MindScribe: Hierarchical EEG Encoding with Contrastive Alignment for Open-Vocabulary EEG-to-Text Translation

Anonymous ACL submission

Abstract

Brain Computer Interface (BCI) is one of the active areas of research with translation of brain activity to natural language as one of it's main challenge. Existing EEG-to-text methods either rely on word-level EEG features segmented by eye-tracking fixations or face difficulty in converting continuous brain activity into discrete text tokens due to modality gap. In this paper, we introduce **MindScribe**, a light-weight end-to-end multimodal sequence-to-sequence architecture that translates raw, continuous EEG signals recorded during different reading tasks, directly into open-vocabulary text. Our model combines a hierarchical EEG encoder along with dual-objective training framework to align the EEG latent space and the pre-trained text embedding space. Evaluated on the combined dataset of ZuCo 1.0 and 2.0 benchmarks, our model achieves 28.20 BLEU-1 and 33.10 ROUGE-1 F1 on the Task-Specific Reading (TSR), and Normal Reading (NR) tasks, establishing competitive performance against existing baselines.

1 Introduction

The ability to decode natural language directly from brain activity can prove to be highly beneficial for people who have lost the ability to speak due to neurological injury or disease. BCIs (Mudgal et al., 2020) that translate neural signals into text have the ability to restore communication for patients with conditions such as locked-in syndrome, Amyotrophic Lateral Sclerosis (ALS), and severe stroke. While invasive and semi-invasive approaches like electrocorticography (ECoG) (Liu, 2023) and stereoelectroencephalography (sEEG) (Herff et al., 2020) have shown impressive results in speech neuroprostheses, they require surgical implantation of electrode arrays, which limits their scalability and accessibility. Non-invasive methods based on electroencephalography (EEG) offer a safer, portable, and more affordable alternative, making them common for clinical use.

Despite these advantages, decoding continuous natural language from EEG signals remains a challenging problem. EEG signals inherently have a low signal-to-noise ratio, high-dimension, and low spatial resolution compared to invasive recordings (Yun, 2024). Early work in EEG-based language decoding (Robinson et al., 2019) focused on classifying brain states into small, closed vocabularies for tasks such as motor imagery and emotion recognition. Moreover, the emergence of pre-trained language models has enabled open-vocabulary EEG-to-text translation. This was first used by (Wang and Ji, 2022) to decode EEG features into text using pre-trained BART model. Subsequent work by (Duan et al., 2023) introduced discrete codex representations using a quantized variational encoder to remove the dependency on eye-tracking markers.

However, apart from architectural optimization and performance improvement, there are numerous other challenges that still needs to be addressed. First, most existing methods rely heavily on event markers for processing word-level EEG features which limit their application in the real-world. Second, the modality gap between continuous neural time-series and discrete text tokens is huge: the EEG signal is a fixed-channel time series sampled at high frequency, while the target output is a sequence of discrete word tokens. Without explicit mechanisms to bridge this gap, the decoder can learn to ignore the EEG input altogether and instead rely on the learned language model. Third, the limited size of available EEG-to-Text datasets (e.g., the ZuCo 1.0 (Hollenstein et al., 2018) and ZuCo 2.0 (Hollenstein et al., 2020) corpus) contains only hundreds of sentences from a handful of subjects making overfitting a serious concern.

In this work, we introduce **MindScribe**, an architecture designed to address the above mentioned challenges. Our paper is arranged as follows:

Section 2 reviews existing approaches in EEG-to-Text and discusses recent advances in cross-modal alignment techniques. Section 3 presents our proposed model architecture, detailing the hierarchical EEG encoder, parameter-efficient BART decoder, and dual-objective training framework. Section 4 shows the experimental setup, datasets, baselines, and evaluation metrics. Section 5 presents our main results and compares performance against existing methods. Section 6 provides ablation studies to validate design choices followed by conclusion and discussion. Our main contributions are:

- We propose a light-weight hierarchical EEG encoder to extract continuous representations from raw EEG input signal.
- We introduce a dual-objective training to bridge modality gap between EEG and text.
- We adopt LoRA-based parameter-efficient fine-tuning to adapt a pre-trained BART decoder to continuous EEG inputs.

