frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 26 August 2025
DOI 10.3389/frobt.2025.1625731

Check for updates

OPEN ACCESS

EDITED BY
Farzan M. Noori,
University of Oslo, Norway

REVIEWED BY
Usman Ghafoor,
Institute of Space Technology, Pakistan
Hsiang-Yun Chien,
Apple, United States

*CORRESPONDENCE
Tidiane Camaret Ndir,
✉ tidiane.camaret.ndir@uniklinik-freiburg.de

RECEIVED 09 May 2025
ACCEPTED 11 August 2025
PUBLISHED 26 August 2025

CITATION
Camaret Ndir T, Schirrmeister RT and Ball T (2025) EEG-CLIP: learning EEG representations from natural language descriptions.
Front. Robot. AI 12:1625731.
doi: 10.3389/frobt.2025.1625731

EEG-CLIP: learning EEG representations from natural language descriptions

Tidiane Camaret Ndir[1,2]*, Robin T. Schirrmeister [1,2] and Tonio Ball [2,3]

[1]Medical Physics, Department of Diagnostic and Interventional Radiology, Medical Center—University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany, [2]Neuromedical A.I. Lab, Department of Neurosurgery, Medical Center—University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany, [3]BrainLinks-BrainTools, IMBIT (Institute for Machine-Brain Interfacing Technology), University of Freiburg, Freiburg im Breisgau, Germany

Deep networks for electroencephalogram (EEG) decoding are often only trained to solve one specific task, such as pathology or age decoding. A more general task-agnostic approach is to train deep networks to match a (clinical) EEG recording to its corresponding textual medical report and *vice versa*. This approach was pioneered in the computer vision domain matching images and their text captions and subsequently allowed to do successful zero-shot decoding using textual class prompts. In this work, we follow this approach and develop a contrastive learning framework, EEG-CLIP, that aligns the EEG time series and the descriptions of the corresponding clinical text in a shared embedding space. We investigated its potential for versatile EEG decoding, evaluating performance in a range of few-shot and zero-shot settings. Overall, we show that EEG-CLIP manages to non-trivially align text and EEG representations. Our work presents a promising approach to learn general EEG representations, which could enable easier analyses of diverse decoding questions through zero-shot decoding or training task-specific models from fewer training examples. The code for reproducing our results is available at https://github.com/tidiane-camaret/EEGClip.

KEYWORDS

electroencephalogram (EEG), contrastive learning, multimodal representation, zero-shot classification, clinical text processing, neural time series, transfer learning

1 Introduction

Recent advances in machine learning have led to deep neural networks being commonly applied to electroencephalogram (EEG) data for a variety of decoding tasks (Roy et al., 2019). While deep learning models can achieve state-of-the-art performance on specialized EEG tasks, their learned representations can often only be used for one specific task. Most EEG analyses focus on training task-specific models for one type of classification or regression problem (Heilmeyer et al., 2018). However, medical EEG recordings are often accompanied by rich unstructured annotations in the form of free text reports written by neurologists and medical experts—a potentially valuable source of supervision that remains largely untapped.

In the computer vision domain, Contrastive Language-Image Pre-training or CLIP (Radford et al., 2021) leverages text-image pairing to learn visual representations that effectively transfer across tasks. CLIP has demonstrated remarkable zero-shot generalization capabilities by learning to align images with natural language descriptions, enabling classification of previously unseen categories and adaptation to novel visual tasks without additional training.

Inspired by CLIP, we propose EEG-CLIP: a contrastive learning approach to align EEG time-series data with corresponding clinical text descriptions in a shared embedding space. This work explores two central questions: (i) how clinical text reports can be effectively incorporated into EEG representation learning, and (ii) whether this multimodal approach enables more generalizable representations that transfer across diverse EEG decoding tasks.

We demonstrate EEG-CLIP's potential for versatile EEG decoding through extensive evaluation on few-shot and zero-shot learning tasks. Our results show that EEG-CLIP achieves strong zero-shot classification performance and consistently outperforms previous transfer learning approaches and task-specific models when labeled data are scarce. This presents a promising direction for EEG analysis by enabling zero-shot inference through natural language queries and more efficient training of specialized models with limited annotations.

Remark: Recently and after the completion of the study presented in this manuscript, (Gijsen and Ritter, 2024), also proposed EEG-language models that align EEG data with clinical reports for pathology detection. Their work explores multiple alignment strategies, including a multiple instance learning extension for flexible matching between EEG segments and text portions. Their approach is primarily focused on pathology detection and classification of epileptiform activity, while our EEG-CLIP study examines model performance on diverse decoding objectives including age, gender, and medication prediction, providing further insights into the versatility of language-supervised EEG representations.

2 Related work

2.1 Deep-learning based EEG decoding

Deep learning has revolutionized EEG analysis by enabling end-to-end decoding directly from raw signals without hand-crafted features. Convolutional neural networks (CNNs) have shown particular promise, with recent advances like batch normalization and exponential linear units boosting performance to match or exceed traditional methods like filter bank common spatial patterns (FBCSP) (Schirrmeister et al., 2017) These architectures automatically learn hierarchical representations that capture relevant spectral and spatial patterns in EEG data.

Various neural network architectures have been applied to EEG tasks, from shallow CNNs for efficient processing to recurrent networks for capturing temporal dependencies. Recent comparative studies have demonstrated that specialized deep learning models can outperform traditional approaches on standard benchmarks such as BCI Competition datasets. Beyond classification, newer approaches like EEG-to-text decoding leverage advanced neural architectures combined with probabilistic modeling to translate neural activity into human-readable text (Lévy et al., 2025), expanding the potential applications for EEG-based interfaces.

Multimodal approaches have also shown promise, such as Khan et al. (2018) who combined EEG with fNIRS to capture complementary neurophysiological signals. While they used traditional signal processing rather than deep learning, their work highlights the value of integrating EEG with additional information sources, a principle that motivates our EEG-text alignment approach.

While task-specific models dominate current approaches, our work explores a more general representation learning framework that leverages the rich information in clinical text reports to develop versatile EEG embeddings useful across multiple decoding tasks.

2.2 Contrastive learning for multimodal alignment

Self-supervised contrastive learning has recently emerged as a powerful approach to learning general visual representations. Models like CLIP are trained to align the image embeddings x_i and the corresponding text embeddings y_i by minimizing contrast loss $\mathcal{L}$:

$$\mathcal{L} = \sum_{i=1}^{N} -\log \frac{\exp\left(\text{sim}\left(x_i, y_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(x_i, y_j\right)/\tau\right)}$$

where *sim*(.) is a measure of similarity. This objective brings the matching image-text pairs closer and separates the mismatched pairs in the learned embedding space.

CLIP was trained on a large dataset of 400 million image-text pairs from diverse Internet sources with unstructured annotations. Through this natural language supervision, CLIP developed versatile image representations that achieve strong zero-shot inference on downstream tasks by querying the aligned embedding space.

The success of CLIP highlights the promise of contrastive learning approaches and the use of readily available text data to learn transferable representations of other modalities.

3 Methods

3.1 Dataset

The Temple University Hospital EEG corpus (Obeid and Picone, 2016) contains over 25,000 EEG recordings collected from over 14,000 patients between 2002 and 2015. The large number of EEG recordings make this a valuable training dataset for deep learning models to learn to decode information such as pathology or age from the EEG and be able to generalize to unseen EEG recordings. The TUH Abnormal dataset (TUAB) is a demographically balanced subset with binary labels indicating pathological or nonpathological diagnosis of each recording. It is partitioned into training (1,387 normal and 1,398 abnormal files) and evaluation (150 normal and 130 abnormal files) sets. It contains a variety of pathological conditions.

TABLE 1 Report section characteristics in the TUAB dataset. Distribution of medical report sections showing highest coverage in diagnostic sections (Impression, Description, Clinical History) with substantial text content, versus limited coverage in specialized sections.

