Enhancing EEG-to-Text Decoding through Transferable Representations from Pre-trained Contrastive EEG-Text Masked Autoencoder

Anonymous ACL submission

Abstract

Reconstructing natural language from non-invasive electroencephalography (EEG) holds great promise as a language decoding technology for brain-computer interfaces (BCIs). However, EEG-based language decoding is still in its nascent stages, facing several technical issues such as: 1) Absence of a hybrid strategy that can effectively integrate cross-modality (between EEG and text) self-learning with intra-modality self-reconstruction of EEG features or textual sequences; 2) Under-utilization of large language models (LLMs) to enhance EEG-based language decoding. To address above issues, we propose the **C**ontrastive **EEG**-**T**ext **M**asked **A**uto**e**ncoder (**CET-MAE**), a novel model that orchestrates compound self-supervised learning across and within EEG and text through a dedicated multi-stream encoder. Furthermore, we develop a framework called E2T-PTR (**E**EG-**to**-**T**ext decoding using **P**retrained **T**ransferable **R**epresentations), which leverages pre-trained modules alongside the EEG stream from CET-MAE and further enables an LLM (specifically BART) to decode text from EEG sequences. Comprehensive experiments conducted on the popular text-evoked EEG database, ZuCo, demonstrate the superiority of E2T-PTR, which outperforms the state-of-the-art in ROUGE-1 F1 and BLEU-4 scores by 8.34% and 32.21%, respectively. These results indicate significant advancements in the field and underscores the proposed framework's potential to enable more powerful and widespread BCI applications. [1]

1 Introduction

Decoding natural language from non-invasive brain recordings with electroencephalography (EEG) is an emerging topic that holds promising benefits for patients suffering from cognitive impairments or language disorders. Thanks to the burgeoning development of pre-trained large language models (LLMs) (Zhao et al., 2023a), the potential of using an open vocabulary to decode human brain activity has been gradually unlocked. Specifically, through the commendable text understanding and generation capabilities of cutting-edge LLMs (Touvron et al., 2023; Ouyang et al., 2022), translating complex spatio-temporal EEG signals into nuanced textual representations, which is known as EEG-to-Text, is achievable. Compared to conventional paradigms of brain-computer interfaces (BCIs), such as motor imagery (MI) (Al-Saegh et al., 2021), steady-state visual evoked potential (SSVEP) (Wang et al., 2017), and P300 (Cecotti and Graser, 2011), EEG-to-Text can convey much more intended commands from the human brain to computers, and thus presents a more extensive range of applications. Its potential as a novel and powerful BCI paradigm marks a significant advancement in the field of BCIs.

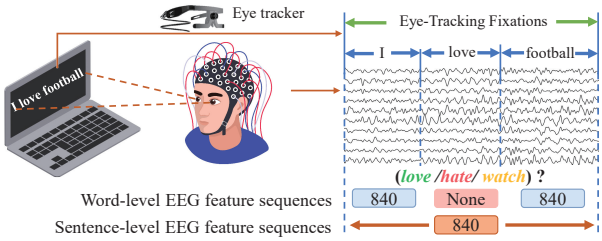

Figure 1: **Text-evoked EEG Recording in ZuCo datasets.** Participants' EEG and eye-tracking data are simultaneously recorded during natural reading to capture text-evoked brain activity.

Several existing EEG-to-Text studies (Li et al., 2022a; Chien et al., 2022) were focused on developing specialized pre-trained models for EEG only, aiming to extract universal semantic representations from the human brain. However, the pre-trained model bridging EEG and text has been ignored, which may be important to enhance the representation learning for inter-modality conver-

[1] Our code is available for reproducibility at the link: https://anonymous.4open.science/r/CET-MAE-E544 .

sion (Bai et al., 2023). This motivates us to develop a hybrid model to orchestrate compound pre-trained representations across and within EEG and text. This endeavor faces the core challenge: *How to bridge the semantic gap between EEG and text while establishing an implicit mapping in the latent representation space?* Responding to this challenge, we focus on self-supervised learning (SSL), because of its great capability in multimodal representation learning (Chen et al., 2024). Contrastive learning is one of the important SSL strategies, learning semantic-level representations across modalities (as CLIP does for language and image) (Radford et al., 2021). Masked modeling methods exhibit significant capability of intra-modality self-reconstruction, such as BERT (Devlin et al., 2019) in nature language processing and masked autoencoder (MAE) (He et al., 2022) in computer vision.

Inspired by the above prevailing SSL strategies, we propose a novel pre-trained model to align EEG and text, Contrastive EEG-Text Masked Autoencoder (CET-MAE), as shown in Figure2(a). CET-MAE integrates contrastive learning and masked signal modeling through a dedicated multi-stream encoder. It effectively learns pre-trained representations of EEG and text by balancing the latent embeddings represented by self-reconstruction and the semantic-level aligned embeddings of text tokens and text-evoked EEG features. In terms of masked signal modeling, CET-MAE implements a high mask ratio (specifically, 75%) on both EEG and text data, presenting a meaningful challenge for the model to handle an increased amount of missing information during the reconstruction phase. This setting not only enhances the model's understanding of individual modality but also facilitates cross-modal interactions and support.

Furthermore, to make the most of LLMs' capability in language understanding and generation as well as to fully use pre-trained representations learned by CET-MAE, we introduce a new EEG-to-Text decoding framework, EEG-to-Text using Pre-trained Transferable Representations (E2T-PTR). E2T-PTR utilizes pre-trained modules alongside the EEG stream from CET-MAE and further adopts the BART (Lewis et al., 2020) to decode language from EEG sequences. By transferring the pre-trained representations from CET-MAE, E2T-PTR significantly enhances EEG-to-Text decoding, surpassing both the baseline and SOTA methods.

Our main contributions are summarised below:

- Introducing CET-MAE, the first pre-trained EEG-text model for EEG-based language decoding. CET-MAE integrates the self-reconstruction of text and EEG features with semantic alignment, forming a multi-stream SSL framework for both intra-modality and cross-modality representation learning.

- Developing a new EEG-to-Text framework via E2T-PTR. The new E2T-PTR framework can leverage CET-MAE's pre-trained EEG representations and the capabilities of LLMs (BART) for text generation.

- Conducting extensive EEG-to-Text experiments on three, four, and five reading tasks in ZuCo. Our experiments are more comprehensive than previous works by using more data and including more methods for comparison. Results show that our framework surpasses previous works, and, thus, sets new SOTA standards.

2 Related Works

2.1 Self-supervised Representations Learning

Multimodal self-supervised representation learning aims to explore the interactions between different modalities to produce semantically generalizable representations for downstream tasks.

In recent years, there have been substantial progresses across various modalities, such as vision-language pre-training (Zhao et al., 2023b; Lin et al., 2023). A range of existing methods rely on contrastive learning, which can effectively draw closer to the global representations of matched pairs in latent spaces with semantic-level self-supervised constraints. But contrastive learning sometimes tends to overlook the self-information of individual modalities, particularly at more granular levels. On the other hand, multimodal masked signal modeling integrates cross-modality self-learning with intra-modality self-reconstruction, focusing on reconstructing one modality from another. This approach may help the model learn the associations between modalities. However, it may lead to an excessive emphasis on fine-grained details, potentially weakening the overall cross-modality correlation and causing issues such as insensitivity to whether the inputs are matched pairs. A series of recent works, such as CMAE (Huang et al., 2023), CAV-MAE (Gong et al., 2023) and SimVTP (Ma et al., 2022), have already successfully integrated

both contrastive learning and masked signal modeling so that their complement advantages can be utilized.

