EEG2TEXT: Open Vocabulary EEG-to-Text Decoding with EEG Pre-Training and Multi-View Transformer

Anonymous ACL submission

Abstract

Deciphering the intricacies of the human brain has captivated curiosity for centuries. Recent strides in Brain-Computer Interface (BCI) technology, particularly using motor imagery, have restored motor functions such as reaching, grasping, and walking in paralyzed individuals. However, unraveling natural language from brain signals remains a formidable challenge. Electroencephalography (EEG) is a non-invasive technique used to record electrical activity in the brain by placing electrodes on the scalp. Previous studies of EEG-to-text decoding have achieved high accuracy on small closed vocabularies, but still fall short of high accuracy when dealing with large open vocabularies. We propose a novel method, EEG2TEXT, to improve the accuracy of open vocabulary EEG-to-text decoding. Specifically, EEG2TEXT leverages EEG pre-training to enhance the learning of semantics from EEG signals and proposes a multi-view transformer to model the EEG signal processing by different t spatial regions of the brain. Experiments show that EEG2TEXT has superior performance, outperforming the state-of-the-art baseline methods by a large margin of up to 5% in absolute BLEU and ROUGE scores. EEG2TEXT shows great potential for a high-performance open-vocabulary brain-to-text system to facilitate communication.

1 Introduction

Recent advances in brain-computer interface (BCI) technology have demonstrated exciting progress in restoring the capabilities of patients with paralysis, such as reaching (Hochberg et al., 2012), grasping (Aflalo et al., 2015; Bouton et al., 2016), and walking (Lorach et al., 2023). The heart of BCI is its ability to accurately decode complex brain signals. Despite the advances in decoding brain signals related to motion, decoding brain signals related to speech remains a formidable challenge. Previous research translating speech-related brain signals to text (brain-to-text) primarily relies on electrocorticography (ECoG), an invasive electrophysiological monitoring method that uses electrodes placed directly on the exposed brain surface to record activity from the cerebral cortex. ECoG offers higher temporal and spatial resolution than traditional non-invasive scalp electroencephalography (EEG), with a significantly better signal-to-noise ratio. However, the invasive nature of ECoG is undesirable for BCI applications. EEG, though offering lower signal quality than ECoG, is non-invasive and widely available, making it ideal for BCI if its noisy signals can be accurately decoded.

Previous studies of EEG-to-text decoding (Herff et al., 2015; Sun et al., 2019; Anumanchipalli et al., 2019; Makin et al., 2020; Panachakel and Ramakrishnan, 2021; Moses et al., 2021; Nieto et al., 2022) have achieved high accuracy on small closed vocabularies, but still fall short of high accuracy when dealing with large open vocabularies. These approaches primarily target high accuracy ($> 90\%$) but are often confined to small closed vocabularies and struggle to decode semantically similar words beyond training sets. Recent studies broaden the scope from closed to open-vocabulary EEG-to-text decoding (Wang and Ji, 2021; Willett et al., 2023; Tang et al., 2023; Duan et al., 2023), drastically expanding the vocabulary size by over 100-fold, from several hundred to tens of thousands of words. Notably, two of these studies (Wang and Ji, 2021; Duan et al., 2023) leverage a pre-trained large language model BART (Lewis et al., 2019), and represent the state-of-the-art for open vocabulary brain-to-text decoding. However, these studies are in their nascent stages and are challenged by their limited accuracy.

To improve the accuracy of EEG-to-text decoding with open vocabularies, we propose a novel EEG-to-text decoding method based on transformers. First, we introduce a Convolutional Neural

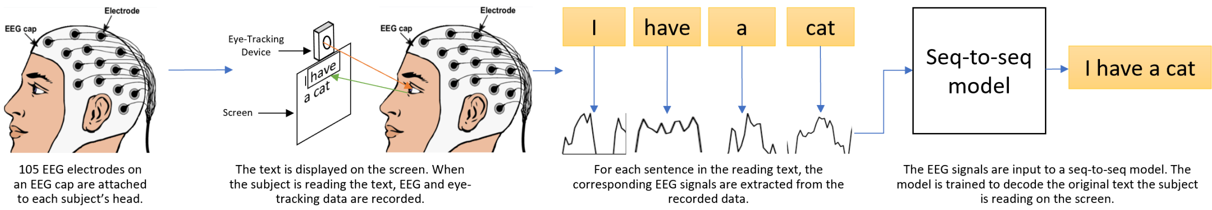

Figure 1: The overall framework of open-vocabulary EEG-to-text translation. The first sub-figure comes from (Nagel and Spüler, 2018).

Network (CNN) module before the base transformer model to enhance the model's ability to handle long EEG signals. Second, we conduct pre-training of the transformer model by reconstructing randomly masked EEG signals from the input data. This pre-training step helps our transformer model better learn the semantics of EEG signals. Last, we propose a multi-view transformer architecture, where each single-view transformer is the pre-trained model from the previous step, to model the EEG signal processing by different spatial regions of the brain. Experiments show that EEG2TEXT has superior performance, outperforming the state-of-the-art baseline methods by a large margin of up to 5% in absolute BLEU and ROUGE scores. EEG2TEXT shows great potential for a high-performance open-vocabulary brain-to-text system to facilitate communication. We will open-source our code and dataset to facilitate future studies of EEG-to-text translation.