Record section	Count of non-empty entries	Average word count
Impression	2,971	16
Description of the record	2,964	70
Clinical history	2,947	26
Medications	2,893	4
Introduction	2,840	31
Clinical correlation	2,698	31
Heart rate	1,458	2
Findings	887	16
Reason for study	713	2
Technical difficulties	684	3
Events	569	8
Condition of the recording	116	30
Past medical history	19	8
Type of study	16	3
Activation procedures	9	3

Each recording contains additional labels: “age” (integer), “gender” (“M” or “F”), and “report” (string), a medical report written in natural language. The report is divided in 15 sections, listed in Table 1.

3.2 EEG data preprocessing

We preprocess the EEG data, following the preprocessing steps from Schirrmeister et al. (2017):

- Select a subset of 21 electrodes present in all recordings.
- Exclude the first minute of the recordings, and only use the first 2 min after that
- Clip the amplitude values to the range of ± 800 μ V to reduce the effects of strong artifacts.
- Resample the data to 100 Hz to further speed up the computation.
- Divide by 30 to get closer to unit variance

3.3 Architecture and training details

The EEG-CLIP model is composed of two main components: an EEG encoder and a text encoder. These encoders are designed to process EEG recordings and medical reports respectively, as depicted in Figure 1.

For the EEG encoder, we use a convolutional neural network (CNN), Deep4 (Schirrmeister et al., 2017), whose architecture is optimized for the classification of EEG data. The Deep4 Network features four convolution-max-pooling blocks, using batch normalization and dropout, followed by a dense softmax classification layer. This enables the model to learn hierarchical spatial-temporal representations of the EEG signal. The output is flattened and passed to a fully-connected layer to derive a 128-dimensional embedding.

For the text encoder, we leverage pretrained text encoders based on the BERT architecture (Devlin et al., 2019). Such transformer-based models have shown state-of-the-art performance on a variety of natural language processing tasks. The advantage of these pretrained models is that they provide rich linguistic representations that can be effectively transferred to downstream tasks through finetuning.

The EEG and text embeddings are then fed into MLP projection heads, consisting of 3 fully-connected layers with ReLU activations. The final layer outputs a 64-dimensional projection of the embedding for contrastive learning. This architecture allows the model to learn alignments between EEG windows and corresponding medical report sentences in a shared embedding space. The contrastive loss enables the useful semantic features to be captured.

We train EEG-CLIP using the Adam optimizer with a learning rate of 5×10^{-3} and weight decay of 5×10^{-4}. The model is trained for 20 epochs with a batch size of 64. We use the same training/testing split as in the TUAB dataset. Each recording is split in windows of length 1,200, corresponding to a 12 s period, and with a stride of 519, which ensures all timesteps are predicted without any gap by our Deep4 model.

3.4 Evaluation methods

Unlike models trained for a specific downstream task, EEG-CLIP has to learn broadly useful representations that capture semantic relationships between EEG signals and text. As such, evaluation methods must aim to quantify the general quality and transferability of the learned representations.

Using the labels and medical reports provided in the TUAB dataset, we select 4 decoding tasks:

- “Pathological”: decode whether the recording was diagnosed as normal or pathological
- “Age”: decode whether the age of the patient is smaller or equal, or greater than 50
- “Gender”: decode the declared gender of the patient
- “Medication”: decode whether the medical report contains at least one of the 3 most common anticonvulsant medications (“keppra”, “dilantin” and “depakote”)

We then design multiple methods to evaluate the model on, listed in the following.

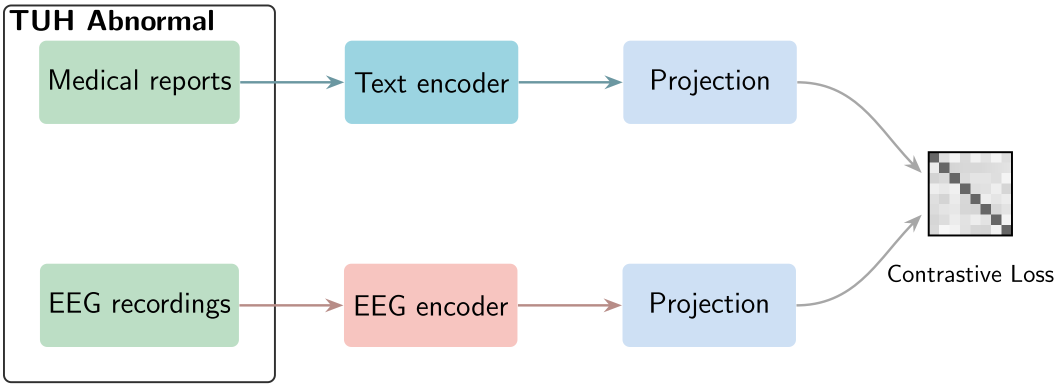

FIGURE 1
Model architecture of EEG-CLIP. The figure illustrates the dual-encoder architecture with an EEG encoder processing EEG time series data and a text encoder (pretrained BERT-based model) processing clinical reports. Both modalities are projected into a shared 64-dimensional embedding space through MLP projection heads. The contrastive loss optimizes for alignment between matching EEG-text pairs while pushing non-matching pairs apart, enabling the model to learn cross-modal representations that capture semantic relationships between neurophysiological patterns and their clinical descriptions.

3.4.1 Classification

We compare EEG-CLIP against two baseline models to contextualize its performance:

- Task-specific model (upper bound): A Deep4 CNN trained end-to-end from random initialization directly on each target task. This provides an upper bound since the entire model can optimize specifically for the task.
- Alternative task transfer model (lower bound): A Deep4 CNN first trained from random initialization on an unrelated task (e.g., age classification), then transferred to the target task (e.g., pathology detection) by freezing the encoder and training only a new classification head. This tests whether simple transfer learning from any EEG task provides useful features.

Both baselines use the same architecture as EEG-CLIP's EEG encoder but differ in their training approach: EEG-CLIP uses contrastive learning with text supervision, while baselines use standard supervised learning with task labels. Figure 2 illustrates the different training strategies.

3.4.2 Zero-shot classification

We also perform zero-shot evaluation, using the embeddings of class-specific text prompts as class prototypes for the trained EEG-CLIP model. For a given classification task, we define a typical prompt sentence for each class (see Table 2) and calculate the distance of an EEG recording to those sentences in the shared embedding space. This allows us to measure the classification performance of EEG-CLIP without any training on the classification task labels.

3.4.3 Classification in a low-data regime

To further evaluate the generalization capability of the learned representations, we assess few-shot performance by training classifiers on varying fractions of a small labeled dataset. Specifically:

- We hold out 20% of the TUAB training set exclusively for few-shot training (never seen during EEG-CLIP's contrastive training)
- From this 20% held-out set, we create subsets of sizes: 1/2, 1/5, 1/10, 1/20, and 1/50

All models (EEG-CLIP, task-specific, and alternative task) are trained on these identical data subsets. The key difference is that EEG-CLIP uses representations learned from contrastive training on the separate 60% split, while the task-specific model trains from scratch. This ensures a fair comparison where all models have access to the same limited labeled data, isolating the benefit of pre-training. Figure 3 illustrates the data partitioning strategy.

4 Results

4.1 Evaluation of the learned representations

In this section, we present results evaluating the learned representations from EEG-CLIP across a diverse set of experiments. As a reminder, our evaluation methodology consisted of classification tasks using the full TUAB dataset, zero-shot classification using text prompts, and few-shot classification on a held-out dataset.