Our work draws inspiration from the above SSL methods but with a novel strategy. In the proposed CET-MAE, the utilization of both text and EEG streams not only achieves an explicit contrastive learning objective to capture global coordination but also avoids erroneous learning processes. Meanwhile, the utilization of the joint stream can facilitate the information interaction between modal-specific embeddings to achieve masked signal modeling effectively. To the best of our knowledge, this is the first EEG-to-Text masked autoencoder that attempts to establish transferable representation learning between EEG and text.

2.2 Open Vocabulary EEG-to-Text Decoding

Previous works (Nieto et al., 2022; Kamble et al., 2023) on EEG-to-Text have been severely confined by a limited number of (several or tens of) words in terms of vocabulary size. These closed-vocabulary efforts primarily focused on recognizing low-level linguistic features, such as individual words or syllables. However, these works can hardly capture more complex, high-level semantic and contextual aspects of language.

The development of LLMs has significantly enhanced the field of EEG-based text decoding. The first work using LLM (Wang and Ji, 2022) integrates an additional EEG encoder to align the pre-trained BART for EEG-to-Text, providing important inspiration for subsequent works. C-SCL (Feng et al., 2023) employs curriculum learning to effectively mitigate the discrepancy between subject-dependent and semantic-dependent EEG representations in EEG-to-Text translation. DeWave (Duan et al., 2023) uses a quantized variational encoder to convert continuous EEG signals into discrete sequences, alleviating the reliance on eye fixations. BELT (Zhou et al., 2023) proposes a novel semi-supervised learning framework that integrates contrastive learning into EEG-to-Text decoding. Despite advancements, prior efforts struggled to bridge the complex semantic gap between EEG and text on an open-vocabulary scale. Our proposed CET-MAE aims to tackle this challenge. Additionally, our E2T-PTR framework transfers CET-MAE's representations and leverages the BART to achieve superior text generation outcomes.

3 Methods

3.1 Preliminary

ZuCo benchmark dataset. For our work, we use the ZuCo1.0 (Hollenstein et al., 2018) and ZuCo2.0 (Hollenstein et al., 2023) datasets, which contain the EEG and eye tracking data during five natural reading tasks. The corpus for sentiment reading (SR) task v1.0 comes from the movie reviews. The corpus for the remaining four tasks is sourced from Wikipedia and comprises two versions each of Natural Reading (NR) and Task-Specific Reading (TSR), specifically NR v1.0, NR v2.0, TSR v1.0, and TSR v2.0. The word-level EEG was recorded and aligned by the eye-tracking fixations, and the sentence-level EEG was recorded during the entire reading procedure. We follow the preprocessing and dataset splits established by baseline work (Wang and Ji, 2022).

Natural masking ratios of EEG feature sequences. Our investigation reveals the word-level contextual EEG presentations in ZuCo datasets are severely corrupted due to missing eye-tracking fixations, leading to mismatches between EEG raw data and text, as shown in Figure1. This misalignment leads to fragmented word-level EEG feature sequences, which fails to capture the cohesive semantics of entire sentences and inevitably complicates the representations learning of EEG and text.

Different from previous works, we concatenate the word-level EEG features and the sentence-level EEG features as our EEG feature sequences $\boldsymbol{E}$ as

$$\mathbf{E} = [E_{word1}, E_{word2}, .., E_{wordN}, E_{sentence}]. \quad (1)$$

Incorporating sentence-level EEG features offers several benefits. First, it provides a holistic view of EEG sequences, enriching the interpretation of overall sentence semantics. Secondly, it acts as a form of data augmentation, which can mitigate the issue of data incompleteness, thereby alleviating semantic discrepancies caused by the misalignment between word-level EEG and text. To provide a clearer overview, we have presented the detailed statistics of the natural masking ratio (NMR) of EEG feature sequences under three categories of reading task combinations in Appendix A.

Definitions in EEG-to-Text Decoding. Given a sequence of EEG features $\boldsymbol{E}$ as the input to the model $\boldsymbol{M}$, the aim is to decode the ground-truth word tokens $\boldsymbol{W}$ from open-vocabulary $\boldsymbol{V}$ via $\boldsymbol{M}$. These corresponding EEG-Text pairs $\langle \boldsymbol{E}, \boldsymbol{W} \rangle$ are collected during natural readings.