2 Related Work

Research on converting brain signals to text has a long history in the BCI community. Early approaches focused on closed-vocabulary classification tasks, such as P300 spellers and steady-state visually evoked potential (SSVEP) systems, which select characters from a fixed alphabet using event-related potentials (Lotte et al., 2018). While effective for basic communication, these systems are slow and do not scale to natural language.

The shift toward open-vocabulary EEG-to-Text decoding began with (Wang and Ji, 2022), who proposed a framework that treats the brain as a special text encoder.They were the first to feed word-level EEG features into a pre-trained BART model (Lewis et al., 2019) for sequence-to-sequence decoding. Although this work expanded the vocabulary size, its application could not be extended to more significant sentence-level EEG signals.

To deal with the above limitation, (Duan et al., 2023) proposed DeWave, which uses a quantized variational encoder to produce discrete codex representations from raw EEG waves. DeWave was the first method to support translation from raw EEG signals without word-level markers. However, the translation quality on raw waves remained substantially lower than word-level decoding, and the model still relied on teacher forcing during evaluation. It also made no use of spatial information from different brain regions.

Recognizing the limitations of flat EEG encoders that ignored spatial information, (Liu et al., 2024) proposed EEG2TEXT, which introduced a multi-view transformer architecture that separately encodes EEG signals from twelve distinct spatial brain regions before fusing them through a global transformer, coupled with self-supervised EEG pre-training using masked signal reconstruction. This approach showed improved performance over DeWave. However, this method relied solely on reconstruction-based pre-training and did not use existing methods of semantic alignment like contrastive learning to bridge gap between modalities.

Contrastive learning has emerged as a powerful training strategy for aligning representations across different modalities. The CLIP model (Radford et al., 2021) showed that contrastive pre-training on image and text pairs can produce a shared embedding space across modalities. This approach has been extended to other modality pairs, including audio-text, video-text and more recently, brain signal-text alignment.

In the context of EEG-to-Text, DeWave was one of the first models to use contrastive learning in its training strategy. Following this, (Wang et al., 2024) proposed CET-MAE (Contrastive EEG-Text Masked Autoencoder) where contrastive learning is combined with masked signal modeling enforcing cross-modal semantic alignment between EEG and text while also reconstructing masked tokens within each modality individually. This hybrid self-supervised strategy addresses the limitation of DeWave, which used contrastive learning without any reconstruction objective.

However, a key limitation of CET-MAE is that it ignores the possibility of different EEG-Text pairs sharing similar semantic meanings (false negatives) which introduces noise and degrades representation quality. To address this, (Tao et al., 2025) proposed SEE (Semantically Aligned EEG-to-Text Translation), which introduces a Semantic Matching module that uses a frozen pre-trained BART encoder to compute semantic similarity labels between all pairs in a batch, giving less importance to false negative pairs in the contrastive loss leading to improved cross-modal alignment and superior EEG-to-Text decoding performance.

Even after all the advancements in EEG-to-Text, an underlying problem of the use of teacher-forcing as training strategy persisted throughout. This con-

cern was recently addressed by Jo et al. (2025) where they demonstrated that teacher-forced evaluation can inflate performance metrics by some factor compared to autoregressive generation. Acknowledging the validity of this concern, we evaluate our model with and without teacher forcing to demonstrate true performance.

3 Methods

In this section, we describe the MindScribe's architecture in detail. We discuss the task definition followed by data preprocessing and model components. Figure 1 provides an overview of the complete architecture.

3.1 Task Definition

Given an input sequence of continuous EEG signals $X \in \mathbb{R}^{C \times T}$, where C is the number of EEG channels and T is the number of time steps, our goal is to generate the corresponding text sequence $Y = (y_1, y_2, \ldots, y_N)$. We aim to learn a mapping function $f : X \rightarrow Y$ that maximizes the conditional probability $P(Y|X)$.

3.2 Data Preprocessing

We evaluate our model using the ZuCo v1.0 (Hollenstein et al., 2018) and ZuCo v2.0 (Hollenstein et al., 2020) datasets, on the Normal Reading (NR) and Task-Specific Reading (TSR) tasks. Taking DeWave (Duan et al., 2023) as our base, we perform the following preprocessing on the dataset:

- **Signal Filtering and Normalization:** We apply a 4$^{\text{th}}$-order Butterworth bandpass filter (0.1 Hz to 100.0 Hz) at a sampling rate of 500 Hz to remove noise and artifacts. Each EEG sequence is then Min-Max normalized to scale features within a $[0, 1]$ range.
- **Temporal Chunking:** To process variable-length EEG recordings effectively, we fix the maximum sequence length to 5500 time steps (padding as necessary). We then apply an overlapping windowing strategy (window size W=200, stride S=100) to chunk the continuous signal into discrete sequential tokens.
- **Leak-Free Data Splitting:** To avoid data leakage where overlapping sentences appear in both training and test sets, we enforce a strict, unique-sentence-level split of 80% training, 10% validation, and 10% testing. All EEG samples corresponding to a specific text string are kept within the same subset. The overview of splits are discussed in Table 7.

