BELT: Bootstrapped EEG-to-Language Training by Natural Language Supervision

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, *Member, IEEE*, and Chin-Teng Lin, *Fellow, IEEE*

***Abstract*— Decoding natural language from noninvasive brain signals has been an exciting topic with the potential to expand the applications of brain-computer interface (BCI) systems. However, current methods face limitations in decoding sentences from electroencephalography (EEG) signals. Improving decoding performance requires the development of a more effective encoder for the EEG modality. Nonetheless, learning generalizable EEG representations remains a challenge due to the relatively small scale of existing EEG datasets. In this paper, we propose enhancing the EEG encoder to improve subsequent decoding performance. Specifically, we introduce the discrete Conformer encoder (D-Conformer) to transform EEG signals into discrete representations and bootstrap the learning process by imposing EEG-language alignment from the early training stage. The D-Conformer captures both local and global patterns from EEG signals and discretizes the EEG representation, making the representation more resilient to variations, while early-stage EEG-language alignment mitigates the limitations of small EEG datasets and facilitates the learning of the semantic representations from EEG signals. These enhancements result in improved EEG representations and decoding performance. We conducted extensive experiments and ablation studies to thoroughly evaluate the proposed method. Utilizing the D-Conformer encoder and bootstrapping training strategy, our approach demonstrates superior decoding performance across various tasks, including word-level, sentence-level, and sentiment-level decoding from EEG signals. Specifically, in word-level classification, we show that our encoding method produces more distinctive representations and higher classification performance compared to the EEG encoders from existing methods. At the sentence level, our model outperformed the baseline by 5.45%, achieving a BLEU-1 score of 42.31%. Furthermore, in sentiment classification, our model exceeded the baseline by 14%, achieving a sentiment classification accuracy of 69.3%.**

***Index Terms*— Brain-computer interface, brain-to-language translation, sentiment classification, large language model, contrastive learning, vector quantization.**

Manuscript received 9 December 2023; revised 17 July 2024; accepted 23 August 2024. Date of publication 27 August 2024; date of current version 11 September 2024. This work was supported in part by Australian Research Council (ARC) under Discovery Grant DP210101093 and Grant DP220100803, in part by the University of Technology Sydney (UTS) Human-Centric Artificial Intelligence (AI) Centre funding sponsored by GrapheneX (2023–2031), in part by Australia Defence Innovation Hub under Contract P18-650825, in part by Australian Cooperative Research Centres Projects (CRC-P) Round 11 under Grant CRCPXI000007, in part by the U.S. Office of Naval Research Global under Cooperative Agreement ONRG-NICOP-N62909-19-1-2058, in part by the Air Force Office of Scientific Research (AFOSR)—Defence Science and Technology (DST) Australian Autonomy Initiative under Agreement ID10134, in part by the New South Wales (NSW) Defence Innovation Network and the NSW State Government of Australia under Grant DINPP2019 S1-03/09 and Grant PP21-22.03.02. *(Corresponding author: Chin-Teng Lin.)*

The authors are with the Australian AI Centre, Human-Centric AI Centre, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: Jinzhao.Zhou@student.uts.edu.au; yiqun.duan@student.uts.edu.au; yu-cheng.chang@uts.edu.au; yukai.wang@uts.edu.au; chin-teng.lin@uts.edu.au).

Digital Object Identifier 10.1109/TNSRE.2024.3450795

I. Introduction

THE decoding of the user's intention from the noninvasive electroencephalography (EEG) signals has been a fascinating topic. Unlike most existing BCI-based applications such as motor imagery classification [1] and emotion recognition [2], [3], the potential to decode language with a large vocabulary size opens the door to a new paradigm for human-to-human and human-to-machine interaction [4], [5], [6], [7]. Although much effort has been made, decoding natural language from EEG signals remains a formidable challenge. Exemplified by the considerable opportunity for improvement in decoding precision, coherence, and open-vocabulary generalization [8], [9], [10], [11].

Existing solutions for EEG-to-language decoding use a generative approach that combines an EEG encoder with a generative language model (LM) as task-specific decoder [12], [13]. We depict such encoder-decoder structure in Figure 1. In these methods, word-level EEG embeddings are first encoded by an EEG encoder and then used as conditions for the decoder to generate sentences. Although this approach has shown promising outcomes, the limited scale of EEG datasets makes the adoption of a simple model architecture, such as a Transformer, less effective. As evidenced in previous research [14], [15], directly training a generic model on a small dataset can result in a lack of semantics in the learned representations, subsequently affecting generalization capacity and decoding performance [16]. To improve language decoding performance from EEG signals, we aim to develop a more effective encoder to ensure that the generated sentences are conditioned on the right EEG information. Therefore, we consider enhancing the EEG encoder in two key aspects. Firstly, we improve the architecture of the EEG encoder to better exploit inter-channel dependencies within the EEG signals. Secondly, we introduce semantic guidance during the learning process to bootstrap more meaningful EEG

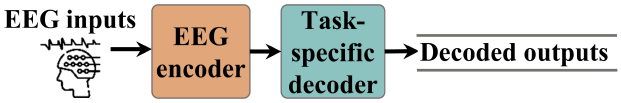

Fig. 1. Overview of the EEG-to-language decoding framework used in our research. The framework consists of an EEG encoder that encodes EEG data and a task-specific decoder that decodes the information from the encoded EEG representations. Our research focuses on enhancing the encoder of this framework. By generating superior EEG representations, we aim to improve performance across multiple decoding tasks.

representations. These areas are currently under-explored in EEG-to-language decoding research.

In this paper, we propose to improve EEG encoding capability for EEG-to-language decoding tasks using a novel encoder architecture and training method. Specifically, we introduced a novel discrete conformer (D-Conformer) as the EEG encoder to exploit both the global context and local brain dynamics from the EEG signals. Then, we bootstrap the learning of semantic EEG representation by imposing EEG-language alignment. We summarize our approach as Bootstrapped EEG-to-Language Training (BELT). Our BELT approach leverages pre-trained language models (BART [17]) to guide the training of our D-Conformer encoder. We evaluate BELT's effectiveness in enhancing the capacity of learned EEG representations across several tasks, including EEG-to-word classification, EEG-to-sentence decoding, and zero-shot sentiment classification. Additionally, we demonstrate that our bootstrapping scheme can be adapted to specific tasks by selecting different sources of language guidance. To handle various decoding tasks, we can bootstrap the training of the D-Conformer encoder using word-level, sequence-level, or context-level modeling strategies. Our extensive experiments show that BELT achieves performance gains over existing methods. The highlights of this paper can be summarized as follows:

- We propose the D-Conformer as a novel EEG encoder architecture that employs vector quantization and Conformer blocks to enhance the extraction and utilization of EEG information.
- We propose bootstrapping the training of the EEG encoder by aligning its representations with pre-trained language models (LMs). Our bootstrapping method adapts to various decoding tasks by leveraging different sources of language guidance, including word-level, sequence-level, and context-level modeling strategies.
- Extensive experiments and ablations were conducted on decoding tasks including EEG-to-word classification, EEG-to-sentence decoding, and zero-shot sentiment classification to evaluate the effectiveness of the proposed methods. Internal comparison between ablated models and external comparison with existing methods show that our proposed approach improves performance across these EEG decoding tasks.