2 Task Definition

Our task involves decoding corresponding text from EEG signals (Figure 1). The data acquisition process involves 1) attaching an EEG cap to each subject's head, 2) displaying the text (reading materials) on a screen, and 3) recording the EEG and eye-tracking (for verification and calibration of the EEG signals) data while the subject is reading the text. The EEG signals are further extracted from the recorded data and fed as input to a decoding model to predict the original text the subject was reading on the screen.

Formally, this task can be formulated as a sequence-to-sequence machine translation task:

$$P(Y|X) = \arg\max_{Y} \prod_{t=1}^{T'} P(y_t|y_{<t}, X) \quad (1)$$

where T' represents the length of the target sentence Y; y_t represents the word or token at position t in the target sentence Y; $y_{<t}$ represents the words or tokens preceding position t in the target sentence Y; X represents the input EEG data; and $P(y_t|y_{<t}, X)$ is the conditional probability of generating word y_t given the previous words $y_{<t}$ and the input EEG data X. Our goal is to maximize the probability $P(Y|X)$ of generating the target sentence given the input EEG data.

3 Methodology

3.1 Baseline Model

Our baseline model (Wang and Ji, 2021) takes the word-level EEG features as the input to a transformer model followed by a pre-trained BART model for text decoding. The raw EEG signals are typically stored as a two-dimensional array with one dimension for time and the other for channels (the number of electrodes used to collect EEG signals). Each value in this two-dimensional array corresponds to the signal strength collected at the corresponding time for the corresponding channel. In the baseline model, the word-level EEG features are extracted from eight independent frequency bands from the raw EEG signals. The above eight word-level EEG features are simply concated across all the channels as input to the decoder framework.

The baseline model faces the following challenges: 1) the reliance on eye-tracking calibration for word-level EEG feature extraction introduces error propagation and lacks generalizability to scenarios such as inner speech decoding (Martin et al., 2018; Nalborczyk et al., 2020), 2) there is room for improvement in EEG representation learning through self-supervised pre-training, and 3) the lack of spatial resolution modeling ignores the varying importance of different brain regions in language processing. To overcome these challenges, we propose a novel framework, EEG2TEXT, that achieves superior performance for open-vocabulary EEG-to-text translation.

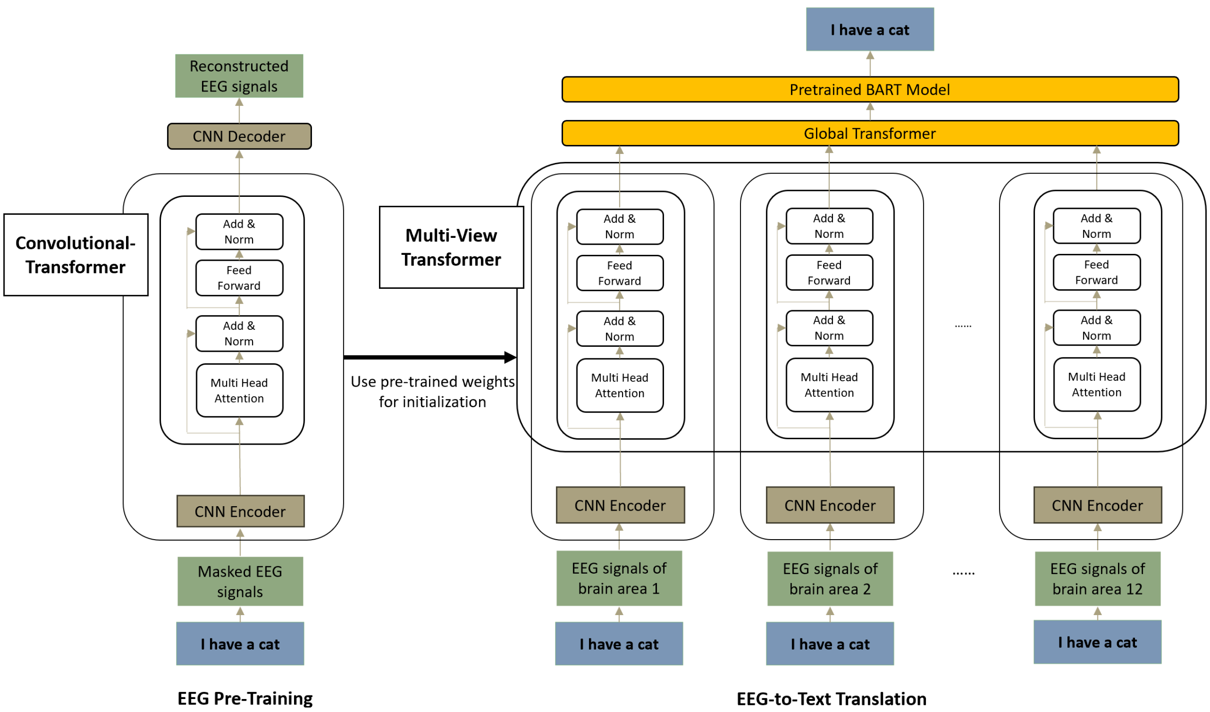

Figure 2: The overall framework of EEG2TEXT. It takes the sentence EEG signals as input and decodes the original text as output. EEG2TEXT includes major steps of 1) a base convolutional transformer model, 2) pre-training for EEG encoding, and 3) a multi-view transformer for different spatial regions of the brain.