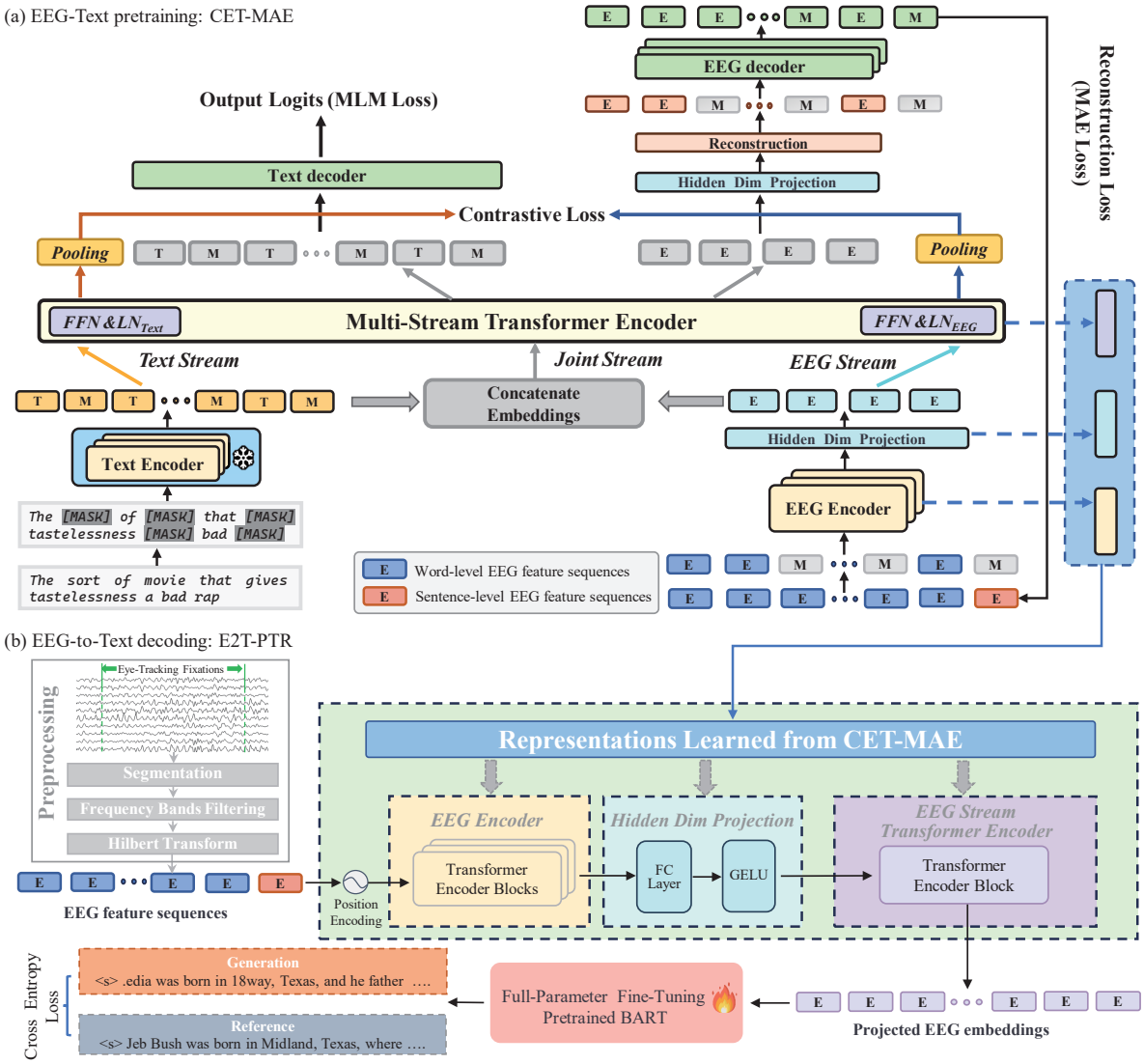

Figure 2: Illustration of the proposed EEG-text pre-training model (CET-MAE) and EEG-to-Text decoding framework (E2T-PTR). (a) **CET-MAE Model:** CET-MAE features modality-specific autoencoders with a masking strategy for text and EEG features, complemented by a multi-stream transformer encoder that orchestrates self-reconstruction and cross-modality semantic alignment, enhancing representation learning for EEG semantic decoding. (b) **E2T-PTR Framework:** E2T-PTR transfers both word- and sentence-level EEG representations extracted from CET-MAE's pre-trained modules, further facilitating text generation through the BART.

During the testing phase, the model $\boldsymbol{M}$ operates with an implicit understanding of the ground-truth word tokens $\boldsymbol{W}$. Its primary objective remains to decode the EEG feature sequences $\boldsymbol{E}$ to generate an output that closely matches tokens $\boldsymbol{W}$. This involves the model generating the sequence of words with the highest probability within the probability distribution $\boldsymbol{P}$ of the $\boldsymbol{V}$.

3.2 EEG-Text Masking

We perform random masking on the text tokens, followed by processing with BERT. For EEG masking, we adopted the following settings. Word-level EEG feature sequences are randomly masked, while sentence-level EEG feature sequences are compulsorily masked. This aims to force the model to fully reconstruct the contextual semantics within the sentence-level EEG feature sequences.

3.3 CET-MAE Encoder

As illustrated in Figure 2(a), the CET-MAE model needs to extract the embeddings of text and EEG separately and then feed the embeddings into the multi-stream transformer encoder to learn the cross-modal representations.

Text encoder. We utilize the pre-trained encoder-decoder model BART as the text encoder. Due to the suitable capabilities in natural language

understanding and generation (Li et al., 2022b), we opt to freeze weights of the BART [2] encoder to maintain its high-level language comprehension from the last hidden states. Firstly, the text tokens are converted into high-quality text embeddings with positional encoding by BART. The learnable embeddings are then used to replace the masked word tokens.

EEG encoder. The EEG encoder is designed as a Multi-layer Transformer Encoder) (Vaswani et al., 2017) to capture the temporal relationships from EEG sequences with spatial and frequency features in each token. A learnable linear projection layer is employed to transform the EEG embeddings from the EEG encoder, aligning their dimensions with those of the text embeddings.

Multi-stream Transformer encoder. The pivotal design of this module lies in the integration of EEG, text, and the joint streams. We implement the dual-modality streams for EEG-text contrastive learning, especially using a specialized head for each modality. It is equipped with the layer normalization (LN) and the feed-forward network (FFN) enabling the production of embeddings that preserve their unique properties(Gong et al., 2023). Notably, we control the learning process to ensure that learnable vectors at masked positions do not enter into the text stream, thereby preventing the inclusion of misleading contrastive feedback. Equally crucial for the two reconstruction tasks, the joint stream is utilized to facilitate the integration of the embeddings from both text and EEG modalities. This design aims to deepen the interaction and enhance the cooperation between EEG and text, fostering a more effective learning synergy.

3.4 CET-MAE Decoder

We apply a lightweight Transformer encoder as the EEG decoder. For EEG reconstruction tasks, EEG embeddings are first mapped to the original dimensions through a learnable linear projection layer. Subsequently, EEG embeddings with learnable masked tokens are inserted back into their original positions. The final EEG embeddings added to the positional embeddings are fed into the EEG decoder. Since the text encoder has already encoded the masked tokens and captured their positional information within the text, we employ a learnable linear projection layer as the text decoder to predict the masked text tokens.

[2] https://huggingface.co/facebook/bart-large

3.5 CET-MAE Training Objectives

CET-MAE is pre-trained by three objectives: (1) Masked Text Modeling (L_T): it aims to predict the masked text tokens by utilizing hybrid representations that integrate information from both textual and EEG embeddings. (2) Masked EEG Modeling (L_E): it learns to reconstruct the original EEG feature sequences, especially predicting masked word- and sentence-level features based on hybrid representations, where the error is measured by mean square error (MSE). (3) EEG-Text Contrastive Learning (L_{CL}): it involves a process where the corresponding EEG and text representations are computed by separate global average pooling layers. The objective is to bring the aligned pairs (matched EEG and text embeddings) closer together while pushing unpaired ones further apart. Our goal L is minimizing is the summation of these three learning objectives:

$$L = \lambda_T \cdot L_T + \lambda_E \cdot L_E + \lambda_{CL} \cdot L_{CL} \quad (2)$$

3.6 E2T-PTR Framework

The proposed E2T-PTR is illustrated in Figure 2(b). It can be summarized into the following key points.

Word-sentence level input tokens. We add the sentence-level EEG features as our input tokens. As detailed in 3.1, concatenating the sentence-level EEG feature sequences as the last token can effectively alleviate the incoherent contextual semantics due to gaps in word-level EEG features.

Effective transfer capability. We investigate how to effectively transfer the cross-modality representations learned from the CET-MAE to downstream tasks such as EEG-to-Text decoding. The E2T-PTR employs a synergy of the following critical components: the EEG encoder, the linear projection layer, and the EEG-stream transformer encoder, all of which are integral components as outlined within the CET-MAE. For the LLM backbone, we also apply the BART which excels at natural language generation tasks.

Fine-tuning strategy. We fine-tune all parameters of E2T-PTR during the training phase. The weights of CET-MAE are first loaded into the EEG encoder, the linear projection layer, and the EEG-stream transformer encoder. As the linguistic backbone of E2T-PTR, the BART is also fully fine-tuned to improve its ability to generate fine-grained text tokens from EEG embeddings.

Method	Training Sample	BLEU-N(%) N=1	N=2	N=3	N=4	ROUGE-1(%) P	R	F
EEG2Text (Wang and Ji, 2022)	10710	40.1	23.1	12.5	6.8	31.7	28.8	30.1
DeWave (Duan et al., 2023)	10710	41.35	24.15	13.92	8.22	33.71	28.82	30.69
E2T-PTR (proposed)	10710	**42.09**	**25.13**	**14.84**	**8.99**	**35.86**	**30.01**	**32.61**
C-SCL (Feng et al., 2023)	14567	35.91(—)	25.91(—)	21.31(—)	18.89(—)	—	—	—
C-SCL*	14407	34.87(44.14)	25.32(31.61)	**21.17(25.67)**	**18.98(22.51)**	36.97	34.31	35.51
E2T-PTR (proposed)	14407	**34.92(44.31)**	**25.43(31.67)**	21.00(25.52)	18.59(22.22)	**37.15**	33.93	35.39
EEG2Text*	18791	58.06	49.98	46.21	44.13	52.31	48.76	50.41
E2T-PTR (proposed)	18791	**59.20**	**50.77**	**46.82**	**44.63**	**53.76**	**50.03**	**51.77**

Table 1: Comparison of our E2T-PTR framework with previous methods on the ZuCo dataset for three and four reading tasks. * means that our reproduced results. Results enclosed in parentheses are calculated following the approach of EEG2Text, which includes retaining consecutive repeated words in the generated text.