3.3 Hierarchical EEG Encoder

The encoder is designed to extract spatial, temporal, and contextual features from C-channel raw EEG sequences through the following 4 stages.

Spatial Filtering: This purpose of this block is to learn a linear combination of channels that best capture specific brain activity patterns. The raw EEG signals are first projected through a 1D convolutional layer with kernel size 1 acting as a trainable spatial filter. This is followed by batch normalization and GELU activation function. The output H_{spatial} is of shape $(4, T')$ where 4 is the number of output filters and T' is the reduced time dimension.

$$\mathbf{H}_{\text{spatial}} = \text{GELU}(\text{BN}(\text{Conv1D}_{k=1}(\mathbf{X}))) \quad (1)$$

Temporal Downsampling: This block maps high-frequency EEG signals to dense continuous embeddings suitable for the language model. It consists of four convolutional block, where each block consists of a 1D convolutional layer with kernel size 3 and stride 2, followed by batch normalization, GELU activation, and dropout. The output H_{temporal} is of shape $B \times d_{\text{model}} \times T/16$.

$$\mathbf{H}_{\text{bn}} = \text{BN}(\text{Conv1D}^{(4)}_{k=3,s=2}(\mathbf{H}_{\text{spatial}})) \quad (2)$$

$$\mathbf{H}_{\text{temporal}} = \text{Dropout}(\text{GELU}(\mathbf{H}_{\text{bn}})) \quad (3)$$

Bidirectional LSTM Contextualization: To capture long-range contextual dependencies, we apply a bidirectional LSTM (BiLSTM) with a hidden dimension of 384 in each direction, yielding a 768-dimensional output at each time step:

$$\mathbf{H}_{\text{context}} = \text{BiLSTM}(\mathbf{H}_{\text{temporal}}) \quad (4)$$

Multi-Head Self-Attention: Finally, a multi-head self-attention mechanism with 8 attention heads is applied to the contextualized representation enabling the model to attend to the important temporal segments of the EEG signal:

$$\mathbf{H}_{\text{EEG}} = \text{MultiHead}(\mathbf{H}_{\text{context}}, \mathbf{H}_{\text{context}}, \mathbf{H}_{\text{context}}) \quad (5)$$

The output $\mathbf{H}_{\text{EEG}} \in \mathbb{R}^{L \times 768}$ represents the final continuous EEG representation, where L is the reduced sequence length after temporal downsampling, and $d_{\text{model}} = 768$ matches the hidden dimension of BART-base.

