Aligning Semantic in Brain and Language: A Curriculum Contrastive Method for Electroencephalography-to-Text Generation

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

***Abstract*— Electroencephalography-to-Text generation (EEG-to-Text), which aims to directly generate natural text from EEG signals has drawn increasing attention in recent years due to the enormous potential for Brain-computer interfaces. However, the remarkable discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation poses a great challenge to this task. To mitigate this, we devise a Curriculum Semantic-aware Contrastive Learning strategy (C-SCL), which effectively recalibrates the subject-dependent EEG representation to the semantic-dependent EEG representation, thereby reducing the discrepancy. Specifically, our C-SCL pulls semantically similar EEG representations together while pushing apart dissimilar ones. Besides, in order to introduce more meaningful contrastive pairs, we carefully employ curriculum learning to not only craft meaningful contrastive pairs but also make the learning progressively. We conduct extensive experiments on the ZuCo benchmark and our method combined with diverse models and architectures shows stable improvements across three types of metrics while achieving the new state-of-the-art. Further investigation proves not only its superiority in both the single-subject and low-resource settings but also its robust generalizability in the zero-shot setting. Our codes are available at: https://github.com/xcfcode/contrastive_eeg2text.**

***Index Terms*— Brain–computer interface, computational neurolinguistics, contrastive learning, curriculum learning.**

I. Introduction

DEVASTATING neurological conditions such as spinal cord injuries or neuromuscular disorders can suddenly lead to people losing their ability to communicate [9], [33]. Such patients may still have intact language and cognitive skills, but injuries might hinder them from expressing themselves [11]. Fortunately, Brain-computer interfaces (BCIs) can restore language abilities to such patients by decoding neural activities into the natural language (Brain-to-Text), which can drastically improve their quality of life [4]. To pursue this goal, various Brain-to-Text works are proposed, building upon either *invasive brain recordings*, such as electrocorticography (ECoG) [2], [24], [25], or *non-invasive brain recordings*, such as functional magnetic resonance imaging (fMRI) [37] and electroencephalography (EEG) [34]. Amongst, EEG shows its superior benefits in portability and cost-effectiveness in real-world applications, thus EEG-to-Text generation gains a lot of research interest recently [10], [34]. Fig. 1 depicts the EEG-to-Text generation task flow.

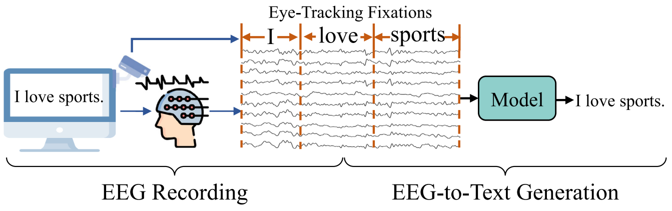


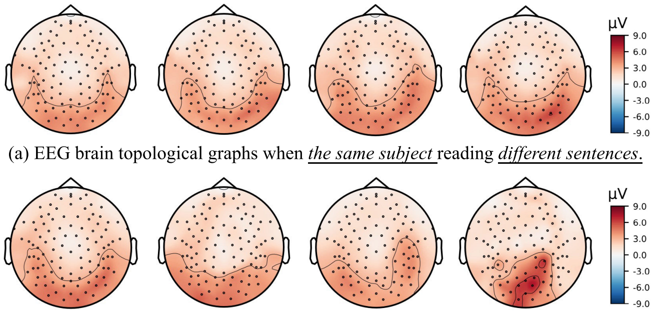
Fig. 1. Illustration of the EEG-to-Text generation. The left part shows the EEG recording process, in which one subject reads a sentence on the screen while recording their EEG signals. Concurrently, the eye-tracking device permits defining exact word boundaries via fixations. Given recorded EEG signals, the task aims to generate the sentence that stimulated those EEG signals.

However, we claim that existing studies neglect the discrepancy between the *subject-dependent EEG representation* and the *semantic-dependent text representation*, which inevitably degrades EEG-to-Text model performance. To explain why it becomes a crucial challenge for this task, we present brain topological graphs to intuitively visualize the discrepancy under two situations. Firstly, as shown in Fig. 2(a), EEG representations elicited by the same subject skewed towards being similar, no matter what the sentence stimulus is, demonstrating the same subject is prone to favour similar cognitive patterns in the face of different sentence stimuli. Secondly, on the contrary, Fig. 2(b) reveals that different subjects act variably even disparately in terms of the same sentence stimulus. These observations are in line with findings in previous studies, including neuroscience [1] as well as some machine learning research areas, such as emotion classification [32] and visual recognition [21]. On this account, such subject-dependent EEG representation negatively impacts the performance of the EEG-to-Text model from two perspectives. On the one hand, it introduces a "many-to-one"

Manuscript received 1 July 2023; revised 22 August 2023; accepted 7 September 2023. Date of publication 12 September 2023; date of current version 9 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0106502, in part by the National Natural Science Foundation of China (NSFC) under Grant 62276078, in part by the Key Research and Development Program of Heilongjiang under Grant 2022ZX01A32,and in part by the International Cooperation Project of PCL under Grant PCL2022D01. *(Corresponding author: Xiaocheng Feng.)*

The authors are with the Research Center for Social Computing and Information Retrieval, Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China (e-mail: xiachongfeng@ir.hit.edu.cn; xcfeng@ir.hit.edu.cn; qinb@ir.hit.edu.cn; tliu@ir.hit.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3314642



(a) EEG brain topological graphs when *the same subject* reading *different sentences*.

(b) EEG brain topological graphs when *different subjects* reading *the same sentence*.

Fig. 2. Brain topological graph of the sentence-level EEG representation (averaged word-level EEG representations). (a) Four topological graphs denote EEG representations elicited by the same subject in response to four different sentences. (b) Four topological graphs describe EEG representations elicited by four different subjects corresponding to the same sentence.

generation problem (multiple EEG signals correspond to the same sentence), which is challenging for training current sequence-to-sequence generation models. On the other hand, it largely hinders good cross-subject generalizability since transferring original subject-dependent EEG representation to unseen subjects is intractable.

To address this issue, we propose a novel **C**urriculum **S**emantic-aware **C**ontrastive **L**earning strategy (C-SCL), which can effectively recalibrate the original subject-dependent EEG representation into our desirable semantic-dependent EEG representation so that it can be better adapted to the EEG-to-Text generation task. In detail, the core part of our C-SCL is the **S**emantic-aware **C**ontrastive **L**earning strategy (SCL), which aims to maximize the similarities of EEG representations across subjects *w.r.t.* the identical sentence stimulus (positive pairs) while minimizing the similarities of EEG representations *w.r.t.* the different sentence stimuli (negative pairs). Note that the critical ingredient for successful contrastive learning is to construct hard positive and negative pairs. However, based on the random selection, we witness that nearly 45.93% of total constructed contrastive pairs already satisfy the final objective, in which positive pairs are similar and negative pairs are dissimilar. Therefore, we manufacture contrastive pairs in different difficulties by pre-computing similarities between numerical EEG signals (e.g., hard positive pairs initially have low similarity while hard negative pairs have high similarity) and drawing support from curriculum learning to not only introduce hard contrastive pairs but also enable a progressive learning process by learning from easy pairs to hard pairs. With the integration of curriculum learning, we finalize our **C**urriculum **S**emantic-aware **C**ontrastive **L**earning strategy (C-SCL).

We conduct experiments on the ZuCo benchmark [18], [19] and assess the generation performance via three types of metrics. The experimental results achieving state-of-the-art performance demonstrate the efficacy of our proposed method across various models and architectures and indicate the necessity of curriculum learning. Further investigation empirically shows its benefits in both the single-subject setting and low-resource settings as well as its robust generalizability in the zero-shot setting. In summary: (a) We take the first step to mitigate the challenge of the discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation for the EEG-to-Text generation task; (b) We devise a curriculum semantic-aware contrastive learning strategy that succeeds in yielding the semantic-dependent EEG representation; (c) We conduct extensive experiments on the ZuCo benchmark that demonstrate the effectiveness of our method and its robustness and superior generalizability.