(1)	Ground Truth: He **was** first *appointed* to fill the Senate **seat of** *Ernest Lundeen* who had **died in** office.
	EEG2Text: **was** a *elected* to the the position **seat** *in* the *Hemy* in died **died in** 18 in
	E2T-PTR: **was** the *elected* to the the position **seat of** *John Hemy*, resigned resigned **in office**.
(2)	Ground Truth: Jeb **Bush was born in** Midland, **Texas**, where **his father was** running an oil drill **company**.
	DeWave: uan **Bush was** a in 18way, **Texas**, in he **father was** an insurance refinery **company**.
	E2T-PTR: uan **Bush was born in** Newway, **Texas**, and **his father was** a a insurance company **company**.
(3)	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 ...
	E2T-PTR: the's from Yale **school**, **he** went **in the** UniversityAmericancleities de Reyrado Corazon" **(University of the Sacred Heart)** in Spain Francisco, Puerto **he** studied a Bachelor.ors ...

Table 2: EEG-to-Text decoding results. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match. We evaluate the translation performance of the same test sentences reported in EEG2Text, DeWave.

4 Experiments

4.1 Datasets and Evaluation

We pre-trained our CET-MAE models under three, four, and five reading tasks in ZuCo v1.0 and ZuCo v2.0. For fairness, we assessed the performance of E2T-PTR for the EEG-to-Text task under the identical dataset scale used during the pre-training phase. We adopt the BLEU and ROUGE-1 scores for evaluating the EEG-to-Text generation performance. More details are presented in Appendix B.

EEG Mask Ratio (%)	Text Mask Ratio (%)	BLEU-N (%) N=1	N=2	N=3	N=4
25	25	**42.14**	25.02	14.55	8.62
50	25	41.74	24.75	14.39	8.52
50	50	41.80	24.69	14.25	8.40
75	50	41.93	25.02	14.72	8.81
75	75	42.09	**25.13**	**14.84**	**8.99**

Table 3: The performance of our E2T-PTR framework under different combinations of CET-MAE mask ratios rising from 25% to 50% , and to 75% across three reading tasks.

4.2 Implementation Details

The CET-MAE model features a robust EEG encoder with transformer encoder blocks (6 layers, 2048 hidden dimensions, and 8 attention heads). The EEG decoder is a lightweight transformer encoder of 1 layer with 8 heads. The multi-stream transformer encoder is designed with 1 layer, a 4096 hidden dimension, and 16 attention heads. The mask ratios for EEG feature sequences and textual tokens are set at 75% (which can achieve the best results based on trial-and-error). For the CET-MAE pertaining objective L, we set λ_T=0.1, λ_E=1, λ_{CL}=0.01. This setting is refined through experiments to balance the gradients of each loss in the overall training objective, ensuring that the model learns effectively from each task. We pre-train the CET-MAE model from scratch for 100 epochs. Subsequently, we fine-tune the E2T-PTR model for EEG-to-Text tasks over 50 epochs, employing a batch size of 32 and utilizing the AdamW optimizer. More details are provided in Appendix C.

4.3 Main Results

Table 1 shows the performance of our E2T-PTR framework on the ZuCo benchmarks. In three reading tasks, E2T-PTR achieves BLEU-1 to BLEU-4 SOTA scores of 42.09%, 25.13%, 14.84%, and 8.99%, respectively. Moreover, it outperforms best in ROUGE-1 Precision, Recall, and F1 scores compared to recent works. Notably, without removing repetitive generated word tokens, E2T-PTR surpasses C-SCL in BLEU-1 and BLEU-2 scores across four reading tasks. Particularly under the five reading tasks with 18791 training samples, E2T-PTR scores 59.20%, 50.77%, 46.82%, and 44.63% in BLEU-1 to BLEU-4, significantly exceeding the baseline work EEG2Text.

Table2 presents a comparative analysis of the decoding results between our model and other models under three reading tasks. Our model E2T-PTR demonstrates an enhanced ability to generate more complete grammatical structures, which is evident from the reduced error rates and increased semantic coherence in the decoded sentences, exemplified by expressions such as "**his father was**" and "**Bush was born in**". Our model also excels in decoding common and proper nouns, such as "**office**" and "**University of the Sacred Heart**". It also adeptly produces semantically similar words, such as, "*appointed*" vs "*elected*", and "*Ernest Lundeen*" vs "*John Hemy*". Intriguingly, upon expanding our training samples to 1.75 times (10710 to 18791), we observe an obvious improvement in the translation quality of the model, especially concerning fine-grained recognition. As shown in Table 4, our model is capable of generating sentences that not only exhibit complete syntactic structures but also cover comprehensive details, such as "**(March 12, 1922 - October 21, 1969)**" and "**(1891-1927)**". However, it's noteworthy that the model faced challenges in decoding named entities, particularly human names, such as misinterpreting "Robert Henry" as "Emerson" or "Barrymore" as "aldmore", a phenomenon not limited to these instances. More comprehensive results are included in the Appendix D.

Our investigation delved into the transfer performance of CET-MAE across varying EEG and text masking ratios under three reading tasks. Table 3 details the performance shifts under different combinations of masking ratios rising from 25% to 50%, and to 75%. We discovered that the CET-MAE model excels at the higher masking ratios of 75%, starkly contrasting with the traditional 15% mask ratio suggested in BERT. This result is consistent with recent findings in multi-modal masked models (Ma et al., 2022; Geng et al., 2022), suggesting that inter-modal interactions may promote performance improvement. We further ponder this phenomenon and suggest that, in terms of CET-MAE structure, it appears to be suited for reconstructing masked EEG features and predicting masked word tokens. In terms of the masking strategy, forcefully masking sentence-level EEG embeddings can better compel the model to learn global semantic information. Furthermore, we discuss the overall masking ratio for the EEG, the natural EEG masking ratio under three reading tasks is 32.51% as mentioned in Appendix A. Therefore, the total masking ratio for the EEG is 83.13% [3] (32.51% of natural + 50.62% of CET-MAE masked.

4.4 Ablation Studies

Table 5 details the ablation experiments, affirming the effectiveness of each component in our approaches for EEG-to-Text generation quality. First, sentence-level EEG features positively impact BLEU scores, notably BLEU-1, underscoring their importance in capturing essential semantic information for improved text generation. Second, CET-MAE, focusing on masked signal modeling and contrastive learning between EEG and text, is fundamental. Integrating CET-MAE with the baseline framework (Wang and Ji, 2022) significantly boosts BLEU scores, especially BLEU-4. Third, combining E2T-PTR with CET-MAE enhances performance across metrics, particularly Precision, Recall, and F1 score of ROUGE-1, showcasing E2T-PTR's role in effectively transferring CET-MAE's learned representations.

4.5 Transfer Performance of SSL Models

We further pre-train and compare the transfer performance of the following SSL models: 1) Contrastive EEG-Text (CET) learning model: The CET that has no reconstruction objective. For a fair comparison, we implement CET using the same encoder architecture (modal-specific encoders + multi-stream encoder) with CET-MAE but remove the reconstruction task (L_E and L_T). We use this model to investigate the impact of contrastive learning. 2) EEG-text masked autoencoder (ET-MAE) model: The ET-MAE has the same architecture as

[3] Overall Masking Ratio = NMR + (1 - NMR) × CET-MAE Masking Ratio.