3.2 Convolutional Transformer for Sentence-Level EEG Encoding

Instead of using the word-level EEG features crafted based on the eye-tracking data, we directly use the sentence-level EEG signals as input to our model. Using sentence-level EEG signals offers several advantages over word-level EEG features. It provides richer information without error propagation from the eye-tracking data and exhibits better generalizability to other tasks, such as inner speech decoding, where acquiring eye-tracking data is infeasible.

However, the sentence-level EEG signals pose a challenge due to their excessive length (24K timestamps), potentially overloading laboratory-level GPUs if directly input into the transformer layer. Traditional Transformer models (Vaswani et al., 2017) (max input length: 512 tokens) and their long-input variations, such as Longformer (Beltagy et al., 2020) (max input length: 4096 tokens) and BigBird (Zaheer et al., 2020) (max input length: 4096 tokens) cannot deal with our long EEG data. Recently, there are some new architectures, specifically designed for extremely long sequence data (Fu et al., 2022; Poli et al., 2023; Gu and Dao, 2023) up to one million input tokens. Inspired by these models, we introduce a convolutional transformer model that incorporates a CNN module for compressing raw EEG signals. Utilizing CNN-Transformer for modeling long sequences has been proven effective in previous EEG signal processing tasks (Song et al., 2022). So we choose this CNN-Transformer as the base architecture to develop our models. This CNN module comprises two convolutional layers, adept at both temporal and spatial (or channel) compression. We also compared two input formats of the sentence-level EEG signals: 1) the raw signals, and 2) the spectrogram of the signals. The spectrogram of a signal (Appendix Figure A1) is a two-dimensional image, where the x-axis represents time, the y-axis represents frequency, and the image pixel value represents the magnitude of the signal at each time-frequency pair. The sentence-level EEG signals are then input into the CNN module to obtain compressed EEG signals, which are then fed into the transformer model for subsequent feature extraction and text translation.

3.3 Transformer Pre-Training for an Enhanced EEG Encoding

To enhance the semantic understanding of the EEG signals, we propose EEG pre-training on the sentence-level EEG signals for brain-to-text translation. There is one recent work, LaBraM (Jiang et al., 2024), on pre-training diverse EEG data across different tasks. However, their input only compasses sparse EEG channels (less than 64) and short signals (less than 14 seconds), while our sentence EEG signals for text translation are collected from dense EEG channels (105) and composed of much longer lengths (48 seconds). Therefore, our input EEG signal significantly exceeds the input length limit of the pre-trained LaBraM model.

We propose a self-supervised pre-training of our convolutional transformer model for parameter initialization (Figure 2). Inspired by the masked language model pre-training strategies (Devlin et al., 2018; Joshi et al., 2019; Liu et al., 2019), we formulate our self-supervised pre-training objective as follows:

$$\theta^* = \arg\max_{\theta} \sum_{(i,j)\in\mathcal{D}} \log P(M|C;\theta), \quad (2)$$

where M represents the masked tokens; C represents the context or surrounding tokens; θ^* represents the optimal model parameters; θ represents the model parameters being optimized; $\mathcal{D}$ represents the training data, where (i,j) are pairs of sentences or sentence fragments; and $P(M|C;\theta)$ is the probability of predicting the masked tokens.

During the self-supervised pre-training stage, we add a convolutional decoder module on top of the convolutional transformer encoder to decode the input EEG signals. The input is the sentence-level EEG signals masked with different strategies and the output is the sentence-level EEG signals reconstructed by the CNN decoder. Specifically, we compared three different masking strategies for the sentence-level EEG signals as follows:

- **Masked Token Prediction** (Devlin et al., 2018): randomly masking 15% of all the tokens.
- **Continuous Masked Token Prediction** (Joshi et al., 2019): randomly masking a sequence of consecutive tokens until a total of 15% of all the tokens are masked.
- **Re-Masked Token Prediction** (Liu et al., 2019): re-randomizing the masking of 15% of all the tokens for each training epoch.

It is important to highlight that our self-supervised pre-training step allows for seamless integration of EEG data from diverse tasks, including image recognition. In our experiments, we further incorporated an image EEG dataset (Gifford et al., 2022) during pre-training, aiming to showcase the model's adaptability to EEG signals from multimodal data and explore the potential for enhanced translation performance through the combination of EEG signals from diverse data modalities.

Brain Regions	Corresponding Electrodes
C	E36, E104, Cz, E30, E105, E41, E103, E7, E31, E35, E80, E106, E110
F	E4, E27, E123, E24, E124, E33, E122, E11, E19, E20, E118
O	E70, E83, E75, E74, E82
P	E52, E92, E60, E58, E64, E96, E95, E85, E51, E97, E62, E50, E53, E59, E61, E69, E78, E86, E89, E91, E101
T	E114, E45, E108, E44, E39, E43, E115, E120
FP	E22, E9, E15
AF	E23, E3, E26, E2, E16, E10, E18
CP	E37, E87, E42, E93, E47, E98, E55, E54, E79
FC	E13, E112, E29, E111, E28, E117, E6, E5, E12
FT	E121, E34, E116, E38
PO	E67, E77, E65, E90, E72, E66, E71, E76, E84
TP	E100, E46, E102, E57, E40, E109

Table 1: 12 brain regions with corresponding channels.

