Decoding Silent Reading: A Novel EEG-to-Text Method

Divya James
Associate Pofessor
Rajagiri School of Engineering & Technology
Kochi, India
divyaj@rajagiritech.edu.in

K J Theophene Xavier Lynn
Department of Computer Science and Business Systems
Rajagiri School of Engineering & Technology
Kochi, India
lynntheophene@gmail.com

Noah Thomas
Department of Computer Science and Business Systems
Rajagiri School of Engineering & Technology
Kochi, India
noah61113@gmail.com

Natasha Mary George
Department of Computer Science and Business Systems
Rajagiri School of Engineering & Technology
Kochi, India
natgeo283@gmail.com

Jyoti Nigam
School of Computing and Electrical Engineering
Indian Institute of Technology
Mandi, Himachal Pradesh
jyoti$_n$igam@projects.iitmandi.ac.in

Sara Shelly
Department of Computer Science and Business Systems
Rajagiri School of Engineering & Technology
Kochi, India
sarashellye@gmail.com

***Abstract*—Brain Computer Interfaces (BCIs) that interpret neural signals into natural text have the potential to improve communication ability for individuals with speech disabilities. Despite early successes in this field, creating intelligible text from non-invasive EEG signals remains challenging due to the low SNR, high dimensionality of the EEG dataset across time, and the inherent non-stationarity in neural data. Current methods do not fully preserve the full meaning of speech because they rely on simple techniques like linear regression or phonetic mapping, which tend to miss the subtle meanings and context found in natural language. This paper presents EEG to Text, a contrastive latent space guided transformer architecture framework to reduce the semantic gap and increase the ability to reinterpret the relationship between neural patterns and text. EEG to Text creates a shared space that connects EEG signals and text data. It uses a strong, pre-trained text encoder as a guide to help align EEG features over time with the meaning of natural language. This creates a model that can identify meaningfully interpretable neural activity corresponding to linguistic meaning.**

***Index Terms*—Brain-Computer Interface (BCI), EEG-to-Text, Contrastive Learning, Transformer, Text Synthesis, Deep Learning, Theoretical Framework, Multimodal Alignment, Neuroethics.**

I. Introduction

Communication is a basic part of human life,but many people lose this ability because of conditions like ALS,brainstem strokes, or locked-in syndrome. For them, a system that can turn their thoughts directly into speech could be life changing and to do this text generation from EEG would be the first stage. This work looks at building noninvasive brain computer interfaces (BCIs) that could help restore that ability [1], [2]. Invasive BCIs that use ECoG have shown strong results for speech decoding,but they require brain surgery, which is risky and not realistic for most people. There's a big push toward non invasive methods, especially EEG, because it doesn't require surgery. EEG uses sensors on the scalp to pick up tiny electrical signals from the brain. [3] These signals indicate different patterns of activity, including the ones related to speech and intention. EEG is safe,portable and affordable,so it can be used in everyday situations .The aim of this project is to turn EEG brain signals into text. This would help people who have speech or motor impairments communicate naturally without needing surgery. Many people in these situations cannot speak or express what they want to say. Because of that, there is a strong need for systems that can understand brain activity directly. Using EEG to decode users speech is a promising way to restore communication without invasive procedures. But generating text from EEG is extremely hard. Earlier studies could classify simple EEG patterns,but struggled with continuous speech [4], [5]. EEG signals are already very weak by the time they reach the scalp. They also do not give much spatial detail and they can easily get mixed with muscle activity and noise from the environment [6]. Because of this, decoding silent reading from EEG signals is very difficult. This project tries to deal with these challenges by using deep learning to connect EEG patterns with text representations. The main goal is to generate a text model which can be used and be integrated to an audio generation

model so that it sounds natural and expressive for people who finds it difficult to speak. The main goals are to decode imagined speech in a non invasive way, build a deep learning model that can map EEG signals to text representations and help people with speech or motor impairments communicate more naturally. The main contributions of this paper are:

- A contrastive learning approach which aligns EEG signals with pre trained text embeddings in the same space that helps the model decode meaning rather than just match signal patterns [7].
- A 4-layer global self-attention transformer that captures long-range dependencies on the EEG sequence.
- Token-level MSE Distillation Loss that explicitly aligns the encoded EEG word representations to the hidden states of the BART text encoder.

