SOnET: Towards Open-vocabulary EEG-to-text Decoding via Neural Signature Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Decoding text from non-invasive brain signals such as electroencephalogram (EEG) has demonstrated potential as a promising brain-computer interface (BCI) application, especially for individuals with communication difficulties. However, key challenges remain in EEG representation learning, particularly differences in individual neural signatures, hindering the development of generalizable EEG-to-text translation systems. We propose SOnET, a novel open-vocabulary EEG-to-text decoding framework that addresses the long-standing challenge of individual neural signature variability by disentangling the neural signatures of the subjects in a split-latent space using a contrastive loss within an autoencoder. Subsequently, the learned signature embeddings from the autoencoder are fused with the EEG embeddings via cross-attention to enable EEG-to-text decoding that accounts for individual neural signature variability. Additionally, we employ a perplexity loss that adaptively scales the gradient updates based on prediction uncertainty during training. Experimental evaluation shows that our proposed model achieves a 15.3% relative improvement in BLEU-1 and a 17.3% relative improvement in ROUGE-1 F1-score over the strongest baseline on three reading tasks from the ZuCo EEG-text dataset. Moreover, on the ROAMM dataset, SOnET yields a 13.6% relative gain in BLEU-1 and a 2.4% improvement in ROUGE-1 F1-score. Detailed ablation highlights the contribution of each component of our proposed model. Our code can be accessed at `https://anonymous.4open.science/r/SOnET-C2F8/`.

1 Introduction

Brain-computer interfaces (BCIs) are a rapidly evolving research field, especially driven by the recent advancements in artificial intelligence (AI), enabling direct interaction between humans and computers (Zhang et al., 2020; Asgher et al., 2023; Subasi et al., 2025). Decoding text from non-invasive brain signals like electroencephalogram (EEG) is a promising area within BCIs, with many practical applications (Wang & Ji, 2022; Murad & Rahimi, 2024; Qiu et al., 2025). In particular, people with communication difficulties, such as stroke survivors, Amyotrophic Lateral Sclerosis (ALS) patients, and individuals with locked-in syndrome, can greatly benefit from EEG-to-text translation systems (Majumdar et al., 2023; Zhang et al., 2024a; Schalk et al., 2024). These systems can also enable hands-free and voice-free control for accessibility, allowing direct communication between brain signals and computer devices. An overview of EEG-to-text systems is shown in Figure 1.

There have been a few advancements in open-vocabulary EEG-to-text decoding in recent years, including works such as (Wang & Ji, 2022; Feng et al., 2023; Zhou et al., 2024; Wang et al., 2024b), which integrate EEG encoders and large language models (LLMs) into the architectural pipeline. However, a critical limitation in the prior works is the implicit use of *teacher forcing* during evaluation, a technique that assumes the availability of ground truth tokens of the previous time-steps to predict the current token in a sequence-to-sequence model (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025). In real-life scenarios, the ground-truth tokens are not readily available to these systems, making teacher forcing impractical. Moreover, a fundamental and long-standing issue in brain signal decoding, often overlooked in prior work, is the variability across subjects (Bomatter & Gouk, 2026; Nørskov et al., 2023; Mishra et al., 2025; Murad & Rahimi, 2024; Zhang et al., 2024b). Specifically, EEG

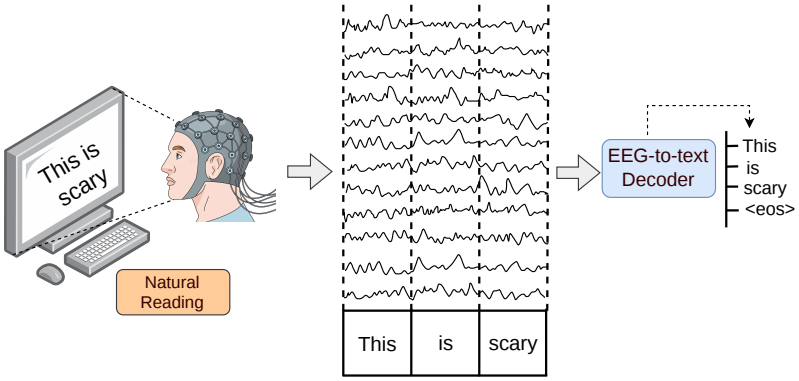

Figure 1: Overview of EEG-to-text translation systems. In the data collection process (Hollenstein et al., 2018; 2020; Sun et al., 2026), subjects are instructed to read natural sentences shown on a computer monitor while their EEG is recorded. This data is used to train EEG-to-text models. During inference, the trained model translates unseen EEG into text.

signals can vary significantly between individuals due to physiological and cognitive differences, making it challenging to develop a robust and generalizable EEG-to-text model. Figure 2 illustrates this issue of individual neural signature variability using brain topographical maps. The brain activity patterns for four sample words show considerable differences across different subjects, highlighting the challenge that neural signature variability poses in the development of robust BCIs.

In this work, we propose a novel method **SOnET** (Neural **S**ignature Representation Learning for **O**pen-vocabulary **E**EG-to-**T**ext Decoding), which addresses individual neural signature variability in EEG signals and enhances the generalization capability of the system. Our approach consists of two main components: a neural signature representation learning module and an EEG-to-text decoder. The neural signature representation learning module is designed as an autoencoder that disentangles the latent space into subjects' signature-specific and residual components using contrastive losses. Next, the EEG-to-text decoder integrates the signature embeddings from the signature representation learning model into the EEG embeddings through cross-attention mechanisms. This enables the decoder to incorporate subjects' neural signature-related features and guide the text generation process accordingly. Additionally, we incorporate a perplexity loss to dynamically scale the gradient updates during training based on model confidence. We test the performance of our method both with and without teacher forcing to evaluate its performance and further conduct extensive ablation studies to demonstrate the effectiveness of each component in our system.

To summarize, our contributions are as follows:

- We introduce SOnET, a novel open-vocabulary EEG-to-text decoding framework that explicitly addresses subjects' neural signature variability by disentangling signature-specific and residual features using a contrastive autoencoder. It then leverages the signature embeddings from the autoencoder to train the EEG-to-text decoder. Experiments demonstrate that our method surpasses the performance of prior works and sets a new state-of-the-art. Moreover, contrary to subject-dependent approaches that require training separate models for each subject, our SOnET model is designed to generalize across multiple subjects using a *single* model.
- We adopt a perplexity loss function for EEG-to-text decoding, which adaptively scales gradients based on model confidence. This encourages larger gradient updates on uncertain predictions in the early stages of training, while later stabilizing the process with smaller updates as training progresses. Ablation studies demonstrate that perplexity loss outperforms standard cross-entropy for EEG-to-text decoding with a relative improvement of approximately 14% in BLEU-1 score.
- We perform extensive evaluation with and without teacher forcing, highlighting a performance gap between commonly reported results with teacher forcing and results obtained under realistic inference

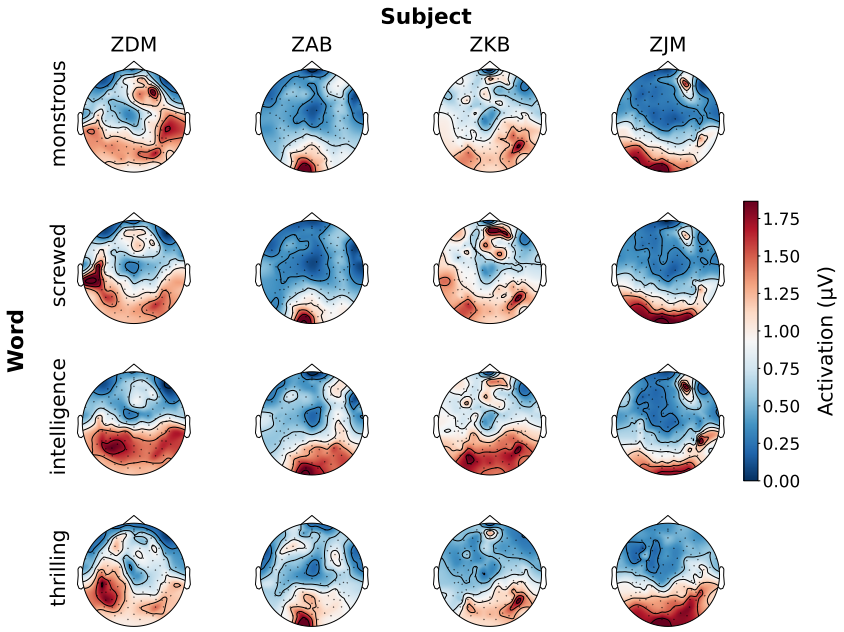

Figure 2: Brain topographical maps for EEG activation patterns across four subjects (*ZDM, ZAB, ZKB* and *ZJM*) and four words (*monstrous*, *screwed*, *intelligence* and *thrilling*). These maps illustrate distinct neural signatures for different individuals.

conditions. To facilitate open-source research, we make our code publicly available at `https://anonymous.4open.science/r/SOnET-C2F8/`.