The goal of this pre-training step is to have the convolutional transformer learn meaningful concepts such as context, relationships, and semantics present in sentence-level EEG signals during this pre-training process. After pre-training, the parameters are saved and used as the initial parameters for the final multi-view transformer model.

3.4 Multi-View Transformer for Different Spatial Regions of the Brain

Another important feature of our model is the novel multi-view transformer decoder architecture we introduced that encodes different regions of the brain with a different convolutional transformer (Figure 2). The multi-view transformer model takes into account the fact that different brain regions potentially play different roles in language processing. This spatial modeling therefore can improve the model performance, but has been overlooked in previous work.

We partition the 105 channels into 12 groups based on their spatial location under the guidance of Geodesic Hydrocel system's technical note (Luu and Ferree, 2005) (Table 1). Geodesic Hydrocel system (Electrical Geodesics, Eugene, Oregon) is an electrode net design used in our main dataset ZuCo (Hollenstein et al., 2018) to record EEG data. In the technical note, the majority of the 105 channels have been matched with the channels in the traditional 10-10 EEG system (Chatrian et al., 1985). The 10-10 EEG system explicitly names channels according to the brain regions they correspond to, such as F: Frontal lobe; O: Occipital lobe. Based on the naming rule, the matched channels have been categorized accordingly. For the remaining unmatched channels, we find the channel with the closest L2 distance to it and classify them into the same category.

After the partition of the electrodes, we create a multi-view transformer model including 12 convolutional transformers at the bottom level, where each convolutional transformer encodes the EEG signals from the electrodes in that region. On top of the 12 convolutional transformers, we add a global transformer to unify the information from different brain regions. The combined information from the global transformer is further fed into the BART model for text decoding.

In summary, the multi-view transformer envisions multiple parallel convolutional transformer models where each captures different aspects of EEG signals combined from different spatial regions of the brain regions. This approach enhances the spatial resolution of the model and further improves the text decoding performance.

4 Experiment

4.1 Experimental Setup

Dataset We utilize both the ZuCo (Hollenstein et al., 2018) and Image-EEG (Gifford et al., 2022) for pre-training and use ZuCo to train the multi-view transformer and BART model for text decoding. Details of both datasets are listed below.

- **ZuCo** (Hollenstein et al., 2018) contains EEG and eye-tracking data from 12 healthy adult native English speakers engaged in natural English text reading for 4 - 6 hours. This dataset covers two standard reading tasks and a task-specific reading task, offering EEG and eye-tracking data for 21,629 words across 1,107 sentences and 154,173 fixations.
- **Image-EEG** (Gifford et al., 2022) is a large and rich dataset containing high temporal resolution EEG signals of images of objects on natural backgrounds. The dataset included 10 participants, each performing 82,160 trials across 16,740 image conditions.

Baselines We compare EEG2TEXT with two baseline models for open-vocabulary EEG-to-text translation.

- **Baseline (EEGtoText)** (Wang and Ji, 2021) uses word-level EEG signals as input to a transformer model followed by a pre-trained BART model for decoding. EEGtoText is the first paper that proposed the open-vocabulary EEG-to-text translation task.
- **DeWave** (Duan et al., 2023) introduces a discrete codex encoding after the transformer layer, and uses both word-level EEG features and the raw EEG signals as input. DeWave is the most recent related work and it only included EEGtoText (Wang and Ji, 2021) as its baseline.

We use BLEU and ROUGE scores as evaluation metrics and conduct parameter study. The details can be found in Appendix A and Appendix B.

4.2 Results

Main Results Table 2 shows our main experimental results. The baseline method (Wang and Ji, 2021) achieves a moderate performance in text decoding with BLEU scores. DeWave (Duan et al., 2023) slightly improved the performance across all metrics, demonstrating the effectiveness of discrete encoding. EEG2TEXT improved the text decoding performance by a large margin due to several technical innovations. First, a single convolutional transformer achieved slightly lower BLEU scores (BLEU-1: -1.3%; BLEU-2: -0.5%; BLEU-3: -0.2%; BLEU-4: -0.0%) but higher ROUGE-1 scores (F1-score: +3.7%; Precision: +2.4%; Recall: -0.9%) compared to DeWave. Second, EEG2TEXT with pre-training further enhanced the BLEU scores (BLEU-1: +1.8%; BLEU-2: +1.9%; BLEU-3: +1.8%; BLEU-4: +1.6%) and ROUGE-1 scores (F1-score: +4.2%; Precision: +2.4%; Recall: +0.0%) compared to DeWave. Pre-training proved effective in enhancing text generation by providing a strong initialization foundation for our model. Third, EEG2TEXT with multi-view transformers achieved the highest scores across all metrics, with a significant increase in the BLEU scores (BLEU-1: **+4.7%**; BLEU-2: **+5.6%**; BLEU-3: **+6.0%**; BLEU-4: **+5.9%**) and ROUGE-1 scores (F1-score: **+8.5%**; Precision: **+6.8%**; Recall: **+4.2%**) compared to DeWave. EEG2TEXT excelled in gen-

Methods	BLEU-N				ROUGE-1		
	N = 1	N = 2	N = 3	N = 4	F	P	R
Baseline (Wang and Ji, 2021)	0.401	0.231	0.125	0.068	0.301	0.317	0.288
DeWave (Duan et al., 2023)	0.413	0.241	0.139	0.082	0.288	0.337	0.306
EEG2TEXT (Convolutional Transformer)	0.400	0.236	0.137	0.082	0.325	0.361	0.297
EEG2TEXT (+ Pre-training)	0.445	0.274	0.175	0.117	0.341	0.383	0.310
EEG2TEXT (+ Multi-View Transformer)	**0.460**	**0.297**	**0.199**	**0.141**	**0.373**	**0.405**	**0.348**

Table 2: Performance comparison of EEG2TEXT with baseline methods.