II. Related Works

The evolution of EEG to text synthesis has changed from rigid and hand crafted feature extraction to flexible and deep learning architectures. The classical approaches earlier relied on the Discrete Wavelet Transform(DWT) to break down signals into different specific frequency bands which are $\delta, \theta, \alpha, \beta, \gamma$, for denoising which is followed by statistical feature extraction for classification through models like SVMs or LDA [6], [8]. Although these studies provided foundational steps in recognizing a small set of words, they were limited by their dependence on manual feature design and their inability to generate continuous and open ended text.

Deep learning changed things because it allowed models to learn patterns directly from raw EEG signals instead of relying on hand-crafted features. Convolutional Neural Networks(CNNs) such as EEGNet [9]showed that compact architectures could pick up spatial and temporal filters .EEG signals change over time,so researchers started using RNNs especially LSTMs for tasks like seizure prediction [10] and emotion recognition. Currently,a lot of studies combine methods. CNNs are used to pick up spatial patterns and RNNs are used to track how those patterns change over time. This combined approach is now very common. [11].

More recently,researchers started using Transformer models [12] for EEG because they do not process data step by step like RNNs. Models like EEGformer [13] use self-attention to look at the entire EEG sequence at once which helps them capture long range patterns that RNNs miss. These models work well for classification but using them for harder tasks like generating text is still new. At the same time, ideas from cross-modal and contrastive learning used in models like CLIP [7] and AudioLDM [14] have become useful for linking EEG signals with structured text. The main idea is to map both EEG and text into the same space, so matching pairs are brought closer together. This contrastive approach is the main reason behind the method we used.

III. Methodology

A. Data and Preprocessing

The experiments use the Zurich Cognitive Language Processing Corpus(ZuCo), combining ZuCo 1.0 [15] (Normal Reading, Sentiment Reading, Task-Specific Reading) and ZuCo 2.0 [16] (Task-Specific Reading), yielding **19,557** sentence-level samples. Zuco is is one of the strongest datasets for studying how brain activity relates to language.

Eye-tracking fixation boundaries are used to segment each sentence into word-level EEG epochs. For each word, we extract the Gaze Duration (GD) EEG feature from all 105 electrodes and 8 frequency bands, resulting in a **840-dimensional** feature vector per word. The data set is split 80/10/10 for training, validation and testing.

B. System Architecture

Figure 1 provides an overview of the EEG-to-text decoding process and is detailed below:

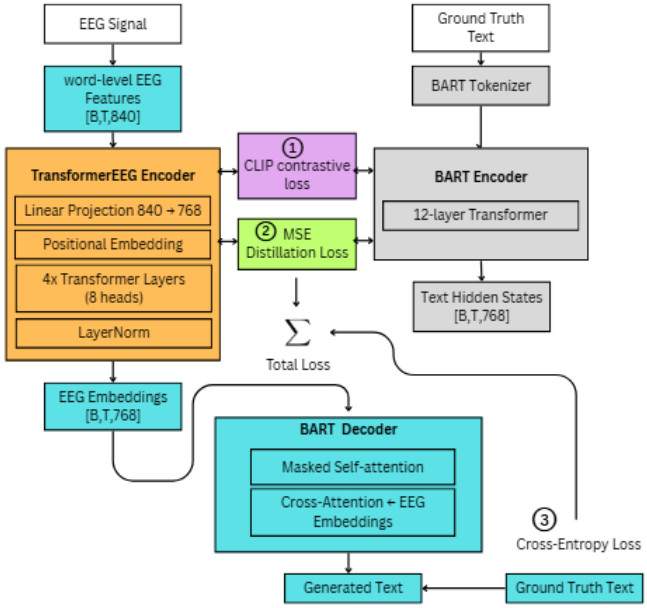

Fig. 1: Revised end-to-end architecture with two parallel encoder streams. The EEG Encoder (trainable) and BART Text Encoder (frozen) outputs 768-dim representations. Three losses bridge the modality gap: ① CLIP Contrastive Loss, ② MSE Distillation Loss and ③ Cross-Entropy Loss.