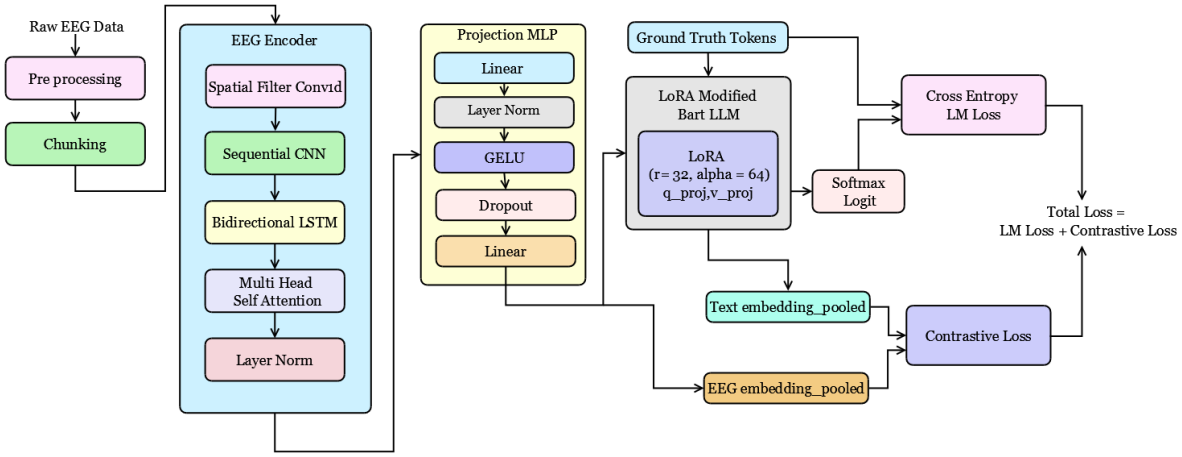

Figure 1: **MindScribe Architecture Overview.** The proposed framework consists of three main components: (a) The hierarchical EEG encoder combining spatial filtering, temporal downsampling via sequential CNN, bidirectional LSTM contextualization, and multi-head self-attention. (b) This is followed by a 2-layer projection MLP and LoRA-adapted BART-base decoder. (c) Dual-objective loss with CLIP-style contrastive alignment loss to align EEG signals and discrete text tokens.

3.4 Cross-Modal Projection Layer

To bridge the modality gap between the continuous EEG representations and the discrete text embedding space of the language model, we pass the encoder output through a 2-layer Multi-Layer Perceptron (MLP) mapping network with a GELU activation and a dropout rate of 0.2.

3.5 Parameter-Efficient Language Decoding

For the decoding component, we use the pre-trained BART-base model (Lewis et al., 2019). To preserve the linguistic priors embedded in BART while adapting it to accept continuous EEG embeddings as input, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022). Through LoRA, we inject low-rank matrices into the query and value projections of both the self-attention and cross-attention layers. For a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times d}$, the adapted weight is:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{BA} \qquad (6)$$

where $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ are the low-rank matrices with rank $r = 32$, and a scaling factor $\alpha = 64$. All other parameters of BART remain frozen during training.

When EEG recordings are different lengths, we pad shorter ones with zeros. But these artificial zeros can confuse the model. We solve this by marking which parts are real data (attention mask) and only using those real parts when computing losses. This way, padding doesn't interfere with learning.

3.6 Dual-Objective Training with Contrastive Alignment

The model is optimized using a dual-loss formulation that combines auto-regressive language modeling loss with contrastive alignment loss.

Language Modeling Loss The primary training objective is the standard cross-entropy loss over the target text tokens, conditioned on the EEG representation H_{EEG} obtained from the encoder:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{N} \log P(y_t | y_{<t}, \mathbf{H}_{\text{EEG}}) \qquad (7)$$

where y_t is the t-th target token and N is the sequence length.

Contrastive Alignment Loss To bridge the modality gap between EEG and text, we introduce a contrastive alignment loss that forces the EEG latent space to the pre-trained semantic space of BART. We first compute a sentence-level EEG representation by mean-pooling the encoder output $\mathbf{H}_{\text{EEG}}$ over the time dimension. This pooled representation is then projected through a non-linear projection head (a two-layer MLP with ReLU activation) into a shared latent space denoted as g_{proj}:

$$\mathbf{z}_{\text{EEG}} = g_{\text{proj}}(\text{MeanPool}(\mathbf{H}_{\text{EEG}})) \qquad (8)$$

Similarly, we obtain text embeddings by passing the ground truth text through BART's text encoder and applying the same mean-pooling operation:

$$\mathbf{z}_{\text{text}} = \text{MeanPool}(\text{BART}_{\text{enc}}(\mathbf{y})) \tag{9}$$

The contrastive loss is a symmetric cross-entropy loss over cosine similarities within each mini-batch, following the formulation used in CLIP:

$$E = \log \frac{e^{\text{sim}(\mathbf{z}^i_{\text{EEG}}, \mathbf{z}^i_{\text{text}})/\tau}}{\sum_{j=1}^{b} e^{\text{sim}(\mathbf{z}^i_{\text{EEG}}, \mathbf{z}^j_{\text{text}})/\tau}} \tag{10}$$

$$T = \log \frac{e^{\text{sim}(\mathbf{z}^i_{\text{text}}, \mathbf{z}^i_{\text{EEG}})/\tau}}{\sum_{j=1}^{b} e^{\text{sim}(\mathbf{z}^i_{\text{text}}, \mathbf{z}^j_{\text{EEG}})/\tau}} \tag{11}$$

$$\mathcal{L}_{\text{CL}} = -\frac{1}{2b} \sum_{i=1}^{b} [E + T] \tag{12}$$

where b is the batch size, $\text{sim}()$ is cosine similarity, and τ is a learnable temperature parameter.