Methods	BLEU-N				ROUGE-1		
	N = 1	N = 2	N = 3	N = 4	F	P	R
Spectrogram + Transformer	0.386	0.220	0.121	0.067	0.306	0.342	0.306
Spectrogram + Convolutional Transformer	0.374	0.209	0.112	0.061	0.302	0.339	0.274
EEG signal + Convolutional Transformer	**0.400**	**0.236**	**0.137**	**0.082**	**0.325**	**0.361**	**0.297**

Table 3: Ablation study of different input formats of the EEG signals.

erating coherent, contextually relevant, and high-quality text.

Convolutional Transformer We first compare different input representations of the EEG signals to see how the representation affects the performance of a base convolutional transformer model. In this ablation study, we compare the raw EEG signals with their spectrograms using the fast Fourier transform (Cochran et al., 1967) to convert the original one-dimensional time array into a two-dimensional time-frequency matrix. The results are shown in Table 3. Using the raw EEG as the input consistently led to better performance than using the spectrogram as the input. The spectrogram only keeps the magnitude information and ignores the phase information of the raw EEG signal. The superior performance of the raw EEG signal suggested that the phase information might be important for decoding. Therefore, the raw EEG signals are used as the input in subsequent experiments.

EEG Pre-Training We then conducted ablation experiments to compare the effectiveness of three pre-training strategies: 1) Masked Token Prediction (Devlin et al., 2018), 2) Continuous Masked Token Prediction, and 3) Re-Masked Token Prediction (Liu et al., 2019). The results are shown in Table 4. The Re-Masked Token Prediction (Liu et al., 2019) exhibits the best performance among all the three masking strategies. One potential reason is that the convolutional transformer model can learn more diverse semantic information by masking different tokens in each training epoch during pre-training.

In the above study, we focused on identifying the optimal pre-training strategy among the three without incorporating image-EEG data (Gifford et al., 2022). As an additional component, we introduced image-EEG data to assess the compatibility of our model with EEG signals from multi-modal inputs. Leveraging our self-supervised pre-training strategy, we directly incorporated image-EEG data into the pre-training phase to enable the model to glean knowledge from diverse sources. The results, detailed in Table 5, demonstrate that adding image-EEG data significantly enhances translation performance for both the single convolutional transformer and the multi-view transformer.

Multi-View Transformer Finally, we compare different training strategies of the multi-view transformer to demonstrate the effectiveness of the multi-view transformer and find the best training strategy. The image-EEG data was not included in this ablation study. Specifically, we compared three training strategies as follows:

- **Only Global Transformer**: Fixing the parameters of all 12 convolutional transformer modules and training only the global transformer for text decoding.
- **Global Transformer + One Convolutional Transformer**: During each training epoch, randomly activate and train one convolutional transformer with the global transformer while fixing the parameters of the remaining 11 convolutional transformers.
- **Global Transformer + Three Convolutional Transformers**: During each training epoch, randomly activate and train three convolutional transformers with the global transformer while

Methods	BLEU-N				ROUGE-1		
	N = 1	N = 2	N = 3	N = 4	F	P	R
Masked Token Prediction	0.409	0.242	0.141	0.087	0.325	0.357	0.300
Continuous Masked Token Prediction	0.411	0.243	0.137	0.078	0.319	0.352	0.294
Re-Masked Token Prediction	**0.431**	**0.260**	**0.157**	**0.098**	**0.330**	**0.361**	**0.306**

Table 4: Ablation study of different pre-training strategies of the EEG signals.

Methods	BLEU-N				ROUGE-1		
	N = 1	N = 2	N = 3	N = 4	F	P	R
Single-View without image-EEG	0.431	0.260	0.157	0.098	0.330	0.361	0.306
Single-View with image-EEG	0.445	0.274	0.175	0.117	0.341	0.383	0.310
Multi-View without image-EEG	0.447	0.280	0.180	0.123	0.357	0.389	0.331
Multi-View with image-EEG	**0.460**	**0.297**	**0.199**	**0.141**	**0.373**	**0.405**	**0.348**

Table 5: Ablation study of adding image-EEG data into pre-training.

fixing the parameters of the remaining nine convolutional transformers.

We have a large dataset with 2K batches to ensure each individual Transformer is trained sufficiently.

The results in Table 6 demonstrate that activating three convolutional transformers together with the global transformer achieves the best performance. This suggests further improvement may be attainable by increasing the number of activated convolutional transformers during each training epoch if more GPU resources are available.

