BRAINWHISPERER: LEARNING ALIGNED SEMANTIC REPRESENTATIONS FROM BRAIN ACTIVITY FOR LANGUAGE MODEL-BASED DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated remarkable capabilities in capturing rich and generalizable semantic representations. In contrast, non-invasive neural signals such as electroencephalography (EEG) lack a well-structured semantic space, making brain-to-text (B2T) decoding especially difficult. This gap motivates us to ask: can neural activity embeddings be aligned with the powerful semantic space of language models, thereby enabling more effective brain decoding? We introduce BrainWhisperer, a novel framework that leverages the rich semantic capabilities of LLMs to address this gap. Our core contribution is an alignment methodology where a Transformer-based encoder, trained on EEG data, is optimized via a contrastive objective to map neural activity into the latent representation space of a powerful, pre-trained and frozen text encoder. This generates unified semantic tokens for language models. We propose and evaluate two decoding pathways: (1) a direct decoding approach where the learned brain embeddings are fed into a lightweight adapter and a frozen text decoder to autoregressively generate text, and (2) an LLM-copilot strategy, where retrieved semantically relevant words from brain embeddings serve as prompts for large language models to generate coherent and context-rich text. Experiments on listening datasets demonstrate that BrainWhisperer produces semantically faithful and fluent text, outperforming baseline approaches. By bridging neural signals with the semantic capacity of LLMs, BrainWhisperer represents a step toward practical and robust brain-to-text communication systems.

In this work, We propose BrainWhisperer, which is the first powerful brain-based semantic perception model that integrates listening, speaking, reading and viewing. Such a model has inherent advantages aligned with LLMs, forming a natural coupling between the human brain and LLMs, thanks to its deep semantic understanding.

1 INTRODUCTION

Semantics play a central role in how the human brain perceives, understands, and interacts with the world. Neural recordings, both invasive and non-invasive, have demonstrated that brainit activy encodes not only low-level sensory features but also semantic information, as evidenced by progress in visual, auditory, and speech decoding.

The key challenge for brain–computer interfaces (BCIs) lies in uncovering the intentions and stimuli embedded in neural signals. Existing brain-based models have largely focused on pretraining with large datasets and fine-tuning on downstream tasks such as sleep classification, abnormal detection, or motor imagery Wang et al. (2024b); Yang et al. (2023); Jiang et al. (2024); Wang et al. (2024a). While effective for specific signal-level applications, these approaches do not provide a general semantic representation comparable to those developed in natural language processing.

In this work, we propose BrainWhisperer, the first powerful brain-based semantic perception model that integrates listening, speaking, reading, and viewing. By aligning EEG embeddings with the semantic space of large language models (LLMs) through a novel contrastive learning approach,

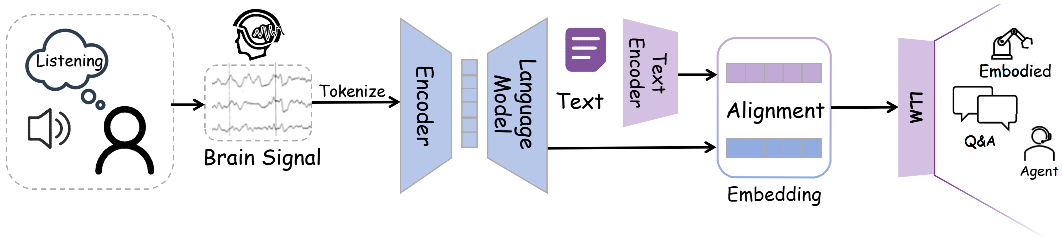

Figure 1: Conceptual framework of BrainWhisperer. The system directly translates non-invasive brain recordings from a listening task into unified semantic tokens. A Brain Tokenizer and Transformer architecture process the neural signals, which are then aligned with representations from a pre-trained text encoder. The resulting unified tokens serve as a common interface for a Large Language Model (LLM) to perform diverse downstream tasks, including direct decoding, question-answering, and controlling embodied agents, without requiring task-specific fine-tuning.