Combined Loss. The total training loss is a weighted sum of the two objectives:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{CL}} \tag{13}$$

where $\lambda = 1.0$ controls the relative weight of the contrastive loss.

3.7 Noise Augmentation and Control Baselines

Given that deep neural models trained on small datasets are prone to memorizing sensor artifacts and exploiting statistical shortcuts, we employ two complementary noise strategies.

Gaussian Noise Augmentation. During training with real EEG data, we add small Gaussian noise ($\sigma = 0.05$) to the input signals. This prevents the model from memorizing specific signal patterns and reduces its sensitivity to sensor artifacts:

$$\mathbf{X}_{\text{aug}} = \mathbf{X} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.05^2) \tag{14}$$

4 Experimental Setup

We compare MindScribe with following baselines:

- **Baseline** (Wang and Ji, 2022): The open-vocabulary EEG-to-Text model, which uses a Transformer encoder to process word-level EEG features and decodes with BART.
- **DeWave** (Duan et al., 2023): Uses a quantized variational encoder to produce discrete codex representations from raw EEG, along with a pre-trained language model via contrastive learning.

4.1 Evaluation Metrics

BLEU Score BLEU (Bilingual Evaluation Understudy) measures how well generated text matches reference translations by counting matching word sequences. Higher BLEU-N values indicate improved alignment with ground truth.

ROUGE Score ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures how much of the target text is recovered by the model's output. Higher scores suggest the model captures more content words and phrases from the reference.

4.2 Implementation Details

Model training uses SGD optimization with a learning rate of 5×10^{-5} and warmup over the first 10% of training iterations, followed by cosine decay. We set batch size to 16 and train for 30 epochs. CNN layers use dropout of 0.2, while other components use 0.1. Gaussian noise ($\sigma = 0.05$) augments the EEG input during training. All experiments run on a single NVIDIA A100 GPU (40GB). Full model training completes in approximately 4 hours. Additional hyperparameters are listed in Table 6 (Appendix).

5 Results

Table 1 compares our model with the baselines using BLEU-N and ROUGE-1 metrics. All models are evaluated with teacher-forced decoding. To assess real-world performance, we additionally report results for our model without teacher forcing, keeping all other settings identical.

Our proposed model outperforms both baselines across all evaluation metrics. It achieves the highest BLEU-N scores at all *n*-gram with notable gains at N=3 and N=4, showing better text generation capabilities. The model also shows consistent improvement on all three ROUGE-1 variants (R, P, F), indicating stronger overlap with reference translations in terms of both recall and precision.

6 Ablation Studies

To show the importance of our choice to incorporate contrastive learning and LoRA, we conduct an ablation study focusing on two critical components: model's parametric distribution and fine-tuning performance with and without LoRA and model's performance on different contrastive learning weights. To show our model's true capabilities, we discuss the ablations on without teacher-forcing setting.

Table 1: Performance comparison of EEG-to-Text models on BLEU-N (in %) and ROUGE-1 (in %) where N=[1, 2, 3, 4]. Evaluation on NR and TSR tasks of ZuCo 1.0 and 2.0. MindScribe shows improved scores in all the metrics.

Methods	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**R-P**	**R-R**	**R-F**
Baseline (Wang and Ji, 2022)	13.07	5.78	2.55	1.10	15.22	18.08	16.36
DeWave (Duan et al., 2023)	21.09	10.69	5.88	3.04	22.01	29.95	24.68
MindScribe - wo/tf	20.70	6.70	2.40	0.90	19.70	15.30	16.40
MindScribe - w/tf	**28.20**	**13.70**	**7.10**	**3.70**	**32.00**	**35.00**	**33.10**

Table 2: Parameter distribution for different LoRA settings. Use of LoRA decreases total trainable parameters.

Architecture	**Encoder Params**	**Proj Params**	**LLM Params**	**Total Trainable**	**% Train.**
Fine-Tuning without LoRA	12.89M	1.18M	139.42M	153.49M	100.00%
Fine-tuning with LoRA	12.89M	1.18M	1.76M	15.84M	10.21%

6.1 Effect of LoRA Adaptation

In this study, we examine whether LoRA can maintain good performance while using fewer parameters. We compared our proposed LoRA-based approach against a full fine-tuning configuration (using pre-trained BART without LoRA). Table 2 shows the parameter distribution across both configurations proving that LoRA results in **89.68%** decrease in total number of trainable parameters. On the other hand, Table 3 presents the performance comparison of the model with and without LoRA setting. Fine-tuning without LoRA results in decrease in model performance across all metric.