4.1.1 Classification performance

Table 3 shows EEG-CLIP's classification performance across four tasks. With logistic regression, EEG-CLIP achieves balanced accuracies of 0.826 for pathological status, 0.713 for age, and 0.687 for gender. A 3-layer MLP classifier further improves results to 0.847, 0.747, and 0.702 respectively, indicating non-linear relationships in the embedding space. The performance gap between EEG-CLIP + MLP and task-specific models remains small (0.004 for pathological, 0.039 for age, 0.050 for gender) despite the latter's end-to-end

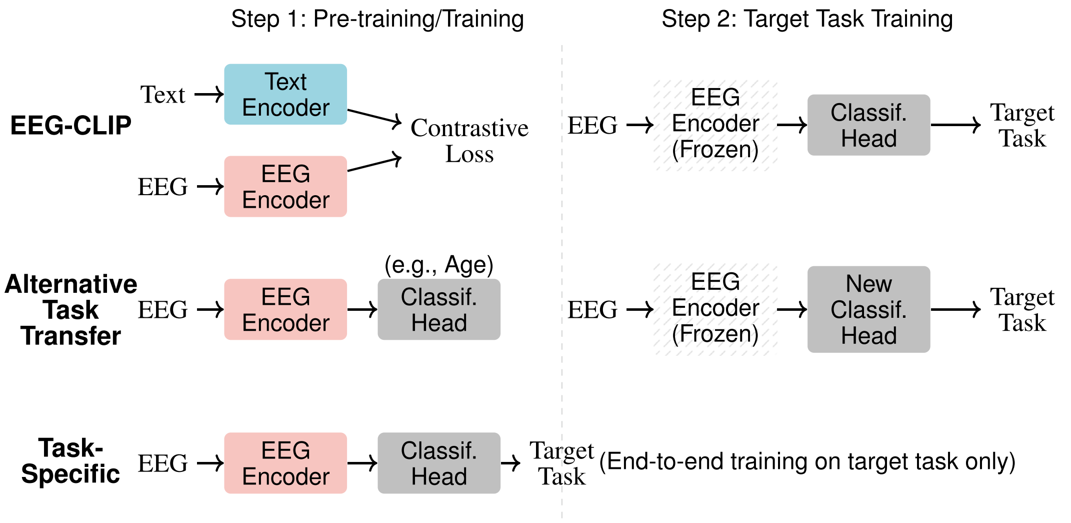

FIGURE 2
Training approaches comparison. Three methods for EEG classification: (1) EEG-CLIP uses contrastive pre-training with text before target task fine-tuning with frozen encoder, (2) Alternative Task Transfer pre-trains on an unrelated task then adapts to target task with frozen encoder, (3) Task-Specific trains directly on target task end-to-end.

TABLE 2 Text prompts for zero-shot classification. Concise natural language prompts representing each class for four classification tasks, enabling classification through EEG-text similarity without task-specific training.

Task	Prompt A	Prompt B
Pathological	"This is a normal recording"	"This is an abnormal recording"
Age	"The patient is under 50 years old"	"The patient is over 50 years old"
Gender	"The patient is male"	"The patient is female"
Medication	"No anti-epileptic drugs were prescribed to the patient"	"Anti-epileptic drugs were prescribed to the patient"

optimization advantage. Most importantly, EEG-CLIP consistently outperforms alternative task pretraining by 10.6% for pathological, 6.2% for age, and 3.5% for gender classification. These quantitative results demonstrate that text-EEG contrastive learning produces more transferable representations than single-task supervised learning.

4.1.2 Zero-shot classification performance

For zero-shot classification, we evaluate EEG-CLIP's ability to classify EEG recordings without any task-specific training. We compute similarities between EEG embeddings and text prompts (see Table 2) in the shared embedding space. As shown in Table 4, EEG-CLIP achieves remarkable zero-shot performance on the pathological task (0.755), demonstrating strong alignment between EEG signals and their textual descriptions. Performance on age classification (0.642) is also substantially above chance, while gender (0.567) and medication (0.532) show more modest scores. These results are particularly encouraging as they represent classification without any labeled training data, relying solely on the semantic alignment learned during contrastive training. The strong pathology detection performance suggests that diagnostic language in the medical reports is effectively aligned with corresponding neurological patterns in the EEG signals. This zero-shot capability could be especially valuable in clinical settings where labeled data for new tasks is scarce or unavailable.

4.1.3 Few-shot classification performance

On the pathological task, EEG-CLIP achieves 0.710 balanced accuracy on the held-out set. This approaches the 0.781 performance of a model trained from scratch with the same limited data. For age classification, EEG-CLIP even outperforms the specialized model. The medication task proves most challenging in the few-shot setting. However, all models struggle to exceed 0.6 accuracy, suggesting intrinsic difficulty of the binary prediction from small samples. The detailed results are presented in Table 5.

Critically, EEG-CLIP substantially outperforms models pretrained on alternative tasks across all but one experiment. This demonstrates the concrete value of pretraining on aligned data, even when fine-tuning data is scarce.

As shown in Figure 4, EEG-CLIP (green lines) maintains relatively stable performance across increasingly smaller fractions of the training set, from $\frac{1}{2}$ down to $\frac{1}{50}$ of the

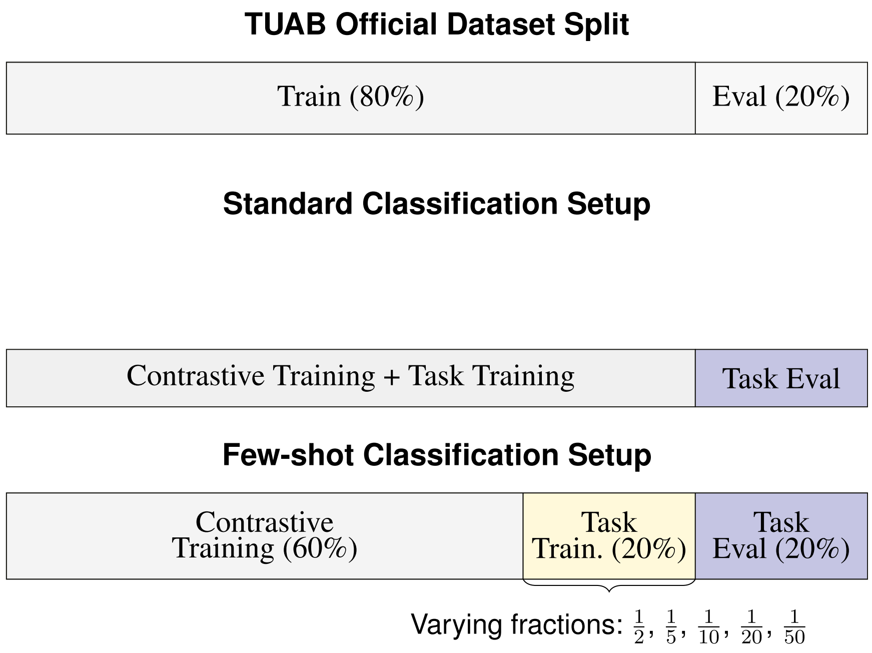
TUAB Official Dataset Split
Train (80%)
Eval (20%)
Standard Classification Setup
Contrastive Training + Task Training
Task Eval
Few-shot Classification Setup
Contrastive Training (60%)
Task Train. (20%)
Task Eval (20%)
Varying fractions: $\frac{1}{2}$, $\frac{1}{5}$, $\frac{1}{10}$, $\frac{1}{20}$, $\frac{1}{50}$

FIGURE 3
Experimental data partitioning strategies for EEG-CLIP. The top section shows the official TUAB dataset split. The middle section illustrates the standard classification setup where the training portion is used for contrastive learning between EEG signals and text descriptions. The bottom section visualizes the few-shot learning approach: 60% is used for EEG-CLIP's contrastive pre-training (without task labels), 20% serves as the few-shot training set (from which varying fractions are sampled), and 20% is held for evaluation. In few-shot experiments, all compared models use only the same subsets from the 20% Task Train split.

TABLE 3 Classification performance comparison (balanced accuracy). EEG-CLIP approaches task-specific performance while substantially outperforming alternative task pretraining, demonstrating effective text-supervised representation learning.

Task	EEG-CLIP		Task-specific	Alternative task
	LogReg	MLP		
Pathological	0.826	0.847	**0.851**	0.741 *(age)*
Age	0.713	0.747	**0.786**	0.685 *(pathological)*
Gender	0.687	0.702	**0.752**	0.667 *(pathological)*
Medication	0.633	0.615	**0.685**	0.573 *(pathological)*

Bold values indicate the best performance for each task.

original dataset. For pathology detection (top left), EEG-CLIP maintains strong performance even with minimal data ($\frac{1}{20}$), outperforming both baselines as data becomes extremely scarce. Age classification (top right) shows EEG-CLIP consistently outperforming other approaches across all data regimes. For gender and medication tasks (bottom panels), all models show performance degradation with reduced data, but EEG-CLIP demonstrates greater robustness to extreme data reductions ($\frac{1}{50}$).