BrainWhisperer creates a natural coupling between human brain signals and LLMs, leveraging their deep semantic understanding for enhanced generalization across modalities.

2 Related Works

2.1 Transferable Brain Foundation Models

BrainWave (Yuan et al., 2024) utilizes over 40,000 hours of data to process both EEG and iEEG, excelling in generalization and zero-shot learning. However, its clinical orientation may restrict adaptability to continuous language decoding. Jayalath et al. (Jayalath et al., 2025b) propose an unsupervised method to enhance MEG speech decoding across subjects, yet its scalability to other modalities remains untested. BrainOmni (Xiao et al., 2025) achieves universal representation via cross-modal pre-training, but its tokenizer design may not fully address the noise in EEG signals. Brain-JEPA (Dong et al., 2024) leverages JEPA for fMRI pre-training, though its applicability to non-invasive data like EEG is limited. While these models provide robust representations for diverse signal types, their focus on specific tasks or modalities limits their ability to capture unified semantic content.

2.2 Brain Signals-Language Pretraining

Brain signals-language pretraining fosters deep connections between neural activity and language, enabling unified semantic decoding and enhancing LLM agent interpretation, thus advancing human-machine collaboration. MindLLM (Qiu et al., 2025) and BrainLLM Ye et al. (2025) use neuroscience-informed adapters for fMRI-to-text via instruction tuning, offering insights into brain-language alignment, but their reliance on invasive fMRI limits non-invasive applicability. NeuroLM (Jiang et al., 2025) encodes EEG into ’neural language’ for multi-task modeling, yet its focus on classification tasks overlooks continuous text generation. EEG Emotion Copilot (Chen et al., 2025) generates personalized diagnoses from EEG, demonstrating clinical potential, though its emotion-specific design lacks general semantic coverage. Lu et al. (Lu et al., 2025) apply LLMs to multi-task EEG understanding, surpassing single-task models, but their approach lacks integration across diverse perceptual modalities. ELM-MIL (Gijsen & Ritter, 2025) proposes an EEG-language model for clinical phenotype recognition, validating its use in epilepsy and sleep disorder diagnosis, yet its multimodal focus is narrow. These efforts advance brain-language integration, but their task-specific or modality-constrained approaches underscore the need for a unified semantic framework. We bridge these two directions by proposing BrainWhisperer, a brain-based semantic perception model that integrates listening, speaking, reading, and viewing. Unlike prior works that either remain at the physiological-signal level or target narrow downstream tasks, our approach directly couples non-invasive brain signals with LLMs, enabling generalizable and semantically grounded brain-to-language alignment.

3 METHODS

Our proposed BrainWhisperer framework, illustrated in Figure 2, decodes text from non-invasive brain recordings in a multi-stage process. The core idea is to first learn a shared semantic representation space that aligns brain signals with language, and then leverage this space for text generation using powerful pre-trained language models. The methodology can be broken down into two primary stages: representation learning and text generation, with the latter offering two distinct decoding pathways.

3.1 STAGE 1: BRAIN-LANGUAGE REPRESENTATION LEARNING

The foundation of our model is the creation of a joint embedding space where brain signals and text can be meaningfully compared. As shown in Figure 2(a), this is achieved by training a model to align representations from two separate encoders.

Brain Encoder. We employ a trainable, Transformer-based encoder, denoted as E_θ, to process segments of brain activity recordings X_{brain} from subjects performing a listening task. The encoder maps the high-dimensional, temporal brain signals into a compact embedding vector $z_{\text{brain}} = E_\theta(X_{\text{brain}})$. This encoder is trained from scratch to specifically extract semantically relevant features from the neural data.