2 Related Work

2.1 Decoding Language from Brain Signals

Open-vocabulary vs. closed-vocabulary. Prior studies have used brain signals for diverse BCI applications, such as the prediction of neurological and psychological disorders (Liu et al., 2025; Klepl et al., 2023), decoding of emotions (Zhang et al., 2022; Zhang & Etemad, 2023; Ahmadzadeh Nobari Azar et al., 2024), and motor imagery (El Ouahidi et al., 2024; Bhalerao & Pachori, 2025; Zhao et al., 2025; She et al., 2025), among others. Despite the significant progress of BCIs in many applications, decoding language from EEG remains a challenging problem since language processing involves highly complex and rapid neural activity in the brain (Murad & Rahimi, 2024; Zhang et al., 2024b). Most efforts in this regard have been constrained to closed-vocabulary settings, where EEG is mapped to only a fixed and small set of phonemes/words. For instance, imagined speech recognition from EEG signals was performed to decode words within constrained vocabularies, such as decoding four Spanish words (Nieto et al., 2022) or fifteen English words (Kamble et al., 2023). This restriction to small vocabularies significantly limits the practical application of these systems in real-world scenarios.

In contrast, the first study on open-vocabulary EEG-to-text decoding from non-invasive EEG signals was conducted in (Wang & Ji, 2022), where the model was able to generate arbitrary words or sentences that were not explicitly part of the training set. They introduced a framework with a multi-layer transformer encoder to process EEG signals and a pre-trained sequence-to-sequence LLM (BART model) to decode text (Wang & Ji, 2022; Lewis et al., 2020; Vaswani et al., 2017). In their work, they used the ZuCo dataset containing EEG and text pairs in an open-vocabulary setup to train the model (Hollenstein et al., 2018; 2020).

Following the first notable work on open-vocabulary EEG-to-text translation (Wang & Ji, 2022), few studies were conducted adopting the state-of-the-art methods using the same ZuCo dataset (Feng et al., 2023; Wang et al., 2024b; Duan et al., 2024; Zhou et al., 2024; Xi et al., 2023). Contrastive learning has been predominantly used in these studies to align EEG representations and text representations (Feng et al., 2023; Wang et al., 2024b; Duan et al., 2024; Zhou et al., 2024). Specifically, to mitigate the discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation, contrastive learning is employed to pull the semantically similar EEG samples together while pushing the dissimilar ones in (Feng et al., 2023). Moreover, to make the model learn progressively, curriculum learning is adopted for making the contrastive pairs (Feng et al., 2023). Another study proposed a hybrid EEG-to-text decoding method by combining cross-modality self-learning between EEG and text with intra-modality self-reconstruction of EEG or text features by employing an autoencoder and contrastive loss (Wang et al., 2024b). However, these studies used frequency-domain EEG features, which required segmenting the continuous raw EEG into word-level features following an eye-fixation procedure. To address this issue, a quantized variational encoder was used to discretize the raw EEG, and pre-trained LLMs have been used for aligning EEG and text representations using contrastive learning for EEG-to-text translation (Duan et al., 2024). Although this method eliminates the need for an eye-fixation procedure to segment the continuous EEG, the performance of this method with raw EEG signals is found to drop significantly.

Similar to (Duan et al., 2024), a study proposes discretizing the EEG to make the EEG embeddings more resilient to variations and aligning the discrete EEG embeddings with text embeddings using contrastive learning to extract semantic information (Zhou et al., 2024). However, contrary to (Duan et al., 2024), they used a discrete Conformer encoder (D-Conformer) to address the shortcomings of existing EEG encoders by capturing both local and global EEG patterns. Moreover, frequency-domain EEG features were fed to the D-conformer encoder of their framework. Another work proposes two distinct models to individually process functional magnetic resonance imaging (fMRI) and EEG data (Xi et al., 2023). In their EEG-to-text translation model, the EEG time series is reconstructed through an encoder-decoder framework, and a pre-trained LLM is used for text translation. However, their work cannot be extensively compared with prior studies, as they randomly split only the ZuCo v1.0 dataset, and their code is not publicly available for reproduction using both the ZuCo v1.0 and ZuCo v2.0 datasets employed in earlier works (Feng et al., 2023; Wang et al., 2024b; Duan et al., 2024; Zhou et al., 2024).

Invasive vs. non-invasive brain signals. Several studies have utilized invasive brain signals, such as electrocorticography (ECoG) for brain-to-text decoding (Willett et al., 2021; Makin et al., 2020; Pandarinath et al., 2017; Morgan et al., 2025; Zada et al., 2025; He et al., 2025). While ECoG offers superior temporal and spatial resolution with a high signal-to-noise ratio compared to non-invasive EEG signals, its reliance on surgical procedures restricts its applicability to healthy individuals (Ho et al., 2025).

Limitations. Despite the recent developments in EEG-to-text decoding as discussed earlier, there are key limitations. First, the previous works rely on the assumption of teacher forcing during evaluation, in which the ground truth tokens for the previous time-steps are used to predict the token at the current time-step (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025). Although this experimental setting eliminates accumulation error and makes the task easier given the difficulty of EEG-to-text decoding, in real-world scenarios, the availability of ground-truth tokens is unrealistic. Second, none of the studies above addressed the issue of neural signature variability, which is a well-known challenge in BCI applications. To the best of our knowledge, our work is the first to address the individual neural signature variability for open-vocabulary EEG-to-text translation.

2.2 Neural Signature Variability in BCI

Neural signature variability in BCI refers to the differences in brain activity among individuals performing the same cognitive task (Kim et al., 2024). This is a well-known issue in BCI applications, caused by biological differences, cognitive styles, and other neurophysiological factors (Nørskov et al., 2023; Murad & Rahimi, 2024). Two main approaches have been proposed in the literature to address neural signature variability: clustering of subject-specific features and transfer learning-based methods. For example, one study reduced variability by clustering subjects based on the similarity of their EEG signals in an embedding space using a deep learning metric for motor imagery BCI (Kim et al., 2024). Another work combined

adversarial training, which makes subject information indistinguishable, with Inter-subject Feature Distance Minimization (IFDM) using Wasserstein distance to reduce neural signature variability, enabling robust subject-independent EEG-based driver fatigue classification (Hwang et al., 2021).

Other studies have leveraged data from source subjects to improve model generalization on a target subject, often through domain adaptation techniques that align feature distributions or generate transferable representations across subjects (Yin et al., 2024). A co-adaptive transfer learning framework was proposed that first trains a generalized model using selected cross-subject data, then fine-tunes it with subject-specific transfer learning and online adaptive updates across sessions to address both individual neural signature variability and inter-session variability for motor imagery decoding (Zhang et al., 2025). However, this method is less effective in practice, as it requires separate models for different subjects.

Pre-training transformer-based models on large amounts of unlabeled data from numerous subjects is a promising method to handle neural signature variability. Several studies have explored this direction, including a mask-based dual self-supervised learning method (Wang et al., 2024a), a patch-based masked EEG reconstruction approach (Wang et al., 2025), a vector-quantized neural spectrum prediction-based method (Jiang et al., 2024), and a topology-agnostic framework (Yi et al., 2023), among others. A recent line of work explores disentangled EEG representation learning, where latent representations are separated into subject-specific and task-specific components (Nørskov et al., 2023; Mishra et al., 2025). While these approaches focus on learning robust EEG representations for subject identification, we employ neural signature representation learning in the context of EEG-to-text decoding.

3 Method

3.1 Problem Definition

Let $E \in \mathbb{R}^{n \times d \times c}$ denote a sequence of EEG features, where n is the sequence length, d is the feature dimension, and c is the number of channels, and $S = \{s_1, s_2, \ldots, s_T\}$ denote the ground-truth text sequence of length T, where each s_t is a token from the vocabulary. The task of EEG-to-text translation can be framed as learning a conditional language model $P(S'|E)$ that maps E to natural language, with the objective of maximizing the likelihood of producing the text sequence $S' \approx S$.

3.2 Solution Overview

Our SOnET framework adopts a two-stage training process: (1) training the neural signature representation learning model to extract the signature embeddings, and (2) training the EEG-to-text decoder model leveraging those signature-specific embeddings from the first stage using a cross-attention module (Figure 3). Additionally, we employ a perplexity loss to train the EEG-to-text decoder. The key components of our method include the signature representation learning model and EEG-to-text decoder, which are discussed in this section.