II. Related Work

A. Decoding Language From Human Brain Signals

Existing research on brain-to-language decoding includes both invasive [18], [19] as well as noninvasive approaches [6], [20], [21]. Compared to an invasive approach that requires sensor array implantation, non-invasive methods are less risky and more accessible. Among non-invasive techniques, EEG offers higher temporal resolution than magnetoencephalography (MEG) [6] and functional magnetic resonance imaging (fMRI) [21], making it particularly suitable for linguistic applications. Consequently, our research focuses on language decoding from EEG. Due to the underlying neural processes involved in speech production, pioneers mainly focus on decoding subword units [22], [23], [24]. For instance, [25] proposed to extract auto-regressive coefficients as features for imagined syllable classification with a k-nearest neighbor (KNN) classifier. [26] leverages the Hilbert transform to extract features and classify the syllables using a Bayesian classifier. To decode higher-level semantics, numerous studies have dedicated efforts to word-level classification using EEG signals [27], [28], [29], [30], [31]. For instance, [32] have evaluated various convolutional neural network (CNN) architectures for decoding imagined speech from EEG, showcasing the potential of deep learning approaches in improving classification accuracy. However, most of these studies have trained and evaluated their models on a dataset comprising a vocabulary of only 4 to 10 words, which can be insufficient for conveying daily communication [33]. To decode EEG into sentences with a larger vocabulary size, [34] proposes a novel Adaptive Graph Attention Convolutional Network (AGACN) to decode sentences or phrases from EEG signals, achieving high classification accuracy and demonstrating the feasibility of decoding silent reading with complex semantics from EEG signals. More recently, EEG decoding methods have predominantly employed end-to-end generative approaches leveraging large language models as decoders. For instance, EEG-to-Text [13] pioneered open-vocabulary decoding of EEG signals into sentences, establishing an initial performance benchmark while DeWave [35] advanced decoding performance by performing raw wave decoding. Different from their works, we focus on improving the EEG encoder architecture and leveraging various language supervision strategies, including word-level, sentence-level, and context-level strategy, to guide the training of the EEG encoder to achieve better performance in various language decoding tasks.

B. Learning Representations by Natural Language Supervision

Training effective representations for EEG signals is critical to achieving high decoding performance. Recent deep learning methods have shown that language modalities can guide the training of semantically aligned multimodal representations, as demonstrated in visual-language [15], [36], video-language [37], and audio-language [38], [39], [40]. Unlike conventional end-to-end supervised learning, the use of natural language supervision introduces additional semantic information and zero-shot generalization capacity to the representation space of non-language modalities. Current methods for leveraging natural language supervision often involve jointly training both the non-text encoder and the

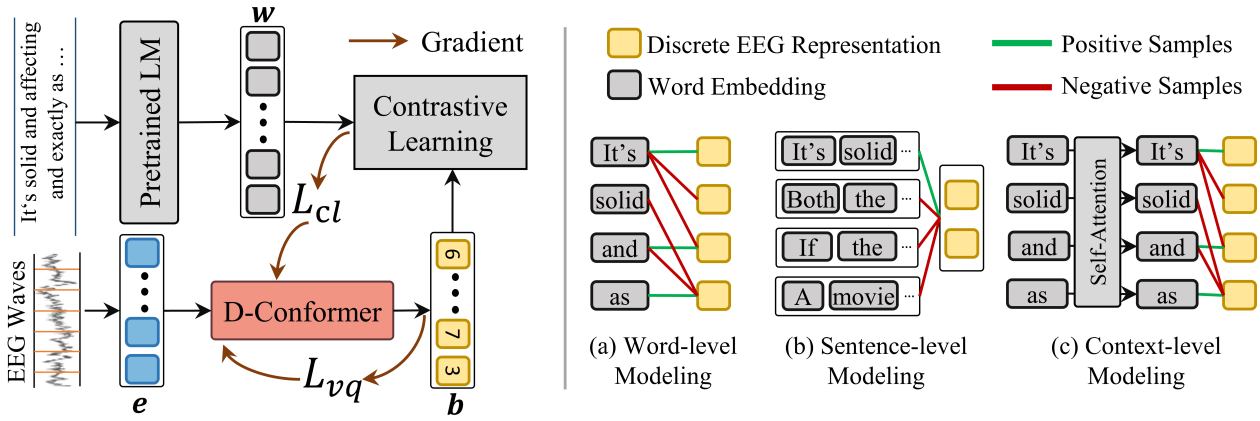

Fig. 2. The overall framework of the proposed approach. After segmenting and applying frequency-domain transform in the preprocessing step, we obtain a sequence of frequency-domain word-level EEG embedding **e** for each word. Then, we use a D-Conformer to encode **e** into discrete EEG representations **b**. These discrete EEG representations will be used as inputs to a subsequent decoder model as conditions for language decoding. To learn semantic EEG representation, we bootstrap the learning of language-aligned EEG representations by training the model using a contrastive objective between **b** and the language representation **w**. A pre-trained LM is used to generate these language representations. A total of three strategies are designed to bootstrap the learning of language-aligned EEG representations for different tasks. Including the (a) word-level strategy where we sample positive and negative word representations for each word-level EEG representation to provide dense supervision information, (b) sentence-level strategy where we only provide sentence-level supervision to the whole EEG sequence, and (c) context-level strategy where we provide word-level as well as context information supervision for each EEG representation.

text encoder [15], [37], [41]. For instance, CLIP [15] jointly trains a text encoder and an image encoder using contrastive learning between images and captions. Similarly, VideoCLIP [42] trains encoders for video and text modalities using a contrastive objective between video frames and their descriptions. However, these methods require large-scale multimodal data pairs to train both encoders from scratch, leading to high training costs. To address this, another branch of research leverages frozen pretrained unimodal models and performs cross-modal alignment instead of training both encoders from scratch [36]. In the field of EEG-to-language decoding, the incorporation of language guidance at the beginning of training an EEG encoder remains unexplored. The absence of a pretrained EEG encoder on a large dataset also hinders the adoption of this approach. Furthermore, the most effective modeling strategies for aligning EEG and language modalities are unclear. To address this research gap, we propose training an EEG encoder using guidance from a pretrained text encoder. To determine the optimal modeling methods, we design and compare sentence-level and context-level modeling strategies with word-level strategies, investigating their impact on subsequent decoding tasks.

III. Method

The proposed BELT method comprises the D-Conformer for EEG encoding and a bootstrapping scheme for training the D-Conformer. The overall framework of BELT is illustrated in Fig. 2. After preprocessing, the D-Conformer encodes the EEG signals into discrete representations, each corresponding to the brain dynamics for a word. To bootstrap the training of semantic EEG representations, we use a pretrained language model to provide supervision through a contrastive learning objective. To optimize performance across different decoding tasks, we designed various bootstrapping strategies leveraging word-level, sequence-level, and context-level supervision.