TABLE 4 Zero-shot classification performance (balanced accuracy). EEG-CLIP achieves strong performance for pathology detection (0.755) and age classification (0.642) using only text prompts, demonstrating effective EEG-text alignment.

Task	Accuracy
Pathological	0.755
Age	0.642
Gender	0.567
Medication	0.532

Taken together, these quantitative results provide strong evidence for the quality and transferability of the multi-modal representations learned by EEG-CLIP. Performance across the range of evaluation paradigms demonstrates that EEG-CLIP successfully encodes general semantic relationships between EEG and text. This enables the model to generalize to new tasks and datasets without task-specific fine-tuning. The recent ELM-MIL approach by Gijsen and Ritter (2024) achieves superior performance (87.11% balanced accuracy) through Multiple Instance Learning extensions that address fine-grained EEG-text alignment. While our approach achieves 84.7%, both works demonstrate that multimodal language supervision significantly outperforms EEG-only baselines, validating this research direction.

TABLE 5 Few-shot learning performance (balanced accuracy). EEG-CLIP outperforms models trained from scratch on age classification, demonstrating representation transferability when labeled data is scarce.

Task	EEG-CLIP + MLP	Task-specific	Alternative task
Pathological	0.710	**0.781**	0.531 *(age)*
Age	**0.712**	0.621	0.631 *(pathological)*
Gender	0.550	**0.648**	0.512 *(pathological)*
Medication	0.551	0.575	**0.598** *(pathological)*

Bold values indicate the best performance for each task.

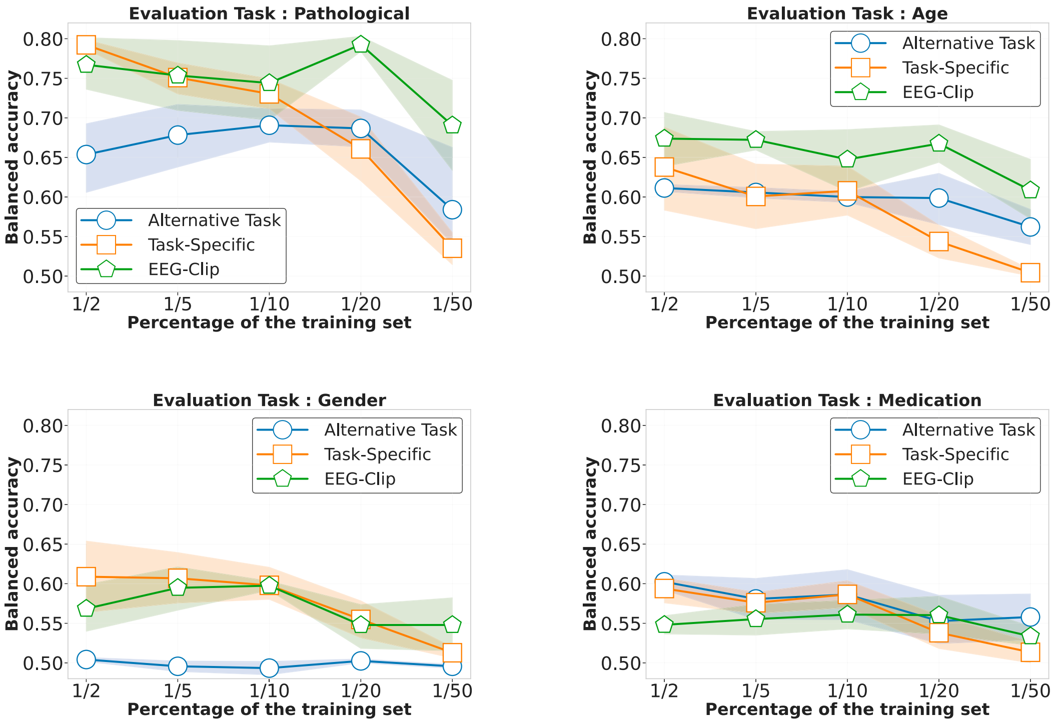

FIGURE 4
Performance comparison in low-data regimes across four decoding tasks. Each panel shows balanced accuracy as a function of training set size (from $\frac{1}{2}$ to $\frac{1}{50}$ of the full dataset) for three approaches: EEG-CLIP (green), task-specific models trained from scratch (orange), and models pretrained on alternative tasks (blue). For pathology detection (top left), EEG-CLIP maintains strong performance even with minimal data ($\frac{1}{20}$), outperforming both baselines as data becomes scarce. Age classification (top right) shows EEG-CLIP consistently outperforming other approaches across all data regimes. For gender (bottom left) and medication (bottom right) tasks, all models show performance degradation with reduced data, but EEG-CLIP demonstrates greater robustness to data scarcity, particularly at extreme reductions ($\frac{1}{50}$). Shaded regions indicate 80% confidence intervals.

4.2 Impact of the report sections on the representations

To analyze how different report sections influence EEG representation learning, we conducted systematic ablation experiments by training separate models with single section inputs. Each model utilized identical methodology except for the textual input, which was restricted to specific report sections (impression, description, history, etc.). We also experimented with randomly sampling sub-strings from each section during training, but this approach led to decreased performance compared to using complete sections.

As shown in Figure 5, while using all report sections yielded the best overall performance across tasks, certain section-specific models demonstrated unexpected strengths. Notably, the heart rate section model achieved superior accuracy in gender classification despite its brevity (average 2 words per report), suggesting cardiac pattern descriptions capture gender-specific physiological differences. Similarly, a model trained exclusively on technical difficulties notes showed enhanced sensitivity to pathological recordings, likely by learning to associate recording artifacts with abnormal brain activity.

The three primary sections (impression, description, and clinical history) provided the strongest individual contributions

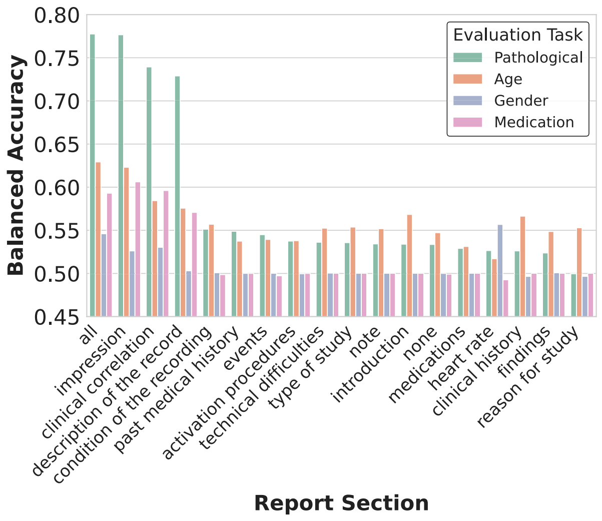

FIGURE 5
Impact of report section selection on representation quality. The heatmap visualization reveals how training on different clinical report sections affects decoding performance across tasks. While using all sections yields the best overall performance, specialized text categories show task-specific advantages. Notably, impressions and descriptions contribute most significantly to pathology detection, while sections like heart rate unexpectedly provide signal for gender classification, suggesting physiological correlations.

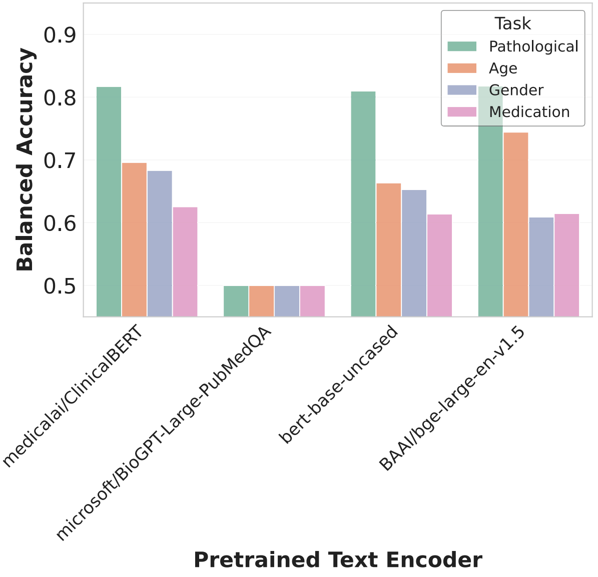

FIGURE 6
Effect of text encoder selection on multimodal alignment. The bar chart compares balanced accuracy across four decoding tasks using different text encoders. Medical/ClinicalBERT shows strong overall performance, while Microsoft/BioGPT-Large-PubMedQA performs poorly across all tasks. BERT-base-uncased maintains competitive pathology detection accuracy, and BAI/bge-large-en-v1.5 excels particularly at pathology and age classification. These results highlight how encoder choice significantly impacts cross-modal representation quality.