3.3 Stage 1: Neural Signature Representation Learning Model

EEG signals vary significantly across individuals for the same stimulus, which is a well-known issue in BCI (Nørskov et al., 2023; Mishra et al., 2025). This variability makes it difficult to develop a single generalizable model that can map EEG to text across many subjects. To address this issue, we introduce a signature representation learning model in our SOnET framework that implicitly provides our EEG-to-text model with subjects' cognitive signature-specific features extracted from EEG, allowing the EEG-to-text model to adapt its training to individual differences. At inference time, our EEG-to-text model leverages these signature embeddings so that it can generate text tailored to that individual's brain signal patterns. To enable this mechanism in our SOnET framework, we first train our signature representation learning model with EEG from multiple subjects. Next, we train the EEG-to-text model leveraging the signature embeddings while keeping our signature representation learning model frozen. Our approach for signature representation learning is inspired by prior works such as (Nørskov et al., 2023; Mishra et al., 2025). Specifically, we use a

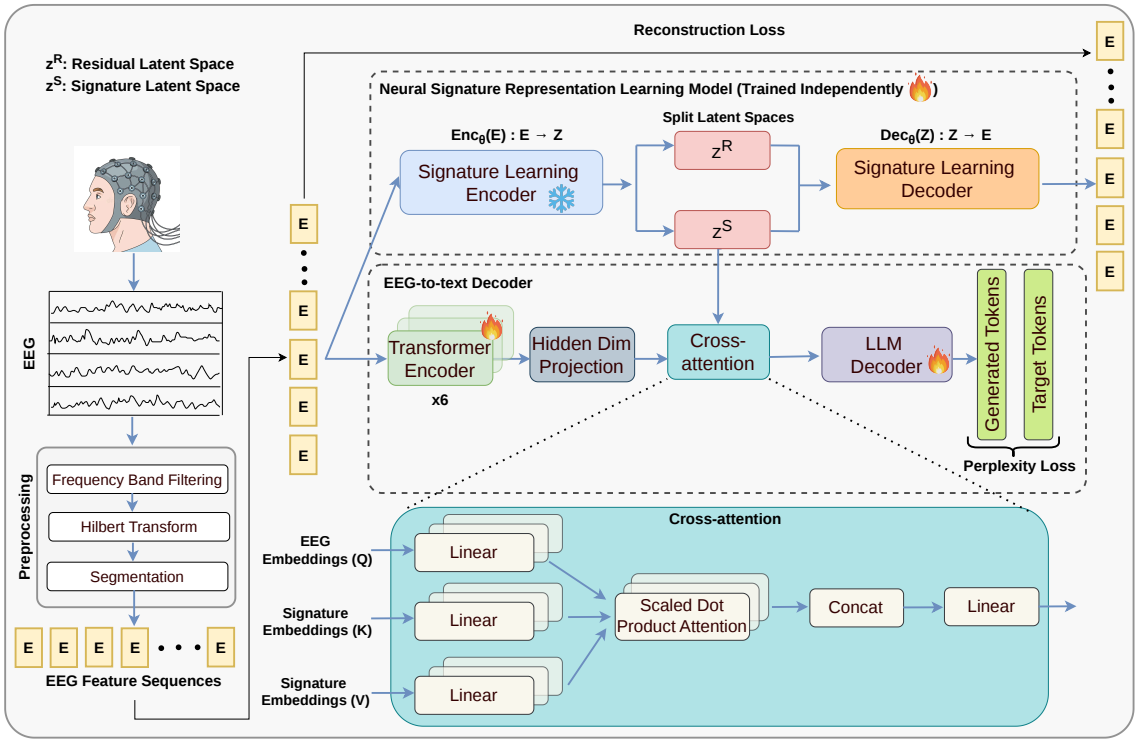

Figure 3: The architecture of our proposed SOnET framework.

variational autoencoder (VAE) with a contrastive learning objective to disentangle the signature embeddings from the EEG.

VAE backbone. We design the signature representation learning model of our SOnET framework with a VAE architecture, a probabilistic generative model well-suited for learning structured latent representations (Kingma et al., 2019). At first, we perform the Z-score normalization on the the input EEG E by transforming E into a normalized representation, $E' = (E - \mu)/\sigma$, where, μ and σ are the means and standard deviations computed from the EEG sample. This ensures better generalization instead of overfitting to specific noisy EEG patterns. We feed E' into an encoder $Enc_\theta(E')$ that maps E' into a latent vector Z and then we attempt to reconstruct E' from Z using our decoder $Dec_\phi(Z)$. This way, the latent vector Z captures a compressed representation of the EEG in a continuous latent space.

EEG data is inherently temporal and often contains long-range dependencies between brain activity patterns (Linkenkaer-Hansen et al., 2001; Smith et al., 2019). To capture these dependencies more effectively, our encoder and decoder in the VAE contain both convolutional and transformer blocks. While the convolutional blocks reduce the dimensionality of the EEG features, the transformer comprises of multi-head self-attention blocks, which capture both short and long-term temporal correlations in the EEG signal through contextualizing each time point with respect to the full sequence (Vaswani et al., 2017). Thus, we ensure that the latent representation Z is context-aware, meaning it encodes not only local features, but also the temporal dynamics of brain activity patterns.

Split-latent representations. In the signature representation learning model of our SOnET framework, we decompose the latent space Z of our VAE into two distinct subspaces: (1) signature-related latent space (z^S) that encodes characteristics unique to subjects' signature, and (2) residual latent space (z^R) that captures all other information in the EEG signal that is not directly tied to signature, for instance, those related to specific types of text used to elicit EEG responses. Our encoder outputs this split representation as $(z^S, z^R) = Enc_\theta(E')$, and our decoder recombines them to reconstruct the input: $\hat{E} = Dec_\phi(z^S, z^R)$. This structured latent design enforces a factorization of information in the EEG representation space, preventing

signature-related features from leaking into residual features and vice versa. We discuss the method for disentangling the EEG representation space below.

Contrastive learning for disentanglement. To ensure that z^S indeed contains signature-related features, we adopt contrastive learning in the signature representation learning component of our SOnET framework. More precisely, we perform the disentanglement in the latent space by ensuring that embeddings from the same subject should be closer in signature latent space, while embeddings from different subjects should be farther apart in the same latent space. By explicitly enforcing this objective in the signature latent subspace, signature-related features can be disentangled from residual variations. We do the same in the residual latent space in which embeddings from the same type of text are closer to each another, while embeddings from different types of texts are further apart.

We employ a multi-class N-pair sampling strategy (Sohn, 2016), where each batch contains EEG segments from K distinct subjects, with two instances per subject. This yields exactly one positive pair for each subject (same subject, different segments). We first measure the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss, which is a type of contrastive loss widely used in representation learning (Chen et al., 2020). NT-Xent loss is derived from the InfoNCE loss (Oord et al., 2018), scaled by a temperature parameter τ (Chen et al., 2020). Then, we optimize the symmetric cross-entropy of NT-Xent loss, which is equivalent to the CLIP loss proposed by (Radford et al., 2021). Let $\mathbf{Z}_A \in \mathbb{R}^{C\times K}$ and $\mathbf{Z}_B \in \mathbb{R}^{C\times K}$ denote the latent representation matrices for K paired samples with a latent feature dimension of C. Then, the NT-Xent loss $L_{\text{NT-Xent}}$ and the CLIP loss L_{CLIP} are defined as,

$$L_{\text{NT-Xent}}(\mathcal{H}; Z', Z'', k) = -\log \frac{\exp(\frac{\text{sim}(z'_k, z''_k)}{\tau})}{\sum_{i=1}^{K} \mathbb{1}_{[i\neq k]} \exp(\frac{\text{sim}(z'_k, z''_i)}{\tau})} \tag{1}$$

$$L_{\text{CLIP}}(\mathcal{H}; Z^A, Z^B) = \frac{1}{K}\sum_{k=1}^{K}\Big(L_{\text{NT-Xent}}(\mathcal{L}; Z^A, Z^B, k) + L_{\text{NT-Xent}}(\mathcal{L}; Z^B, Z^A, k)\Big) \tag{2}$$

where, $\mathbb{1}_{[c]} \in \{0, 1\}$ is an indicator function, producing 1 if and only if the condition c is true, and $\text{sim}(z'_i, z''_j)$ denotes the cosine similarity between two embeddings. $\mathcal{H} = \{z^S, z^R\}$ denotes the latent space as the loss is applied separately on the signature latent space z^S and the residual latent space z^R, yielding $L_{CLIP}(z^S; ., .,)$ and $L_{CLIP}(z^R; ., .,)$, respectively. These contrastive objectives refine the latent space to be more separable with respect to signatures of the subjects' and disentangled from residual variations.

Overall, by training a VAE with transformer-based temporal encoding, and latent space decomposition using contrastive learning, the signature representation learning component of our SOnET framework captures the neural stylistic features that are useful for the downstream EEG-to-text translation task.