Case Study Table 7 shows our case study results. In the first sentence, the baseline model accurately translates "good," whereas EEG2TEXT, in addition, accurately captures the first half of the sentence with "movie" (synonymous with "film"). Additionally, EEG2TEXT correctly translates the second half of the sentence with "disaster movie" corresponding to "monstrous one" in the original sentence. In the second sentence, EEG2TEXT accurately captured "won Nobel Prize in Chemistry," while the baseline produced incorrect information, stating "Pulitzer Prize" and the wrong field, "Literature." In the third sentence, both EEG2TEXT and the baseline correctly identified "book" and "Pulitzer Prize." However, EEG2TEXT, in addition, correctly identified the field as "Biography," while the baseline erroneously outputted "Fictionography."

In addition, we conducted an interesting case study to show that EEG2TEXT has the ability of zero-shot image-to-text translation. Details can be found in Appendix D.

5 Related Work

Brain Computer Interface The landscape of brain-to-speech and brain-to-text decoding encompasses three principal approaches grounded in the features they capture: motor imagery-based, overt speech-based, and inner speech-based. These methods explore a variety of brain signals, including electroencephalogram (EEG), electrocorticography (ECoG), and functional magnetic resonance imaging (fMRI). Despite these endeavors, existing approaches exhibit limitations concerning vocabulary size, articulation dependence, speed, and device compatibility. Motor imagery-base systems, exemplified by point-and-click (Pandarinath et al., 2017) mechanisms and imaginary handwriting (Willett et al., 2021), show high accuracy but modest typing rates. Overt speech-based techniques for decoding speech offer expedited communication rates. However, they require either physical vocal tract movement (Herff et al., 2015; Anumanchipalli et al., 2019; Makin et al., 2020) or mental articulation imagination (Moses et al., 2021; Willett et al., 2023). This engenders language dependency and pronunciation variations across languages. Another line of research tackles articulation dependency by decoding imagined speech (Nieto et al., 2022) or reading text (Sun et al., 2019; Panachakel and Ramakrishnan, 2021). Our work follows this line of decoding reading text directly from EEG signals.

EEG-to-Text Translation Prior investigations into EEG-to-text translation, as documented in the literature (Herff et al., 2015; Sun et al., 2019; Anumanchipalli et al., 2019; Makin et al., 2020; Panachakel and Ramakrishnan, 2021; Moses et al., 2021; Nieto et al., 2022), have demonstrated com-

Methods	BLEU-N				ROUGE-1		
	N = 1	N = 2	N = 3	N = 4	F	P	R
Only Global Transformer	0.411	0.243	0.143	0.089	0.324	0.356	0.298
+ One Convolutional Transformer	0.440	0.273	0.171	0.111	0.348	0.381	0.322
+ Three Convolutional Transformers	**0.447**	**0.280**	**0.180**	**0.123**	**0.357**	**0.389**	**0.331**

Table 6: Ablation study of different training strategies of the multi-view transformer.

(1)	Ground Truth: It's not a particularly **good film**, but neither is it a **monsterous** one.
	Baseline Output: was a a bad **good** story, but it is it **bad bad**. one.
	EEG2TEXT output: 's a a **great** romantic **movie**, but it is not the **disaster** movie one.
(2)	Ground Truth: He won a **Nobel Prize in Chemistry** in 1928
	Baseline Output: was the Pulitzer Prize for Literature in 18.
	EEG2TEXT Output: won **Nobel Prize in Chemistry** for 1935 for
(3)	Ground Truth: The book was awarded the 1957 **Pulitzer Prize for Biography**.
	Baseline Output: first is published the Pulitzer **Pulitzer Prize** for Fictionography.
	EEG2TEXT Output: book is a the **Pulitzer Prize for Biography** and.

Table 7: Case study of the output sentences comparing EEG2TEXT and the baseline method (Wang and Ji, 2021).

mendable accuracy when applied to limited and closed vocabularies. Nevertheless, these studies encounter challenges in attaining comparable levels of accuracy when confronted with more extensive and open vocabularies. New investigations have expanded their focus from closed-vocabulary EEG-to-text translation to encompass open-vocabulary scenarios (Wang and Ji, 2021; Willett et al., 2023; Tang et al., 2023; Duan et al., 2023). The two research studies most similar to our work are a baseline method (Wang and Ji, 2021) and DeWave (Duan et al., 2023). The baseline method proposes a framework utilizing transformer and pre-trained BART language models, which establish baseline performance of open-vocabulary EEG-to-text translation. DeWave employs a quantization encoder to derive discrete encoding and aligns it with a pre-trained language model for the open-vocabulary EEG-to-text translation. The limitations of both the baseline method and DeWave lie in their reliance on eye-tracking calibration for word-level EEG feature extraction that introduces error propagation and lacks generalizability to scenarios such as inner speech decoding. EEG2TEXT improves the open-vocabulary EEG-to-text translation performance as well as enhancing the generality by requiring only sentence-level EEG signals as input.

EEG Pre-Training Recent work, such as BrainBERT (Wang et al., 2023), BENDR (Kostas et al., 2021), MAEEG (Chien et al., 2022) and LaBraM (Jiang et al., 2024), has been done on EEG signal pre-training that greatly inspired EEG2TEXT.