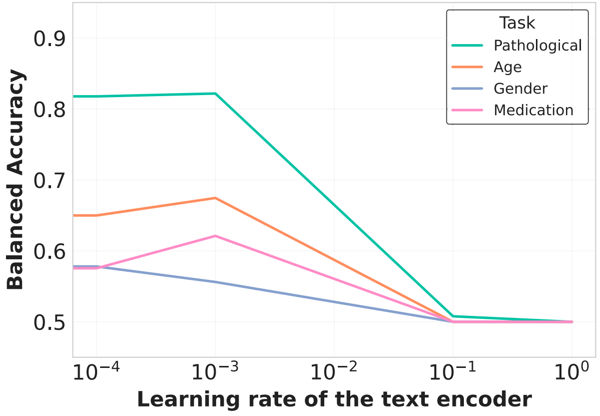

FIGURE 7
Learning rate impact on text encoder fine-tuning. The plot demonstrates how varying the learning rate ratio between the text encoder and the overall model affects representation quality. A moderate learning rate (10^{-3} × the main learning rate) balances preservation of pretrained knowledge with adaptation to the EEG domain, optimizing cross-modal alignment without catastrophic forgetting of linguistic structures.

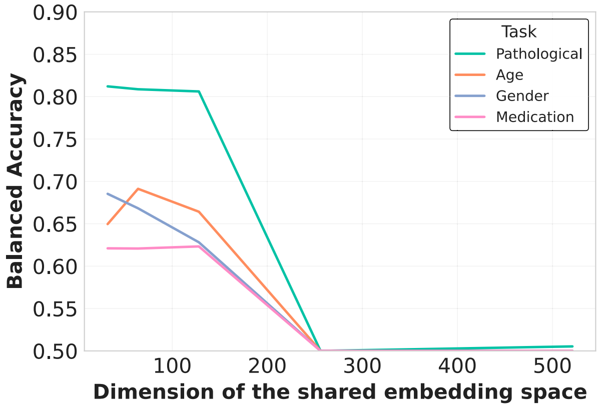

FIGURE 8
Embedding dimensionality analysis for multimodal representation. Counterintuitively, smaller embedding dimensions (32-128) consistently outperform larger ones (256-512) across all tasks. This suggests that compact shared embedding spaces better distill essential cross-modal patterns by enforcing more precise alignment between neurophysiological signals and their textual descriptions, while filtering out modality-specific noise.

to performance, aligning with their higher word counts and prevalence across the dataset (Table 1). However, even sections with limited representation, such as medication lists, contributed unique predictive signals for specific tasks.

These findings reveal how specialized clinical descriptions, even when isolated, can help models detect task-relevant physiological patterns in EEG data. While combining all sections remains optimal for general-purpose representations, our analysis demonstrates the potential value of targeting specific report sections when developing specialized decoders or when working with incomplete clinical documentation.

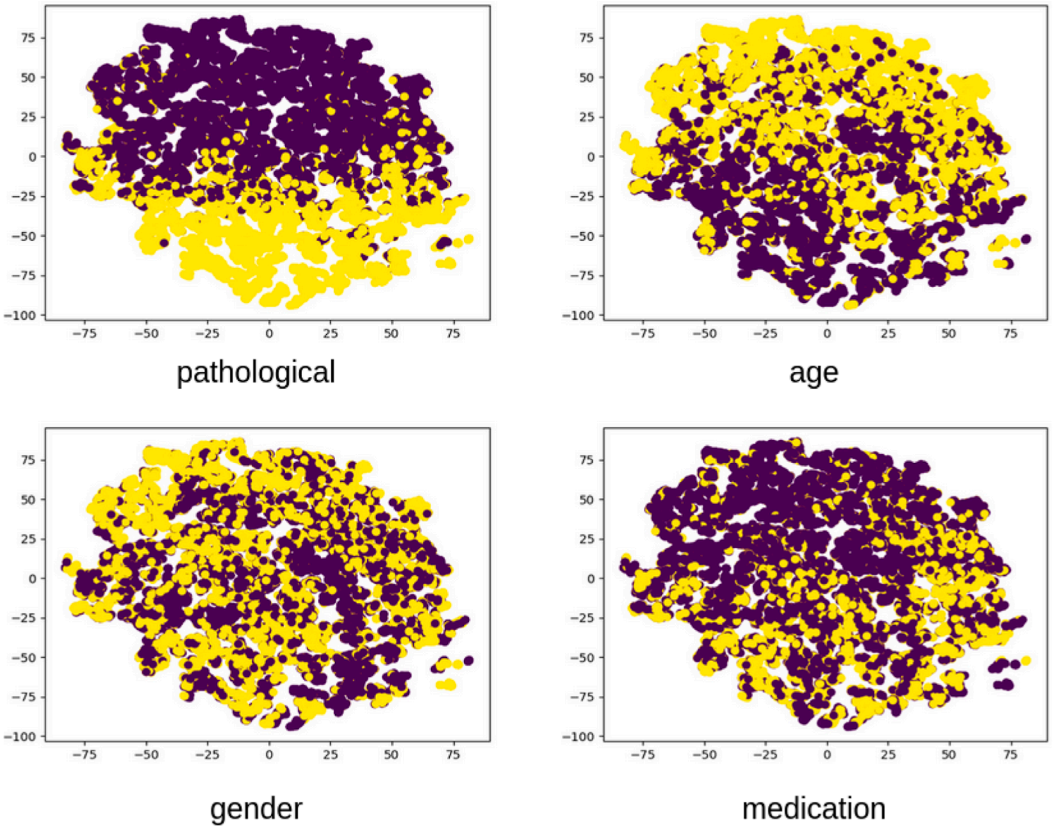

FIGURE 9
Visualized embedding space using t-SNE dimensionality reduction. The 2D projection of EEG embeddings from the evaluation set reveals clear clustering patterns corresponding to clinically relevant attributes. Pathological recordings (yellow) form distinct regions from normal recordings (purple), demonstrating that the unsupervised contrastive learning effectively captures diagnostically relevant features. Secondary clusters corresponding to age and gender are also visible, indicating the multifaceted nature of the learned representations.

4.3 Study on parameter importance

4.3.1 Pre-training of the text encoder

We experimented with several variants of pre-trained language models as text encoders, available publicly on the Hugging Face Hub (Wolf et al., 2020), as shown in Figure 6. These included BERT-base-uncased (Devlin et al., 2019), a general domain model trained on Wikipedia and BookCorpus; ClinicalBERT (Huang et al., 2019), a model finetuned on clinical notes; BioGPT-Large-PubMedQA (Luo et al., 2022), tailored for biomedical text; and BGE-Large (Xiao et al., 2023), a model trained on scientific papers and designed for generation tasks.

ClinicalBERT demonstrated the strongest overall performance across tasks, highlighting the advantage of domain-specific pre-training for clinical text processing. Interestingly, the general-purpose BERT-base-uncased maintained competitive performance on pathology detection despite lacking medical specialization. BGE-Large showed particular strength in pathology and age classification tasks, while BioGPT-Large-PubMedQA consistently underperformed across all evaluations. These results emphasize how encoder architecture and pre-training domain significantly impact the quality of cross-modal representations in our EEG-text alignment framework.

The learning rate ratio between the text encoder and EEG encoder also proved critical, as shown in Figure 7. Optimal performance was achieved when the text encoder was fine-tuned at 10^{-3} times the learning rate of the EEG encoder, balancing adaptation of pre-trained linguistic knowledge while preserving domain-specific understanding.