3.4 Stage 2: EEG-to-text Decoding

Given the EEG features E, our EEG-to-text decoder in the SOnET framework maximizes the conditional probability of the predicted sentence S',

$$P(S'|E) = \prod_{t=1}^{T} P(s'_t \in V \mid E, s'_{t-1}) \tag{3}$$

where, $S' = (s'_1, s'_2, \ldots, s'_T)$ represents the predicted text sequence, s'_t is the predicted token at time step t, T is the length of the sequence, and V is an open-vocabulary set containing English tokens. This formulation follows the standard left-to-right auto-regressive decoding used in neural sequence generation (Sutskever et al., 2014; Vaswani et al., 2017). Our EEG-to-text decoder consists of three primary components: (1) A multi-layer transformer encoder that captures the temporal dependencies and contextual relationships within the EEG sequence and projects them into a latent space. (2) A cross-attention module that fuses

contextualized EEG representations with signature embeddings obtained from a trained neural signature representation learning model from stage 1. (3) A pre-trained sequence-to-sequence (seq-to-seq) decoder that maps the fused EEG representations into a textual output sequences.

Contextualized encoding of EEG features. We begin by feeding the input EEG features E into the multi-layer transformer encoder. The transformer-based encoder projects the EEG features into a latent embedding space through a learned projection matrix W_e and produces contextualized EEG representations, $h_f = \text{Enc}(E)^T W_e$. Here, h_f encodes the EEG data, such that each time step in the EEG sequence is represented in the context of its preceding and following elements. We choose the transformer-based encoder since EEG signals often contain long-range dependencies (Linkenkaer-Hansen et al., 2001; Smith et al., 2019), and transformers, with their multi-head self-attention mechanism, are particularly well-suited to handling such relationships (Vaswani et al., 2017).

Fusion of contextualized EEG embeddings and signature embeddings. In parallel, we feed the same input EEG data E into the frozen encoder of our trained signature representation learning model (from stage 1), which produces a signature embedding, h_s. We extract h_s to leverage it in the second stage of our EEG-to-text translation pipeline, since this enables the EEG-to-text decoder to adapt its generation process to the unique characteristics of a subject or a group of subjects with similar brain signal responses. This method can reduce neural signature variability in the mapping from EEG to language. To this end, our SOnET method fuses h_f and h_s using a cross-attention mechanism, which allows the model to query signature embeddings when processing each contextualized EEG representation. Formally, the cross-attention operation is defined as,

$$h_x = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

where, $Q = h_f$ (contextualized EEG features), and $K = V = h_s$ (signature-specific features). The scaling factor $\sqrt{d_k}$ normalizes the dot-product similarity scores to stabilize training. The output of this module, h_x, represents the embeddings that are both temporally contextualized and adapted to subject-specific neural signature patterns.

Decoding into text using an LLM. Finally, the fused representations h_x are fed into a seq-to-seq decoder, which transforms them into a probability distribution over the vocabulary V at each time step, $P(s'_t \in V) = \text{Dec}(h_x)^T W_o$. Here, W_o is the learned projection matrix mapping the decoder's output hidden states to the vocabulary logits. Our seq-to-seq decoder is initialized from the weights of a pre-trained LLM, which provides two key advantages: (1) it offers strong prior knowledge of grammar, syntax, and semantics, reducing the need for extensive EEG-text parallel training data and (2) it allows for better generalization to unseen linguistic topics and text materials.

Teacher forcing. An important consideration in training seq-to-seq models is whether to use teacher forcing, which is a strategy where the ground truth token s_{t-1} is used as input at each step instead of the model's own previous prediction s'_{t-1} during training (Williams & Zipser, 1989). Teacher forcing can prevent divergence by providing the decoder with accurate context. However, it can also lead to exposure bias, where the model performs poorly when it must rely on its own predictions during inference. Prior EEG-to-text studies have frequently adopted teacher forcing during evaluation (Wang & Ji, 2022; Duan et al., 2024; Wang et al., 2024b; Feng et al., 2023; Zhou et al., 2024). In our work, we evaluate both strategies with and without teacher forcing to assess their impact on decoding performance.

Overall, our EEG-to-text decoder integrates temporal modeling using the multi-layer transformer encoder, cross-attention module that fuses signature-related features with EEG embeddings, and linguistic priors from a large language model into a unified architecture.

Objective. For training the EEG-to-text decoder in the second stage, we adopt a perplexity-based loss function for our EEG-to-text decoder. Formally, for a sequence of N predicted tokens $s'_1, \ldots, s'_N$, and ground truth tokens $s_1, \ldots, s_N$, perplexity is defined as,

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} P(s_i) \log P(s'_i \mid s'_{<i})\right) \tag{5}$$

Similar to cross-entropy, perplexity loss quantifies the divergence between the predicted probability distribution $P(s'_i \mid s'_{<i})$ and the true distribution $P(s_i)$. However, unlike cross-entropy, which penalizes each misprediction linearly, the exponential form amplifies the penalty for uncertain predictions in the early stages of training. This encourages larger gradient updates when the model is less confident, while progressively stabilizing as predictions improve.

Table 1: Dataset Statistics. SR: Normal Reading (Sentiment), NR: Normal Reading (Wikipedia), TSR: Task Specific Reading (Wikipedia) for the ZuCo dataset. (Hollenstein et al., 2018; 2020).

Dataset	**Reading Task**	**Training Sample**	**Validation Sample**	**Testing Sample**
ZuCo	SRv1.0 + NRv1.0 + NRv2.0	10710	1332	1407
	SRv1.0 + NRv1.0 + NRv2.0 + TSRv1.0	14407	1790	1799
ROAMM	-	13253	1640	1696

Table 2: Summary of model parameters.

Hyper-parameter	**signature rep. model**	**EEG-to-text**
Batch size	256	32
Random seed	1	312
Num. of transformer layers	4	6
Embedding size	-	2048
Att. heads in cross att.	-	16
In channel size	56	-
Out channel size	-	-
Dropout	0.3	0.3
Max epochs	200	70
Warmup epochs	30	30
Optimizer	Adam	SGD
Weight decay	-	-
Learning rate (Base LR)	0.0001	0.00002
Pred LR ($\times$ Base LR)	0.1$\times$	0.1$\times$
Centering momentum	0.9	0.9

4 Experimental Setup

4.1 Datasets

We use two open vocabulary EEG-to-text datasets, including Zurich Cognitive Language Processing Corpus (ZuCo) and Reading Observed At Mindless Moments (ROAMM) to benchmark our proposed model (Hollenstein et al., 2018; 2020; Sun et al., 2026). To the best of our knowledge, ZuCo and ROAMM are the **only publicly available** corpora that provides simultaneous EEG and English text in an open-vocabulary setting. Prior work on open-vocabulary EEG-to-text decoding was restricted to using the ZuCo dataset solely for benchmarking model performance as it was the only dataset for this task till the recent release of ROAMM (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025).

ZuCo. There are two versions of the ZuCo dataset including ZuCo v1.0 and ZuCo v2.0, which are publicly released in 2018 and 2020, respectively (Hollenstein et al., 2018; 2020). ZuCo v1.0 includes recordings from 12 native English-speaking subjects (5 female, 7 male) aged between 25 and 51 years. ZuCo v2.0 includes data from 18 subjects (10 female, 8 male) aged between 23 and 52 years. In both versions, EEG signals are recorded at a sampling rate of 500 Hz using a 128-channel Geodesic Hydrocel system. Cz channel is used as the reference channel. A bandpass filter of 0.1 to 100 Hz is applied to the EEG data to remove unwanted noise and artifacts.

Each of ZuCo v1.0 and ZuCo v2.0 includes three subsets based on different reading tasks: (1) sentiment reading (SR), where subjects read movie review sentences labeled as positive, negative, or neutral, (2) normal reading (NR), involving subjects reading factual Wikipedia sentences with comprehension questions, and (3) task-specific reading (TSR), where subjects searched for predefined semantic relations (e.g., award, employer) within sentences after reading them. We use both versions of the ZuCo dataset in this study and experiment with two different dataset settings with 10710 (three reading tasks) and 14407 (four reading tasks) training samples by combining the subsets following prior work on EEG-to-text decoding (Table 1) (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025).

ROAMM. ROAMM contains EEG data during naturalistic reading with annotations of mind wandering episodes and is suitable for both mind wandering detection and EEG-to-text translation tasks (Sun et al., 2026). EEG is collected from 44 subjects aged between 18 to 64 years (mean 22.6±7.8), including 30 females, 9 males, and 5 non-binary. All subjects are reported to be fluent in English. EEG signals are recorded at a sampling rate of 2048 Hz using a 64-channels BioSemi ActiveTwo system. The dataset is released with a sampling rate of 256 Hz. There are in total 50.3 hours of EEG recordings, among which 2.2 hours of EEG are annotated as mind wandering. We exclude the mind-wandering episodes for EEG-to-text translation to eliminate potential noise. Text materials are collected from five articles selected from Wikipedia: *Pluto (the dwarf planet)*, *the Prisoner's Dilemma*, *Serena Williams*, *the History of Film*, and *the Voynich Manuscript*.