The remainder of this section is organized as follows: Section III-A introduces the preprocessing steps crucial for converting raw EEG signals into the proper input format for the D-Conformer. Section III-B provides a detailed description of the D-Conformer encoder, including its Conformer building blocks and the vector quantizer used to discretize the EEG representation. Section III-C presents our bootstrapped training scheme. Lastly, Section III-D explains how we applied the D-Conformer model to various decoding tasks and details the final training objectives.

A. EEG Signal Preprocessing

In the preprocessing step, the EEG signals are transformed into word-level embeddings using frequency-domain transformation. First, the EEG recordings are segmented according to the eye-tracking fixation on each word. Following the preprocessing pipeline in previous works [13], [43], the segmented EEG signals are band-pass filtered into eight frequency bands: theta1 (4-6Hz), theta2 (6.5-8Hz), alpha1 (8.5-10Hz), alpha2 (10.5-13Hz), beta1 (13.5-18Hz), beta2 (18.5-30Hz), gamma1 (30.5-40Hz), and gamma2 (40-49.5Hz). The Hilbert transform is then applied to each channel. Finally, word-level EEG embeddings are obtained by averaging the frequency band power within each frequency band. In the remainder of this paper, we denote the word-level EEG embedding as **e** and the corresponding word as **w**.

B. Discrete Conformer for EEG Encoding

After preprocessing the EEG data into word-level embeddings, we introduce the D-Conformer EEG encoder to extract discrete representations. The proposed D-Conformer consists of a number of Conformer blocks and a vector quantizer. They are explained in Sections III-B.1 and III-B.2 respectively.

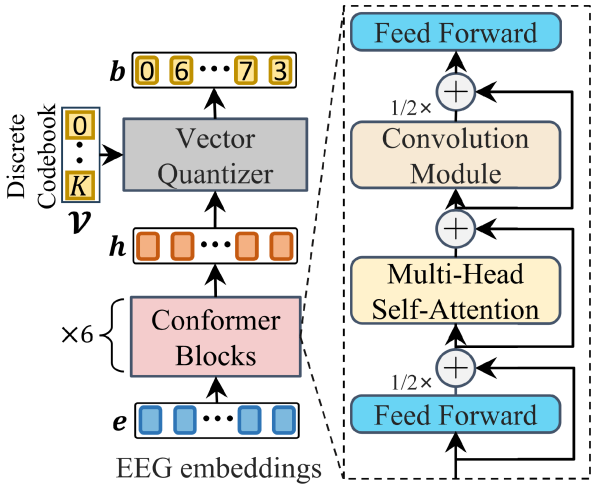

Fig. 3. The detailed structure of the proposed D-Conformer architecture for EEG encoding. The D-Conformer encoder uses EEG embeddings *e* as input and outputs discrete EEG representations *b*. Our D-Conformer is comprised of 6 Conformer blocks and a vector quantizer. Each Conformer block contains a convolution module for exploiting local patterns within the EEG embeddings and a multi-head attention layer for exploiting global information from all input EEG embeddings. The outputs of the Conformer model are fed to a vector quantizer where each continuous EEG representation *h* is replaced by a discrete codebook code from the discrete codebook *V*. Finally, we obtain the discrete EEG representation *b* for each word from the D-Conformer.

Fig. 4. The detailed structure of the convolution module used in the conformer blocks.

1) Conformer Block: A critical step in EEG encoding is handling the input signal's multi-channel characteristics and temporal dynamics. Thus, exploiting the local patterns within channels or the change of patterns when reading through a sentence is crucial for effective pattern extraction in linguistic tasks [44], [45]. However, traditional transformer models lack the mechanisms to effectively capture the local patterns [46]. To overcome this limitation, we proposed to utilize a Conformer block in our encoder to extract local patterns within each EEG embedding as well as the contextual information among the EEG embeddings in a sentence simultaneously.

As depicted in Fig. 3, our D-Conformer is comprised of six conformer blocks. Each conformer block contains four modules including a feed-forward layer, a convolution module, a multi-head self-attention layer, and another feed-forward layer. The convolution module is depicted in Fig. 4, which is in turn comprised of two pointwise convolution layers and a depthwise convolution layer. Detailed configuration of the convolution module is listed in Table I. The first pointwise convolution layer of the convolution module uses the gated linear unit (GLU) as the activation function. A batch normalization layer and a swish activation function were also used after the depthwise convolution layer. Overall, the Comformer blocks take the EEG embeddings $\mathbf{e}$ as input and output the continuous EEG representation $\mathbf{h}$.

TABLE I
DETAILED CONFIGURATION OF THE CONVOLUTION MODULE

Layer	Kernel	Stride	In Channel	Out Channel
Layer Norm	-	-	840	840
Pointwise Conv.	1	1	840	2×840
Depthwise Conv.	31	1	840	840
Batch Norm	-	-	840	840
Pointwise Conv.	1	1	840	840
Dropout	-	-	-	-

2) Vector Quantizer: To achieve representations that better conserve the important features of the word-level EEG segmentations, we further quantize the EEG representations into discrete codes using a vector quantizer. After word-level segmentation, each EEG embedding has been associated with a unique language symbol (e.g., a word). Hence, it can be a more natural fit to represent these EEG symbols in discrete representations [47]. Specifically, a vector quantizer $z_e(\mathbf{h})$ is added after the conformer blocks to map each word-level EEG representation $\mathbf{h}$ into a discrete code $\mathbf{b}$ by finding the nearest discrete element $\mathbf{v}$ from a codebook $\mathbf{V} \in \mathbb{R}^{K \times D}$. The codebook $\mathbf{V}$ contains K discrete embeddings $\{\mathbf{v}_1, \cdots, \mathbf{v}_K\}$, with each embedding being a vector of size D. A fully connected layer is used to adjust the size of $\mathbf{h}$ to D for calculating and comparing the distances between $\mathbf{v}_i$ and $\mathbf{h}$. The codebook of the vector quantizer is randomly initialized when building the D-Conformer. Mathematically, the vector quantizer finds the nearest discrete code for each EEG representation by applying the following nearest neighbor lookup algorithm:

$$\mathbf{b} = \mathbf{z}_e(\mathbf{h}) = \mathbf{v}_k, k = \arg\min_j \|\mathbf{h} - \mathbf{v}_j\|_2^2, \tag{1}$$

where $\|\mathbf{h}_j - \mathbf{v}_j\|_2^2$ denotes the Euclid distance between $\mathbf{h}$ and a codebook embedding $\mathbf{v}_j$. We use L_{vq} (Equation 2) to train the discrete codebook $\mathbf{V}$. The $\mathcal{L}_{vq}$ comprises two terms. The first term is the codebook loss for updating the codebook $\mathcal{V}$ and the second term is the commitment loss to keep the output $z_e(\mathbf{h})$ close to input $\mathbf{h}$.