6.2 Effect of Contrastive Learning

To measure how effectively our contrastive loss forces the model to decode the actual EEG signal, we ablated the contrastive loss weight (λ) across three settings: $\lambda = 0.0$ (no contrastive alignment), 0.1 (weak alignment), and 1.0 (our proposed configuration). Table 4 shows the result of model performance on different contrastive loss weights.

The ablation reveals three distinct behaviors. At $\lambda = 0.0$, the model achieves a superficially reasonable BLEU-1 score of 20.4 but suffers from posterior collapse: without the contrastive constraint, the decoder ignores the EEG latent space and instead generates generic phrasing based on language model priors. This results in suppressed ROUGE-F1 (15.6), indicating poor semantic understanding of the model.

At $\lambda = 0.1$, the contrastive loss is too weak compared to the generation loss. The model starts to align EEG and text, but the language model still dominates. This causes the model to generate fluent text that sounds correct but doesn't match the actual EEG content. Performance drops significantly (BLEU-1 falls to 0.131), even though ROUGE-Recall remains high (0.190). This mismatch indicates that the model produces wordy outputs that share some words with the reference but miss the true meaning.

At $\lambda = 1.0$, balancing the contrastive loss equally with the generation loss solves the mode collapse problem. This forces the EEG embeddings to align properly with text embeddings before generating words, achieving the best performance (BLEU-1: 0.207, ROUGE-F1: 0.164). This demonstrates that strong alignment between modalities is essential for successful brain-to-text translation. The equal weighting prevents the model from ignoring the EEG signal and forces it to extract actual semantic information from the brain activity.

7 Conclusions and Key Takeaways

We introduced MindScribe, a lightweight architecture for EEG-to-text translation that achieves BLEU-1 of 28.20 (with teacher forcing) and 20.70 (without teacher forcing) on the ZuCo benchmark. Our key findings demonstrate that (1) LoRA-based parameter efficiency reduces trainable parameters by 89.68% without performance degradation, and (2) aggressive contrastive alignment ($\lambda = 1.0$) is essential for preventing the model from ignoring EEG inputs and relying solely on language priors. The dual-objective training framework effectively bridges the modality gap between continuous neural signals and discrete text tokens. Our work pro-

Table 3: Model performance in different LoRA settings. Performance decreases when fine-tuned without LoRA.

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	R-P	R-R	R-F
Fine-Tuning without LoRA	18.60	5.80	2.20	1.0	17.50	14.70	15.10
Fine-Tuning with LoRA	20.70	6.70	2.40	0.90	19.70	15.30	16.40

Table 4: Impact of different contrastive loss weights (λ) on model performance.

λ	BLEU-1	BLEU-2	BLEU-3	BLEU-4	R-P	R-R	R-F
0.0	20.40	6.70	2.50	1.00	18.40	14.70	15.60
0.1	13.10	3.80	1.40	0.50	9.80	19.0	12.20
1.0	20.70	6.70	2.40	0.90	19.70	15.30	16.40

vides practical insights for developing deployable BCI with improved parameter efficiency and explicit cross-modal alignment mechanisms.

8 Limitations

Our model achieves BLEU-1 of 20.7 without teacher forcing, which remains substantially lower than traditional language translation systems, highlighting the inherent difficulty of decoding from noisy EEG signals. The model is evaluated exclusively on the ZuCo dataset with limited participants (12-18 subjects), raising concerns about generalizability across diverse populations. Despite using noise augmentation and dropout, the small dataset size makes overfitting a concern. Finally, the model still requires eye-tracking fixation markers for sentence-level segmentation, limiting applicability to truly marker-free scenarios such as inner speech decoding or real-time assistive communication.