Dataset splitting strategy. For both ZuCo and ROAMM, following prior EEG-to-text decoding works (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025), we split the data for each subject such that 80% of the samples from the corresponding subject are used for training, 10% for validation, and the remaining 10% for testing. Therefore, our dataset splitting strategy ensures that the EEG data corresponding to the same sentence appear in only one of the splits.We also perform a cross-subject generalization analysis using ZuCo, where 18 subject-specific EEG-to-text models are independently trained on individual subjects and evaluated on the remaining unseen subjects.

4.2 Data Pre-processing

We use a sampling rate of 500 Hz and 256 Hz for ZuCo and ROAMM, respectively. For ZuCo, there remain 105 EEG channels after discarding the channels lying on the neck and face from the original 128-channel EEG setup. For pre-processing the continuous EEG data from both ZuCo and ROAMM, we first extract the frequency band-specific features. For this, we use a band-pass filter to convert the raw EEG into eight different frequency bands: theta1 (4–6 Hz), theta2 (6.5–8 Hz), alpha1 (8.5–10 Hz), alpha2 (10.5–13 Hz), beta1 (13.5–18 Hz), beta2 (18.5–30 Hz), gamma1 (30.5–40 Hz), and gamma2 (40–49.5 Hz). We then apply a Hilbert transform to each band which allows us to compute the instantaneous amplitude (envelope) and phase to capture the oscillatory dynamics of EEG more accurately (Kschischang, 2006). Finally, we segment EEG into word-level chunks using word-level gaze durations obtained through an eye-tracking fixation process, as provided in the ZuCo and ROAMM datasets (Hollenstein et al., 2018; 2020).

4.3 Implementation Details

We conduct all the experiments using an Nvidia RTX8000 GPU with 48 GB of memory. In our SOnET framework, both the encoder and decoder of our *signature representation learning* model consist of two layers of convolutional blocks and four transformer layers with 16 attention heads. We train our signature representation learning model with either ZuCo or ROAMM train set at first. In our *EEG-to-text translation* module, the encoder contains six layers of transformer with an embedding size of 2048 and eight attention heads to encode the high-level EEG features. As for the *seq-to-seq decoder*, we utilize the T5 model (Raffel et al., 2020), which treats any task as a text-to-text problem and is pre-trained on a large amount of text. Specifically, we choose the large variant of T5 with 770 million trainable parameters. We use T5 since prior work reported improved performance of T5 as a decoder compared to BART and PEGASUS (Jo et al., 2025). Moreover, the goal of our study is to address neural signature variability by proposing a novel neural signature representation learning model, rather than conducting extensive experiments with the text decoders. We set the embedding size of the decoder to 1024. For our cross-attention block in the EEG-to-text translation module, we use an embedding size of 1024 with 16 attention heads. To train SOnET, we use a stochastic

gradient descent (SGD) optimizer. The batch size is set to 32, and the learning rate is 0.00002. The key hyperparameters are shown in Table 2. We use PyTorch framework to implement our model (Paszke et al., 2019).

4.4 Baselines

Prior work on EEG-to-text translation includes **(1)** BrainBART-Large (Wang & Ji, 2022), **(2)** BrainT5-Large (Jo et al., 2025), **(3)** BrainPEGASUS-Large (Jo et al., 2025), **(4)** BELT (Zhou et al., 2024), **(5)** E2T-PTR (Wang et al., 2024b), **(6)** DeWave (Duan et al., 2024), and **(7)** C-SCL (Feng et al., 2023). Among them, BrainBART-Large, BrainT5-Large, and BrainPEGASUS-Large use the same transformer-based encoder, however, they differ in the decoder architecture used for transforming EEG embeddings into text. All prior works were implemented under the assumption of teacher forcing during inference. We use BrainBART-Large, BrainT5-Large, and BrainPEGASUS-Large for comparison with SOnET since their code is publicly available, making it possible to remove teacher forcing to show the true translation performance. Their hyperparameters were further optimized after removal of the teacher forcing feature to obtain the best possible results for the baselines. Since the code for BELT, DeWave, and C-SCL is not open-sourced and some important scripts for E2T-PTR reproduction were missing, we could not use them for comparison. However, we still report their teacher forcing performance to highlight the inflated results in prior works.

4.5 Metrics

To compare our proposed model with the existing models, we use two commonly used evaluation metrics in text generation, namely, Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004). BLEU is precision-based, meaning it calculates what fraction of the generated n-grams match the reference, while ROUGE is recall-based, measuring how much of the reference text is recovered in the generated output. In our work, we report the corpus-level BLEU score in which all the sentences in the test set are concatenated and BLEU score is computed. This approach is found to be more stable than computing sentence-level BLEU scores (Papineni et al., 2002).

5 Results

5.1 Comparison with Prior Works

We compare SOnET to existing EEG-to-text translation models without the teacher forcing criteria and present the results in Table 3. On ZuCo, we report the results under two different dataset settings: three reading tasks (10,710 training samples) and four reading tasks (14,407 training samples). On three reading tasks, SOnET achieves the best performance with the highest BLEU scores and ROUGE scores, surpassing prior works (Wang & Ji, 2022; Jo et al., 2025). Compared to the strongest baseline (BrainT5-Large), SOnET achieves significant improvements in BLEU scores with approximately 15.3% relative improvement in BLEU-1, 36.3% in BLEU-2, 46.1% BLEU-3, and 59.8% in BLEU-4. For ROUGE, we observe gains of 11.2% in precision, 24.3% in recall, and 17.3% in F1 over the baseline BrainT5-Large model. These results demonstrate SOnET's superior ability to reconstruct both individual words and longer n-gram sequences from EEG. We observe the same trend of better performance from SOnET compared to baselines when evaluated on four reading tasks. Compared to BrainT5-Large, SOnET achieves notable relative improvements, with 8.1%, 16.9%, 13.1%, and 15.6% gains in BLEU-1 to BLEU-4, respectively, and modest gains of 2.2%, 4.6%, and 3.7% in ROUGE-1 precision, recall, and F1.

On ROAMM, SOnET also demonstrates superior EEG-to-text decoding performance compared to prior works. As shown in Table 3, SOnET achieves the highest BLEU scores across all n-gram settings, outperforming BrainBART-Large, BrainT5-Large, and BrainPEGASUS-Large. Compared to the strongest baseline (BrainT5-Large), SOnET achieves approximately 13.6% relative improvement in BLEU-1, 22.8% in BLEU-2, and 104.8% in BLEU-3. While BrainT5-Large fails to generate meaningful 4-gram matches (BLEU-4 = 0.00), SOnET achieves a BLEU-4 score of 1.10, indicating improved reconstruction of longer word sequences

Table 3: Evaluation results ***without*** using teacher forcing during inference on ZuCo and ROAMM datasets.

Dataset	Model	Training Sample	BLEU-N (%) N=1	N=2	N=3	N=4	ROUGE-1 (%) P	R	F
ZuCo	BrainBART-Large	10710	14.71	4.11	1.66	0.85	15.93	13.59	13.84
	BrainT5-Large	10710	15.70	5.42	2.06	0.92	14.56	12.54	12.70
	BrainPEGASUS-Large	10710	9.55	2.58	0.95	0.37	15.16	9.52	11.05
	SOnET	10710	**18.10**	**7.39**	**3.01**	**1.47**	**16.19**	**15.59**	**14.90**
	BrainBART-Large	14407	13.34	1.52	0.14	0.00	9.51	**22.17**	12.85
	BrainT5-Large	14407	17.14	6.33	2.45	1.09	14.58	13.82	13.38
	SOnET	14407	**18.52**	**7.40**	**2.77**	**1.26**	**14.90**	14.45	**13.87**
ROAMM	BrainBART-Large	13253	14.40	1.41	0.00	0.00	9.20	**15.11**	11.03
	BrainT5-Large	13253	15.13	4.60	1.05	0.00	**13.84**	12.30	12.33
	BrainPEGASUS-Large	13253	13.95	3.34	0.81	0.00	13.27	11.61	11.70
	SOnET	13253	**17.18**	**5.65**	**2.15**	**1.10**	13.48	12.78	**12.63**

Table 4: Impact of teacher forcing during inference on ZuCo.

Model	Training Sample	BLEU-N (%) N=1	N=2	N=3	N=4	ROUGE-1 (%) P	R	F
BrainBART-Large	10710	40.12	23.18	12.61	6.80	31.69	28.84	30.10
BrainT5-Large	10710	46.03	28.23	17.35	10.55	28.14	33.84	30.53
BrainPEGASUS-Large	10710	41.18	23.42	13.12	7.22	30.32	34.83	32.31
BELT	10710	42.31	25.26	14.81	8.73	36.06	29.86	32.57
E2T-PTR	10710	42.09	25.13	14.84	8.99	35.86	30.01	32.61
DeWave	10710	41.35	24.15	13.92	8.22	33.71	28.82	30.69
SOnET	10710	47.05	29.32	18.28	11.42	34.46	28.77	31.18
BrainT5-Large	14407	45.43	27.40	16.49	9.82	31.86	26.18	28.54
BrainPEGASUS-Large	14407	41.10	23.04	13.10	7.34	33.30	29.26	31.04
C-SCL	14407	34.87	25.32	21.17	18.98	36.97	34.31	35.51
E2T-PTR	14407	34.92	25.43	21.00	18.59	37.15	33.93	35.39
SOnET	14407	47.11	29.29	18.29	11.41	33.75	28.47	30.75

from EEG signals. For ROUGE-1, SOnET achieves overall balanced precision and recall scores, leading to an F1-score with a 2.4% improvement over BrainT5-Large.