$$\mathcal{L}_{vq} = \|sg\left[z_e(\mathbf{h})\right] - \mathbf{v}\|_2^2 + \beta \left\|z_e(\mathbf{h}) - sg[\mathbf{v}]\right\|_2^2, \tag{2}$$

where $sg\left[\cdot\right]$ denotes the stop-gradient operation for the straight-through gradient estimation process [48]. During the forward pass, $sg\left[\cdot\right]$ is equivalent to an identical function and passes zero partial gradients, constraining its operand to be a non-updated constant. Coefficient β is a coefficient that controls the impact of the commitment term, we set β to 0.3 in our experiments. This commitment term helps constrain the EEG representations $\mathbf{h}$ from the Conformer model to be compatible with the discrete codes from the codebook. Due to the non-stationarity nature of EEG signals, quantizing the EEG representation could reduce the variations by replacing the EEG representation $\mathbf{h}$ with its discrete counterpart $\mathbf{b}$ and consequently increase the EEG encoder's robustness against subject-specific noise and perturbations while conserving key information from the EEG embeddings.

C. Bootstrapped EEG-to-Language Training for D-Conformer

Nonetheless, the D-Conformer alone cannot guarantee the extraction of semantic EEG representations without an appropriate learning method. To address this, we employ EEG-language alignment to bootstrap the learning of the D-Conformer. Considering that different decoding tasks may emphasize different types of information, we designed strategies to use word-level, sequence-level, and context-level language representations to adapt our method to various tasks. Specifically, we utilize a pretrained Bart model [17] to provide these language representations for corresponding sentences, which are then aligned with the D-Conformer's outputs. This alignment process allows the D-Conformer to extract semantic information from EEG signals.

1) Word-Level Bootstrapping Strategy: For tasks that focus on decoding precision, we consider providing dense language guidance on the word level. In this strategy, we align the discrete EEG representations $\mathbf{b}$ to the word embedding $\mathbf{w}$. The sampling method between the word embeddings and the EEG representations is illustrated in Fig. 2.(a), where we sample the corresponding word as positives and others as negatives. Let $\mathcal{M} = \{\mathbf{w}_1, \cdots, \mathbf{w}_n; \mathbf{b}_1, \cdots, \mathbf{b}_n\}$ represent a mini-batch containing n EEG-word representation pairs, we will use the following contrastive term $\mathcal{L}_{cl}^{w}$ during training:

$$\mathcal{L}_{cl}^{w} = \underset{i \le n}{E}\left[-log\frac{f(\mathbf{b}_i, \mathbf{w}_i)}{f(\mathbf{b}_i, \mathbf{w}_i) + \sum_{i \neq j} f(\mathbf{b}_i, \mathbf{w}_j)}\right]$$
$$f(\mathbf{b}_i, \mathbf{w}_j) = \exp(f_e(\mathbf{b}_i)^T f_w(\mathbf{w}_j))/\tau, \quad (3)$$

where f_e and f_w are linear layers that align the input dimensions for the discrete EEG representation and the word embedding, respectively. τ is a temperature hyperparameter. We apply masking to words outside the vocabulary set of the language model, as well as for padded elements in the input sequences so that they do not affect the training process.

2) Sentence-Level Bootstrapping Strategy: Unlike the word-level strategy, which enhances precision in decoding tasks by focusing on individual words, sentence-level representations from an LM emphasize the topical information of the entire sentence. This approach is particularly beneficial for tasks such as sentiment classification from EEG signals. To generate sentence-level EEG representations, we add a global pooling layer to the outputs of both the D-Conformer and the word embeddings, thereby obtaining sequence-level representations from both EEG and text modalities, as demonstrated in previous works [37], [41], [49], [50]. We denote the sentence-level EEG representation and the sentence representation after pooling the word-level representations as $\bar{\mathbf{b}}$ and $\bar{\mathbf{w}}$. As depicted in Fig. 2.(b), we treat all sentences other than the ground truth as negative samples. Thus, the contrastive term for sentence-level bootstrapping $\mathcal{L}_{cl}^{s}$ can be expressed as follows:

$$\mathcal{L}_{cl}^{s} = \underset{i \le n}{E}\left[-log\frac{f(\bar{\mathbf{b}}_i, \bar{\mathbf{w}}_i)}{f(\bar{\mathbf{b}}_i, \bar{\mathbf{w}}_i) + \sum_{i \neq j} f(\bar{\mathbf{b}}_i, \bar{\mathbf{w}}_j)}\right], \quad (4)$$

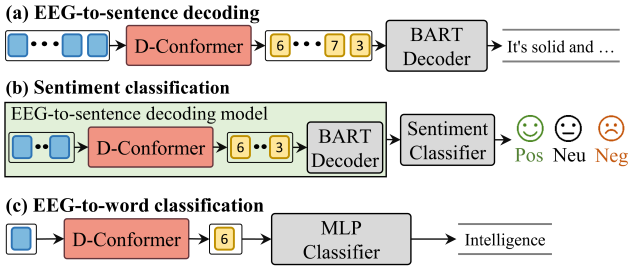

Fig. 5. Various decoding tasks from EEG signals using the proposed D-Conformer.

where i, and j are indexes sampled from a mini-batch containing n pairs of EEG sequences and sentences.

3) Context-Level Bootstrapping Strategy: The context-level strategy is designed to introduce guidance from the deeper representation of a pretrained LM, mirroring how humans gradually grasp sentence meaning by assimilating semantic cues from multiple words. Illustrated in Fig. 2.(c), the context-level modeling strategy aligns the EEG encoder's representation space with a specific transformer block from the LM encoder. We denote the context-level word representation as $\mathbf{c} = LM(\mathbf{w})$ which is output by the Transformer layer of a LM. We use the same sampling strategy between the EEG and words in the word-level modeling strategy to obtain the following contrastive term:

$$\mathcal{L}_{cl}^{c} = \underset{i \le n}{E}\left[-log\frac{f(\mathbf{b}_i, \mathbf{c}_i)}{f(\mathbf{b}_i, \mathbf{c}_i) + \sum_{i \neq j} f(\mathbf{b}_i, \mathbf{c}_j)}\right] \quad (5)$$

D. Decoding Tasks and Training Objectives

When combined with different decoders or classifiers, discrete representations from our D-Conformer can be utilized for multiple tasks, including EEG-to-sentence decoding, zero-shot sentiment classification, and EEG-to-word classification. The overall model structure for each task is depicted in Figure 5.