II. Preliminaries

In this section, we first describe the task formulation and then introduce the ZuCo benchmark.

A. Task Formulation

Given a sequence of word-level EEG features $\boldsymbol{E}$, EEG-to-Text generation task aims at producing a sentence $\mathcal{S}$ via a model θ, where $\boldsymbol{E}$ consists of $|\boldsymbol{E}|$ features $[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{|\boldsymbol{E}|}]$ and $\mathcal{S}$ consists of $|\mathcal{S}|$ tokens $[s_1, s_2, \ldots, s_{|\mathcal{S}|}]$. $\boldsymbol{e} \in \mathbb{R}^n$ symbolizes a word-level EEG feature vector and θ denotes the parameters of a sequence-to-sequence model. Each sequence of EEG features $\boldsymbol{E}$ is associated with a subject $p_i \in \mathbb{P}$, $\mathbb{P}$ being a set of subjects. During the training phase, EEG-Text pairs come from various subjects and the learning objective. At the test phase, sentences are totally unseen. Besides, the train, valid and test sets maintain the same set of subjects $\mathbb{P}$.

B. ZuCo Benchmark

We use the ZuCo dataset, which is a corpus of EEG signals and eye-tracking data during natural reading. The reading materials are collected from movie reviews and Wikipedia articles. Specifically, following Wang and Ji [34], we utilize the combination of both ZuCo [18] and ZuCo 2.0 [19] to form our final ZuCo benchmark. For each EEG-text pair in the dataset, EEG signals are composed of a sequence of word-level EEG features $\boldsymbol{E}$. For each word-level feature $\boldsymbol{e}$, 8 frequency bands are recorded and denoted as the following: theta1 (4-6Hz), theta2 (6.5-8Hz), alpha1 (8.5-10Hz), alpha2 (10.5-13Hz), beta1 (13.5-18Hz) beta2 (18.5-30Hz) and gamma1 (30.5-40Hz) and gamma2 (40-49.5Hz). Each band of the feature has a fixed dimension of 105[1]. We concatenate all 8 bands of features to construct the final word-level feature vector with a dimension of 840 ($\boldsymbol{e} \in \mathbb{R}^{840}$). Additionally, all features are Z-scored as done by Willett et al. [35]. We further split the dataset into the train, valid and test (80%,10%,10%) parts following Wang and Ji [34]. Note that each part of the dataset maintains the same subject set with no overlapping sentences. Table I shows the statistics of the ZuCo benchmark[2].

III. Method

In this section, we thoroughly introduce our curriculum semantic-aware contrastive learning strategy (C-SCL) step by step, including (1) semantic-aware contrastive learning, (2) curriculum learning, (3) the backbone model Brain-Translator and (4) the overall learning procedure.

[1] More details on the data preprocessing steps can be found in the source publication [18], [19].

[2] We omit EEG signals that contain *NaN* values following Wang and Ji [34]. Therefore, different subjects may associate with different sentence sets, as shown in Table I.

TABLE I

STATISTICS FOR THE ZUCO BENCHMARK. "# PAIRS" MEANS THE NUMBER OF EEG-TEXT PAIRS, "# UNIQUE_SENT" REPRESENTS THE NUMBER OF UNIQUE SENTENCES, "# SUBJECT" DENOTES THE NUMBER OF SUBJECTS AND "AVG.WORDS" MEANS THE AVERAGE NUMBER OF WORDS OF SENTENCES

	Train	Valid	Test
# pairs	14567	1811	1821
# unique_sent	1061	173	146
# subject	30	30	30
avg.words	19.89	18.80	19.23

A. Semantic-Aware Contrastive Learning

1) Motivation: The critical ingredient of training a superior model for EEG-to-Text generation is reducing the discrepancy between the *subject-dependent EEG representation* and the *semantic-dependent text representation*. To this end, we draw support from contrastive learning [16], which is skilled at recalibrating the representation space, and propose our semantic-aware contrastive learning strategy (SCL) by pulling semantically similar EEG representations together (positive pairs) and pushing apart dissimilar ones (negative pairs). Note that through employing the semantic embedded within EEG signals as a supervisory signal to direct the optimization of EEG representations, we implicitly achieve a joint model of EEG signals and textual semantics, thereby deriving *semantic-dependent EEG representation*.

2) Positive Pairs: One important question in contrastive learning is how to construct positive pairs $(\boldsymbol{E}_i, \boldsymbol{E}_i^+)$. Towards achieving our goal of learning semantic-dependent EEG representations, given an anchor EEG representation $\boldsymbol{E}_i$ with its corresponding sentence $\mathcal{S}_i$, we randomly choose one EEG $\boldsymbol{E}_i^+$ from the positive set $\mathbb{E}_i^+$, in which all EEG signals correspond to the same sentence stimulus $\mathcal{S}_i$ across different subjects, as shown in Fig. 3(a). Such positive pairs will promote clustering of semantically similar EEG signals.

3) Negative Pairs: Practically speaking, original in-batch negative samples insufficiently provide weak supervision for contrastive learning. To alleviate this problem, Gao et al. [13] verify that introducing specially designed negative pairs can further promote the learning process. Inspired by this conclusion, given the anchor EEG representation $\boldsymbol{E}_i$ elicited by p_i with its corresponding sentence $\mathcal{S}_i$, we construct the negative pair $(\boldsymbol{E}_i, \boldsymbol{E}_i^-)$, where $\boldsymbol{E}_i^-$ satisfies two conditions[3]: (1) $\boldsymbol{E}_i^-$ corresponds to sentences except for $\mathcal{S}_i$ and (2) $\boldsymbol{E}_i^-$ is elicited by subjects except for p_i. All $\boldsymbol{E}_i^-$ that satisfy both two conditions form the negative set $\mathbb{E}_i^-$, as shown in Fig. 3(b).

B. Curriculum Learning

1) Motivation: Recall that our final learning objective is to make the EEG representations corresponding to the same sentence similar while making the EEG representations corresponding to semantically different sentences also dissimilar.

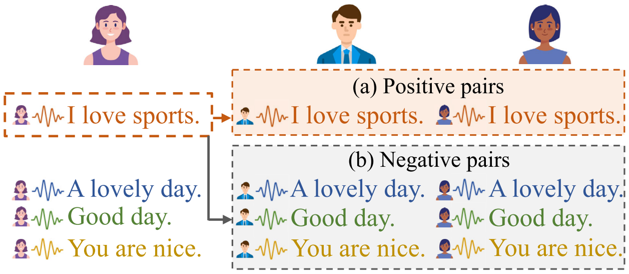


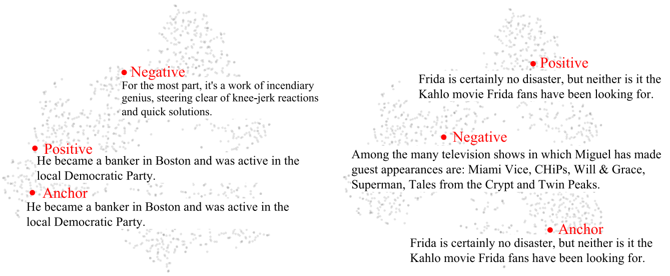
Fig. 3. Illustration of our semantic-aware contrastive learning strategy. (a) Positive pairs derive from EEG signals corresponding to the same sentence elicited by different subjects. In contrast, (b) Negative pairs come from EEG signals elicited by different subjects corresponding to different sentences.