5.2 Impact of Teacher Forcing

By comparing the performance of EEG-to-text decoding in Tables 3 and 4, we can observe that with teacher forcing, models yield a BLEU-1 score of as high as 47.05% while without teacher forcing, the best performance remains at 18.10% on three reading tasks (10710 training samples). Similar performance gaps exist for four reading tasks. These contrasts highlight the significant influence that teacher forcing has on the overall performance. Teacher forcing allows the model to bypass errors by continuously providing it with previous ground-truth tokens at each decoding step to predict the current token, thereby creating an artificially idealized prediction environment. In other words, when this technique is applied, error propagation is entirely suppressed, meaning that early mistakes do not have the opportunity to propagate through the sequence. This explains the disproportionately high BLEU scores with teacher forcing. Conversely, when teacher forcing is removed, the model must handle the full complexity of sequential dependencies, resulting in a much more challenging evaluation of its generative ability. While teacher forcing is assumed in prior EEG-to-text decoding works given the complexity of decoding language from noisy brain signals (Wang & Ji, 2022; Feng et al., 2023; Zhou et al., 2024; Wang et al., 2024b), it does not represent real-world deployment scenarios. To the best of our knowledge, SOnET is the first proposed EEG-to-text translation framework that eliminates the implicit use of teacher forcing and reports the true capabilities of EEG-to-text systems (Wang & Ji, 2022; Wang et al., 2024b; Zhou et al., 2024; Duan et al., 2024; Feng et al., 2023; Jo et al., 2025).

Table 5: Ablation study performed using ZuCo. Results are reported without applying teacher forcing.

signature representations	Perplexity loss	BLEU-N (%)				ROUGE-1 (%)		
		N=1	N=2	N=3	N=4	P	R	F
✓	✓	**18.10**	**7.39**	**3.01**	**1.47**	**16.19**	**15.59**	**14.90**
✓	✗	15.82	5.45	2.02	0.95	15.59	13.20	13.46
✗	✓	14.04	4.90	1.88	0.86	15.49	12.25	12.92
✗	✗	13.59	4.18	1.54	0.66	14.17	11.44	11.89

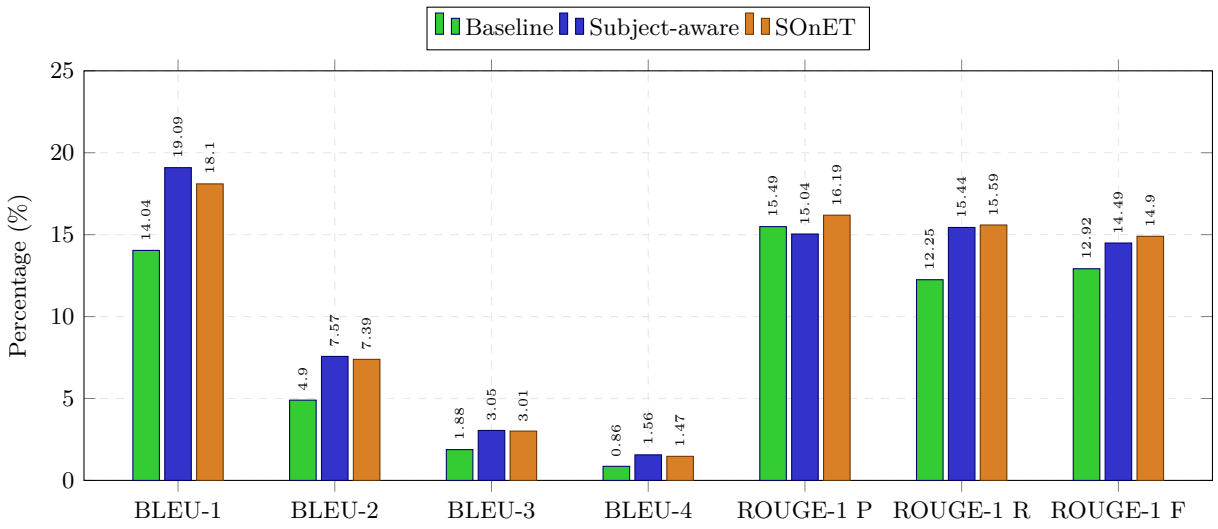

Figure 4: Impact of modeling individual neural signature variability using ZuCo.

5.3 Ablation Study

To understand the contribution of key components in SOnET, we perform an ablation study by removing two important modules: the cross-attention mechanism that fuses the signature representations with the EEG embeddings, and the perplexity loss (replacing it with the standard cross-entropy loss following prior work (Wang & Ji, 2022)). We conduct these experiments without teacher forcing, using training samples from three reading tasks in ZuCo. In Table 5, we observe that when cross-entropy loss is used (instead of our proposed perplexity loss), we observe a significant decline in performance, with only 15.82% BLEU-1 and 13.46% ROUGE-1 F1-score. This improved performance of perplexity loss is due to the dynamic adaptation of the gradient updates based on model confidence. At the beginning of the training, when the model is uncertain at predicting tokens, perplexity loss penalizes the model with larger gradient updates to avoid local minima in the loss curve. On the other hand, when the training starts to stabilize and prediction tends to improve, this loss ensures smaller gradient updates. Without the signature representation module, SOnET achieves only a 14.04% BLEU-1 score and a 12.92% ROUGE-1 F1-score. This demonstrates that the signature representation component of our model contributes significantly to decoding performance, by enabling the model to adapt to neural signature-specific variabilities in EEG signals. Lastly, without both signature representation module and perplexity loss, our model achieves only a 13.59% BLEU-1 score and an 11.89% ROUGE-1 F1-score. This is the lowest performance across all ablations, highlighting the importance of both components in our proposed framework.

5.4 Impact of Modeling Neural Signature Variability

Neural signature variability is a well-known challenge in EEG analysis (Bomatter & Gouk, 2026; Nørskov et al., 2023; Mishra et al., 2025; Murad & Rahimi, 2024; Zhang et al., 2024b), as neural responses to the same linguistic stimulus can differ substantially across individuals. This variability is illustrated in Figure 2, where

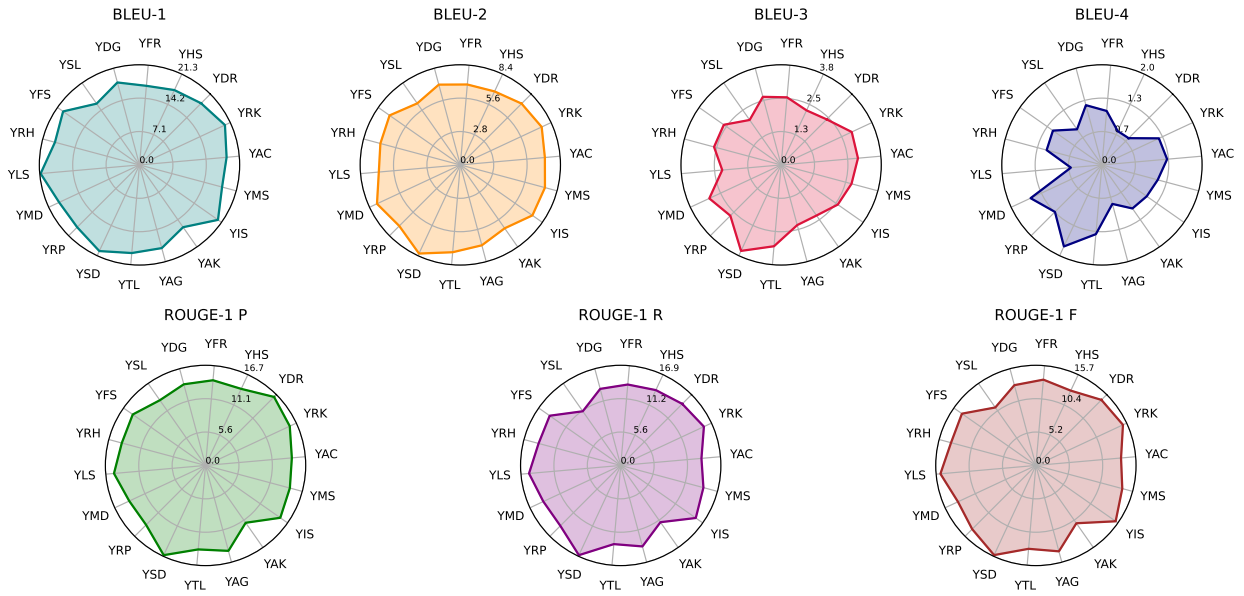

Figure 5: Individual subject-specific evaluation on ZuCo to study performance variation across different subjects. Each model is trained on an individual subject and tested on 17 unseen subjects. Average number is reported per subject.

the brain topographies exhibit distinct spatial activation patterns across subjects despite corresponding to the same word. Such differences suggest that subject-dependent neural characteristics can act as a source of distribution shift, making it difficult for a single EEG-to-text model to generalize across individuals. We conduct an experiment using subject representations from a pre-defined codebook that maps subject IDs to unique representations and fuse them with EEG using the same cross-attention mechanism as SOnET for EEG-to-text decoding. Our hypothesis is that EEG-to-text translation performance should improve over the baseline that does not consider individual neural signature variability. We report the results in Figure 4. We observe that both SOnET and the subject-aware approaches outperform the baseline. Comparing SOnET and the subject-aware approach, both yield similar overall performances. However, the subject-aware approach performs slightly better than SOnET in terms of BLEU scores, while SOnET performs slightly better in terms of ROUGE scores. Note that the subject-aware approach requires subject identity information during inference, whereas our approach does not require any subject-related information during inference, since the neural signatures of subjects are automatically learned through the autoencoder in the SOnET framework. This experiment highlights the importance of addressing the individual neural signature variability for EEG-to-text translation.

