EEG-LANGUAGE PRETRAINING FOR HIGHLY LABEL-EFFICIENT PATHOLOGY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal language modeling constitutes a recent breakthrough which leverages advances in large language models to pretrain capable multimodal models. The integration of natural language during pretraining has been shown to significantly improve learned representations, particularly in computer vision. However, the efficacy of multimodal language modeling in the realm of functional brain data, specifically for advancing pathology detection, remains unexplored. This study pioneers EEG-language models (ELMs) trained on clinical reports and 15000 EEGs. We propose to combine multimodal alignment in this novel domain with timeseries cropping and text segmentation. This also enables an extension based on multiple instance learning to alleviate misalignment between irrelevant EEG or text segments. Our results indicate that models learn richer representations from being exposed to a variety of report segments, including the patient's clinical history, description of the EEG, and the physician's interpretation. Compared to models exposed to narrower clinical text information, we find such models to retrieve EEGs based on clinical reports (and vice versa) with substantially higher accuracy. Particularly in regimes with few annotations, we observe that ELMs can significantly improve pathology detection compared to EEG-only models, as demonstrated by both zero-shot classification and linear probes. The integration of multiple instance learning further improves performance across tasks. In sum, these results highlight the potential of integrating brain activity data with clinical text, suggesting that ELMs represent significant progress for clinical applications.

1 INTRODUCTION

Medical neuroimaging such as electroencephalography (EEG) has not yet benefited to the same extent as other domains from the considerable advances deep learning has brought about. While EEG sees widespread clinical use for pathology detection, in particular for epilepsy (Binnie & Stefan, 1999; Jing et al., 2020) as well as sleep disorders (Malhotra & Avidan, 2013), available annotated data is scarce. As the impressive scaling properties of deep learning are now well described (Kaplan et al., 2020; Smith et al., 2023), self-supervised learning (SSL) is a promising direction by enabling pretraining with unlabeled data and thereby increasing available training sample sizes (Hadsell et al., 2006; Chen et al., 2020). Various such approaches have shown initial success when applied to EEG. These include methods relying on data-augmentations (Mohsenvand et al., 2020; Yang et al., 2021), the temporal ordering of EEG data (Banville et al., 2021), as well as masking and reconstruction (Jiang et al., 2024). However, these are hindered by the difficulty of creating appropriate data augmentations and, especially reconstruction techniques, by low signal-to-noise. Thus, progress in the medical context has lagged, likely further exacerbated by the modality displaying high similarity between pathologies.

Meanwhile, important further progress was made in computer vision by leveraging natural language as a signal during pretraining (Radford et al., 2021). Specifically, contrastive approaches which aim to align embeddings of image-text pairs have shown to yield representations powerful for downstream tasks in radiology (Zhang et al., 2022a; 2023). Given that success in radiology is also believed to be bottlenecked by the availability of labeled data and the reliance on fine-grained information (Zhang et al., 2022a), this joint modeling approach is a particularly interesting and novel application for the challenging problem of medical EEG. Fortunately, this is made possible by the clinical reports of physicians which accompany hospital EEG recordings and contain information about the patient and recording itself (Obeid & Picone, 2016).

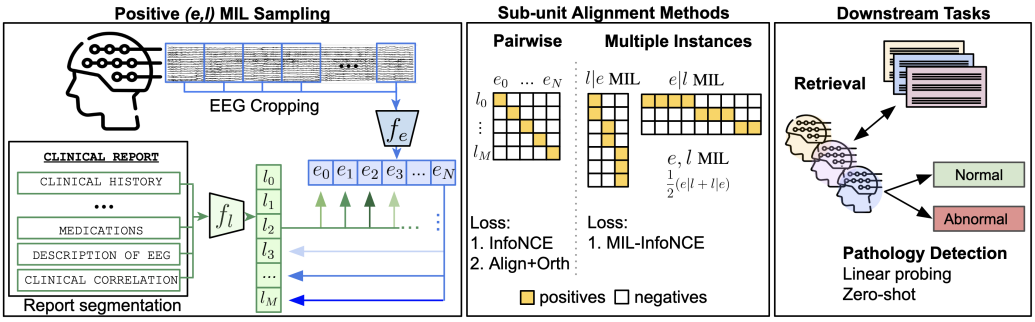

Figure 1: Overview of the methodology. (Left) The ELM-MIL approach allows flexible multimodal alignment by cropping EEG and segmenting medical reports. We sample multiple positives in a cross-modal fashion, such that each EEG crop can be aligned with any number of segments from the paired report ($l|e$). Vice versa, text can be aligned selectively to crops across the EEG recording ($e|l$), illustrated by the differently shaded arrows. (Middle) An overview of investigated methods by visualizing the cross-modal similarity matrices. (Right) To evaluate models, we perform bidirectional retrieval analyses and use both linear probing and zero-shot classification for pathology detection.