BrainBERT converts intracranial recordings to spectrograms, masks multiple continuous bands of random frequencies and time intervals from spectrograms, and reconstructs the spectrogram. BENDR uses a convolutional layer to convert the raw EEG signals to embedding features, which are masked by using masked token prediction (Devlin et al., 2018) and reconstructed. MAEEG uses the same input, convolutional layer, and masking strategy as BENDR while MAEEG's reconstruction goal is the raw EEG signals. LaBraM segments the EEG signal into channel patches and is pre-trained by predicting the masked EEG channel patches. EEG2TEXT directly masks the raw EEG signals with the pre-training objective to reconstruct the raw EEG signals. EEG2TEXT also experimented with various masking strategies and incorporated EEG signals for the pre-training process.

6 Conclusion

In this work, we proposed a novel EEG-to-text decoding model, EEG2TEXT that takes raw EEG signals as input and leverages EEG pre-training and a multi-view transformer to enhance the decoding performance. EEG2TEXT achieved superior performance for open-vocabulary EEG-to-text decoding. Future work includes expanding the model's capabilities to EEG signals from diverse multi-modal data.

7 Ethics Statement

This research strictly followed the highest ethical standards and best practices as outlined in the ACL Code of Ethics. ZuCo (Hollenstein et al., 2018) and Image-EEG (Gifford et al., 2022) datasets we used are open-source datasets that follow CC-By Attribution 4.0 International license, ensuring there were no concerns regarding privacy, confidentiality, or personal information. Data and pre-trained models are used under a specified license that is compatible with the conditions under which access to data was granted. The data is sufficiently anonymized to make identification of individuals impossible to ensure compliance with ethical guidelines. Moreover, we carefully considered the broader impacts and potential applications of our work to prevent any inadvertent harm or misuse. Therefore, we believe this research is ethically sound.

8 Limitations

In this paper, we proposed a novel EEG-to-text decoding model called EEG2TEXT. Despite our efforts, the model still has some limitations.

Reliance on pre-trained models Our architectural framework relies on the pre-trained model, BART, which may make biased decisions influenced by its pre-training data. While our experiments have not shown explicit performance issues due to biases, we must recognize that this observation may be limited to the specific dataset and pre-trained model we used. It is essential to stay vigilant and continue exploring methods to mitigate and correct potential biases that could arise when using pre-trained models.

GPU requirements Since our multi-view model needs to train multiple convolutional transformers at the same time, there is a certain requirement for the scale of GPU. Laboratory-level GPU can only support the training of a small number of convolutional transformer models at the same time. To train all convolutional transformers at the same time to fully realize the potential of the multi-view model, researchers need a GPU with great performance.

References

Tyson Aflalo, Spencer Kellis, Christian Klaes, Brian Lee, Ying Shi, Kelsie Pejsa, Kathleen Shanfield, Stephanie Hayes-Jackson, Mindy Aisen, Christi Heck, et al. 2015. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science*, 348(6237):906–910.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Chad E Bouton, Ammar Shaikhouni, Nicholas V Annetta, Marcia A Bockbrader, David A Friedenberg, Dylan M Nielson, Gaurav Sharma, Per B Sederberg, Bradley C Glenn, W Jerry Mysiw, et al. 2016. Restoring cortical control of functional movement in a human with quadriplegia. *Nature*, 533(7602):247–250.

Gian Emilio Chatrian, Ettore Lettich, and Paula L Nelson. 1985. Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities. *American Journal of EEG technology*, 25(2):83–92.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. 2022. Maeeg: Masked auto-encoder for eeg representation learning. *Preprint*, arXiv:2211.02625.

William T Cochran, James W Cooley, David L Favin, Howard D Helms, Reginald A Kaenel, William W Lang, George C Maling, David E Nelson, Charles M Rader, and Peter D Welch. 1967. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *Preprint*, arXiv:2309.14030.

Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. 2022. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9:217.

Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick Van Der Smagt, et al. 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Weibang Jiang, Liming Zhao, and Bao liang Lu. 2024. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. 2021. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Henri Lorach, Andrea Galvez, Valeria Spagnolo, Felix Martel, Serpil Karakas, Nadine Intering, Molywan Vat, Olivier Faivre, Cathal Harte, Salif Komi, et al. 2023. Walking naturally after spinal cord injury using a brain–spine interface. *Nature*, pages 1–8.

Phan Luu and Thomas Ferree. 2005. Determination of the hydrocel geodesic sensor nets' average electrode positions and their 10–10 international equivalents. *Inc, Technical Note*, 1(11):7.

Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582.

Stephanie Martin, Iñaki Iturrate, José del R Millán, Robert T Knight, and Brian N Pasley. 2018. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, 12:422.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227.

Sebastian Nagel and Martin Spüler. 2018. Modelling the brain response to arbitrary visual stimulation patterns for a flexible high-speed brain-computer interface. *PloS one*, 13(10):e0206107.

Ladislas Nalborczyk, Romain Grandchamp, Ernst HW Koster, Marcela Perrone-Bertolotti, and Hélène Lœvenbruck. 2020. Can we decode phonetic features in inner speech using surface electromyography? *PloS one*, 15(5):e0233282.

Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. 2022. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific Data*, 9(1):52.

Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. 2021. Decoding covert speech from eeg-a comprehensive review. *Frontiers in Neuroscience*, 15:392.

Chethan Pandarinath, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. 2017. High performance communication by people with paralysis using an intracortical brain-computer interface. *Elife*, 6:e18554.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2022. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7047–7054.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. 2023. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*.