1) EEG-to-sentence Decoding: As depicted in Figure 5(a), we aim to generate the target sentence $\mathcal{S}$ using a sequence of word-level EEG representations $\mathcal{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_L\}$, with L being the maximum length of the input EEG sequence. These EEG inputs will be first encoded into discrete representations by the D-Conformer model and the discrete EEG representations will in turn used as input to a LM decoder. Following the settings proposed in [13] and [35], we use a pretrained BART model [17] as decoder in this task. For training our model, we train the model for end-to-end EEG-to-sentence generation using the machine translation loss $\mathcal{L}_{tr} = -\sum \log p(\mathcal{S}|\mathcal{E})$. Additionally, we employ L_{vq} for training the discrete codebook and L_{cl} to bootstrap the learning of the semantic representation space. The final loss function can be written as follows:

$$\mathcal{L} = \mathcal{L}_{tr} + \alpha\mathcal{L}_{cl}^{r} + \lambda\mathcal{L}_{vq}, \quad (6)$$

where α and λ are coefficients used to control the weighting of the bootstrapping term and the codebook training term.

The bootstrapping strategy is determined by $r \in \{w, s, c\}$, corresponding to word-level, sentence-level, and context-level strategies, respectively.

The performance of the translation task is measured using the bilingual evaluation understudy (BLEU) score [51] and the recall-oriented understudy for gisting evaluation (ROUGE) score [52]. The ROUGE scores are a set of metrics (precision, recall, and F1-score) used to evaluate the unigram performance between the target and generated sentences. On the other hand, the BLEU scores assess the quality of generated text by comparing n-gram matches between the target and the generated sentence. The BLEU-N score is calculated as $BLEU-N = bp \times \exp(\sum_{n=1}^{N} \varsigma_n \log p_n)$, $bp = \exp(1 - \frac{l_t}{l_g})$, where $bp = \exp(1 - \frac{l_t}{l_g})$ is the brevity penalty term, l_t and l_g are the lengths of the generated translation and the closest target translation, respectively, and $\varsigma_n = 1/N$ is the weight assigned to each n-gram precision score. We evaluate BLEU-1,2,3,4 scores in this paper.

2) Zero-shot Sentiment Classification: Building on the previous EEG-to-sentence decoding model, we can perform zero-shot sentiment analysis using a sentiment classifier that has not been trained on the same dataset as the EEG-to-sentence decoding model, as exemplified in [13]. The zero-shot sentiment classification pipeline is depicted in Figure 5(b), where the EEG-to-sentence decoding model and the sentiment classifier are trained individually using different datasets. For the EEG-to-sentence decoding model, we use Equation 6 for training as described in the previous section. For the sentiment classifier, we experiment with pretrained BART [17] and XLNet [53], fine-tuning them using the Stanford Sentiment Treebank (SSTB) dataset [54]. The objective function for training the sentiment classifier is defined as $\mathcal{L}_{ss} = -\sum y \log(p(\hat{y}|\mathcal{S}))$, where $\hat{y}$ denotes the sentiment prediction for an input text sample from the SSTB sentence-sentiment pairs $\langle \mathcal{S}, y \rangle$, and y is the target sentiment label. This sentiment classifier is then used to classify the generated sentences from the EEG-to-sentence decoding model and output a sentiment prediction. To evaluate the model's performance on sentiment classification, we calculate both micro and macro metrics, including accuracy, precision, recall, and F1 scores.

3) EEG-to-word Classification: Unlike the sentence decoding task, where we leverage the power of a pretrained language model as the decoder, we also evaluate the encoder model at the word level to determine whether the proposed encoding architecture can encode more precise information from EEG compared to previous methods. To do this, we select the 500 most frequently occurring words from the sentences in the training, evaluation, and testing splits for training and evaluating word-level classification performance. As depicted in Figure 5.(c), we use a multi-layer perceptron (MLP) classifier to output the probability distribution over the selected vocabulary set using the discrete EEG representations from the D-Conformer encoder. Given the challenging nature of this task, we use top-10 accuracy as our evaluation metric, following [6]. To train our model for classification, we employ a cross-entropy loss, denoted as $\mathcal{L}_{ce} = -\sum z \log p(\hat{z}|\mathbf{e})$, where z and $\hat{z}$ denote the ground truth word and the word prediction, respectively. When incorporating the L_{vq} and L_{cl} terms, the full training loss is written as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{cl}^{r} + \lambda \mathcal{L}_{vq} \tag{7}$$

TABLE II
DATASET STATISTICS FOR EACH DECODING TASK

Decoding Task	Training Samples	Validation Samples	Testing Samples
EEG-to-sentence decoding	10710	1332	1407
Sentiment classification	3609	467	456
EEG-to-word classification	84815	9732	10432

IV. EXPERIMENT

A. Dataset

We use the ZuCo dataset [43], [55] to conduct our experiments and evaluate our proposed model. The ZuCo dataset comprises EEG data recorded during a natural reading task, supplemented by eye-tracking data for word-level segmentation. It includes 105 EEG channels, with EEG waves denoised and filtered into eight frequency bands after segmentation. A more complete preprocessing details can be found in the dataset paper [43]. The dataset features two reading tasks: normal reading (NR) and task-specific reading (TSR). In the NR task, text passages are sourced from online movie reviews and Wikipedia. The TSR task provides ground-truth sentiment labels in three categories: positive, neutral, and negative. For fair comparison with existing methods, we follow the data division approach outlined in [13], splitting the data into training, validation, and testing sets with proportions of approximately 80%, 10%, and 10%, respectively. The distribution of training samples for each tasks is detailed in Table II. In particular, for word-level classification tasks, we select the 500 most frequently occurring words from the sentences. These sentences are obtained from sentence decoding task across each dataset split.

B. Implementation Details

We use a D-Conformer encoder with 6 Conformer blocks in our experiments, each comprising 8 attention heads. The size of the output EEG representations is set to 840. For the vector quantizer, we set the number of discrete codebook embeddings K to 1024 with the codebook embedding size D to be 1024. All models are trained on Nvidia A40 GPUs. During training, we use a learning rate of $5e$-6 and a batch size of 64. For the loss function, we set α to 0.9 and λ to 1.0. We train our model for a total of 60 epochs, with the best model selected based on validation set performance before evaluation on the test set. We use the SGD optimizer [56] for all training.

C. EEG-to-Sentence Decoding Performance

The performance for the EEG-to-sentence decoding task is presented in Table III. We primarily report results using word-level bootstrapping, as it achieves the best performance for this task. A comparison of different bootstrapping strategies will be presented and discussed in Section IV-F.2.