(1)	Ground Truth: Robert Henry **Dee (born May 18, 1933 in Quincy, Massachusetts) is a former three-sport letterman at Holy Cross College who was one of the first players signed by the Boston Patriots in 1960.**
	E2T-PTR: Emerson **Dee (born May 18, 1933 in Quincy, Massachusetts) is a former three-sport letterman at Holy Cross College who was one of the first players signed by the Boston Patriots in 1960.**
(2)	Ground Truth: Barrymore **married Katherine Corri Harris (1891-1927), an actress who starred in the 1918 film The House of Mirth, on September 1, 1910 and divorced in 1916.**
	E2T-PTR: aldmore **was Katherine Corri Harris (1891-1927), an actress who starred in the 1918 film The House of Mirth, on September 1, 1910 and divorced in 1916.**

Table 4: EEG-to-Text decoding example results on test sentences under five reading tasks. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match.

Sentence-level EEG feature sequences	CET-MAE	E2T-PTR	Training Sample	BLEU-N (%)				ROUGE-1 (%)		
				N=1	N=2	N=3	N=4	P	R	F
×	×	×	10710	41.16	23.99	13.49	7.68	34.68	28.96	31.45
✓	×	×	10710	41.63	24.48	13.96	8.06	35.13	29.27	31.83
✓	✓	×	10710	41.88	24.85	14.52	8.74	35.26	29.50	32.02
✓	✓	✓	10710	**42.09**	**25.13**	**14.84**	**8.99**	**35.86**	**30.01**	**32.61**

Table 5: The results of ablation experiments on CET-MAE and E2T-PTR structures under three reading tasks. We verified the effectiveness of each component and used BLEU-N (%) and ROUGE-1 (%) as the evaluation metrics.

Metrics (%)	Our SSL Models		
	CET	ET-MAE	CET-MAE
BLEU-1	41.77	41.80	**42.09**
BLEU-2	24.68	24.72	**25.13**
BLEU-3	14.33	14.43	**14.84**
BLEU-4	8.60	8.53	**8.99**
ROUGE-1 P	35.59	35.06	**35.86**
ROUGE-1 R	**30.11**	29.31	30.01
ROUGE-1 F	32.51	31.82	**32.61**

Table 6: Evaluating transfer performance across CET, ET-MAE, and CET-MAE under three reading tasks.

CAV-MAE but the contrastive loss (L_{CL}) is set to 0. The masking strategy is the same as CET-MAE. We use this model to examine the effectiveness of masked signal modeling. 3) Our proposed CET-MAE is detailed in Section 3.

To ensure fairness, CET and ET-MAE are pre-trained with the same pipeline as CET-MAE. We assess their EEG-to-Text transfer performance using the E2T-PTR framework. Results in Table 6 demonstrate CET-MAE's superiority over two other SSL models (CET and ET-MAE) across most evaluation metrics. Specifically, CET-MAE achieves improvements of 0.32%, 0.45%, 0.51%, and 0.39% in BLEU-1 to BLEU-4, respectively, compared to CET. Against ET-MAE, CET-MAE records increases of 0.29%, 0.41%, 0.41%, and 0.46% for these metrics, respectively. The trend of enhancement is consistent in ROUGE-1 metrics as well.

5 Conclusion

This study contributes to the development of EEG-based language decoding by introducing an effective EEG-text pre-trained model, CET-MAE, and a highly capable and LLM-empowered EEG-to-Text decoding framework, E2T-PTR. CET-MAE uses a multi-stream architecture to incorporate both intra- and cross-modality SSL within one unified system: 1) Intra-modality streams explore representative embeddings that reflect the intrinsic characteristics of EEG or text sequences, leveraging masked modeling with a mask ratio of up to 75%; 2) Inter-modality stream provides dual-modal representations to enhance intra-modality reconstruction and constrains the encoder to maximize semantic consistency between text and its corresponding EEG sequences. E2T-PTR transfers pre-trained EEG representations and leverages BART's capabilities for text generation from these consistent and representative features. Extensive experiments on the latest text-evoked EEG dataset, ZuCo, demonstrate the superiority of this work in both qualitative and quantitative assessments. Our work in improving EEG-based language decoding holds great significance, as it has the potential to revolutionize BCI technology and enhance the quality of life for individuals with communication impairments.

6 Limitation

The limitations of our study are summarized as follows:

Dataset Scale: The performance of both the CET-MAE model and the E2T-PTR framework is constrained by the scale of currently available datasets. We are in the process of developing our datasets to fully exploit the potential of our models and frameworks.

Teacher Forcing: While our results are pushing the open vocabulary EEG-to-Text decoding performances to a new SOTA, they still depend on the implicit use of teacher forcing, a common precondition in recent studiess (Wang and Ji, 2022; Duan et al., 2023; Feng et al., 2023; Zhou et al., 2023; Xi et al., 2023). This reliance on teacher forcing could be constraining the full capabilities of the LLMs. Our future work will aim to reduce dependence on teacher forcing by exploring the autoregressive capabilities of LLMs.

Exploration of LLMs: We plan to explore more advanced LLMs to enhance our EEG-to-Text decoding capabilities. This will involve testing new models and techniques to improve performances and uncover deeper insights from EEG data.

7 Ethics Statement

In this work, we do not generate new EEG data, nor do we perform experiments on human subjects. We use the publicly available ZuCo v1.0 and ZuCo v2.0 datasets without any restrictions. We do not anticipate any harmful applications of our work.

References

Ali Al-Saegh, Shefa A. Dawwd, and Jassim M. Abdul-Jabbar. 2021. Deep learning for motor imagery EEG-based classification: A review. *Biomedical Signal Processing and Control*, 63:102172.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. 2023. DreamDiffusion: Generating High-Quality Images from Brain EEG Signals.

H Cecotti and A Graser. 2011. Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):433–445.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. 2024. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. 2022. MAEEG: Masked Auto-encoder for EEG Representation Learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2023. DeWave: Discrete EEG Waves Encoding for Brain Dynamics to Text Translation.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. 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.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. 2022. Multimodal Masked Autoencoders Learn Transferable Representations.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. 2023. Contrastive Audio-Visual Masked Autoencoder. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291.

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. 2023. The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data. *Frontiers in Psychology*, 13:1028824.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. 2023. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ashwin Kamble, Pradnya H. Ghare, Vinay Kumar, Ashwin Kothari, and Avinash G. Keskar. 2023. Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition. *IEEE Transactions on Instrumentation and Measurement*, 72:1–9.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. 2022a. A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning. *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6–14.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022b. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11002–11010.

Yuanze Lin, Chen Wei, Huiyu Wang, Alan Yuille, and Cihang Xie. 2023. Smaug: Sparse masked autoencoder for efficient video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2459–2469.

Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. 2022. SimVTP: Simple Video Text Pre-training with Masked Autoencoders.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. 2017. A Benchmark Dataset for SSVEP-Based Brain–Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752.

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.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. 2023. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13277–13291. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A Survey of Large Language Models.

Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. 2023b. MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1528–1538. ACM.

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2023. BELT:Bootstrapping Electroencephalography-to-Language Decoding and Zero-Shot Sentiment Classification by Natural Language Supervision.

A Natural Masking Ratio of Datasets

To provide a clear perspective, we present the detailed statistics of the NMR of EEG feature sequences for three categories of reading task combinations in Table7.