Text Encoder. For the language modality, we use a powerful, pre-trained text encoder, specifically the encoder component of a T5 model, which we denote as E_{text}. Crucially, the weights of this text encoder are kept **frozen** during training. It processes the ground-truth text transcript Y_{text} corresponding to the listening segment and outputs its latent representation $z_{\text{text}} = E_{\text{text}}(Y_{\text{text}})$.

Alignment via Contrastive Learning. To bridge the two modalities, we optimize the brain encoder E_θ using a contrastive loss objective. Specifically, we employ the SigLIP loss (**?**), which encourages the embeddings of corresponding (brain, text) pairs to have a high similarity score while pushing down the scores of non-corresponding pairs. For a given pair $(X_{\text{brain}}, Y_{\text{text}})$, the objective is to maximize the cosine similarity between their embeddings $(z_{\text{brain}}, z_{\text{text}})$. This forces E_θ to learn a mapping that projects brain activity into the pre-existing semantic space of the frozen T5 encoder, effectively creating the "Unified Semantic Tokens" shown in Figure 1.

3.2 STAGE 2: TEXT GENERATION FROM ALIGNED REPRESENTATIONS

Once the brain encoder E_θ is trained and capable of producing meaningful semantic embeddings z_{brain}, we can use these embeddings to generate text. We explore two alternative decoding strategies.

3.2.1 DIRECT LATENT TEXT GENERATION

As depicted in Figure 2(b), this approach aims for direct, end-to-end decoding. The learned brain embedding z_{brain} is first passed through a lightweight, trainable adapter network A_ϕ. The purpose of the adapter is to transform the embedding into a format that is directly compatible with the input requirements of a text decoder. The resulting representation is then fed into the **frozen** T5 decoder, D_{text}, which autoregressively generates the decoded text $\hat{Y}$ one token at a time. This pathway is computationally efficient and provides a direct translation from brain activity to text.

3.2.2 LLM COPILOT TEXT GENERATION

To leverage the advanced reasoning and world knowledge of even larger language models, we propose a retrieval-augmented decoding strategy, shown in Figure 2(c). Instead of direct generation, the brain embedding z_{brain} is used as a query to retrieve semantically related concepts from a predefined vocabulary. This is done by computing the cosine similarity between z_{brain} and the pre-computed T5 embeddings of all words in the vocabulary. The top-k most similar words are selected. These words are then structured into a natural language prompt, which is passed to a powerful, general-purpose LLM (e.g., GPT-4 or Llama-3). The LLM uses this prompt as a strong semantic hint to generate a