To examine the learning efficiency, we conduct one preliminary experiment by running SCL for 10 epochs on the ZuCo train set, resulting in 145670 (14567 × 10) contrastive triples. However, we find that 66906 triples already satisfy the final objective. In other words, 45.93% ($\frac{66906}{145670} = 45.93\%$) of the positive and negative pairs satisfy the condition that EEG representations with respect to the same sentence are already similar and semantically different EEG representations are already dissimilar without needing contrastive learning, which severely reduces the effectiveness of the learning process. To overcome this problem, we employ curriculum learning to not only introduce hard contrastive pairs but also ensure the model learning efficiency, thus finalizing our **C**urriculum **S**emantic-aware **C**ontrastive **L**earning strategy (C-SCL). Compared with SCL that randomly selects a positive sample and negative sample from $\mathbb{E}_i^+$ and $\mathbb{E}_i^-$ respectively, C-SCL selects samples in an easy-to-hard order.

2) Curriculum Criterion: How to determine the ordering? Recall that our goal is to introduce hard contrastive pairs, where positive pairs are initially far away from each other while negative pairs are oppositely similar. Therefore, we pre-calculate the cosine similarity between two EEG representations and craft contrastive pairs of varying difficulties by taking the similarity into consideration. Specifically, given an anchor EEG representation $\boldsymbol{E}_i$, for positive pair construction, we calculate similarities between the $\boldsymbol{E}_i$ and all $\boldsymbol{E}_i^+ \in \mathbb{E}_i^+$ and then sort the $\mathbb{E}_i^+$ in the descending order, resulting in $\grave{\mathbb{E}}_i^+$. On the contrary, for the negative set $\mathbb{E}_i^-$, we sort it in the ascending order and attain $\acute{\mathbb{E}}_i^-$. Both hard positive and negative samples *w.r.t.* the anchor $\boldsymbol{E}_i$ are located at the end of the $\grave{\mathbb{E}}_i^+$ and $\acute{\mathbb{E}}_i^-$, respectively. In other words, samples in the $\grave{\mathbb{E}}_i^+$ and $\acute{\mathbb{E}}_i^-$ are now in an easy-to-hard order.

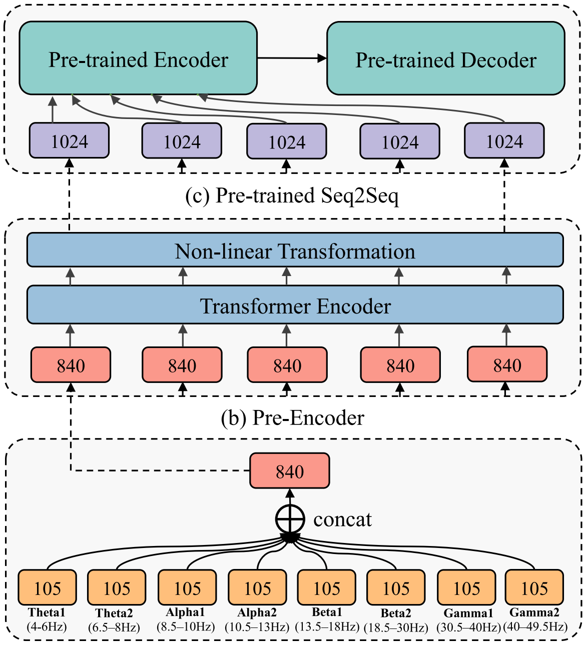
3) Curriculum Level: What are the curriculum levels? We conduct preliminary experiments by setting up the number of curriculum levels from 2 to 5 and finally decide to split the $\grave{\mathbb{E}}_i^+$ and $\acute{\mathbb{E}}_i^-$ into 3 levels due to their better performance. In detail, we split the sorted $\grave{\mathbb{E}}_i^+$ into three equal-length parts, including $[\mathbb{E}_i^{\mathtt{easy}+}, \mathbb{E}_i^{\mathtt{medium}+}, \mathbb{E}_i^{\mathtt{hard}+}]$ and $\acute{\mathbb{E}}_i^-$ into three equal-length parts, $[\mathbb{E}_i^{\mathtt{easy}-}, \mathbb{E}_i^{\mathtt{medium}-}, \mathbb{E}_i^{\mathtt{hard}-}]$. In other words, we obtain curriculums of different difficulty according to the length of sorted $\grave{\mathbb{E}}_i^+$ and $\acute{\mathbb{E}}_i^-$. Fig. 4 shows two examples of contrastive pairs of different difficulties. We can clearly find the easy pair

[3] In our preliminary experiments, we consider both two conditions and only the first condition, the results show that considering both conditions can achieve better results.


• Negative
For the most part, it's a work of incendiary genius, steering clear of knee-jerk reactions and quick solutions.
• Positive
He became a banker in Boston and was active in the local Democratic Party.
• Anchor
He became a banker in Boston and was active in the local Democratic Party.
• Positive
Frida is certainly no disaster, but neither is it the Kahlo movie Frida fans have been looking for.
• Negative
Among the many television shows in which Miguel has made guest appearances are: Miami Vice, CHiPs, Will & Grace, Superman, Tales from the Crypt and Twin Peaks.
• Anchor
Frida is certainly no disaster, but neither is it the Kahlo movie Frida fans have been looking for.

(a) Easy contrastive pair (b) Hard contrastive pair

Fig. 4. Contrastive pairs of different difficulties.



(a) Word-Level EEG Feature Construction

Fig. 5. Illustration of the BRAINTRANSLATOR. The number in rectangles denotes the dimension of the vector.

already satisfies the condition: *positive pairs are similar while negative pairs are dissimilar*. In contrast, the hard contrastive pair instead follows the condition: *positive pairs are dissimilar while negative pairs are similar*.

4) Curriculum Scheduler: When to update the curriculum? We adopt a One-Pass scheduler with a linear pace [3] to progressively train the model in an easy-to-hard order. *One-Pass scheduler* means that training the model only once per curriculum, while *linear pace* ensures that each curriculum takes the same amount of training time. In detail, when reaching the hard level, given an anchor EEG $\boldsymbol{E}_i$, we select the positive sample and negative sample from $\mathbb{E}_i^{\texttt{hard+}}$ and $\mathbb{E}_i^{\texttt{hard-}}$, respectively.

C. Backbone Model

Our backbone model BRAINTRANSLATOR inherits a typical Encoder-Decoder framework, which first encodes a sequence of word-level EEG features $\boldsymbol{E}$ to distributed representations and then generates the target sentence $\mathcal{S}$ with the decoder. The overall architecture is shown in Fig. 5. BRAINTRANSLATOR takes word-level EEG features as input and produces the corresponding sentence. It mainly consists of three parts: (a) *Word-Level EEG Feature Construction* that concatenates features of different bands of one word to form the final word-level EEG feature. (b) *Pre-encoder* that transforms original EEG features into the pre-trained Seq2Seq embedding space, and (c) *Pre-trained Seq2Seq* that takes a sequence of transformed embeddings and produces the final output sentence. Formally speaking, the overall model is formulated as:

$$\begin{aligned}
\boldsymbol{E}^{N_1} &= \texttt{Pre-Encoder}(\boldsymbol{E}) \\
&\underset{n=1}{\overset{N_1}{:=}} \textsc{Ffn}\left(\textsc{Att}(\boldsymbol{E}^{n-1})\right) \\
\boldsymbol{X}^{N_1+N_2} &= \texttt{Pre-trained Encoder}(\boldsymbol{E}^{N_1}) \\
&\underset{n=1}{\overset{N_2}{:=}} \textsc{Ffn}\left(\textsc{Att}(\boldsymbol{X}^{n-1})\right) \\
\boldsymbol{Y}^{M} &= \texttt{Pre-trained Decoder}(\boldsymbol{Y}^0, \boldsymbol{X}^{N_1+N_2}) \\
&\underset{m=1}{\overset{M}{:=}} \textsc{Ffn}\left(\textsc{Att}\left(\textsc{Att}(\boldsymbol{Y}^{m-1}), \boldsymbol{X}^{N_1+N_2}\right)\right) \qquad (1)
\end{aligned}$$

where $\underset{n=1}{\overset{N}{:=}}$ denotes N identical encoding layers and $\underset{m=1}{\overset{M}{:=}}$ denotes M decoding layers. $\boldsymbol{Y}^0$ describes the shifted right version of $\mathcal{S}$, $\textsc{Ffn}(\cdot)$ represents a position-wise feed-forward network, and $\textsc{Att}(\cdot)$ represents a multi-head attention.