9 Ethical Considerations

EEG-based brain-to-text systems have potential to restore communication for patients with severe motor impairments such as locked-in syndrome and ALS. The non-invasive nature of EEG makes this technology more accessible and deployable compared to invasive alternatives. However, several ethical concerns must be addressed. The ability to decode brain signals raises privacy concerns regarding unwanted mental surveillance, particularly if extended to inner speech decoding. Limited system accuracy may lead to miscommunication and harm if deployed prematurely in clinical settings without adequate fail-safe mechanisms. Additionally, training data biases could result in disparate performance across demographic groups, exacerbating existing healthcare disparities. Clear regulatory frameworks, informed consent protocols, and diverse dataset collection efforts are essential before widespread deployment.

Future work should focus on developing truly marker-free architectures that can decode continuous EEG streams without eye-tracking fixations, enabling real-time brain-to-text communication. Scaling pre-training through self-supervised learning on large unlabeled EEG datasets could improve generalization and reduce overfitting. Multimodal fusion approaches combining EEG with complementary modalities such as fNIRS or MEG could overcome spatial resolution limitations. Additionally, user studies with target patient populations and human evaluation of fluency and semantic accuracy would provide more ecologically valid assessment of system performance in real-world deployment scenarios.

References

Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin teng Lin. 2023. Dewave: Discrete encoding of EEG waves for EEG to text translation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Christian Herff, Dean J. Krusienski, and Pieter Kubben. 2020. The potential of stereotactic-eeg for brain-computer interfaces: Current progress and future directions. *Frontiers in Neuroscience*, Volume 14 - 2020.

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*, 5:180291.

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

Marseille, France. European Language Resources Association.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chang Liu. 2023. Application of ecog and electrode in bci. *Theoretical and Natural Science*, 3:410–416.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. 2024. Eeg2text: Open vocabulary eeg-to-text translation with multi-view transformer. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1824–1833.

F Lotte, L Bougrain, A Cichocki, M Clerc, M Congedo, A Rakotomamonjy, and F Yger. 2018. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005.

Shiv Kumar Mudgal, Suresh K Sharma, Jitender Chaturvedi, and Anil Sharma. 2020. Brain computer interface advancement in neurosciences: Applications and issues. *Interdisciplinary Neurosurgery*, 20:100694.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Neethu Robinson, Seong-Whan Lee, and Cuntai Guan. 2019. Eeg representation in deep convolutional neural networks for classification of motor imagery. pages 1322–1326.

Yitian Tao, Yan Liang, Luoyu Wang, Yongqing Li, Qing Yang, and Han Zhang. 2025. See: Semantically aligned eeg-to-text translation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. 2024. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7278–7292.

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

Seokho Yun. 2024. Advances, challenges, and prospects of electroencephalography-based biomarkers for psychiatric disorders: a narrative review. *Journal of Yeungnam Medical Science*, 41:261–268.

A Architecture Details

Table 5 provides a detailed summary of the hierarchical EEG encoder architecture, including the output dimensions at each stage.

Table 5: Hierarchical EEG encoder layer details.

Stage	Operation	Output
Input	—	$105 \times T$
Spatial	Conv1D (k=1) + BN + GELU	$105 \times T$
Temporal 1	Conv1D (k=3, s=2)	$256 \times T/2$
Temporal 2	Conv1D (k=3, s=2)	$512 \times T/4$
Temporal 3	Conv1D (k=3, s=2)	$768 \times T/8$
Temporal 4	Conv1D (k=3, s=2)	$768 \times T/16$
BiLSTM	Hidden=384×2	$768 \times T/16$
Attention	8 heads	$768 \times T/16$

B Training Hyperparameters

Table 6 lists the key hyperparameters used in our experiments.

Table 6: Training hyperparameters.

Hyperparameter	Value
Learning rate	1×10^{-4}
Optimizer	SGD
Batch size	16
Training epochs	30
LoRA rank (r)	32
LoRA alpha (α)	64
CNN dropout	0.2
Other dropout	0.1
Gaussian noise σ	0.05
Contrastive weight (λ)	1.0
Temperature (τ) init	0.07
Warmup ratio	0.1
LR schedule	Cosine decay

C Dataset split information

We combined the Natural Reading (NR) and Task Specific Reading (TSR) from ZuCo 1.0 and ZuCo