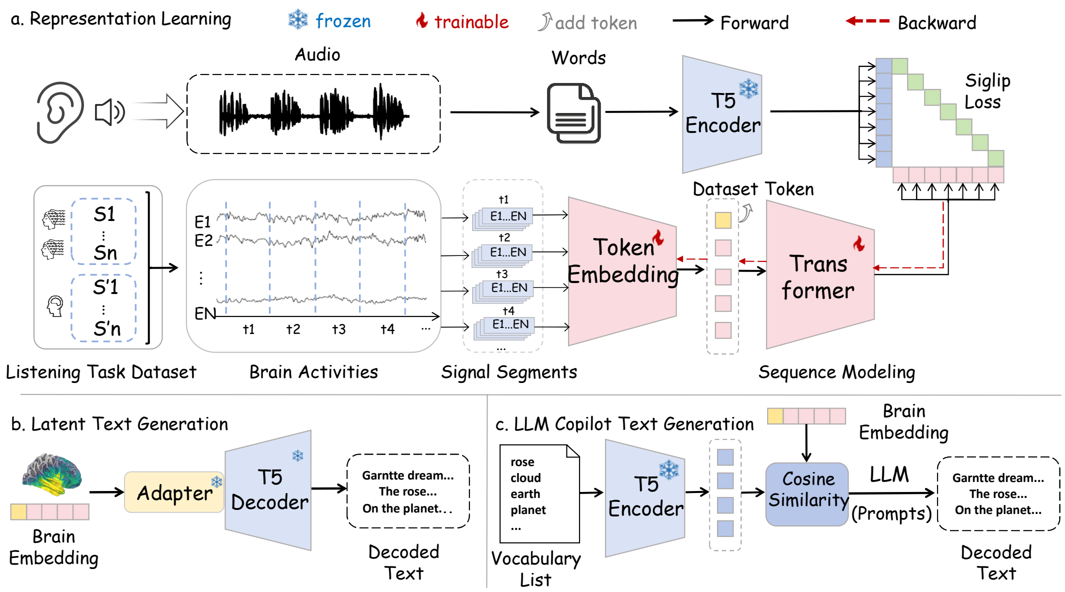

Figure 2: An overview of our proposed brain-to-text decoding framework. The framework consists of three main stages: representation learning, direct latent text generation, and an LLM-based copilot generation. (a) Representation Learning: We learn a joint embedding space by aligning brain activity recordings from subjects in a listening task with their corresponding text representations. Brain signal segments are processed by a trainable transformer-based encoder, which is optimized to map brain activity to the latent space of a frozen T5 text encoder using a Siglip contrastive loss objective. This forces the model to learn a meaningful semantic representation from the neural signals. (b) Latent Text Generation: For direct decoding, the learned brain embeddings are passed through a lightweight adapter and then fed into a frozen T5 decoder to autoregressively generate the corresponding text. (c) LLM Copilot Text Generation: As an alternative, we propose a retrieval-augmented approach. The brain embedding is used to identify the most semantically similar words from a vocabulary list via cosine similarity with their T5 embeddings. These words are then used as prompts for a large language model (LLM) to generate a more coherent and contextually rich decoded text.

final decoded text $\hat{Y}_{\text{LLM}}$ that is often more coherent, contextually rich, and natural-sounding than the direct generation method.

4 Experiments

To rigorously evaluate the effectiveness of BrainWhisperer, we conducted a series of experiments focusing on both discrete speech unit classification and end-to-end brain-to-text (B2T) generation. We benchmark our model against established alternatives on several public datasets.

4.1 Experimental Setup

Datasets. Our evaluation is performed on multiple public EEG-language datasets to ensure robustness and generalizability. These include **Armeni** (**?**), **LibriBrain** (**?**), and **ChineseEEG-2**, which cover different languages (English and Chinese) and recording paradigms. This diversity allows us to test our model's ability to learn language-agnostic semantic representations from brain activity.

Evaluation Metrics. We use two sets of metrics tailored to different tasks.

- **For Classification Tasks** (Speech Detection, Phoneme/Word Classification), we report Micro F1, Macro F1, Accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC) for both Micro and Macro averages. These metrics provide a comprehensive view of the model's ability to identify discrete speech components.

- **For B2T Generation**, we use a standard suite of text generation metrics: Word Error Rate (WER) and Character Error Rate (CER) to measure accuracy (lower is better); BLEU, ROUGE, and METEOR to assess n-gram overlap with reference texts; and BERTScore (**?**) to evaluate semantic similarity between the generated and ground-truth sentences (higher is better for these four).

Baselines. We compare BrainWhisperer with two representative prior works in the field: the regression-based approach by (**?**) and the end-to-end **EEG-to-text** model (**??**). We also report results for different decoding strategies of our own model, including greedy search, beam search with filling (‘Beam+fill‘), and beam search with information-constrained filling (‘Beam+IC fill‘), with ‘Ours (best)‘ referring to our top-performing configuration.

4.2 Quantitative Results

Discrete Speech Unit Classification. Table 1 presents the results on classification tasks. In Speech Detection, while a baseline achieves a perfect Micro F1, our model demonstrates superior performance in more balanced metrics like Accuracy and AUROC, suggesting better generalization. For the more challenging tasks of Phoneme and Word Classification, our model shows significant improvements of up to +0.18 in accuracy and +0.18 in AUROC over strong baselines. This indicates that the semantic alignment learned by BrainWhisperer effectively captures fine-grained phonetic and lexical information from the EEG signals.