D. Learning Procedure

The overall training process follows a two-step manner. The first is the C-SCL that aims to pre-train the pre-encoder. The second is the language modelling that aims to jointly optimize the whole EEG-to-text generation model.

Firstly, we adopt our C-SCL to train the pre-encoder. Formally, given anchor $\boldsymbol{E}_i$ and one specific curriculum level $\texttt{c\_level}$, we have the contrastive triple $(\boldsymbol{E}_i, \boldsymbol{E}_i^{\texttt{c\_level+}}, \boldsymbol{E}_i^{\texttt{c\_level-}})$ (Algorithm 1 shows the construction process). After the transformation of the pre-encoder, we can get $(\boldsymbol{h}_i, \boldsymbol{h}_i^+, \boldsymbol{h}_i^-)$, where $\boldsymbol{h}_i$ is the averaged vector of the outputs of the pre-encoder. Following the contrastive framework in Gao et al. [13], we minimize the cross-entropy loss ℓ_i defined by (N is the mini-batch size):

$$\begin{aligned}
&\ell_i(\boldsymbol{E}_i, \boldsymbol{E}_i^{\texttt{c\_level+}}, \boldsymbol{E}_i^{\texttt{c\_level-}}) \\
&= -\log \frac{e^{\texttt{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^+)/\tau}}{\sum_{j=1}^{N}\left(e^{\texttt{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^+)/\tau} + e^{\texttt{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^-)/\tau}\right)} \qquad (2)
\end{aligned}$$

where τ is a temperature hyperparameter[4]. $\texttt{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ is the cosine similarity. Note that our C-SCL works in an online manner, which means both positive and negative pairs are constructed dynamically along with the training process (*pairs are decided during training*) rather than constructing them offline (*pairs are decided before training*). This increases the distribution of contrastive pairs, thus improving training efficiency. Accordingly, the overall learning objective of the first step is:

$$\mathcal{L}_{\texttt{step1}} = \sum\nolimits_{\texttt{c\_level}\in[\texttt{easy},\texttt{medium},\texttt{hard}]} \sum\nolimits_{\boldsymbol{E}_i\in\mathbb{E}} (\ell_i) \qquad (3)$$

[4] The key for successful EEG contrastive training is the tiny τ, we show our parameter search results in the V-C Result section.

Algorithm 1 Contrastive Pairs Construction for Specific Curriculum Level

```
Input: EEG $\boldsymbol{E}_i$ with its corresponding subject $p_i$ and
       sentence $\mathcal{S}_i$; a dict $f_s : \mathcal{S}_i \rightarrow \mathbb{E}_{\mathcal{S}_i}$ maps $\mathcal{S}_i$ to a
       set of EEG signals $\mathbb{E}_{\mathcal{S}_i}$; a dict $f_p : p_i \rightarrow \mathbb{E}_{p_i}$
       maps $p_i$ to a set of EEG signals $\mathbb{E}_{p_i}$; a set of
       all sentences $\mathbb{S}$; curriculum level c_level;
Output: a contrastive triple
       $(\boldsymbol{E}_i, \boldsymbol{E}_i^{\texttt{c\_level+}}, \boldsymbol{E}_i^{\texttt{c\_level-}})$.
Function C_SCL($\boldsymbol{E}_i$, c_level):
    // positive sample
    $\mathbb{E}_i^+ = f_s(\mathcal{S}_i) \backslash \boldsymbol{E}_i$;
    $\grave{\mathbb{E}}_i^+$ = cur_cri($\boldsymbol{E}_i$, $\mathbb{E}_i^+$, descending);
    curriculums = cur_lev($\grave{\mathbb{E}}_i^+$);
    $\boldsymbol{E}_i^{\texttt{c\_level+}}$ =
      cur_sche(curriculums, c_level);
    // negative sample
    $\mathbb{E}_i^- = f_s(\mathbb{S} \backslash \mathcal{S}_i) - f_p(p_i)$;
    $\acute{\mathbb{E}}_i^-$ = cur_cri($\boldsymbol{E}_i$, $\mathbb{E}_i^-$, ascending);
    curriculums = cur_lev($\acute{\mathbb{E}}_i^-$);
    $\boldsymbol{E}_i^{\texttt{c\_level-}}$ =
      cur_sche(curriculums, c_level);
    return $(\boldsymbol{E}_i, \boldsymbol{E}_i^{\texttt{c\_level+}}, \boldsymbol{E}_i^{\texttt{c\_level-}})$;
// curriculum criterion
Function cur_cri($\boldsymbol{E}$, $\mathbb{E}$, order):
    sims = list();
    for $\boldsymbol{E}_j \in \mathbb{E}$ do
        sim_j = cosine_similarity($\boldsymbol{E}, \boldsymbol{E}_j$);
        sims.append(sim_j)
    indices = sims.sort(order);
    return $\mathbb{E}$[indices];
// curriculum level
Function cur_lev($\mathbb{E}$):
    [$\mathbb{E}^{\texttt{easy}}, \mathbb{E}^{\texttt{medium}}, \mathbb{E}^{\texttt{hard}}$] = split($\mathbb{E}$);
    return [$\mathbb{E}^{\texttt{easy}}, \mathbb{E}^{\texttt{medium}}, \mathbb{E}^{\texttt{hard}}$];
// curriculum scheduler
Function cur_sche(curriculums, c_level):
    $\mathbb{E}^{\texttt{select}}$ = select(curriculums,
      curr_level);
    $\boldsymbol{E}$ = random_select($\mathbb{E}^{\texttt{select}}$);
    return $\boldsymbol{E}$;
```

Secondly, based on the contrastive-trained pre-encoder, we jointly fine-tune all the parameters of the BrainTranslator to minimize the cross-entropy loss in a parallel training corpus $(\mathbb{E}, \mathbb{S})$:

$$\mathcal{L}_{\texttt{step2}} = -\sum\nolimits_{(\boldsymbol{E},\mathcal{S})\in(\mathbb{E},\mathbb{S})} \log p(\mathcal{S}\,|\,\boldsymbol{E};\theta) \tag{4}$$

IV. Experiments

A. Baseline Models

We adopt the previous state-of-the-art **BrainBART** [34] as our baseline model, which is composed of the Transformer pre-encoder[5] and the BART pre-trained seq2seq model [22]. Besides, we further employ other two types of widely used pre-trained seq2seq models, including PEGASUS [36] and T5 [30], building upon the Transformer pre-encoder to form **BrainPEGASUS** and **BrainT5** respectively. All the above three models come in two model-size variants, including **Large** and **Base**, leading to six models in total.

B. Evaluation Protocol

Following Wang and Ji [34], we adopt **ROUGE** [23] and **BLEU** [29] for evaluating the EEG-to-Text generation task. Besides, following Metzger et al. [25], we also adopt **Word Error Rate (WER)** as our metric to examine more fine-grained generation performance.