TABLE III
BRAIN-TO-SENTENCE DECODING RESULT

Model	BLEU-N(%)				ROUGE-1(%)		
	N=1	N=2	N=3	N=4	Precision	Recall	F1-Score
EEG-To-Text [13]	40.12	23.18	12.61	6.80	31.70	28.80	30.18
DeWave [35]	41.35	24.15	13.92	8.22	33.71	28.82	30.69
D-Conformer (ours)	**42.31**	**25.26**	**14.81**	**8.73**	**36.06**	**29.86**	**32.57**
w/o word-level bootstrapping	41.57	24.70	14.54	8.51	35.69	29.40	32.14
w/o vector quantier	41.34	24.14	13.90	8.20	33.71	28.80	30.06

TABLE IV
EEG-TO-SENTENCE DECODING EXAMPLES

(1)	target string:	Everything its title **implies**, a standard-**issue** crime drama spat out from the Tinseltown assembly **line**.
	predicted string:	about a **implies** is and movie for**issue**, story. between of the depthsseltown set **line**.
(2)	target string:	The Kid Stays **in the** Picture" **is a** great story, terrifically told **by** the **man who wrote it** but this Cliff Notes edition **is a cheat**.
	predicted string:	The movie"ays **in the** House" **is a** film$_{ld:40}$ film about andally funny **by** a **man who wrote it**$_{ld:91}$. also is Richard version **is a cheat**$_{ld:83}$.
(3)	target string:	Jeb **Bush was born in** Midland, Texas, where his **father was** running an oil drilling **company**.
	predicted string:	**Bush was born in** Newway, Texas, and he **father was**$_{ld:89}$ a a insurance **company** company.
(4)	target string:	When Bush **was** six **years old**, the **family moved to** Houston, Texas.
	predicted string:	he **was** elected **years old**, he **family moved to**$_{ld:97}$ New, Texas.
(5)	target string:	Bush attended **the University of** Texas **at Austin**, **where he** graduated Phi **Beta Kappa** with a Bachelor's **degree** in Latin **American Studies** in 1973, taking only two and **a half years** to **complete** his work, and obtaining generally excellent grades.
	predicted string:	was **the University of** Chicago **at Austin**, **where he** was in **Beta Kappa**$_{ld:78}$ in a **degree** of degree in History **American Studies**. 1968$_{*}$. and a one years **a half years**$_{ld:73}$. **complete**. degree. and was a mediocre grades$_{ld:44}$.
(6)	target string:	At the urging **of his wife**, Columba, a devout Mexican **Catholic**, the Protestant Bush became a Roman **Catholic**.
	predicted string:	the time **of his wife**, hea, he former **Catholic** Catholic, he actor ministerman a **Catholic** Catholic.
(7)	target string:	He is a prominent **member of the** Bush **family**, the younger **brother of President George W. Bush** and the second **son of** *former* **President George** H. **W. Bush** and Barbara **Bush**.
	predicted string:	was a former **member of the** American **family**$_{ld:78}$. and son **brother of President George W. Bush**$_{ld:88}$. the younger **son of** President **President George** W. **W. Bush.** *former* **Bush**$_{ld:72}$.
(8)	target string:	After World **War II**, Kennedy entered **politics** (partly to fill the void of his popular brother, Joseph P. **Kennedy**, Jr., on whom his family had pinned many of **their hopes** but who was killed in the war).
	predicted string:	the **War II**$_{ld:64}$, he was **politics** asasly as serve the gap$_{*}$ left the father father, John **Kennedy**. Kennedy, who.) who the he father had been **their hopes** the hopes). who had assassinated in the$_{ld:60}$ Korean).

[1] **Bold** words indicates excat match and Underline denotes fuzzy match.
[2] *Italics* words indicates match but out of correct grammar order.
[3] We hightlight fuzzy match results based on two criterias: 1) by the levenshtein distance [57] (annotated by the subscript *ld*) between two text sequences, or 2) by semantic similarity (annotated by the subscript $*$).

Overall, our model achieves state-of-the-art BLEU scores of (42.31, 25.26, 14.81, 8.73) and ROUGE-1 precision, recall, and F1-scores of (36.06, 29.86, 32.57). External comparisons show that our model outperforms EEG-to-Text [13] and Dewave [35] on both metrics. Notably, the main differences between these models lie in the design of the EEG encoder. Compared to the EEG-to-Text method, both our method and the Dewave method encode EEG into discrete representations, indicating that encoding EEG signals into discrete codes is more robust to noise than continuous encoding. Additionally, compared to the Dewave method, our Conformer-based encoder further exploits spatial dependencies within EEG inputs, contributing to improved sentence decoding performance.

In Table IV, we present examples of sentence decoding results using the proposed method for qualitative evaluation.

TABLE V
ZERO-SHOT SENTIMENT CLASSIFICATION RESULT ON ZUCO DATASET

Model	Classifier	Micro Acc.(%)	Micro P.(%)	Marco R.(%)	Marco F1(%)
Transformer [13]	Bart_{large}	55.3	62.4	56.5	55.6
D-Conformer_s	Bart_{large}	60.5	57.8	58.3	56.9
D-Conformer_s	**XLNet_{large}**	**69.3**	**68.8**	**68.3**	**68.0**
D-Conformer_w	Bart_{large}	60.0	59.9	57.9	56.5
D-Conformer_w	XLNet_{large}	67.3	66.5	65.7	65.0
D-Conformer_c	Bart_{large}	60.1	62.0	57.9	56.8
D-Conformer_c	XLNet_{large}	63.1	61.4	61.4	60.8

[1] The under script s, w, and x denote the use of the word-level, sentence-level, and context-level bootstrapping strategy during training.
[2] Acc. denotes the accuracy, P. denotes the precision and R. denotes the recall.

TABLE VI
TOP-10 ACCURACY (%) FOR WORD-LEVEL CLASSIFICATION

Encoder	Top-10 Accuracy (%)
EEG-to-Text [13]	20.92
DeWave [35]	24.64
D-Conformer (Ours)	**31.04**
w/o word-level bootstrapping	25.26
w/o vector quantizer	23.82

Fig. 6. Visualization of top-10 word level prediction results.

The examples illustrate semantic similarity by comparing Levenshtein distances between phrases of the target and decoded sentences. Despite the inherent challenges of EEG decoding, our model significantly improves both single-word decoding precision and the semantic similarity of the decoded phrases.

From on the decoded results, we observe that our method is capable of decoding the verbs and nouns that contain the critical information of a sentence. For instance, in sentence (3) “was born in” vs. “was born in” and in sentence (4) “moved to” vs. “move to”, our model correctly decoded the action to be taken in a sequence of EEG signals. This characteristic could be critical in some tele-control applications. In addition, our model also decodes critical concepts such as “Catholic”, “family”, “president”, and “politics”. This suggests that in the future if more training data is available, our proposed model could help convey sophisticated or even philosophical ideas using EEG signals.

When it comes to short phrases, the proposed method tends to decode semantically similar translations. Such as in sentence (7) the second son of former President vs. the younger son of President and in the sentence (8) fill the void vs. serve the gap. We hypothesize this issue is due to two reasons. The first and major reason is that the EEG representation extracted by our model still lacks discriminative power for the subsequent language model to recognize when the EEG signal is collected from a new person or from a different session. Secondly, we hyporheic another reason could be that when reading a sentence, the words instead provoke the reader to paraphrase the words into some meanings or inner sentence that the person is familiar with or can relate to. This process could potentially help the reader understand unfamiliar or complex ideas in the sentence better and quickly. Therefore our model decodes a similar meaning or situation that the reader is actually related to or thinking of instead of the words displayed on the screen.