4.3.2 Projected embedding dimension

Additionally, we analyzed EEG-CLIP model performance across varied hidden dimensionality sizes for the jointly learned EEG and text embeddings, as illustrated in Figure 8. Counter to typical representation learning trends, we found higher decoding accuracy with smaller shared embedding spaces between 32–128 dimensions rather than larger 256 or 512 sizes. A t-SNE visualization of the 64-dimensional embeddings in Figure 9 reveals clear clustering by pathological status, demonstrating effective semantic organization of the learned representation space.

This counterintuitive finding suggests that compressing both modalities into compact unified vectors distills patterns into their most essential characteristics necessary for generalization, while avoiding overfitting to training distribution artifacts that may

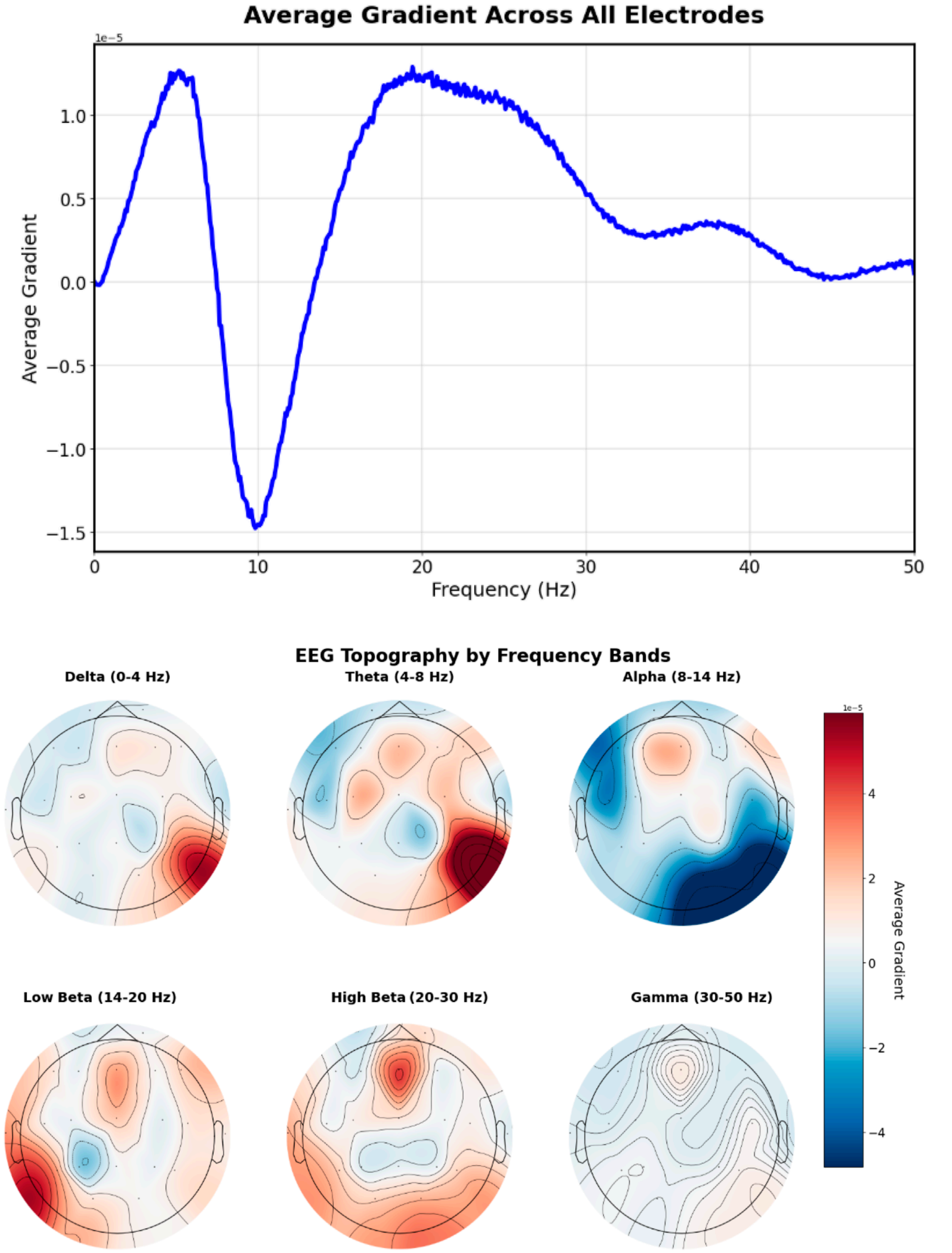

FIGURE 10
EEG-CLIP EEG encoder gradients that direct to the embeddings of “Excessive beta activity” Top: EEG encoder gradients. Bottom: EEG encoder gradients per electrode. Gradients show elevated responses in the beta frequency range (14–30 Hz).

occur in higher-dimensional spaces. The constrained dimensionality may also enforce more direct alignment between descriptive clinical concepts and underlying neurological patterns. These results indicate that EEG-CLIP benefits from lower-complexity manifolds that capture key cross-modal correspondences while filtering out extraneous signals.

4.4 Gradient analysis for model interpretability

In order to provide insights into EEG-CLIP’s decision-making process, we performed gradient-based analysis to visualize which EEG regions contribute most strongly to embeddings aligned with