C. Implementation Details

Our pre-encoder consists of 6 layers, each with 8 heads and a hidden dimension of 2048. The dimension of the input EEG representation is 840. For the contrastive training process, we use Adam with a learning rate of 0.001 with a batch size of 32. τ is set to 0.00001. For the curriculum training process, we train one epoch for each curriculum from easy to hard (easy, medium and hard). For the overall training process, we first load the checkpoint of the contrastive-trained pre-encoder and then fine-tune the whole model using Adam with a learning rate of 2e-5 and batch size of 32. For the generation process, following Wang and Ji [34], we equip our model with greedy decoding to produce final sentences. For all three metrics, we use standard implementations provided by HuggingFace.[6]

V. Results

A. Automatic Evaluation

Table II shows the performance of our SCL and C-SCL on the ZuCo benchmark. In detail, we evaluate our model following two settings: (1) the 10-fold cross-validation setting and (2) the same data split setting with respect to Wang and Ji [34][7]. Overall, we find that SCL can consistently attain strong performance across various baseline models and architectures. With the enhancement of curriculum learning, C-SCL can further boost performance. In detail, our approach achieves state-of-the-art performance across six different architectures. Specifically, when comparing our method to the previous SOTA model (BrainBART-Large), we observe a 1.58-point increase in ROUGE-L and a 2.41-point increase in BLEU-4, and a 2.25-point enhancement in WER, which serves as substantial evidence of the effectiveness of our method. When comparing SCL to C-SCL, our state-of-the-art C-SCL demonstrates comprehensive supremacy across all metrics. In addition to the main observations, our empirical

[5] We also tested Conformer [15] as the pre-encoder. However, the experimental results showed no major difference. Accordingly, we kept using the Transformer pre-encoder in our paper.

[6] https://github.com/huggingface/evaluate

[7] To directly compare our method with the previous state-of-the-art BrainBART-Large under identical experimental conditions, we perform the further analyses utilizing the same data splits according to Wang and Ji [34].

TABLE II

TEST SET RESULTS ON THE ZUCO BENCHMARK UNDER THE 10-FOLD CROSS-VALIDATION SETTING. THE RESULTS ENCLOSED IN PARENTHESES ARE OBTAINED UTILIZING THE IDENTICAL DATASET SPLITS AS THOSE EMPLOYED BY WANG AND JI [34]. ↑ MEANS HIGHER IS BETTER. ↓ MEANS LOWER IS BETTER

	ROUGE(%)↑			BLEU(%)↑				
Model	**R-1**	**R-2**	**R-L**	**B-1**	**B-2**	**B-3**	**B-4**	**WER(%)↓**
BRAINBART-LARGE [34]	37.78 (37.85)	18.63 (18.83)	35.74 (35.92)	34.55 (34.79)	24.11 (24.38)	19.32 (19.58)	16.94 (17.02)	70.42 (70.31)
BRAINBART-LARGE (w/ SCL)	38.57 (38.71)	20.29 (20.05)	36.65 (36.73)	35.51 (35.65)	25.55 (25.74)	21.26 (**21.31**)	18.78 (**18.96**)	69.34 (69.12)
BRAINBART-LARGE (w/ C-SCL)	**39.17 (39.14)**	**20.75 (20.35)**	**37.32 (37.12)**	**36.25 (35.91)**	**26.17 (25.96)**	**21.76 (21.31)**	**19.35** (18.89)	**68.17 (68.48)**
BRAINBART-BASE	36.56 (36.46)	18.05 (17.75)	34.25 (34.23)	33.70 (33.64)	23.64 (23.60)	18.85 (18.78)	16.31 (16.23)	72.72 (73.01)
BRAINBART-BASE (w/ SCL)	36.98 (36.70)	18.31 (17.92)	34.45 (34.55)	34.20 (34.18)	23.88 (24.07)	19.27 (19.31)	16.96 (16.79)	71.77 (72.27)
BRAINBART-BASE (w/ C-SCL)	**37.38 (37.01)**	**18.93 (18.05)**	**34.91 (34.69)**	**35.10 (34.55)**	**24.42 (24.39)**	**20.07 (19.61)**	**17.65 (17.04)**	**70.73 (71.65)**
BRAINPEGASUS-LARGE	37.33 (37.50)	16.05 (16.10)	34.01 (34.27)	34.37 (34.56)	22.32 (22.57)	16.92 (17.07)	14.18 (14.26)	76.41 (76.21)
BRAINPEGASUS-LARGE (w/ SCL)	39.51 (39.34)	18.37 (18.07)	35.85 (35.83)	36.40 (36.35)	24.78 (24.74)	19.46 (19.38)	16.75 (16.62)	74.34 (74.54)
BRAINPEGASUS-LARGE (w/ C-SCL)	**40.26 (40.18)**	**19.38 (19.20)**	**36.96 (36.72)**	**37.25 (37.24)**	**26.03 (25.89)**	**20.78 (20.63)**	**17.98 (17.92)**	**73.27 (73.43)**
BRAINPEGASUS-BASE	36.89 (36.70)	14.40 (14.37)	33.28 (33.23)	33.84 (33.74)	21.34 (21.05)	14.99 (14.80)	11.78 (11.53)	78.08 (78.19)
BRAINPEGASUS-BASE (w/ SCL)	37.03 (36.74)	15.59 (**15.33**)	33.39 (33.29)	34.09 (33.84)	22.24 (**21.88**)	16.75 (**16.38**)	13.81 (**13.60**)	77.47 (77.95)
BRAINPEGASUS-BASE (w/ C-SCL)	**37.69 (37.27)**	**15.63** (15.21)	**33.77 (33.66)**	**34.61 (34.20)**	**22.83** (21.73)	**17.08** (16.26)	**14.31** (13.50)	**76.23 (76.59)**
BRAINT5-LARGE	32.04 (32.17)	12.05 (12.12)	29.54 (29.81)	30.13 (30.43)	19.12 (19.24)	13.24 (13.48)	10.16 (10.32)	83.88 (83.69)
BRAINT5-LARGE (w/ SCL)	32.42 (32.65)	13.67 (14.84)	30.03 (30.33)	30.88 (31.06)	20.35 (20.80)	15.57 (15.87)	12.71 (13.25)	82.83 (82.61)
BRAINT5-LARGE (w/ C-SCL)	**33.31 (32.87)**	**14.95 (14.87)**	**30.94 (30.54)**	**31.27 (31.18)**	**21.01 (20.91)**	**16.69 (15.98)**	**13.63 (13.40)**	**81.78 (81.91)**
BRAINT5-BASE	31.29 (31.12)	8.02 (7.77)	27.71 (27.65)	27.25 (27.05)	13.47 (13.31)	6.51 (6.44)	3.51 (3.38)	86.37 (86.46)
BRAINT5-BASE (w/ SCL)	31.42 (31.37)	8.93 (8.56)	28.50 (**28.17**)	28.43 (28.38)	15.01 (14.90)	8.18 (**8.08**)	4.97 (4.81)	86.02 (86.15)
BRAINT5-BASE (w/ C-SCL)	**31.90 (31.38)**	**9.32 (8.63)**	**29.43** (28.15)	**29.38 (28.46)**	**15.94 (14.95)**	**8.74** (8.06)	**5.53 (4.86)**	**84.79 (85.10)**

results also demonstrate the following two findings. Firstly, BART performs well. Although this finding is exclusively derived from results based on three pre-trained seq2seq models, it still provides the guideline for choosing future backbone seq2seq models for the EEG-to-Text generation task: choosing task-agnostic language models (e.g., BART) rather than task-oriented models (e.g., PEGASUS for summarization and T5 requiring task prompts). Secondly, EEG-to-Text generation also follows the scaling law, which means the generation performance scales up with the increasing number of model parameters[8].

B. Human Evaluation

To further assess the quality of the generated texts, we conduct a human evaluation study. We choose two metrics: *consistency* (EEG representations with respect to the same sentence can be consistently decoded into the same sentence) and *correctness* (the decoded sentence is factually consistent with the reference sentence). Specifically, we employ three evaluators to undertake the human evaluation. Each evaluator is remunerated $40 for this evaluation task. We randomly select 50 unique sentences from the test set and take 5 EEG representations elicited by different subjects for each sentence to conduct the evaluation. For consistency, given 5 EEG representations corresponding to one sentence, we evaluate whether the generated 5 sentences are consistent. For correctness, we evaluate whether 250 generated sentences are factually consistent with the ground truth. For each metric, the score ranges from 1 (worst) to 5 (best). The results are shown in Table III. Firstly, we find that our proposed SCL and C-SCL can achieve better scores in terms of two metrics, with the C-SCL performing the best. Secondly, even our best method still cannot achieve very good results on correctness, which shows that factual inconsistency remains an important challenge for the brain-to-text generation task.