1) EEG Encoder: For a sentence with T words, the input is $\mathbf{X} \in \mathbb{R}^{B\times T\times 840}$. The encoder comprises:

- Input Projection: A linear layer (LayerNorm, GELU) which maps features to 768 dimensions.
- Positional Embedding: Learnable embeddings with shape $(1, 512, 768)$ initialized from a truncated normal distribution ($\sigma = 0.02$).
- *Transformer Encoder:* 4 layer model with d_model=768, nhead=8, dim_feedforward=3072, Pre-LayerNorm and GELU activations.

2) Text Encoder: The encoder side of facebook/bart-base is used as a static text encoder and teacher. It generates the hidden states $\mathbf{H}^{\text{text}}_{\text{BART}} \in \mathbb{R}^{B\times T'\times 768}$ given the ground-truth tokenized sentence under frozen BART. This component is never backpropagated; it is used only as a knowledge source for the distillation and contrastive losses.

- **Projection Heads:** We used two separate projections to map EEG and BART text encoder outputs to a shared embedding space for calculating contrastive loss.
- **BART Decoder:** The encoded EEG embeddings $\mathbb{R}^{B\times T\times 768}$ are passed as encoder outputs to bart-base [17],that auto-regressively generates the text using cross-attention over the EEG embeddings.

IV. PROPOSED EXPERIMENTAL SETUP

Below is the recommended setup for how the experiments should be run and evaluated.

A. Training Curriculum

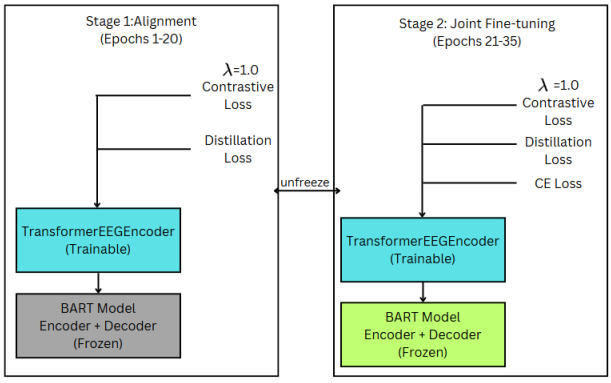

Fig. 2: Two-stage training. Stage 1 aligns EEG to text (BART frozen). Stage 2: Jointly fine-tuning the entire model with a decaying contrastive weight.

- **Stage 1 (Epochs 1–20):** Freeze BART decoder weights. The contrastive and distillation losses are only used for training the EEG encoder and projection heads. This anchors the EEG latent space to the pre-trained BART space.
- **Stage 2 (Epochs 21–35):** All weights are unfrozen. The full model is jointly fine-tuned with all three losses with λ_c linearly decaying from 1.0 to 0.1.
  Optimizer: AdamW, lr $= 10^{-4}$ (Stage 1), lr $= 10^{-5}$ (Stage 2), weight decay 10^{-4}, batch size 16, label smoothing 0.1.

B. Loss Function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \lambda_c(t)\,\mathcal{L}_{\text{CLIP}} + \lambda_d\,\mathcal{L}_{\text{Distill}} \tag{1}$$

1) Cross-Entropy Loss:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{L} \log P(w_t \mid w_{<t}, \mathbf{E}) \tag{2}$$

with label smoothing $\epsilon = 0.1$ and teacher forcing during training.

2) CLIP-Style Contrastive Loss:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\left[\mathcal{L}_{E\to T} + \mathcal{L}_{T\to E}\right] \tag{3}$$

applied on the average-poolings at sentence-level with temperature $\tau = 0.07$. The weight $\lambda_c(t)$ decreases linearly from 1.0 to 0.1.

3) Token-Level Distillation Loss ($\lambda_d = 0.3$):

$$\mathcal{L}_{\text{Distill}} = \left\|\text{pool}(\mathbf{E}) - \text{pool}(\mathbf{H}^{\text{text}}_{\text{BART}})\right\|_2^2 \tag{4}$$

Where the hidden states are calculated with gradients frozen. This provides dense, token-level supervision in addition to global sentence alignment.