5.5 Cross-subject Performance

Prior work in EEG-to-text decoding reports subject-specific performance by training a model using data from a single subject and evaluating it on the remaining subjects in the ZuCo dataset to study performance variation across subjects (Duan et al., 2024; Wang et al., 2024b; Zhou et al., 2024). Following this evaluation protocol, we train 18 separate EEG-to-text decoders, each using data from a single subject while leveraging the signature representations learned from data across all subjects, and evaluate each model on the remaining 17 subjects. Figure 5 presents the results in a radar chart format. The relatively consistent performance across subjects indicates that comparable decoding performance can be achieved across different subject-specific training, suggesting that no single subject provides a disproportionate advantage during training. Notably, BLEU-4 scores remain low across all subjects, highlighting the inherent difficulty of generating longer text sequences solely from EEG signals.

Table 6: EEG-to-text decoding performance of our SOnET model on ZuCo and ROAMM. w/ TF and w/o TF refer to predictions generated with and without teacher forcing, respectively.

Ground truth	After Raymond graduated from high school, he enrolled in the "Universidad del Sagrado Corazon" (University of the Sacred Heart) of San Juan, where he earned a Bachelors Degree in "Radio and Television Production".
BrainT5-Large w/ TF	The Chandler from the school, he wentenrolled in the UniversityBity de la Salvadoroo"tzon" inUniversity of San SouthSacred Heart) and San Francisco, Puerto he was a degree of of in EnglishEnglish Communication Television".".
SOnET w/ TF	He ' from the school, he wentenrolled at the UniversityBity de la Salvadoroo"tzon" inUniversity of the SouthSacred Heart) in San Francisco, Puerto he studied a degree of of in EnglishEnglish Communication Television."".
BrainT5-Large w/o TF	He was a graduate of the University of Illinois at Urbana-Champaign.
SOnET w/o TF	He is a graduate of the University of Maryland, College Park where he received his bachelor's degree in English literature.
Ground truth	Williams teamed with Venus to win the gold medal in doubles at the Sydney Olympics in September.
BrainT5-Large w/ TF	The teamed to win the medal in the at the 2006 Sydney Olympis.
SOnET w/ TF	The teamed to win the medal in thes at the 2006 Sydney Olympics in 2000.
BrainT5-Large w/o TF	In the end, it was not a matter of whether or not there were any significant differences between the two teams.
SOnET w/o TF	This was the first time that a woman had ever won an Olympic gold medal in any sport.
Ground truth	For cooperation to emerge between game theoretic rational players, the total number of rounds N must be random, or at least unknown to the players.
BrainT5-Large w/ TF	The, the be, the developers ande, the game number of games played must be greater, and at least, the players.
SOnET w/ TF	The the, be, the developersorist ande, the game number of games playedept be greater. and the least . the player.
BrainT5-Large w/o TF	The game was won by a team of six players, including the British and American champions.
SOnET w/o TF	The game is played in a semi-staged fashion, with the winner being chosen at random.

5.6 Qualitative Performance Analysis

We present the EEG-to-text decoding performance of SOnET and the baseline (BrainT5-Large) with and without teacher forcing (Table 6). We observe several interesting trends in the model predictions. With teacher forcing, the generated texts more closely resemble the ground truth, and the content is generally more accurate compared to outputs obtained without teacher forcing, as expected. However, despite this higher accuracy, the sentences frequently suffer from misspellings arising from character-level errors. In contrast, without teacher forcing, which better reflects realistic evaluation scenarios, the output text is often able to recover the underlying semantics and identify relevant entities, although it is overall less accurate than its teacher forcing counterpart. For example, in the first example, the generated sentence from SOnET without teacher forcing identifies that the text describes a male student graduating from an educational institution and earning a bachelor's degree (*Ground Truth: 'graduated from high school"* → *SOnET: 'graduate of the University of Maryland, College Park"*, *Ground Truth: "a Bachelors Degree in 'Radio and Television Production"* → *SOnET: "bachelor's degree in English literature"*). BrainT5-Large also captures the graduation-related context but misses the content about the bachelor's degree.

In the second example, both models without teacher forcing do not recover the exact event of Williams and Venus winning an Olympic gold medal in doubles. SOnET preserves a broader semantic association with Olympic gold medal achievements (*Ground Truth: "win the gold medal in doubles at the Sydney Olympics"* → *SOnET: "woman had ever won an Olympic gold medal"*), while BrainT5-Large produces a more generic statement about team differences with weaker alignment to the reference.

In the third example, both models without teacher forcing do not fully reconstruct the original statement regarding cooperation in game-theoretic settings. SOnET retains partial semantic information related to

randomness in repeated games (*Ground Truth: "total number of rounds must be random or ..."* → *SOnET: "winner being chosen at random"*), whereas BrainT5-Large generates a more general description of gameplay that is less aligned with the reference content.

Overall, the predictions exhibit fewer misspellings without teacher forcing. However, this comes at the cost of hallucination, where the model generates fluent but content-divergent outputs with high confidence. We argue that SOnET is able to capture the general domain, but the text decoder operates with increased independence without teacher forcing, which leads to semantically plausible yet incorrect outputs. In contrast, with teacher forcing, the decoder is conditioned on previous ground truth tokens during generation, which limits this independence and encourages closer alignment with the reference sequence, resulting in outputs that are more grounded in the reference but more prone to character-level errors.

6 Limitations and Future Work

There exists several limitations of our work. First, despite the improved performance of SOnET compared to existing models, the performance of EEG-to-text decoding without teacher forcing is still lacking to a great extent, as shown in Table 6, which limits immediate real-world deployment. However, it is important to emphasize that decoding language from complex brain signals is an inherently challenging problem due to the noisy nature of EEG, intra-subject variability (variability for the same subject across different recording sessions), limited dataset sizes, the difficult and expensive data collection procedure, high dimensionality of EEG, attention fluctuations of subjects, poor domain generalizability, and the complexity of language processing in the brain, among others. Therefore, making progress in this challenging field and achieving satisfactory EEG-to-text translation performance requires a holistic approach in AI, neuroscience, and linguistics to address these difficulties. Our research is an incremental step in this direction, and we expect ongoing research to steadily push the boundaries of EEG-to-text decoding.

Second, while in this work we focus on addressing the issue of neural signature variability, we use a T5-large LLM as the text decoder. In future work, we plan to explore more advanced LLMs, better alignment techniques between EEG and language embedding spaces and address the hallucination issue of the decoder.

Third, although our work proposes addressing neural signature variability by modeling the neural signatures of subjects using an autoencoder for EEG-to-text decoding, this method can be applied to a wide range of BCI tasks that involve data from multiple subjects, including emotion recognition, imagined speech decoding, motor imagery classification, sleep staging, and cognitive load estimation, to name a few. We intend to explore diverse BCI tasks adopting our neural signature representation learning technique in the future.

7 Conclusion

We introduce SOnET, an open-vocabulary EEG-to-text decoding framework designed to address individual neural signature variability. We first train a split-latent autoencoder with a contrastive loss for neural signature representation learning and then integrate the signature embeddings from the autoencoder with EEG embeddings in our EEG-to-text decoder through a cross-attention module. Additionally, we adopt a perplexity loss during training the EEG-to-text decoder to scale gradients adaptively based on model confidence. This approach improves overall performance, achieving a 15.3% relative improvement in BLEU-1 and a 17.3% improvement in ROUGE-1 F1-score over the strongest baseline on three reading tasks from the ZuCo dataset. Moreover, on the ROAMM dataset, SOnET shows a relative gain of 13.6% in BLEU-1 and 2.4% in ROUGE-1 F1-score. We conduct a thorough ablation study that shows the importance of each component in our framework. Our findings with SOnET motivate broader applications in BCI across multiple tasks by effectively addressing neural signature variability.