TABLE III

HUMAN EVALUATION RESULTS

	Consistency	**Correctness**
BRAINBART-LARGE	2.40	2.83
BRAINBART-LARGE (w/ SCL)	3.40	3.22
BRAINBART-LARGE (w/ C-SCL)	**4.20**	**3.58**

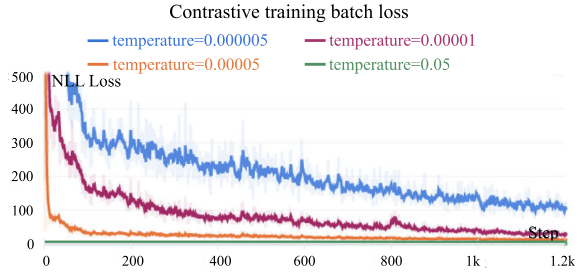


Fig. 6. Contrastive training loss with respect to different τ.

C. Analysis

1) Parameter Search for τ: Fig. 6 shows the contrastive training loss under different τ[9] We can find that setting τ to a small number is critical for successful EEG contrastive training. In contrast, a larger τ value of 0.05 results in an almost 0 loss, indicating that contrastive training is ineffective for settings where $\tau = 0.05$. We attribute this to the fact that the original EEG signals are similar to each other, a small τ can produce more distinguishable EEG representations, thus enabling effective contrastive learning. We conduct

[8] Due to the better results based on the BRAINBART-LARGE model, the following experiments are all based on the BRAINBART-LARGE model.

[9] The figure is obtained via https://wandb.ai/

TABLE IV
RESULTS OF THE DIFFERENT NUMBER OF CURRICULUM LEVELS BASED ON THE BRAINBART-LARGE (W/ C-SCL)

Number of Levels	ROUGE-1↑	BLUE-1↑	WER↓
2	38.67	35.31	68.67
3	**39.14**	**35.91**	**68.48**
4	38.49	35.07	68.91
5	37.84	34.63	69.25

TABLE V
ABLATION STUDY FOR EXPLICIT NEGATIVE PAIRS

	R-1↑	B-1↑	WER↓
BRAINBART-LARGE (W/ SCL)	**38.71**	**35.65**	**69.12**
BRAINBART-LARGE (W/ SCL) (w/o explicit negative)	38.16	34.89	69.88

preliminary experiments and find that setting τ to 0.00001 yields better EEG-to-Text generation performance.

2) Parameter Search for the Number of Curriculum Levels: The search results for varying numbers of curriculum levels are presented in Table IV. In particular, the $\grave{\mathbb{E}}_i^+$ and $\acute{\mathbb{E}}_i^-$ are partitioned equally based on the number of curriculum levels. Subsequently, the model is trained progressively in an easy-to-hard order. After evaluating the performance, the number of curriculum levels is set to 3.

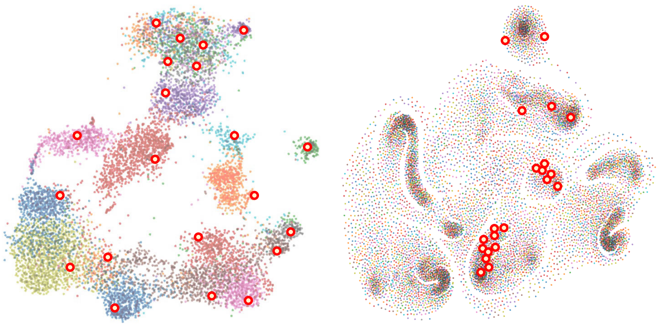
3) Ablation Study for Explicit Negative Pairs: To examine whether explicit negative pairs are necessary. We conduct the ablation study by only considering the in-batch negative pairs without incorporating explicitly crafted negative pairs. Table V shows the results. We can find that explicit negative pairs indeed do good to the contrastive learning, thus are effective and necessary.

4) Comparison with Domain-Adversarial Learning Method: Recall that the key challenge of the brain-to-text generation task is to mitigate the discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation. Accordingly, for a more comprehensive evaluation of our proposed method, we additionally explore one critical method, domain-adversarial learning (DAL) [12], which is also skilled in learning domain-invariant representations, to address this challenge. Specifically, the objective of domain-adversarial learning is to learn EEG representations that are indiscriminate with respect to the same sentence by treating any two EEG representations corresponding to the same sentence as the source domain and the target domain, respectively. The experimental results are presented in Table VI. We can find that both contrastive learning and domain-adversarial learning can mitigate the discrepancy and improve brain-to-text performance. However, DAL underperforms compared to C-SCL, suggesting the method requires careful adaptation to this task. Overall, we believe domain-adversarial learning holds promise as an important research direction for brain-to-text generation. Further efforts are warranted to fully realize its potential.

5) Embedding Visualization: To verify whether our C-SCL can achieve learning of semantic-dependent EEG representations. We give a straightforward comparison via t-SNE between the original EEG representations (Fig. 7(a)) and EEG representations obtained after the transformation of the contrastive-trained pre-encoder (Fig. 7(b)). We can easily observe that our learned EEG representations of the same sentence tend to be closer compared with original desultorily distributed ones. This result coincides with our initial goal. Besides, Fig. 7(a) also shows distinct subject clusters (different colours) while Figure 7(b) reveals subjects distributed more equally. Nevertheless, Fig. 7(b) also shows the EEG representations of the same sentence are not fully clustered. Instead, multiple sub-clusters are formed, which indicates achieving a desirable semantic-dependent EEG representation space is a challenging task. To alleviate this challenge, we envision three potential paths. First, optimize EEG signal preprocessing. We could introduce a new normalization method by introducing the semantic-dependent EEG representation idea during the preprocessing to bias the initial EEG representation. Second, employ pre-training techniques. Pre-trained EEG models could also enhance EEG modelling. Through meticulous pre-training objective design, we could guide the model to learn semantic-dependent EEG representations. Third, leverage joint learning approaches like contrastive learning and domain-adversarial learning to augment the model's learning objective and accomplish enhanced performance.

TABLE VI
RESULTS OF DIFFERENT PRE-TRAINING METHODS

	R-1↑	B-1↑	WER↓
BRAINBART-LARGE	37.85	34.79	70.31
BRAINBART-LARGE (W/ DAL)	38.52	35.49	69.54
BRAINBART-LARGE (W/ SCL)	38.71	35.65	69.12
BRAINBART-LARGE (W/ C-SCL)	**39.14**	**35.91**	**68.48**



(a) Original EEG Representations (b) Contrastive-learned EEG Representations

Fig. 7. T-SNE visualization of sentence-level EEG representations of sentences in the training set, which are (a) original EEG representations and (b) generated by the pre-encoder after C-SCL. Different colours mean different subjects. Each dot represents a sentence. The red box dots represent the EEG representations corresponding to the same sentence "He and his wife had seven children".

6) Single-Subject Setting: Given that the subject-dependent EEG representation poses a great challenge to the EEG-to-Text generation task, in this analysis, we aim to answer one question: *Whether single-subject training is a more suitable way for the EEG-to-Text generation task?* To verify this, we test both mixed-subjects training and single-subject training methods on data from 4 distinct subjects. The results are shown in Fig. 9. Compared with single-subject training, all

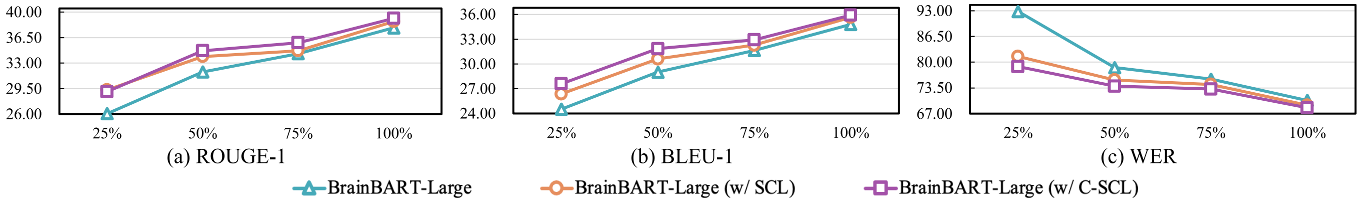


Fig. 8. Results of different training data sizes.