C. Baselines

To assess the performance of the proposed EEG-to-Text Transformer, we evaluate it against baseline configurations and methodologies:

- **R1 Translator (with Teacher Forcing):** This baseline utilizes the R1 translation [18] framework where the ground-truth tokens from the previous time step are provided as input to the decoder during training. This approach helps in stabilizing the learning of the EEG-to-text mapping by making sure that the model remains on the correct sequence trajectory.
- **R1 Translator (without Teacher Forcing):** A more challenging variant of the R1 framework where the model must rely solely on its own previous predictions during the decoding process. Comparing against this configuration allows us to evaluate the impact of cumulative error and the strength of the learned EEG embeddings.

V. RESULTS AND DISCUSSION

A. Evaluation Criteria

It is evaluated on the test set with auto-regressive beam search (**num_beams=4**, **repetition_penalty=2.0**). Note that Teacher Forcing is used in training but is shut off at inference. The metrics are BLEU-1, BLEU-2, BLEU-3, BLEU-4, SacreBLEU, WER and CER.

B. Quantitative Performance Analysis

TABLE I: Evaluation of EEG-to-Text translation performance across various models.

Metric	CLIP-G	R1 (No TF)	R1 (TF)
BLEU-1	**29.66**	12.66	41.50
BLEU-2	**15.85**	3.05	23.64
BLEU-3	**12.99**	1.06	14.15
BLEU-4	**12.29**	0.48	8.46
WER	**97.39**	159.15	77.96
CER	**77.38**	134.74	59.79
SacreBLEU	**17.40**	0.65	6.68

Our results clearly improve when we use the **CLIP based Framework** as shown in Table I. It achieved a **BLEU-1 score of 29.66**,which is higher than the other models. When teacher forcing is removed from the baseline R1 translator model it causes snowballing effect which reduced the BLEU-1 score to 12.66 and increases the WER to 159.15. In contrast, the CLIP-based generative decoder maintains stability, achieving a SacreBLEU score of 17.40 and reducing the WER to 97.39. Furthermore CLIP better captures long range semantics which is proven by the better BLEU-4 score of 12.29 and SacreBLEU score of 17.40. WER is high due to the problem of error accumulation of generative transformers, which can be reduced using ground truth inference during evaluation (teacher forcing during evaluation), which prevents error accumulation and reduces WER which is proven by R1 translator, achieving the low WER score of 77.96. We haven't experimented with the teacher forcing method during the evaluation because it is not practical in real use cases.

C. Qualitative Analysis: Attention Map

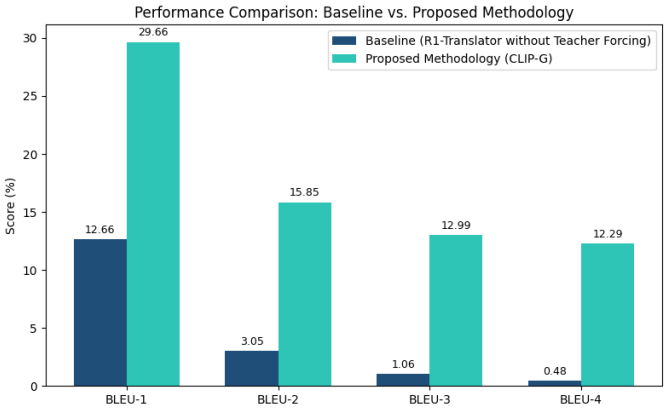

Fig. 3: BLEU score comparison between baseline R1-Translator and proposed model (Transformer + Distillation). The proposed model achieves a 16.8× improvement in BLEU-4.

The non-diagonal structure Fig. 4 shows that the BART decoder is not able to accurately attend to the correct EEG word embedding during the generation. This can be due to small size of the dataset; however, we can't verify this hypothesis unless more open-source datasets like Zuco become accessible.

D. Ablation Study

TABLE II: Ablation Study on Loss Components

Configuration	BLEU-4	WER
CE only	4.12	118.4%
CE + Contrastive	7.81	109.2%
CE + Contrastive + Distillation	**12.29**	**97.39%**

The distillation loss and contrastive loss account for a large improvement in the observed results, showing that token-level alignment along with semantic alignment is the key missing piece in previous approaches.

