of frameworks that can adapt to stimulus and task heterogeneity while capturing shared neural encoding principles. They also suggest that future benchmarking efforts should explicitly account for task attributes, attention labeling granularity, and temporal continuity when comparing decoding or regression performance across datasets.

V. CONCLUSION

This study addresses inconsistencies in publicly available multi-modal EEG datasets by standardizing six benchmarks spanning auditory attention, language comprehension, working memory, and visual perception. We implemented a unified preprocessing pipeline that harmonizes annotation formats, channel structures, sampling rates, and temporal alignment, enabling cross-dataset compatibility and reproducible analyses.

Beyond producing a consolidated resource, our analysis highlights several cross-dataset observations. Common issues—including mismatched annotations, heterogeneous EEG formats, and drifting stimulus–EEG alignment—can be systematically resolved through a unified conversion and alignment framework. The standardized outputs support consistent application of both classification and regression tasks, as demonstrated by baseline within-dataset benchmarks. The pipeline also exposes structural commonalities, such as stimulus-type organization and the availability of continuous envelopes, which facilitate EEG-to-stimulus regression and broader multimodal analyses.

The standardized outputs provide a reliable foundation for future multi-modal neural decoding research, cross-corpus comparisons, and models that transfer knowledge across datasets. We plan to release preprocessing scripts and documentation to support reproducibility and allow the community to extend or integrate additional datasets within the same structure. Such standardized resources may also benefit emerging large-scale EEG foundation models requiring heterogeneous electrode representations, such as HEAR [19].

REFERENCES

[1] G. Zhang, S. Wang, Y. Chen *et al.*, “A large-scale MEG and EEG dataset for object recognition in naturalistic scenes,” *Scientific Data*, vol. 12, Art. 857, 2025.

[2] S. Xue, J. Liu, W. Yang *et al.*, “A multi-subject and multi-session EEG dataset for modeling human visual object recognition,” *Scientific Data*, vol. 12, Art. 663, 2025.

[3] A. Lu, M. Huang, K.-L. Liao, Z. Chen, and R. Yang, “Crafting Usability in Neuro-Narrative Games: The Joint Influence of Imagery Perspectives and Task Sequences in BCI-VR System,” *IEEE Transactions on Games*, 2025.

[4] Y. Xiao, M. Lin, Z. Wang *et al.*, “EEG-AV: A tri-modal EEG-audio-video dataset for emotional analysis in real-world conversations,” *Scientific Data*, 2024.

[5] J. Gao, Y. Liu, B. Yang, J. Feng, and Y. Fu, “CineBrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing,” *arXiv preprint arXiv:2503.06940*, Mar. 2025.

[6] R. Pillalamarri and U. Shanmugam, “A review on EEG-based multi-modal learning for emotion recognition,” *Artificial Intelligence Review*, vol. 58, Art. 131, 2025.

[7] C. Qin, R. Yang, W. You, Z. Chen, L. Zhu, M. Huang, and Z. Wang, “EEGUnity: Open-Source Tool in Facilitating Unified EEG Datasets Towards Large-Scale EEG Model,” *arXiv preprint arXiv:2410.07196*, 2024.

[8] J. Hjortkjær, J. Märcher-Rørsted, S. A. Fuglsang, and T. Dau, “Cortical oscillations and entrainment in speech processing during working memory load,” *European Journal of Neuroscience*, vol. 48, no. 7, pp. 1–12, 2018. [Online]. Available: https://doi.org/10.1111/ejn.13855

[9] Y. Yao, A. Stebner, T. Tuytelaars, S. Geirnaert, and A. Bertrand, “Identifying temporal correlations between natural single-shot videos and EEG signals,” *Journal of Neural Engineering*, vol. 21, no. 1, p. 016018, 2024. [Online]. Available: https://doi.org/10.1088/1741-2552/ad2333

[10] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, “Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech,” *Current Biology*, vol. 28, no. 5, pp. 803–809.e3, 2018. [Online]. Available: https://doi.org/10.1016/j.cub.2018.01.080

[11] M. Momenian *et al.*, “Le Petit Prince Hong Kong (LPPHK): Naturalistic fMRI and EEG data from older Cantonese speakers,” *Scientific Data*, vol. 11, Art. 992, 2024. [Online]. Available: https://doi.org/10.1038/s41597-024-03745-8

[12] S. A. Fuglsang, D. D. E. Wong, and J. Hjortkjær, “EEG and audio dataset for auditory attention decoding,” Zenodo, Mar. 15, 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011

[13] A. Simon, S. Bech, G. Loquet, and J. Østergaard, “EEG data of continuous listening of music and speech,” Zenodo, 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7500806

[14] Z. Chen, R. Yang, M. Huang, Z. Wang, and X. Liu, “Electrode Domain Adaptation Network: Minimizing the Difference Across Electrodes in Single-Source to Single-Target Motor Imagery Classification,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 2, pp. 1994–2008, Apr. 2024.

[15] A. Gramfort *et al.*, “MEG and EEG data analysis with MNE-Python,” *Frontiers in Neuroscience*, vol. 7, p. 267, 2013.

[16] J. H. Friedman, “Regularized Discriminant Analysis,” *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[17] A. J. Lawhern, D. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces,” *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[18] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, “The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli,” *Frontiers in Human Neuroscience*, vol. 10, p. 604, 2016.

[19] Z. Chen, C. Qin, W. You, R. Liu, C. Chu, R. Yang, K. C. Tan, and J. Wu, “HEAR: An EEG Foundation Model with Heterogeneous Electrode Adaptive Representation,” *arXiv preprint arXiv:2510.12515*, 2025.