B Datasets

We utilize the combination of both ZuCo v1.0 and ZuCo v2.0 to form the final ZuCo benchmark. The EEG features are collected with a 128-channel system under the sampling rate of 500Hz. After the noise canceling process, only 105 channels are used. There are 8 frequency bands determined in the ZuCo dataset as follows: 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) and gamma1 (30.5–40 Hz) and gamma2 (40–49.5 Hz). The Hilbert transform is applied in each of these time series. The final features of the EEG are formed by concatenating features from all 8 frequency bands, resulting in a vector with a dimension of 840. For three reading tasks, we pre-train and fine-tune the models on "*SR v1.0 + NR v1.0 + NR v2.0*". For four reading tasks, we choose the combination of "*SR v1.0 + NR v1.0 + NR v2.0 + TSR v1.0*". For five reading tasks, the models are pre-trained and fine-tuned on "*SR v1.0 + NR v1.0 + NR v2.0 + TSR v1.0 + TSR v2.0*". During pre-training, the datasets were split into training and testing sets in a 90% to 10% ratio. During the EEG-to-Text fine-tuning phase, the datasets were further divided into training, validation, and testing sets with an 80%, 10%, and 10% split respectively. The test set samples remained consistent throughout the above two stages. The dataset statistics of EEG-to-Text decoding are detailed in Table 8.

C Implementation Details

Our training hyper-parameters are listed in Table 9. To ensure a fair comparison, we conducted both pre-training and fine-tuning for the EEG-to-Text decoding task using datasets with the same combinations of reading tasks.

Reading Tasks	Missing Pairs	Total words	NMR(%)
SRv1.0+NRv1.0+NRv2.0	90362	277966	32.51
SRv1.0+NRv1.0 +NRv2.0+TSRv1.0	137460	373817	36.77
SRv1.0+NRv1.0+NRv2.0 +TSRv1.0+TSRv2.0	204089	515979	39.55

Table 7: Statistics for natural masking ratios under three, four, and five reading tasks in ZuCo benchmarks.

D Generated Samples

We show more details in EEG-to-Text translation results generated on our models in Table 10, Table 11, and Table12. In our experiments, we aim to select the same sentences from the test sets of three, four, and five reading tasks where feasible. This enables us to directly observe and compare the generated results with the ground truth across different task conditions.

E Subject-independent Performance

As reported in Table1, we present the average BLEU-N and ROUGE-1 scores for all 30 subjects. However, considering the individual variations of brain activities during semantic processing and cognitive operations within different subjects, we further provide individual BLEU-N and ROUGE-1 scores for each subject. We use radar charts shown in Figure3 and Figure4 to visually represent these differences, allowing for an intuitive comparison across subjects. For a detailed numeric breakdown of these variances, refer to Table13 and Table14. We utilize radar charts shown in Figure3 and Figure4 to visually represent these differences, allowing for an intuitive comparison across subjects.

F Impact of the Multi-Stream Design

Our investigation, as detailed in Table 16, reveals the transfer performance of a multi-stream design in the CET-MAE and E2T-PTR frameworks. The multi-stream approach, which provides the specialized handling of text and EEG using separate streams, outperformed a single joint stream design. Notably, in the E2T-PTR framework, leveraging the EEG-specific stream for fine-tuning yielded a marked improvement in EEG-to-Text task performance over a joint modality stream. This modality-focused approach appears to capitalize on the nuanced semantic information inherent in EEG embeddings, resulting in a more sophisticated and contextually relevant latent space. This is substantiated by the observed uptick in BLEU and ROUGE metrics. Our study underscores the criticality of fine-grained, modality-specific processing approaches in the domain of EEG-Text representation learning.

Reading Task	Training Sample	Validation Sample	Testing Sample
SR v1.0 + NR v1.0+NR v2.0	10710	1332	1407
SRv1.0+NRv1.0+NRv2.0+TSRv1.0	14407	1790	1799
SRv1.0+NRv1.0+NRv2.0+TSRv1.0+TSRv2.0	18791	2287	2404

Table 8: Dataset Statistics of the EEG-to-Text decoding. SR: Normal Reading (Sentiment), NR: Normal Reading (Wikipedia), TSR: Task Specific Reading (Wikipedia).

Hyperparameters	**Pre-training**			**Fine-tuning**		
Models	CET-MAE			E2T-PTR		
Reading Tasks	3	4	5	3	4	5
Datasets Splits	9:1			8:1:1		
Epochs	100			50	40	40
Batch Size	32			32		
Learning Rate	5e-7			2e-7	2e-5	2e-5
Optimizer	AdamW, weight decay= 1e-2, betas =(0.9,0.999)					
LR Scheduler	Cosine Annealing, T_max=20					
GPUs	RTX4090					

Table 9: Implementation details in our pre-training and fine-tuning.

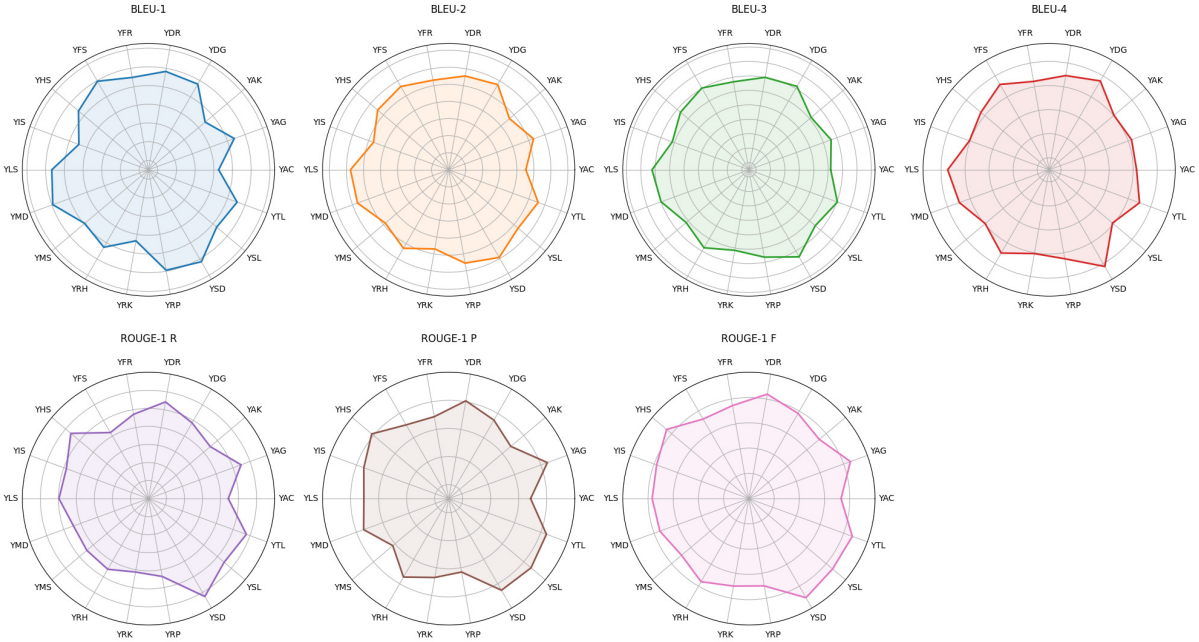

Figure 3: The radar chart of 18 subjects from Subject YAG to YSD on each metric.