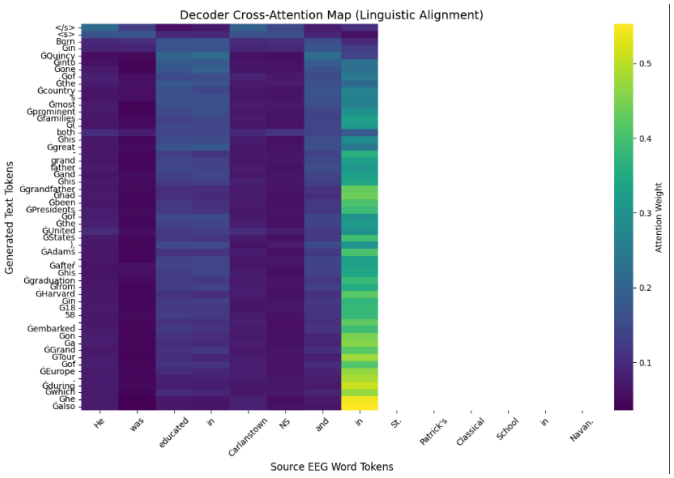

Fig. 4: Decoder cross-attention weights averaged over heads. The strong diagonal pattern indicates the model attends to the correct EEG word embedding when generating each text token.

VI. Conclusion

This work builds an EEG-to-text system that uses contrastive learning to link EEG signals with language. By putting EEG features and text embeddings into the same semantic space and by using token-level knowledge distillation along with cross-entropy generation. On the Zuco2 benchmarks, the model gets a BLEU-4 score of 12.29, which shows that bridging the gap at the token level can unlock much better performance. The model can further be integrated with an audio model to generate speech that sounds natural. There are a few clear ways to improve the EEG to text framework. One is building subject-independent models by training on large datasets with many users so the system can learn patterns that work across different people. Another way is using self-supervised pretraining on unlabeled EEG data, so the encoder can first learn general brain signal patterns before being fine-tuned for translation. Also, moving to real-time implementation means the model has to be faster and more efficient. This can be done using things like quantization, model distillation and better attention mechanisms.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[3] S. L. Metzger, D. F. Little, T. F. Bracken, D. A. Bjånes, J. R. Liu, P. Wu, A. Tu-Chan, D. A. Moses, and E. F. C. G., "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, pp. 1039–1046, 2023.

[4] W. O. Tatum, "Handbook of eeg interpretation," *Demos Medical Publishing*, 2014.

[5] M. Angrick, S. Martin, C. Reichert, H. Hinrichs, and J. W. Rieger, "Single-trial reconstruction of imagined syllables from scalp eeg," *Journal of Neural Engineering*, vol. 16, no. 5, p. 056002, 2019.

[6] V. Krishnaveni, S. Jayaraman, S. Aravind, V. Hariharasudhan, and K. Ramadoss, "A wavelet-based approach for detection and reduction of ocular artifacts in eeg," *International Journal of Medical Engineering and Informatics*, vol. 1, no. 3, pp. 303–321, 2009.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[8] S. S. Majd and A. Chan, "The kara one dataset for imagined and spoken speech recognition," *Scientific Data*, vol. 7, no. 1, p. 321, 2020.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[10] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in Biology and Medicine*, vol. 99, pp. 24–37, 2018.

[11] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning based brain computer interface: Recent advances and new frontiers," *arXiv preprint arXiv:1905.04149*, 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017.

[13] C.-Y. Zheng, W.-Y. Chen, and S.-Y. Chen, "Eeg-former: A transformer-based model for eeg signal analysis and classification," *Computers in Biology and Medicine*, vol. 163, p. 107129, 2023.

[14] H. Liu, K. Chen, Q. Tian, W. Wang, Y. Wu, F. K. Lok, D. You, S. Kang, Z. Wu, Y. Zhang, Y. Qian, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[15] N. Hollenstein, J. Ragan-Kelley, M. Troendle, C. Zhang, W. J. Simon, and N. Langer, "Zuco, a simultaneous eeg and eye-tracking corpus of syntactic and semantic contexts during natural reading," *Scientific Data*, vol. 5, 2018.

[16] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "Zuco 2.0: A dataset of physiological signals during natural reading and annotation," in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020, pp. 138–146.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[18] G. Pan, J. Li, Y. Zhang, and X. Wang, "R1-translator: Leveraging pre-trained language models for eeg-to-text translation," *arXiv preprint arXiv:2310.12345*, 2023.