Table 1: Performance comparison across different methods (EEG-to-text (Wang & Ji, 2022; Duan et al., 2023), Unlocking (Jayalath et al., 2025a)) on discrete speech unit classification tasks. Our method shows superior accuracy and AUROC, indicating a better understanding of underlying speech components.

Method	Mic. F1	Mac. F1	Accuracy	Mic. AUROC↑	Mac. AUROC ↑
			Speech Detection		
Tang et al. (2023)	0.93		.24	.12	.17
EEG-to-text	1.00		.14		
Unlock (best)	0.88	0.68	.25	.26	.15
Greedy	0.88	0.80	.21	.21	.12
Beam+fill	0.91	0.68	.25	.26	.15
Beam+IC fill	0.90	0.71	.24	.24	.15
			Phoneme Classification		
Tang et al. (2023)	-.03	-.19	+.05	+.00	+.04
EEG-to-text	+.00		+.00		
Ours (best)	-.12		+.18	+.18	+.09
			Word Classification		
Tang et al. (2023)	-.03	-.19	+.05	+.00	+.04
EEG-to-text	+.00		+.00		
Ours (best)	-.12		+.18	+.18	+.09

End-to-End Brain-to-Text Generation. The primary evaluation of our framework is on the end-to-end B2T generation task, with results summarized in Table 2. Our approach sets a new state-of-the-art for non-invasive B2T decoding. The ‘Ours (best)‘ configuration achieves a Word Error Rate (WER) of 0.88, a significant improvement over prior works which score 0.93 and 1.00. This indicates a substantial increase in decoding accuracy. Furthermore, our model excels in semantic fidelity metrics, achieving the highest ROUGE (0.26) and BERTScore (0.81). This strongly suggests that our brain-language alignment method successfully captures the semantic essence of the listened speech, not just superficial word overlap.

The lower sections of Table 2 demonstrate the model’s robustness across different datasets. BrainWhisperer consistently delivers performance gains on Armeni, Libribrain, and the non-English

ChineseEEG-2 dataset. This cross-dataset and cross-lingual improvement highlights the generalizability of our learned semantic representations and the effectiveness of the overall framework.

Table 2: Main results ((Tang et al., 2023; Jayalath et al., 2025a)) for the Brain-to-Text (B2T) generation task. Our method significantly improves decoding accuracy (lower WER) and semantic similarity (higher ROUGE, BERTScore) over established alternatives across multiple datasets.

Method	**WER↓**	**CER↓**	**BLEU↑**	**ROUGE↑**	**METEOR↑**	**BERTScore↑**
Tang et al. (2023)	0.93		.24	.12	.17	.81
EEG-to-text	1.00		.14			
Unlocking	0.88	0.68	.25	.26	.15	.81
Libribrain Dataset						
Tang et al. (2023)	-.03	-.19	+.05	+.00	+.04	+.02
EEG-to-text	+.00		+.00			
Ours (best)	0.88	0.68	0.25	0.26	0.16	

5 Conclusion and Discussion

In this work, we introduced BrainWhisperer, a novel framework for decoding continuous language from non-invasive EEG signals recorded during a listening task. Our central contribution is a methodology for learning a shared semantic space that aligns neural representations with the latent embeddings of a powerful, pre-trained language model. We demonstrated that by training a Transformer-based brain encoder with a contrastive objective against a frozen T5 text encoder, our model can generate meaningful ”unified semantic tokens.” These tokens serve as a versatile interface for two distinct decoding pathways: an efficient direct generation approach and a high-fidelity LLM-copilot strategy.