D. Zero-Shot Sentiment Classification Performance

In addition to the EEG-to-sentence decoding task, we also evaluate the performance of zero-shot sentiment classification. Here, we use the sentence-level bootstrapping strategy for training the D-Conformer model. As can be observed from the quantitative results displayed in Table V, our method substantially outperforms the baseline method. We could also observe that finetuning a larger sentiment classifier (XLNet) has a positive impact on all classification metrics. Overall, we observe that when using the same BART classifier, our method gains a 5.2% improvement in accuracy. When replacing the zero-shot sentiment classifier with the XLNet model, we additionally achieve an +8.8% improvement in classification performance.

E. EEG-to-Word Classification Performance

The sentence and sentiment tasks assess the encoder’s ability to capture high-level information, with the powerful language decoder compensating for any missing details to generate coherent sentences. To more accurately evaluate the encoder’s capability to capture word-specific patterns, we use a simpler MLP classifier for the EEG-to-word task, avoiding the influence of a powerful language model decoder. This approach allows us to directly compare the encoder’s ability in learning word-specific features. Since no existing studies have tackled word-level classification using the Zuco dataset, we implemented the encoders from the EEG-to-Text model [13] and the Dewave model [35] for external comparison with our method. Given the lack of contextual information in the word-level task, we exclusively use the word-level strategy for training the D-Conformer model. Evaluation results, presented in Table VI, indicate that our model predicts the correct word with a top-10 accuracy of 31.04%, outperforming other methods. Fig.6 shows decoding results from the test set using linear probing, demonstrating that our word-level bootstrapping training enhances the semantic richness of EEG representations. For instance, when predicting “intelligence”, our model also identifies “school”, “university”, and “college” as highly probable candidates.

F. Ablations Studies

1) Ablation on Encoder’s Design Components: We evaluate the impact of the proposed encoder improvements on EEG

TABLE VII
IMPACT OF DIFFERENT BOOTSTRAPPING STRATEGIES

Model	BLEU-N (%)				ROUGE-1 (%)		
	N=1	N=2	N=3	N=4	P.	R.	F1
D-Conformer$_w$	**42.31**	**25.26**	**14.81**	**8.73**	**36.06**	**29.86**	**32.57**
D-Conformer$_s$	42.23	24.95	14.29	8.14	36.06	29.82	32.54
D-Conformer$_c$	42.20	24.94	14.46	8.37	36.12	29.80	32.56

[1] The under script s, w, and x denote the use of the word-level, sentence-level, and context-level bootstrapping strategy during training.

decoding tasks using translation and word-level prediction, as shown in Tables III and VI. In both tasks, we compare the performance of our model and its ablated versions. Notably, in Table VI, we observe that word-level classification performance decreases when either language guidance or the vector quantizer is removed during training, indicating that these design choices positively impact classification performance. Additionally, the word-level bootstrapping strategy contributes more to prediction accuracy (+5.78%) compared to the use of the vector quantizer (+1.44%). This highlights the importance of learning a semantic representation space in linguistic EEG decoding tasks. For the sentence decoding task, Table III shows a similar trend, but both the vector quantizer and bootstrapping learning contributes to a smaller increase in the BLEU-1 score compared to the improvement seen in the word-level task.

2) Ablation on Bootstrapping Strategies: The impact of different bootstrapping strategies is illustrated in Table VII for the sentence decoding task and Table V for the sentiment classification task. For sentence decoding, the word-level strategy achieves the highest BLEU scores, likely because it provides more fine-grained and precise information for the EEG encoder. In contrast, context-level word embeddings, which contain both word-level and sequence-level context information, perform worse than the word-level and sequence-level strategies. This may be due to the general context information introducing additional noise to the already noisy EEG signals, making it more challenging to train an effective EEG encoder. For the sentiment classification task, sequence-level strategies yield the best results across all sentiment classifiers. Since sentiment classification considers the entire sentence rather than individual words, the sequence-level supervision enables our model to better capture the overall context and sentiment inclination of the entire sequence.

3) Ablation on Bootstrapping Coefficients: We also investigate the impact of a range of bootstrapping coefficients α ranging from 0.05 to 0.9 by the word-level strategy with different EEG encoders. In Figure 8, we present the BLEU-1 scores for a bootstrapped D-Conformer with or without the use of vector quantization (red and blue curves), as well as a comparison between the Conformer encoder and the Transformer encoder. It is noteworthy that the performance curve of the Transformer encoder (green curve) corresponds to the baseline EEG-to-Text model, as they both use the Transformer architecture for their EEG encoder. We could observe that the increase in bootstrapping coefficient comes

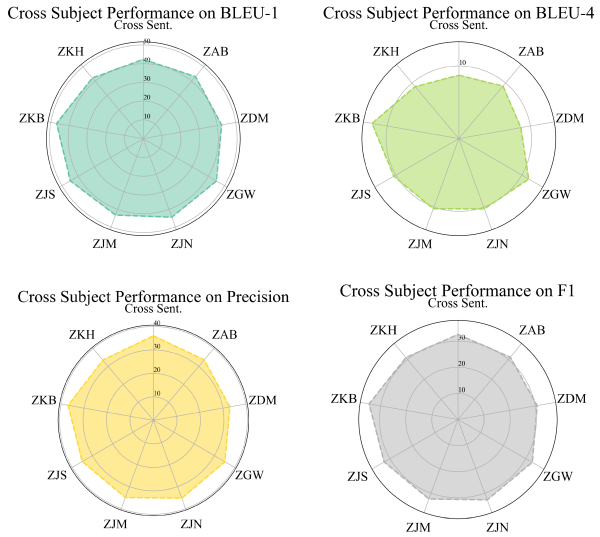

Fig. 7. The cross-subjects performance for translation tasks. Cross Sent. denotes the performance of the cross-sentence setting on each evaluated metric.

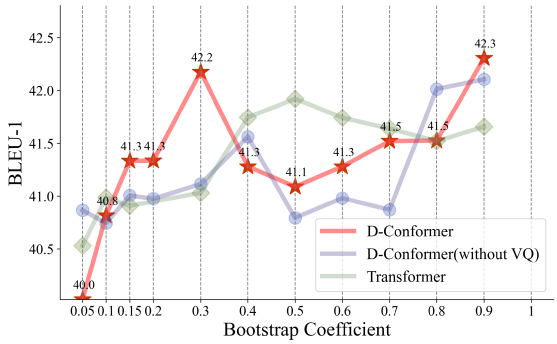

Fig. 8. Comparison of BLEU-1 scores for sentence decoding across varying contrastive coefficients.

with an increase in translation performance from a broad perspective. With the introduction of the conformer block in the D-Conformer (the blue curve) to replace the transformer encoder used in the baseline EEG-to-Text model (the green curve), our method could reach better performance under higher bootstrapping coefficients (0.8 and 0.9). However, we could also observe that the further introduction of the vector quantization method (the red curve) could bring greater sensitivity to the final performance relative to the change of the bootstrapping coefficients.