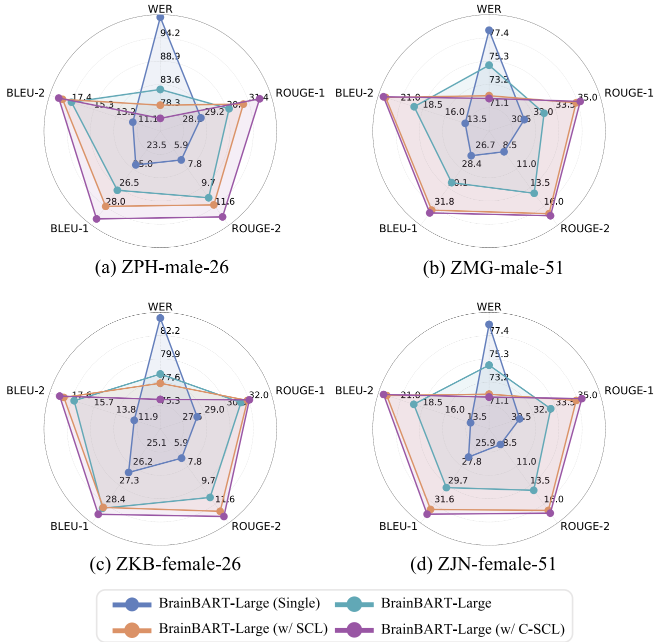


Fig. 9. Results of different methods testing on 4 subjects respectively, including both male and female, youth and middle-aged, e.g., ZPH-male-26 describes the subject identified as ZPH, is male and 26 years old. BRAINBART-LARGE (Single) means that training and testing on the data of a single subject. Others mean that training on the whole data while testing on the data of a single subject.

other three mixed-subjects training methods achieve remarkable improvements, which precisely indicates that it is worth exploring mixed-subjects training methods. Besides, the results also show the effectiveness of our proposed SCL and C-SCL at a more fine-grained level.

7) Low-Resource Setting: To verify the robustness of our methods on varying data sizes, we provide datasets of different sizes to train the pre-encoder using SCL and C-SCL, then fine-tune the whole model. Note that the size of the test set is the same across all experiments. The results are shown in Fig. 8. We can find that the model performance clearly improves with the growing of dataset size in terms of all metrics. Prominently, our methods show great advantages in the low-resource setting. Especially when only using 25% of the dataset, our C-SCL can directly reduce the WER from 92.83% to 78.89%, achieving comparable results compared with using 50% of the dataset.

8) Zero-Shot Setting: To verify the generalizability of our methods, we conduct zero-shot experiments by training on the partial ZuCo dataset, which excludes the data of one selected test subject. The results are shown in Fig. 10. We can see that our methods yield strong performance for unseen ZPH and ZKP respectively. We attribute this good generalizability to the fact that contrastive learning not only learns better representations for currently available subjects but also optimizes a distinguishable representation space that can be easily transferred and adapted to unseen subjects.

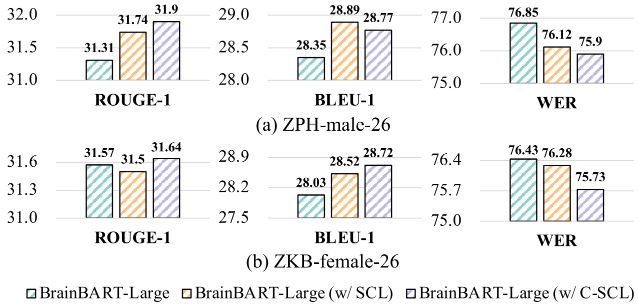


Fig. 10. Zero-shot results by training on data that excluded the final test subject.

TABLE VII
RESULTS OF DIFFERENT CURRICULUM LEVELS

	R-1↑	B-1↑	WER↓
SCL($\mathbb{E}^{\text{easy}}$)	37.89	34.82	70.05
SCL($\mathbb{E}^{\text{medium}}$)	38.21	35.10	69.83
SCL($\mathbb{E}^{\text{hard}}$)	37.92	35.08	70.09
C-SCL($[\frac{\mathbb{E}^{\text{easy}}}{3}, \frac{\mathbb{E}^{\text{medium}}}{3}, \frac{\mathbb{E}^{\text{hard}}}{3}]$)	**38.52**	**35.34**	**69.49**

9) Single-Curriculum Setting: To verify the necessity of curriculum learning for our C-SCL. We individually perform SCL based on contrastive pairs from each curriculum level, including SCL($\mathbb{E}^{\text{easy}}$), SCL($\mathbb{E}^{\text{medium}}$) and SCL($\mathbb{E}^{\text{hard}}$). Then, we select one-third of the data from each curriculum level and conduct C-SCL based on the fixed $[\frac{\mathbb{E}^{\text{easy}}}{3}, \frac{\mathbb{E}^{\text{medium}}}{3}, \frac{\mathbb{E}^{\text{hard}}}{3}]$. Note that all the above contrastive learning datasets keep the same size and the fine-tuning is based on the whole ZuCo train part. The results are shown in Table VII. Firstly, we can find that curriculum learning indeed does good to the model performance. Besides, both SCL($\mathbb{E}^{\text{easy}}$) and SCL($\mathbb{E}^{\text{hard}}$) achieve relatively lower results. We attribute this fact to that easy pairs are insignificant but directly leveraging hard pairs is quite challenging for model learning. In addition to the extrinsic evaluation based on the downstream Brain-to-Text generation task, we further conduct the intrinsic analysis to give an in-depth understanding of the efficiency of our proposed C-SCL compared with the SCL. For each method, during the training process, we calculate the average cosine similarity of EEG representations (obtained after the transformation of the contrastive-trained pre-encoder)

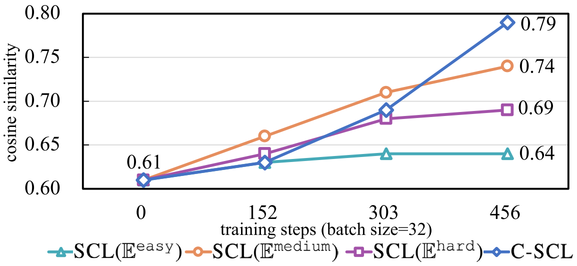


Fig. 11. Average cosine similarity of EEG representations corresponding to the same sentence in the valid set during the training process.

corresponding to the same sentence in the valid set. Specifically, we set four calculation points, which are at the start, one-third, two-thirds, and the end of the full training process, respectively. The results are shown in Fig. 11. Firstly, we find that our C-SCL can learn more similar EEG representations corresponding to the same sentence, which is in line with our learning objective. Secondly, the results coincide with the finding in the Table VII, where C-SCL performs the best, SCL($\mathbb{E}^{\text{easy}}$) and SCL($\mathbb{E}^{\text{hard}}$) perform worse. By means of this intrinsic analysis, we can attribute the success of our C-SCL to the effective learning of semantic-dependent EEG representations.

D. Case Study

Fig. 12 shows the case study. We can find our method generates the same sentence for EEG signals elicited by different subjects based on learned semantic-dependent EEG representations, whereas the baseline produces different ones. Besides, our result is more semantic-related compared with baseline results, which indicates that semantic-dependent EEG representation can enhance the generation performance. However, there still exists a large gap between our generation and the golden reference. We believe future works should pay attention to the following research directions: (1) Strategies by jointly modelling continuous word-level EEG signals and the syntactic structure of sentences, since the current generation still failed to capture the linguistic structure; (2) Strategies to close the gap between the word-level EEG feature and token-level generation, since the current generation still has several spelling errors.