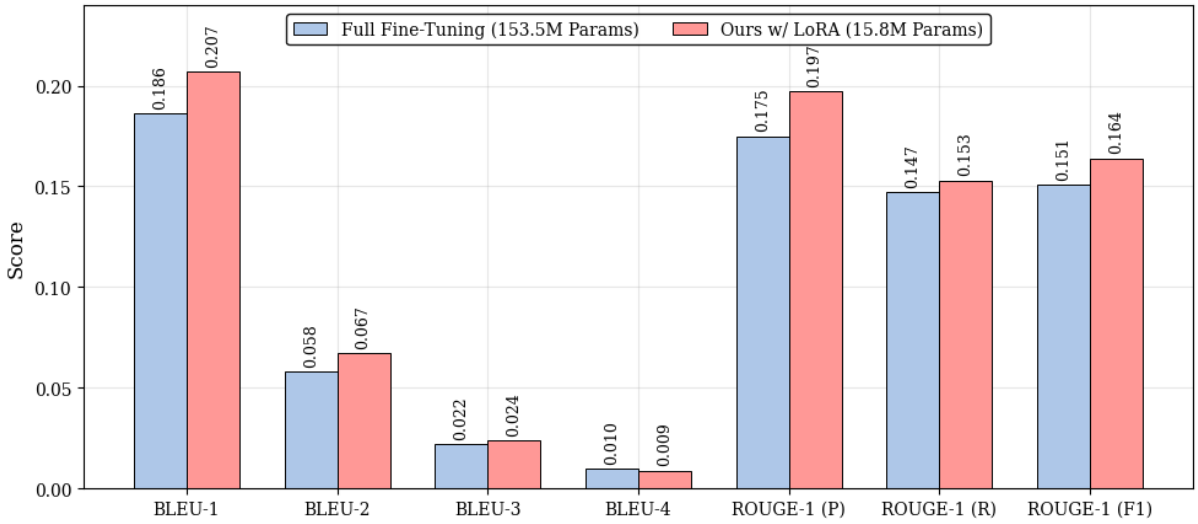

Figure 2: Graph showing Fine-tuning results of our model with and without the use of LoRA adapter

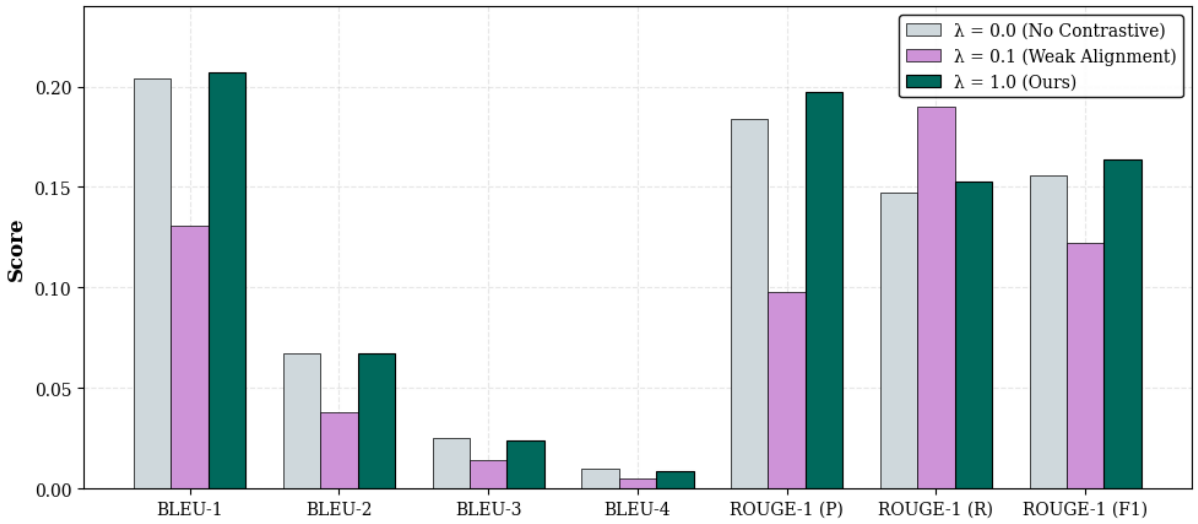

Figure 3: Bar plot showing results on different constrative weight setting.

2.0. We used the 'Matlab' files from the dataset's repository which contained sentence-level processed (bad channels removed) raw EEG signals. It is to be noted that this data split is different from the split showed in the DeWave (Duan et al., 2023) paper which did not provide much details on how they arrived at such split.

Table 7: Dataset Configuration Profile

Split	Total Seq.	Unique Sent.
Training	16,280	917
Validation	1,988	114
Testing	2,168	116