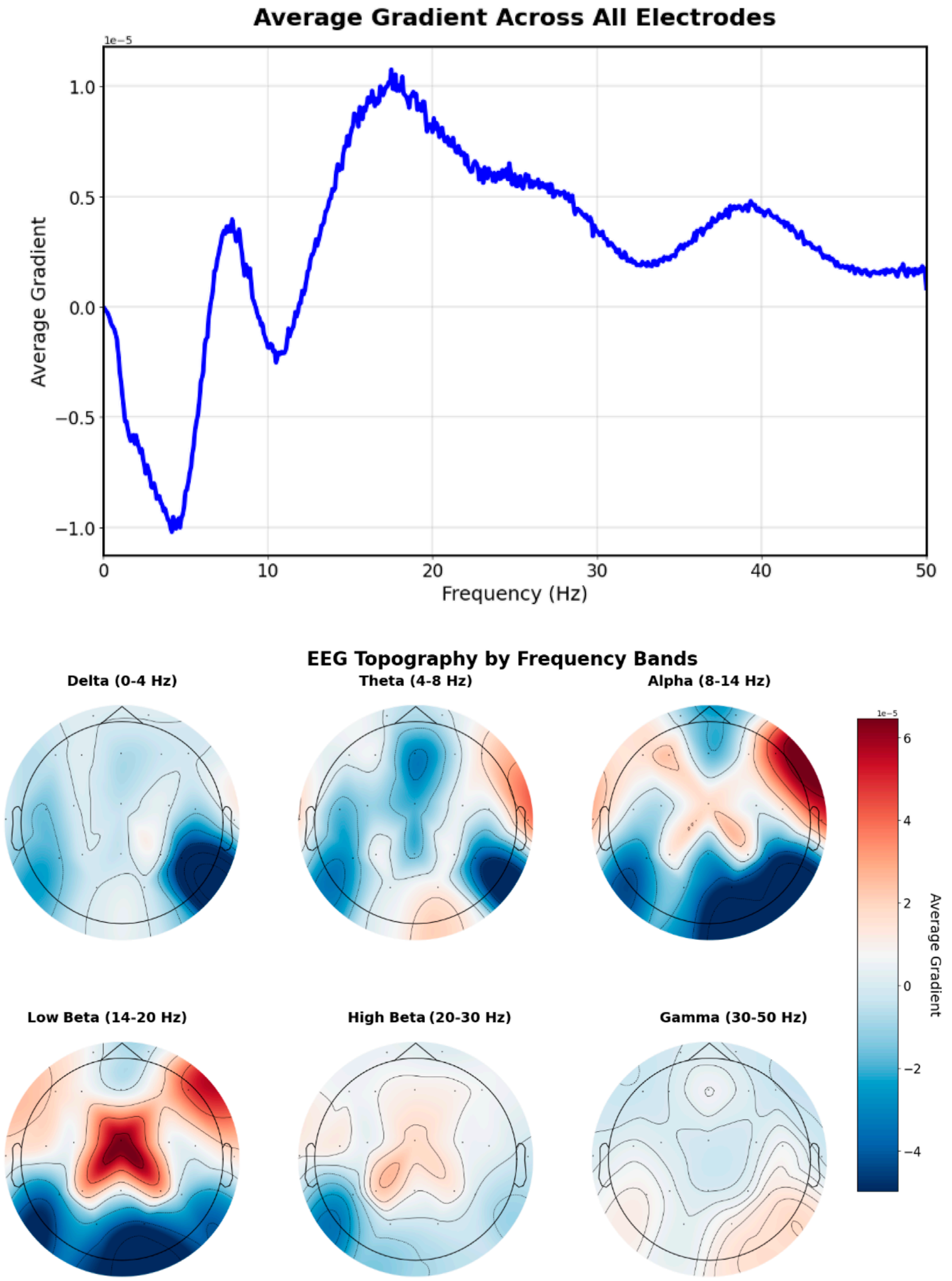

FIGURE 11
EEG-CLIP EEG encoder gradients that direct to the embeddings of "20 Hz beta spindles in central regions". Top: EEG encoder gradients. Bottom: EEG encoder gradients per electrode. Gradients concentrate in the central electrode locations (C3, Cz, C4) with a peak at approximately 20 Hz, demonstrating the model's ability to capture both spatial and frequency-specific features mentioned in the text.

specific textual concepts. We computed gradients of the cosine similarity between EEG embeddings and text embeddings for various prompts with regard to the EEG input that is forwarded through the EEG encoder. We computed those gradients with regard to the input in the frequency domain for easier analysis. Gradients were averaged across the validation set of TUAB.

Analysis of prompts containing specific frequency descriptions reveals characteristic gradient patterns. The prompt "Excessive

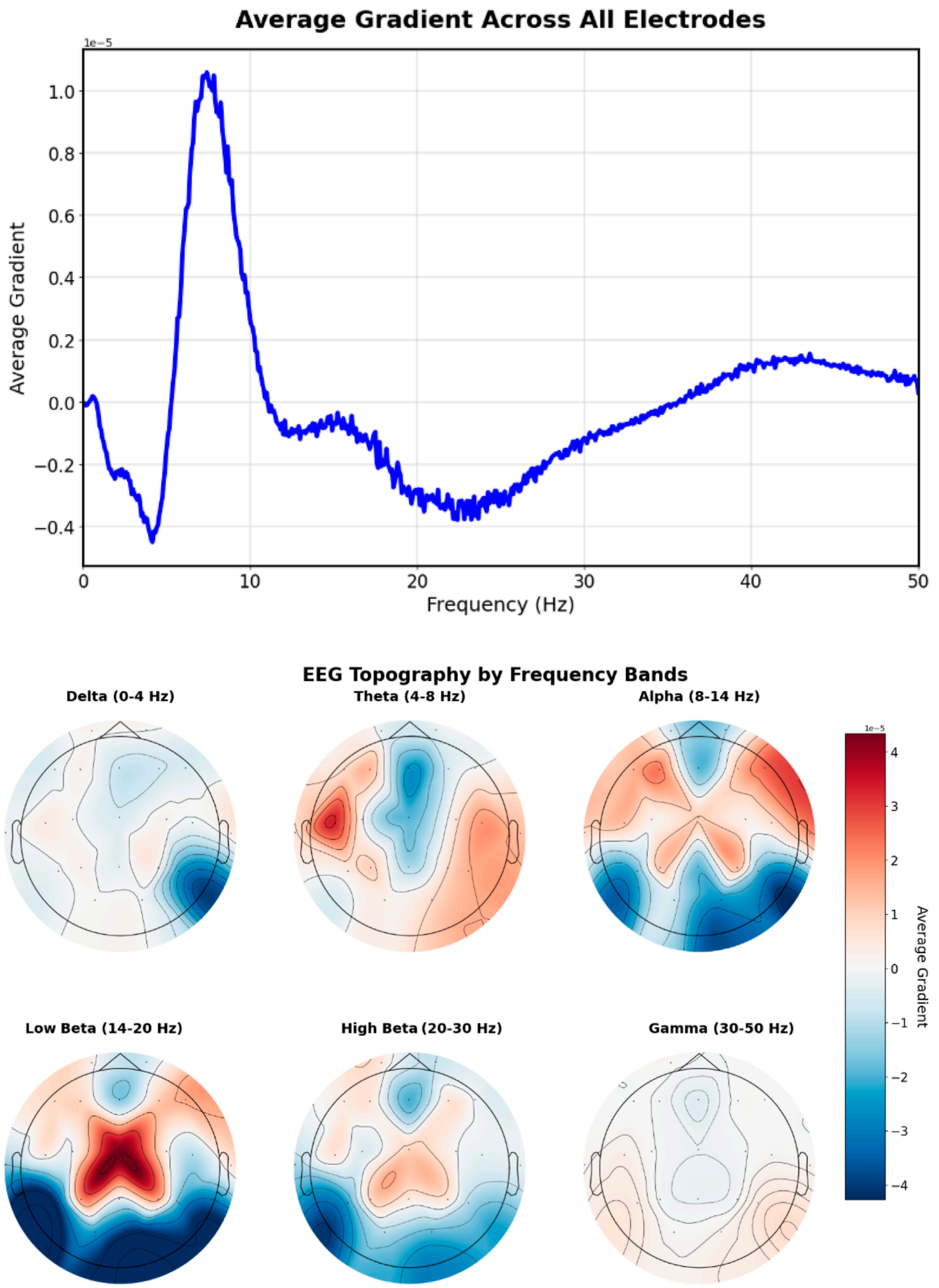

FIGURE 12
EEG-CLIP EEG encoder gradients that direct to the embeddings of “Left temporal sharp waves with a frequency of 6 Hz”. Top: EEG encoder gradients. Bottom: EEG encoder gradients per electrode. Higher magnitudes in left temporal electrodes (T3, T5) with a prominent peak around 6 Hz, indicating successful spatial-frequency alignment with the textual description.

beta activity” (Figure 10) produces elevated gradient responses in the beta frequency range (14–30 Hz).

For the more spatially specific prompt “20 Hz beta spindles in central regions” (Figure 11), gradients concentrate in the central electrode locations (C3, Cz, C4) with a peak at approximately 20 Hz, demonstrating the model’s ability to capture both spatial and frequency-specific features mentioned in the text.