VI. Related Work

A. Brain-to-Text Generation

Brain-to-Text generation is an active area of research at the intersection of artificial intelligence and neuroscience [31] and is closely related to research on simulating human perceptual experiences and reasoning processes [7], [8], [20]. According to the classification criterion of vocabulary size, there are two series of related works: *closed vocabulary* and *open vocabulary* brain-to-text generation. The first line of works generates words in small closed vocabularies [24], [28]. For example, Moses et al. [28] focus on a 50-word vocabulary. While exhibiting promising generation accuracy and speed, expanding access to a larger vocabulary enables effective day-to-day communication. Accordingly,

Baseline: BrainBART	
ZPH	He was educated radio in famous Railroad and the serveded for York City government to prompt workings of the radio.
ZKB	He was born founding in the famous Railroad and and grownwed up in York City leded for its United.
ZJM	He was a member in the Alamo Railroad and lived in York City government to allow from the govered.

Our: BrainBART (w/ C-SCL)	
ZPH ZKB ZJM	He was a member in various famous Railroad and the serveded for President York City Senate to havecede from the United.

Golden	
ZPH ZKB ZJM	He was a participant in the Underground Railroad, and even petitioned the New York state Legislature to secede from the Union in a protest against slavery.

Fig. 12. Generations for EEG signals of different subjects. The EEG signals correspond to the same sentence. “ZPH”, “ZKB”, and “ZJM” are three subjects.

Wang and Ji [34] study the problem of open vocabulary EEG-to-Text decoding task by utilizing pre-trained language models (PLMs) [22]. It brings two benefits: on the one hand, PLMs offer a large vocabulary, on the other hand, PLMs can serve as a bridge between brain signals and linguistic information [26]. In our work, we focus on the open vocabulary paradigm due to the non-invasive nature and widespread application prospects of EEG-based BCIs. Specifically, we pay particular attention to the challenge of the discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation for the EEG-to-Text generation task.

B. Contrastive Learning

Contrastive learning is a technique that aims to make the representation of a given anchor data to be similar to its positive pairs while being dissimilar to its negative pairs. It shows promising results in computer vision [5], [16], [17] and has gained popularity in natural language processing [13], [14]. After witnessing its superiority in the above areas, contrastive learning is attracting the attention of neuroscientists and has been applied to several EEG-based classification tasks [6], [10], [21], [27], [32]. More recently, Shen et al. [32] propose a contrastive learning method to tackle the cross-subject emotion recognition problem. Défossez et al. [10] devise a contrastive learning objective to align representations of brain signals and natural speech. In our work, we devise a novel curriculum semantic-aware contrastive learning strategy (C-SCL), aiming to learn semantic-dependent EEG representations, which effectively reduce the discrepancy between the EEG and text representations.

VII. Conclusion and Future Work

In this paper, we propose a curriculum semantic-aware contrastive learning strategy (C-SCL) to reduce the discrepancy between the subject-dependent EEG representation and the semantic-dependent text representation. The experimental results based on the ZuCo benchmark demonstrate its effectiveness for the EEG-to-Text generation task. Besides, our analyses also verify the robustness and superior generalizability of our C-SCL in the low-resource setting and the zero-shot setting, respectively. Moreover, single-subject setting experiments also point to the necessity of exploring mixed-subjects training methods for the EEG-to-Text generation task.

We believe that forthcoming research endeavours will seek to implement the proposed method in real scenarios. First, building upon the existing C-SCL framework, future work could consider semantic similarity when constructing contrastive pairs and integrate multiple solutions like contrastive learning and domain-adversarial learning to further improve performance. Second, findings from neuroscience research could inform the text decoding stage, associating brain-inspired related words during decoding to mitigate the hallucination problem. Third, collaborating with hospitals would enable deploying the method with actual patients, gauging its effectiveness and robustness. Overall, opportunities remain to refine the technique through semantic-similarity-aware contrastive learning, brain-inspired text decoding, and validation in real-world clinical settings.

References

[1] R. Adolphs, "Neural systems for recognizing emotion," *Current Opinion Neurobiol.*, vol. 12, no. 2, pp. 169–177, Apr. 2002.

[2] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[3] Y. Bengio, J. Louradour, and R. Collobert, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2009, pp. 41–48.

[4] J. S. Brumberg, K. M. Pitt, A. Mantie-Kozlowski, and J. D. Burnison, "Brain-computer interfaces for augmentative and alternative communication: A tutorial," *Amer. J. Speech-Lang. Pathol.*, vol. 27, no. 1, pp. 1–12, 2018.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[6] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020, *arXiv:2007.04871*.

[7] Z.-Q. Cheng, Q. Dai, S. Li, T. Mitamura, and A. Hauptmann, "GSRFormer: Grounded situation recognition transformer with alternate semantic attention refinement," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3272–3281.

[8] Z.-Q. Cheng, X. Wu, S. Huang, J.-X. Li, A. G. Hauptmann, and Q. Peng, "Learning to transfer: Generalizable attribute learning with multitask neural model search," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 90–98.

[9] J. Claassen et al., "Detection of brain activation in unresponsive patients with acute brain injury," *New England J. Med.*, vol. 380, no. 26, pp. 2497–2505, Jun. 2019.

[10] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech from non-invasive brain recordings," 2022, *arXiv:2208.12266*.

[11] S. H. Felgoise, V. Zaccheo, J. Duff, and Z. Simmons, "Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis Frontotemporal Degeneration*, vol. 17, nos. 3–4, pp. 179–183, May 2016.

[12] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Apr. 2016.

[13] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6894–6910.

[14] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 879–895.

[15] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[18] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading," *Scientific Data*, vol. 5, no. 1, pp. 1–13, Dec. 2018.

[19] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 138–146. [Online]. Available: https://aclanthology.org/2020.lrec-1.18

[20] S. Huang, Z.-Q. Cheng, X. Li, X. Wu, Z. Zhang, and A. Hauptmann, "Perceiving physical equation by observing visual scenarios," 2018, *arXiv:1811.12238*.

[21] P. Lee, S. Hwang, J. Lee, M. Shin, S. Jeon, and H. Byun, "Inter-subject contrastive learning for subject adaptive EEG-based visual recognition," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–6.

[22] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–10.

[23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Jul. 2004, pp. 74–81.

[24] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature Neurosci.*, vol. 23, no. 4, pp. 575–582, Apr. 2020.

[25] S. L. Metzger et al., "Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis," *Nature Commun.*, vol. 13, no. 1, pp. 1–15, Nov. 2022.

[26] J. Millet et al., "Toward a realistic model of speech processing in the brain with self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33428–33443.

[27] E. Alsentzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, S. Roy, and S. L. Hyland, "Contrastive representation learning for electroencephalogram classification," in *Proc. Mach. Learn. Health Workshop (MLH)*, Dec. 2020, pp. 238–253. [Online]. Available: https://proceedings.mlr.press/v136/mohsenvand20a.html

[28] D. A. Moses et al., "Neuroprosthesis for decoding speech in a paralyzed person with anarthria," *New England J. Med.*, vol. 385, no. 3, pp. 217–227, 2021.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[30] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[31] N. Savage, "How AI and neuroscience drive each other forwards," *Nature*, vol. 571, no. 7766, pp. 15–17, Jul. 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198494781

[32] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Apr. 4, 2022, doi: 10.1109/TAFFC.2022.3164516.

[33] C. A. Stanger and M. F. Cawley, "Demographics of rehabilitation robotics users," *Technol. Disability*, vol. 5, no. 2, pp. 125–137, Oct. 1996.

[34] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 5, pp. 5350–5358.

[35] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, May 2021.

[36] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11328–11339.

[37] S. Zou, S. Wang, J. Zhang, and C. Zong, "Cross-modal cloze task: A new task to brain-to-word decoding," in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 648–657.