G Impact of the Masking Strategy

The masking strategy is crucial in Masked Autoencoders. For the text, the BERT masking strategy has proven highly effective. For the EEG modality, we introduce a pivotal design that involves mandatory masking of sentence-level EEG feature sequences, as detailed in Section 3.2. We delve into the impact of this strategy on the EEG-to-Text decoding task. Comparative results between random and forced masking strategies are presented in Table 15. The forced masking strategy outperforms the random masking strategy in the EEG-to-Text decoding, highlighting the efficacy of our proposed strategy in compelling the model to reconstruct the contextual semantics within sentence-level EEG feature sequences comprehensively.

(1)	Ground Truth: At the urging of **his wife**, Columba, a devout Mexican **Catholic**, the Protestant Bush became a Roman Catholic.
	E2T-PTR: the time of **his wife**, hea, he former Catholic **Catholic**, he actor pastorman a Catholic **Catholic**.
(2)	Ground Truth: While attending a motorcycle race, **he** met a local girl **named** Columba Garnica Gallo, *whom* **he** eventually **married**.
	E2T-PTR: in the local school, **he** was his man boy **named** Marya,ett,o, *who* **he** later **married**.
(3)	Ground Truth: He then enrolled at Phillips Andover, a **private** boarding **school in Massachusetts** already attended **by his** brother George.
	E2T-PTR: was went in the Academy Mary College where **private** school **school in Massachusetts**. known **by his** father,.
(4)	Ground Truth: He took **a job** in real **estate** with Armando Codina, a 32**-year-old** Cuban **immigrant** and self**-made** American **millionaire.**
	E2T-PTR: was **a job** as the **estate** in theando Iice in who company**-year-old** Italian **immigrant**. former**-made** millionaire **millionaire**.
(5)	Ground Truth: After earning **his** *degree*, Bush went **to work** in an entry level **position** in the international division of Texas Commerce Bank, which was run **by** Ben Love.
	E2T-PTR: the **his** *bachelor* in he became **to work** for the office- **position** at the Department banking of the Instruments.. where was later **by** theitott
(6)	Ground Truth: He later **became** *an* educator, teaching music theory at **the University of** the District **of Columbia**; he was **also** director **of the** District **of** *Columbia* Music Center jazz workshop band.
	E2T-PTR: was **became** *a* American and and at and and **the University of** California West **of Columbia**. and also **also** a **of the** school **of** *Columbia's* Department. department..
(7)	Ground Truth: Bush stayed in Houston with another family to finish the school **year**, **and** spent most **summers** and holidays at **the** *family* estate, known as **the Bush Compound.**
	E2T-PTR: was in the until his family, raise his year **year**. **and** then the of in **summers** there **the** *family's*. including as **the Bush Ranchound**.
(8)	Ground Truth: Robert Henry Dee (**born** May 18, 1933 in Quincy, **Massachusetts**) **is a** former three-sport letterman at Holy **Cross College** who was one **of the** first *players* signed **by the** Boston Patriots in 1960.
	E2T-PTR: Frost, (**born** April 5, 18) New, **Massachusetts**) **is a** retired United-timeport star carrier and the **Cross College**. *played* a **of the** founders African to **by the** University Celtics. the.

Table 10: EEG-to-Text decoding example results on test sentences under three reading tasks. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match.

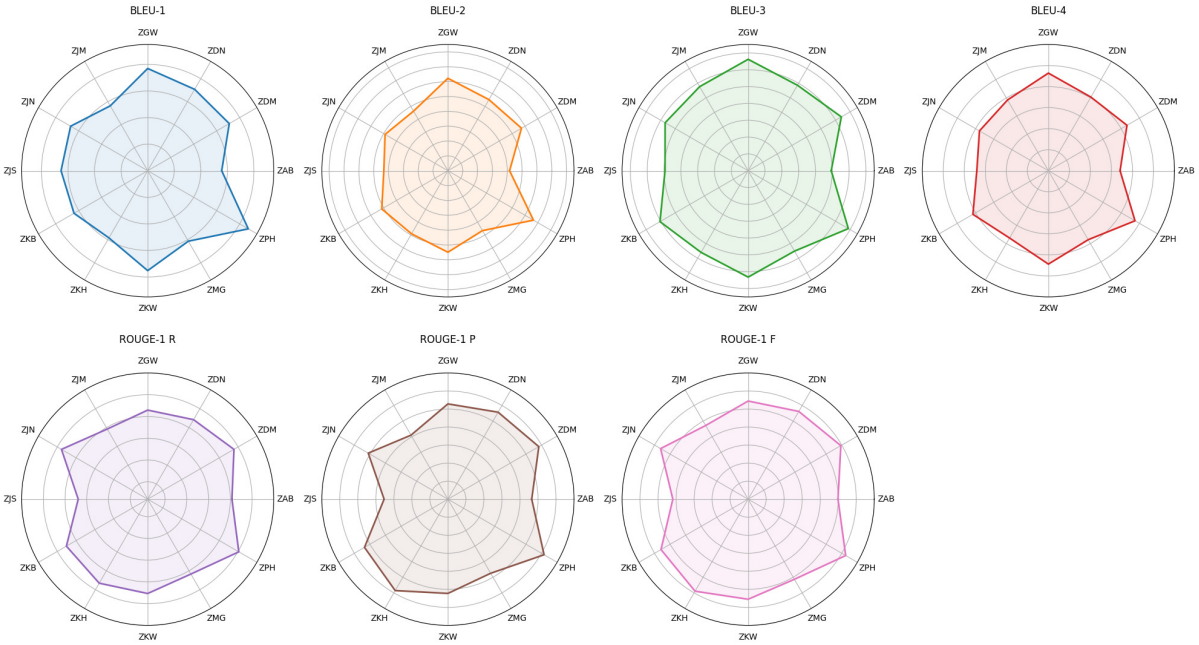

Figure 4: The radar chart of 12 subjects from Subject ZKW-ZJS on each metric.

(1)	Ground Truth: At the urging **of his** wife, Columba, a devout Mexican **Catholic**, the Protestant Bush became a Roman **Catholic**.
	E2T-PTR: the academy **of his** mother, hea, she young **Catholic**-, she young preacher co an Catholic **Catholic** in
(2)	Ground Truth: While attending a motorcycle race, **he met a** local girl **named** Columba Garnica Gallo, *whom* **he** eventually **married**.
	E2T-PTR: serving the Louisiana school he **he met a** man hero **named** Dela Jacksonett.ienne. *who* **he** would struck.
(3)	Ground Truth: He then enrolled at Phillips Andover, a private boarding school in **Massachusetts** already attended **by his** brother George.
	E2T-PTR: was returned in the University Mary College **Massachusetts** public school school **in** the. owned **by his** father,.
(4)	Ground Truth: He took a job in real **estate with** Armando Codina, a 32**-year-old** Cuban immigrant and self-made American **millionaire**.
	E2T-PTR: was many second as the **estate with** theando Feric, where local**-year-old** hotel shipping who hotel-trained millionaire **millionaire** who
(5)	Ground Truth: After earning **his** degree, Bush **went to work** in an entry level position in the international division of Texas Commerce Bank, which was run **by** Ben Love.
	E2T-PTR: a **his** Ph at he **went to work** for the apprentice- role at the Springfield trade of the Instruments. at working he subsequently **by** Jamesoittt
(6)	Ground Truth: He later became an educator, teaching music theory **at the University of** the District **of Columbia**; he was **also** director **of the** District of **Columbia** Music Center jazz workshop band.
	E2T-PTR: was earned president assistant at and English at **at the University of** Wisconsin Arts **of Columbia**, and **also** the a **of the** Special School **Columbia** Library Project. line..
(7)	Ground Truth: Bush stayed **in** Houston with another family to finish the school **year**, **and** spent most **summers and holidays at the** family **estate**, known **as the Bush Compound**.
	E2T-PTR: was **in** Hollywood for his oil, work his term **year**, **and** to the **summers and holidays at the** sprawling **estate**, the **as the Bush Compound**.
(8)	Ground Truth: Robert Henry Dee (born May 18, 1933 in Quincy, **Massachusetts**) **is** a **former** three-sport letterman at Holy *Cross College* **who** was one of the first players signed by **the** Boston Patriots in 1960.
	E2T-PTR: Joseph Bol,born July 22, 1923) Ball, **Massachusetts**) **is** best **former** Republican-timeides quarterbackman **who** the *Cross College*, is elected of the founder " to to **the** University Bruins. 1993.