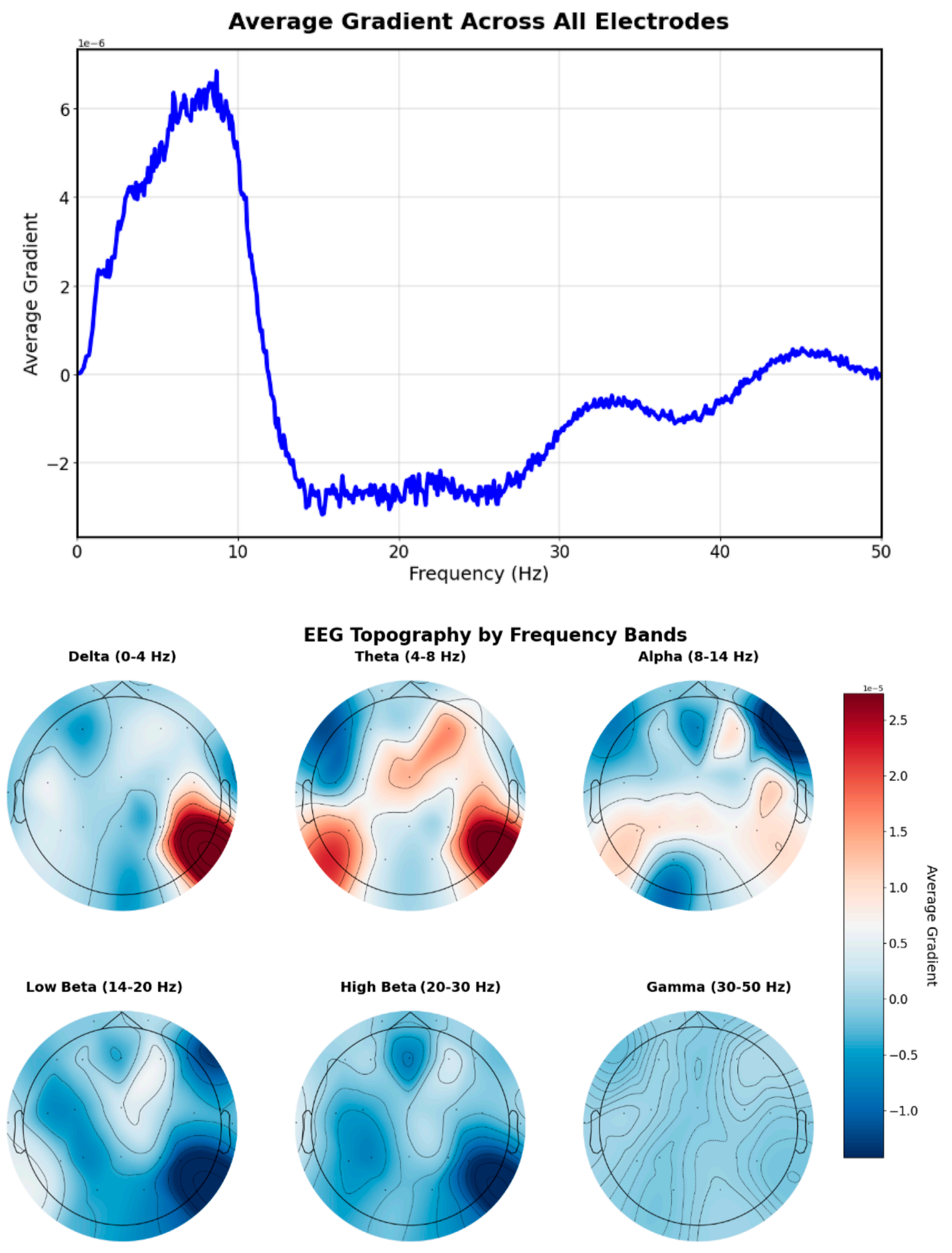

FIGURE 13
EEG-CLIP EEG encoder gradients that direct to the embeddings of “This is a normal recording”. Top: EEG encoder gradients. Bottom: EEG encoder gradients per electrode. Positive peak around 8–9 Hz in the alpha range, indicating the model identifies healthy recordings with increased alpha band activity.

The prompt “Left temporal sharp waves with a frequency of 6 Hz” (Figure 12) produces gradients with lateralized patterns, showing higher magnitudes in left temporal electrodes (T3, T5) with a prominent peak around 6 Hz, indicating successful spatial-frequency alignment with the textual description.

Figure 13 shows the gradient patterns when the model aligns EEG signals with “This is a normal recording” embeddings. The frequency gradient analysis reveals a positive peak around 8–9 Hz in the alpha range, indicating the model identifies healthy recordings with increased alpha band activity. This pattern suggests the model has learned to identify normal EEG patterns through alpha frequency features.

These gradient visualizations provide preliminary evidence that EEG-CLIP learns clinically relevant spatial-temporal patterns rather than purely relying on spurious correlations. The distinct gradient patterns between different prompts suggest the model captures meaningful neurophysiological differences. However, more sophisticated interpretability methods and validation with clinical experts would be needed to fully understand the clinical relevance of these learned representations.

5 Discussion

Our experiments demonstrate that EEG-CLIP successfully learns to align EEG recordings and their clinical text descriptions in a shared embedding space. This approach shows promise for developing more versatile and generalizable EEG representations that can transfer across multiple decoding tasks.

The model achieves strong performance on standard classification tasks (balanced accuracies of 0.847 for pathology, 0.702 for gender, and 0.747 for age) when using an MLP classifier head. Most notably, EEG-CLIP demonstrates zero-shot classification capabilities, achieving 0.755 balanced accuracy on pathology detection using only natural language prompts without any task-specific training. In low-data regimes, our approach shows substantial benefits over models trained from scratch or pretrained on alternative tasks, suggesting efficient capture of generalizable features.

Our ablation studies reveal that while using all report sections yields the best overall performance, specific sections provide distinct advantages for certain tasks. Interestingly, we found that smaller embedding dimensions (32-128) outperformed larger ones, contrary to common intuition in representation learning. This suggests that compressing information into a more compact shared embedding space may better distill essential cross-modal patterns.

Despite these promising results, several limitations warrant discussion. A significant limitation is our primary reliance on TUAB for evaluation, which raises valid concerns about dataset-specific biases and overfitting. Clinical reports in the dataset vary in quality, detail, and structure, potentially limiting the model's ability to learn precise alignments. Additionally, our current approach treats entire EEG recordings and their corresponding reports as aligned pairs, whereas more fine-grained temporal alignment between specific EEG segments and relevant report sections could improve performance, as demonstrated by recent Multiple Instance Learning frameworks (Gijsen and Ritter, 2024). While we implemented validation strategies including held-out evaluation sets, cross-validation for few-shot experiments, and evaluation across four diverse tasks, broader multi-site validation across different hospital systems, recording protocols, patient populations, and clinical practices would strengthen generalizability claims.

Future work could explore methods for aligning specific EEG segments with relevant sentences in clinical reports, scaling to larger and more diverse EEG datasets, incorporating additional data modalities, and developing methods to interpret the learned representations in terms of clinically meaningful EEG patterns. The field would benefit from standardized multi-site evaluation protocols for EEG-language models to better assess generalizability across diverse clinical settings.

In conclusion, EEG-CLIP demonstrates the feasibility of contrastive learning between EEG signals and natural language descriptions for developing more general and transferable EEG representations. While this work represents an initial exploration, the approach opens up new possibilities for leveraging unstructured clinical text to enhance deep learning models for EEG analysis, potentially leading to more flexible and data-efficient tools for neurological assessment and research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

TCN: Software, Investigation, Writing – original draft, Writing – review and editing, Visualization, Methodology, Conceptualization. RS: Writing – original draft, Supervision, Investigation, Conceptualization, Writing – review and editing, Methodology. TB: Funding acquisition, Resources, Writing – review and editing, Supervision.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under SFB 1597 (SmallData), grant no. 499552394, and from the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) through the Renormalized Flows-1IS19077C grant.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. LLMs were used for polishing the writing without altering the scientific content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers) (Minneapolis, MN: Association for Computational Linguistics) 1, 4171–4186.

Gijsen, S., and Ritter, K. (2024). EEG-language modeling for pathology detection. arXiv preprint arXiv:2409.07480.

Heilmeyer, F. A., Schirrmeister, R. T., Fiederer, L. D. J., Völker, M., Behncke, J., and Ball, T. (2018). "A large-scale evaluation framework for eeg deep learning architectures," in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 07-10 October 2018 (IEEE), 1039–1045.

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Khan, M. J., Ghafoor, U., and Hong, K.-S. (2018). Early detection of hemodynamic responses using EEG: a hybrid EEG-fNIRS study. *Front. Hum. Neurosci.* 12, 479. doi:10.3389/fnhum.2018.00479

Lévy, J., Zhang, M., Pinet, S., Rapin, J., Banville, H., d'Ascoli, S., et al. (2025). Brain-to-text decoding: a non-invasive approach via typing. *arXiv preprint. arXiv:2502.17480*. doi:10.48550/arXiv.2502.17480

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., et al. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics.* 23, bbac409. doi:10.1093/bib/bbac409

Obeid, I., and Picone, J. (2016). The temple university hospital eeg data corpus. *Front. Neurosci.* 10, 196. doi:10.3389/fnins.2016.00196

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in International conference on machine learning (PMLR), 8748–8763.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16, 051001. doi:10.1088/1741-2552/ab260c

Schirrmeister, R., Gemein, L., Eggensperger, K., Hutter, F., and Ball, T. (2017). "Deep learning with convolutional neural networks for decoding and visualization of EEG pathology," in 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 02-02 December 2017 (IEEE), 1–7.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Online: Association for Computational Linguistics), 38–45. doi:10.18653/v1/2020.emnlp-demos.6

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. (2023). *C-pack: packaged resources to advance general Chinese embedding.*