4) Ablation on Cross-Subject Performance: In this section, we evaluate the performance in the cross-subject setting, which is a vital indicator for application on unseen subjects during training. Unlike the cross-sentence setting as evaluated in Section IV-C, this section evaluates the performance of unseen subjects. Figure 7 shows the cross-subject translation performance for a total of 8 subjects compared to the cross-sentence result we achieved in the cross-sentence setting. The radar charts in Figure 7 denote the performance is stable across different subjects with subjects achieving BLEU-1 scores ranging from 42.25 to 46.90. However, the variant in longer-gram BLEU-4 score is larger among subjects ranging from

9.34 to 12.11. This difference is mainly due to the word-level strategy we used in the enhancement of single-word decoding precision.

V. Conclusion and Future Work

In this paper, we present BELT, which consists of an innovative D-Conformer architecture for encoding EEG into discrete representations and a bootstrapping training method for learning language-aligned EEG representations. Our experiments show that leveraging supervision from natural language is an effective way to facilitate the learning of semantic EEG representations. This is supported by substantial improvements in various EEG decoding tasks, including EEG-to-word classification, EEG-to-sentence decoding, and sentiment classification. The proposed method also encourages more in-depth exploration and discussion of the pivotal topic of decoding thoughts into text, which could potentially lead to numerous new BCI applications. Despite the progress achieved, there is still room for future improvement in terms of translation precision and fluency without the implicit use of teacher-forcing evaluation. In the future, we plan to collect more language-related EEG data to train a more general EEG encoder and tackle the fundamental problem of data scarcity in this research area.

References

[1] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.

[2] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 344–353, Jun. 2020.

[3] G. Zhang, V. Davoodnia, and A. Etemad, "PARSE: Pairwise alignment of representations in semi-supervised EEG learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2185–2200, Oct. 2022.

[4] J. Thomas Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "A novel deep learning architecture for decoding imagined speech from EEG," 2020, *arXiv:2003.09374*.

[5] L. Cao, D. Huang, Y. Zhang, X. Jiang, and Y. Chen, "Brain decoding using fNIRS," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 14, pp. 12602–12611.

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

[7] Y.-E. Lee, S.-H. Lee, S.-H. Kim, and S.-W. Lee, "Towards voice reconstruction from EEG during imagined speech," 2023, *arXiv:2301.07173*.

[8] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition," *Sci. Data*, vol. 9, no. 1, p. 52, Feb. 2022.

[9] B. van der Berg, S. v. Donkelaar, and M. Alimardani, "Inner speech classification using EEG signals: A deep learning approach," in *Proc. IEEE 2nd Int. Conf. Hum.-Mach. Syst. (ICHMS)*, Sep. 2021, pp. 1–4.

[10] J. Berezutskaya et al., "Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film," *Sci. Data*, vol. 9, no. 1, p. 91, Mar. 2022.

[11] Y. V. Varshney and A. Khan, "Imagined speech classification using six phonetically distributed words," *Frontiers Signal Process.*, vol. 2, 2022. [Online]. Available: https://www.frontiersin.org/journals/signal-processing/articles/10.3389/frsip.2022.760643

[12] X. Feng, X. Feng, B. Qin, and T. Liu, "Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3874–3883, 2023.

[13] 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.

[14] M. B. Sariyildiz, J. Perez, and D. Larlus, "Learning visual representations with caption annotations," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 153–170.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[16] H. L. Dawson, O. Dubrule, and C. M. John, "Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification," *Comput. Geosci.*, vol. 171, Feb. 2023, Art. no. 105284.

[17] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[18] 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.

[19] 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.

[20] J. Tang, A. LeBel, S. Jain, and A. G. Huth, "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nature Neurosci.*, vol. 26, no. 5, pp. 858–866, May 2023.

[21] N. Xi, S. Zhao, H. Wang, C. Liu, B. Qin, and T. Liu, "UniCoRN: Unified cognitive signal ReconstructioN bridging cognitive signals and human language," 2023, *arXiv:2307.05355*.

[22] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Proc. 13th Int. Conf. Hum.-Comput. Interact., New Trends, HCI Int.*, San Diego, CA, USA. Cham, Switzerland: Springer, Jul. 2009, pp. 40–48.

[23] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 1311–1316.

[24] M.-O. Tamm, Y. Muhammad, and N. Muhammad, "Classification of vowels from imagined speech with convolutional neural networks," *Computers*, vol. 9, no. 2, p. 46, Jun. 2020.

[25] K. Brigham and B. V. K. V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy," in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2010, pp. 1–4.

[26] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, Aug. 2010, Art. no. 046006.

[27] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 901–908, Nov. 2013.

[28] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, "Sonification and textification: Proposing methods for classifying unspoken words from EEG signals," *Biomed. Signal Process. Control*, vol. 37, pp. 82–91, Aug. 2017.

[29] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 992–996.

[30] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomed. Eng. Lett.*, vol. 10, no. 2, pp. 217–226, May 2020.

[31] M. R. A. Bejestani, G. R. M. Khani, V. R. Nafisi, and F. Darakeh, "EEG-based multiword imagined speech classification for Persian words," *BioMed Res. Int.*, vol. 2022, pp. 1–20, Jan. 2022.

[32] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, Aug. 2020.

[33] 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, Jul. 2021.

[34] C. Li, Y. Liu, J. Li, Y. Miao, J. Liu, and L. Song, "Decoding bilingual EEG signals with complex semantics using adaptive graph attention convolutional network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 249–258, 2024.

[35] Y. Duan, J. Zhou, Z. Wang, Y. K. Wang, and C.-T. Lin, "DeWave: Discrete encoding of EEG waves for EEG to text translation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, A. Oh, T. Naumann, A. G. Son, K. Saenko, M. Hardt, and S. Levine, Eds. Curran Associates, Inc., 2023, pp. 9907–9918. [Online]. Available: https://proceedings.neurips.cc/paper files/paper/2023/file/1f2fd23309a5b2d2537d063b29ec1b52-Paper-Conference.pdf

[36] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[37] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[38] S. Forsgren and H. Martiros. (2022). *Riffusion—Stable Diffusion for Real-Time Music Generation*. [Online]. Available: https://riffusion.com/about

[39] Q. Huang et al., "Noise2Music: Text-conditioned music generation with diffusion models," 2023, *arXiv:2302.03917*.

[40] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 12449–12460.

[41] Y. Zhong et al., "RegionCLIP: Region-based language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16793–16803.

[42] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," 2021, *arXiv:2109.14084*.

[43] 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," *Sci. Data*, vol. 5, no. 1, pp. 1–13, Dec. 2018.

[44] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[45] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2958–2965.

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

[47] A. Van Den Oord et al., "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. Von Luxburg, S . Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

[48] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[49] X. Hu et al., "Scaling up vision-language pretraining for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17959–17968.

[50] K. Zhou, B. Zhang, W. Xin Zhao, and J.-R. Wen, "Debiased contrastive learning of unsupervised sentence representations," 2022, *arXiv:2205.00656*.

[51] 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.

[52] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[53] Z. Yang, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*.

[54] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Oct. 2013, pp. 1631–1642.

[55] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation," 2019, *arXiv:1912.00903*.

[56] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[57] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, Mar. 2001.