Zhenhailong Wang and Heng Ji. 2021. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. *CoRR*, abs/2112.02690.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

A Evaluation Metrics

We utilize BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-1 evaluation metrics to compare the performance of EEG2TEXT with the baselines.

The BLEU-N scores (N = 1, 2, 3, 4) are used to measure the quality of the generated text, with higher values indicating better performance.

$$\text{BLEU} = \text{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \cdot \log\left(\frac{\text{count}_{\text{clip},n}}{\text{count}_{\text{ref},n}}\right)\right), \tag{3}$$

where BLEU represents the BLEU score; BP represents the brevity penalty; N represents the max n-gram order; w_n represents the n-gram weights; $count_{clip,n}$ represents count of candidate n-grams in reference and $count_{ref,n}$ represents count of reference n-grams.

ROUGE-1 scores, which include F (F1-score), P (precision), and R (recall), are used to evaluate the overlap between generated text and reference text.

$$\text{ROUGE-1} = \frac{\sum_{\text{ref}} \sum_{\text{1-gram}} \min(\text{match}, \text{ref})}{\sum_{\text{ref}} \sum_{\text{1-gram}} \text{ref}}, \tag{4}$$

where ROUGE-1 represents the ROUGE-1 score; match represents the count of matching 1-gram; ref represents the count of 1-gram.

B Parameter Study

We used four A40 GPUs as our computing infrastructure and each training epoch took about 40 minutes. The ZuCo dataset (Hollenstein et al., 2018) split of our experiments are shown in Table 8. The optimal hyper-parameters for our results are listed in Table 9. The value ranges of each hyper-parameter are listed below:

- Batch Size $\in \{4, 8, 16\}$
- Learning Rate $\in \{1\times10^{-6}, 3\times10^{-6}, 5\times10^{-6}, 7.5\times10^{-6}, 8\times10^{-6}, 9\times10^{-6}, 1\times10^{-5}, 2\times10^{-5}, 3\times10^{-5}, 4\times10^{-5}, 5\times10^{-5}, 7.5\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 7.5\times10^{-4}, 1\times10^{-3}\}$
- Epoch $\in \{15\}$

C EEG to Spectrogram

Figure A1 shows a piece of EEG signals and its corresponding spectrogram.

# Train	# Dev	# Test
10967	1392	1444

Table 8: Statistics of ZuCo (Hollenstein et al., 2018), depicting the sizes of the training, testing, and development set.

Methods	Batch Size	Learning Rate
EEG2TEXT (Convolutional Transformer)	4	1×10^{-5}
EEG2TEXT (+ Pre-training)	4	5×10^{-5}
EEG2TEXT (+ Multi-View Transformer)	4	5×10^{-5}

Table 9: Optimal hyper-parameters for EEG2TEXT ablations.

D Zero-Shot Image-to-Text Translation

Figure A2a and A2b show the zero-shot image-to-text translation results. We directly input the EEG signals of image-EEG data into the multi-view transformer model after training, and the output results are image-to-text translation results. The first image contains multiple cars, and the output accurately captures the "car" keyword. The second image contains a fish, and the output captures the "fish" keyword equally accurately.

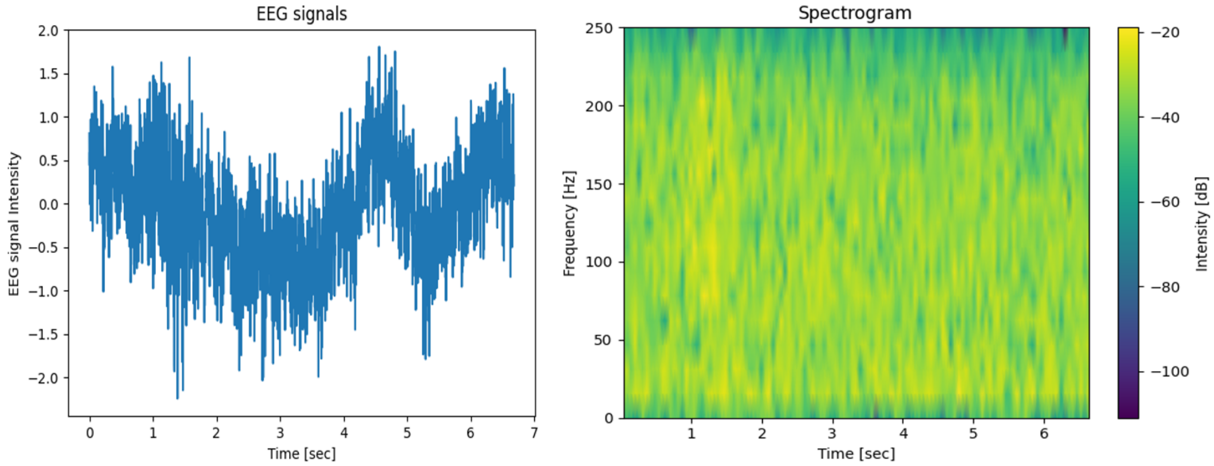

Figure A1: a piece of EEG signals and its corresponding Spectrogram

(a) An image of car. The translation result of EEG2TEXT is: "alog,,. **car**,,,,,,,,,,,,,,,,,,,,,"

(b) An image of car. The translation result of EEG2TEXT is: "**fish**,,... has,,,,,,,,,,,,,,,,,,,,,,,,"

Figure A2: Zero-Shot Image-to-Text Translation.