References

Nasim Ahmadzadeh Nobari Azar, Nadire Cavus, Parvaneh Esmaili, Boran Sekeroglu, and Süleyman Aşır. Detecting emotions through eeg signals based on modified convolutional fuzzy neural network. *Scientific Reports*, 14(1):10371, 2024.

Umer Asgher, Yasar Ayaz, and Redha Taiar. Advances in artificial intelligence (ai) in brain computer interface (bci) and industry 4.0 for human machine interaction (hmi), 2023.

Shailesh Vitthalrao Bhalerao and Ram Bilas Pachori. Essdm: An enhanced sparse swarm decomposition method and its application in multi-class motor imagery–based eeg-bci system. *IEEE Transactions on Automation Science and Engineering*, 2025.

Philipp Bomatter and Henry Gouk. Is limited participant diversity impeding eeg-based machine learning? *Advances in Neural Information Processing Systems*, 38:117447–117474, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yassine El Ouahidi, Vincent Gripon, Bastien Pasdeloup, Ghaith Bouallegue, Nicolas Farrugia, and Giulia Lioi. A strong and simple deep learning baseline for bci motor imagery decoding. *IEEE transactions on neural systems and rehabilitation engineering*, 2024.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883, 2023. doi: 10.1109/TNSRE.2023.3314642.

Tianyu He, Mingyi Wei, Ruicong Wang, Renzhi Wang, Shiwei Du, Siqi Cai, Wei Tao, and Haizhou Li. Vocalmind: A stereotactic eeg dataset for vocalized, mimed, and imagined speech in tonal language. *Scientific Data*, 12(1):657, 2025.

Robert Tsi-Lok Ho, William Xue, Michael Ka-Wing See, David Yuen-Chung Chan, Anderson Chun-On Tsang, Calvin Hoi-Kwan Mak, Sui-To Wong, and Michael Wing-Yan Lee. Advances in clinical brain–computer interfaces for assistive substitution and rehabilitation: A rapid scoping review. *Surgical Practice*, 29(1):35–49, 2025.

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

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146, 2020.

Sunhee Hwang, Sungho Park, Dohyung Kim, Jewook Lee, and Hyeran Byun. Mitigating inter-subject brain signal variability for eeg-based driver fatigue state classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 990–994. IEEE, 2021.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=QzTpTRVtrP`.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating eeg-to-text models through noise-based performance analysis. *Scientific Reports*, 2025.

Ashwin Kamble, Pradnya H Ghare, Vinay Kumar, Ashwin Kothari, and Avinash G Keskar. Spectral analysis of eeg signals for automatic imagined speech recognition. *IEEE Transactions on Instrumentation and Measurement*, 72:1–9, 2023.

Kyungdo Kim, Kwangsoo Kim, and Seung-Bo Lee. Prism: Deep metric learning based personal grouping method to reduce intersubject variability for motor imagery brain–computer interface. *Neurocomputing*, 593:127805, 2024.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Dominik Klepl, Fei He, Min Wu, Daniel J Blackburn, and Ptolemaios Sarrigiannis. Adaptive gated graph convolutional network for explainable diagnosis of alzheimer's disease using eeg data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3978–3987, 2023.

Frank R Kschischang. The hilbert transform. *University of Toronto*, 83:277, 2006.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Klaus Linkenkaer-Hansen, Vadim V Nikouline, J Matias Palva, and Risto J Ilmoniemi. Long-range temporal correlations and scaling behavior in human brain oscillations. *Journal of Neuroscience*, 21(4):1370–1377, 2001.

Mingze Liu, Jie Liu, Mengna Xu, Yasheng Liu, Jie Li, Weiwei Nie, and Qi Yuan. Combining meta and ensemble learning to classify eeg for seizure detection. *Scientific Reports*, 15(1):10755, 2025.

Sharmila Majumdar, Amin Al-Habaibeh, Ahmet Omurtag, Bubaker Shakmak, and Maryam Asrar. A novel approach for communicating with patients suffering from completely locked-in-syndrome (clis) via thoughts: brain computer interface system using eeg signals and artificial intelligence. *Neuroscience Informatics*, 3(2):100126, 2023.

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

Aditya Mishra, Ahnaf Mozib Samin, Ali Etemad, and Javad Hashemi. Subject representation learning from eeg using graph convolutional variational autoencoders. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Adam M Morgan, Orrin Devinsky, Werner K Doyle, Patricia Dugan, Daniel Friedman, and Adeen Flinker. Decoding words during sentence production with ecog reveals syntactic role encoding and structure-dependent temporal dynamics. *Communications Psychology*, 3(1):87, 2025.

Saydul Akbar Murad and Nick Rahimi. Unveiling thoughts: A review of advancements in eeg brain signal decoding into text. *IEEE Transactions on Cognitive and Developmental Systems*, 2024.

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

Anders Nørskov, Alexander Neergaard Zahid, and Morten Mørup. Cslp-ae: A contrastive split-latent permutation autoencoder framework for zero-shot electroencephalography signal conversion. *Advances in Neural Information Processing Systems*, 36:13179–13199, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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. High performance communication by people with paralysis using an intracortical brain-computer interface. *elife*, 6:e18554, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Yingyi Qiu, Han Liu, and Mengyuan Zhao. A review of brain–computer interface-based language decoding: From signal interpretation to intelligent communication. *Applied Sciences*, 15(1):392, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Gerwin Schalk, Peter Brunner, Brendan Z Allison, Surjo R Soekadar, Cuntai Guan, Tim Denison, Jörn Rickert, and Kai J Miller. Translation of neurotechnologies. *Nature Reviews Bioengineering*, 2(8):637–652, 2024.

Qingshan She, Chengjun Li, Tongcai Tan, Feng Fang, and Yingchun Zhang. Improved few-shot learning based on triplet metric for motor imagery eeg classification. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Rachel J Smith, Hernando C Ombao, Daniel W Shrey, and Beth A Lopour. Inference on long-range temporal correlations in human eeg data. *IEEE journal of biomedical and health informatics*, 24(4):1070–1079, 2019.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Abdulhamit Subasi, Saeed Mian Qaisar, Akash Kumar Bhoi, and Parvathaneni Naga Srinivasu. *Artificial Intelligence Applications for Brain–Computer Interfaces*. Elsevier, 2025.

Haorui Sun, Ardyn Vivienne Olszko, Niharika Singh, and David C. Jangraw. Roamm: A benchmark dataset for multimodal human attention decoding and eeg-to-text modeling during naturalistic reading, 2026. URL `https://openneuro.org/datasets/ds007629/versions/1.1.0`.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

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

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.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.393. URL `https://aclanthology.org/2024.acl-long.393/`.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegmamba: An eeg foundation model with mamba. *Neural Networks*, pp. 107816, 2025.

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.

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

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13277–13291, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Kang Yin, Elissa Yanting Lim, and Seong-Whan Lee. Gitgan: Generative inter-subject transfer for eeg motor imagery analysis. *Pattern Recognition*, 146:110015, 2024.

Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, et al. The "podcast" ecog dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135, 2025.

Guangyi Zhang and Ali Etemad. Distilling eeg representations via capsules for affective computing. *Pattern Recognition Letters*, 171:99–105, 2023.

Guangyi Zhang, Vandad Davoodnia, and Ali Etemad. Parse: Pairwise alignment of representations in semi-supervised eeg learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(4): 2185–2200, 2022.

Hongyu Zhang, Le Jiao, Songxiang Yang, Haopeng Li, Xinzhan Jiang, Jing Feng, Shuhuai Zou, Qiang Xu, Jianheng Gu, Xuefeng Wang, et al. Brain–computer interfaces: the innovative key to unlocking neurological conditions. *International Journal of Surgery*, 110(9):5745–5762, 2024a.

Liying Zhang, Yueying Zhou, Peiliang Gong, and Daoqiang Zhang. Speech imagery decoding using eeg signals and deep learning: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 2024b.

Shuailei Zhang, Yi Ding, Muyun Jiang, Ning Tang, Effie Chew, Kai Keng Ang, and Cuntai Guan. Cat-net: A co-adaptive transfer learning network for bci-assisted neurorehabilitation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Xiayin Zhang, Ziyue Ma, Huaijin Zheng, Tongkeng Li, Kexin Chen, Xun Wang, Chenting Liu, Linxi Xu, Xiaohang Wu, Duoru Lin, et al. The combination of brain-computer interfaces and artificial intelligence: applications and challenges. *Annals of translational medicine*, 8(11):712, 2020.

Zhenxi Zhao, Yingyu Cao, Hongbin Yu, Huixian Yu, and Junfen Huang. A novel hybrid brain—computer interface for eeg and emg based on an ensemble learning approach for motor imagery classification tasks. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. Belt: Bootstrapped eeg-to-language training by natural language supervision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.