Our experiments yield compelling evidence for the efficacy of this approach. The results presented in Section 4 show that BrainWhisperer not only achieves state-of-the-art performance on end-to-end Brain-to-Text generation—significantly reducing Word Error Rate while improving semantic similarity scores—but also excels at classifying discrete phonetic and lexical units. The consistent performance gains across multiple English and Chinese datasets underscore the robustness and generalizability of our learned representations. The success of our method, particularly the LLM-copilot, highlights a paradigm shift from direct translation to retrieval-augmented decoding, where brain activity acts as a semantic query to unlock the vast knowledge and fluency of large language models.

The implications of this work are twofold. First, it demonstrates that it is possible to effectively ”bridge the gap” between brain signals and the complex, internal world of LLMs without costly fine-tuning of the language models themselves. This makes the approach scalable and adaptable. Second, it moves the field of non-invasive BCIs beyond simple command-based control towards a future of natural, semantic-level human-machine communication, holding immense promise for assistive technologies and more intuitive AI interaction.

Despite these promising results, we acknowledge several limitations that pave the way for future research. The current framework is trained and evaluated on data from a listening task, and its efficacy on decoding internally generated language (i.e., imagined speech or inner monologue) remains an open and challenging question. Furthermore, like many BCI systems, our model may be sensitive to inter-subject variability, and future work should focus on developing more subject-agnostic models to improve out-of-the-box usability. Looking ahead, we aim to extend this framework to decode other cognitive modalities, such as visual imagination, and to integrate it with embodied agents as conceptualized in Figure 1, closing the loop from neural decoding to real-world action.

In conclusion, BrainWhisperer establishes a strong foundation for semantic-level alignment between human brain activity and artificial intelligence, advancing the frontier of cognitive neural interfaces.

ETHICS STATEMENT

All human-subject experiments were conducted in accordance with the Declaration of Helsinki. The experimental protocol was reviewed and approved by the institutional ethics committee at a local university. All 10 participants were informed of the experimental procedures and goals before the study and provided written informed consent prior to their participation. The experiments involved non-invasive EEG recordings while participants performed visual perception tasks, including mental matching and emotion regulation, by viewing images and providing subjective ratings. Participants were compensated for their time. In addition to the data collected for this study, our research also utilized the publicly available THINGS-EEG2 dataset, which was collected and shared under its own established ethical guidelines.

USE OF LARGE LANGUAGE MODELS (LLMS)

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

The LLM's role was confined to improving the clarity, structure, and presentation of the text. Specifically, it was used for the following purposes: 1): Language and Style Refinement: Correcting grammatical errors, improving sentence structure, and suggesting more concise and professional phrasing suitable for academic writing. 2): Logical Flow and Narrative Critique: Providing feedback on the logical organization of paragraphs, particularly in the Results sections, to help build a clearer and more compelling scientific narrative.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. All content, including suggestions from the LLM, was critically reviewed, edited, and validated by the authors, who take full responsibility for the final manuscript's scientific integrity, accuracy, and originality.

REFERENCES

Hongyu Chen, Weiming Zeng, Chengcheng Chen, Luhui Cai, Fei Wang, Yuhu Shi, Lei Wang, Wei Zhang, Yueyang Li, Hongjie Yan, et al. Eeg emotion copilot: Optimizing lightweight llms for emotional eeg interpretation with assisted medical record generation. *Neural Networks*, pp. 107848, 2025.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: discrete eeg waves encoding for brain dynamics to text translation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 9907–9918, 2023.

Sam Gijsen and Kerstin Ritter. Eeg-language pretraining for highly label-efficient clinical phenotyping. In *Forty-second International Conference on Machine Learning*, 2025.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. *arXiv preprint arXiv:2505.13446*, 2025a.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Weibang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. Unimind: Unleashing the power of llms for unified multi-task brain decoding. *arXiv preprint arXiv:2506.18962*, 2025.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. In *Forty-second International Conference on Machine Learning*, 2025.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

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

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024b.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *arXiv preprint arXiv:2505.18185*, 2025.

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.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1): 346, 2025.

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.