Table 11: EEG-to-Text decoding example results on test sentences under four reading tasks. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match.

(1)	Ground Truth: At the urging **of his** wife, Columba, a devout Mexican **Catholic**, the Protestant Bush became a Roman **Catholic**.
	E2T-PTR: the academy **of his** mother, hea, she young Catholic **Catholic**, she young and accepted a Catholic **Catholic** in
(2)	Ground Truth: While attending **a motorcycle race, he met a local girl named Columba Garnica Gallo, whom he eventually married.**
	E2T-PTR: serving **a motorcycle race, he met a local girl named Columba Garnica Gallo, whom he eventually married.**
(3)	Ground Truth: He then **enrolled at Phillips Andover, a private boarding school in Massachusetts already attended by his brother George**.
	E2T-PTR: was **enrolled at Phillips Andover, a private boarding school in Massachusetts already attended by his brother George**.
(4)	Ground Truth: He took a **job** in real **estate with** Armando Codina, a 32**-year-old** Cuban immigrant and self-made American **millionaire**.
	E2T-PTR: was his **job** with the **estate with** theco Ferela and and firm**-year-old** firm shipping who hotel-trained **millionaire** merchant.
(5)	Ground Truth: After earning **his degree, Bush went to work in an entry level position in the international division of Texas Commerce Bank, which was run by Ben Love**.
	E2T-PTR: a **his degree, Bush went to work in an entry level position in the international division of Texas Commerce Bank, which was run by Ben Love**.
(6)	Ground Truth: He later **became** an **educator, teaching music theory at the University of the District of Columbia; he was also director of the District of Columbia Music Center jazz workshop band**.
	E2T-PTR: was **became** president **educator, teaching music theory at the University of the District of Columbia; he was also director of the District of Columbia Music Center jazz workshop band**.
(7)	Ground Truth: Bush stayed **in** Houston with another family to finish the school **year**, and spent most **summers and holidays at the** family **estate**, known **as the Bush Compound**.
	E2T-PTR: is **in** Hollywood for his company, work his war **year**. **and** enrolled the **summers and holidays at the** sprawling **estate**, the **as the Bush Compound**.
(8)	Ground Truth: Robert Henry **Dee (born May 18, 1933 in Quincy, Massachusetts) is a former three-sport letterman at Holy Cross College who was one of the first players signed by the Boston Patriots in 1960**.
	E2T-PTR: Emerson **Dee (born May 18, 1933 in Quincy, Massachusetts) is a former three-sport letterman at Holy Cross College who was one of the first players signed by the Boston Patriots in 1960.**

Table 12: EEG-to-Text decoding example results on test sentences under five reading tasks. **Bold** words indicate exact match, *Italic* words indicate semantic resemblance, and Underline words indicate error match.

Subjects	YAG	YAK	YMS	YHS	YSL	YRK	YRH	YDR	YIS	YRP	YLS	YTL	YFR	YDG	YAC	YFS	YMD	YSD
BLEU-1	46.23	46.67	45.65	46.12	46.50	46.34	45.90	46.13	45.90	46.45	46.12	46.56	44.75	46.78	46.28	46.51	46.89	45.65
BLEU-2	28.98	28.93	28.80	28.94	29.57	29.10	28.88	29.28	28.78	29.41	28.94	29.63	27.79	29.60	28.70	29.93	29.82	28.52
BLEU-3	18.07	17.74	17.69	17.85	18.32	17.70	17.82	18.76	17.57	18.45	17.64	18.44	16.90	18.22	17.44	18.87	18.52	17.88
BLEU-4	11.27	10.85	11.09	11.04	11.22	10.70	10.89	12.10	10.64	11.82	10.67	11.44	9.88	11.34	10.50	12.09	11.48	11.18
ROUGE1-R	35.21	35.66	35.73	34.99	36.00	35.23	35.86	35.17	34.77	35.37	35.13	35.58	34.24	34.86	35.30	35.62	35.94	35.03
ROUGE1-P	41.55	42.46	42.88	41.91	43.32	41.69	42.36	41.53	41.29	42.20	42.20	42.04	40.27	41.37	42.30	42.84	42.97	41.90
ROUGE1-F1	38.02	38.65	38.87	38.04	39.22	38.09	38.73	37.97	37.66	38.39	38.24	38.45	36.92	37.72	38.41	38.79	39.04	38.05

Table 13: Subject-independent Performance of BLEU-N(%) and ROUGE-1 from Subject YAG to YSD.

Subjects	ZKW	ZPH	ZAB	ZKB	ZMG	ZJN	ZDN	ZJM	ZGW	ZDM	ZKH	ZJS
BLEU-1	37.99	38.49	38.16	38.02	37.97	38.31	37.84	38.05	38.36	38.15	38.19	37.11
BLEU-2	20.83	21.07	20.83	20.89	21.14	20.74	20.81	20.73	21.58	20.92	21.00	20.34
BLEU-3	10.82	11.19	11.14	10.91	11.40	11.16	11.19	10.72	11.75	10.90	11.13	10.48
BLEU-4	5.76	6.01	6.18	5.70	6.34	6.18	6.27	5.55	6.60	5.82	6.29	5.49
ROUGE1-R	25.34	25.21	24.51	25.38	25.44	25.53	25.46	25.27	26.15	25.08	25.78	24.15
ROUGE1-P	30.44	30.43	29.39	30.74	30.55	30.48	30.31	30.27	31.14	30.10	31.02	28.84
ROUGE1-F1	27.55	27.45	26.62	27.67	27.64	27.65	27.53	27.43	28.30	27.24	28.04	26.17

Table 14: Subject-independent performance of BLEU-N(%) and ROUGE-1 from Subject ZKW to ZJS.

Method	Training Sample	Mask Stragety	BLEU-N(%) N=1	N=2	N=3	N=4	ROUGE-1 (%) P	R	F
E2T-PTR	10710	Random Mask	40.27	23.99	13.95	8.17	35.31	29.63	32.11
	10710	Force Mask	42.09	25.13	14.84	8.99	35.86	30.01	35.61

Table 15: Investigating the impact of mask strategy in EEG feature sequences during CET-MAE pre-training.

Model CET-MAE	Model E2T-PTR	Training Sample	BLEU-N(%) N=1	N=2	N=3	N=4	ROUGE-1(%) P	R	F
×	Joint Stream	10710	41.60	24.53	14.19	8.35	35.34	29.57	32.09
✓	Joint Stream	10710	41.61	24.57	14.34	8.52	35.74	29.79	32.37
✓	EEG Stream	10710	42.09	25.13	14.84	8.99	35.86	30.01	32.61

Table 16: We validated the performance impact of multi-stream design on pre-training and downstream tasks. The ✓ indicates the use of a multi-stream design during pre-training, while the × indicates no use.