However, language-EEG pretraining also entails unique challenges. First, datasets are generally smaller than those used in radiology and especially computer vision. Second, the clinical reports tend to be highly heterogeneous. While previous applications have paired natural and medical images with short captions (Radford et al., 2021; Zhang et al., 2022a), EEG reports tend to span multiple paragraphs and include information irrelevant to downstream clinical tasks, potentially hindering the pretraining process. Moreover, they do not contain any temporal information about when events occurred during the recording.

The current work presents the application of aligning functional brain data with medical textual information for the first time by training EEG-language models (ELMs). To overcome the challenging formats of modalities, constituting long timeseries and multiparagraph reports, we propose sub-unit alignment. To address inconsistent relevance of EEG-text pairs, we additionally propose an extension drawing on insights from the field of multiple instance learning (MIL). Furthermore, we investigate how to best handle the heterogeneity of medical EEG reports. Specifically, we perform content-based text segmentation, enabling inference on the relative importance of the different sources of information in the reports. By fixing pretraining data and encoder architectures across comparisons, we enable inference on the utility of different pretraining strategies per se. Our approach allows us to provide the first evidence of considerable retrieval capabilities for clinical reports and EEG. We furthermore test downstream performance of ELMs on classifying normal and pathologically abnormal EEG, which is a widespread clinical task. These tests include zero-shot classification by leveraging the language capabilities to evaluate the flexibility of the approach. Our results constitute considerable increases in pathology detection performance in scenarios with few labels. These are particularly relevant for clinical contexts, which tend to operate with smaller datasets compared to many common areas of deep learning applications.

2 Related Work

- **Self-supervised learning with EEG data.** SSL with EEG data has been predominantly applied to emotion recognition (Zhang et al., 2022b; Wang et al., 2023), motor imagery (Cheng et al., 2020; Rommel et al., 2022), sleep staging (Yang et al., 2021; Rommel et al., 2022), as well as pathology detection. For the latter application, the temporal order of EEG crops was used initially to demonstrate label-efficient representation learning (Banville et al., 2021). Augmentation-based contrastive learning, combined with larger EEG encoders trained on multiple datasets, further improved pathology detection (Mohsenvand et al., 2020). Recent studies have explored the use of transformers (Yang et al., 2024; Jiang et al., 2024), with a focus on scaling while adopting tokenization in an attempt to improve the challenge of effective cross-dataset EEG training.
- **Using EEG for pathology detection.** While SSL shows good performance for pathology detection, it is particularly in contexts with little annotated data that it performs well. When more labeled data

is available, expert-based feature extraction combined with traditional machine learning classifiers are competitive together with supervised deep learning (Roy et al., 2019; Gemein et al., 2020; Western et al., 2021; Kiessner et al., 2023; Darvishi-Bayazi et al., 2024). This trend has also been observed in other EEG applications (Schirrmeister et al., 2017; Lotte et al., 2018). This may indicate that label noise indices a ceiling effect on classification performance (Engemann et al., 2018; Gemein et al., 2020); specifically, the inter-rater reliability of EEG classification into normal or pathologically abnormal by neurologists. If this hypothesis holds, a focus on improving classification with limited labels may be of extra importance.

- **Medical multimodal language modeling.** Medical vision-language modeling aims to guide self-supervised pretraining on medical images using textual information in reports, with performance on a variety of downstream tasks benefiting as a result (Huang et al., 2021; Wang et al., 2022; Zhang et al., 2022a). Due to less available data in the medical domain, using a pretrained language encoder and freezing its weights was found to boost downstream performance while considerably reducing computational cost (Liu et al., 2023a). Nevertheless, this line of work has focused mainly on the ECG, X-ray, CT images, and structural MRI images (Chen et al., 2023; Lalam et al., 2023; Liu et al., 2023b).

- **Multiple instance learning.** MIL has seen only limited exploration for EEG. Initial studies have investigated the framework by casting crops of EEG as instances and training classifiers for emotion recognition (Caicedo-Acosta et al., 2019), motor imagery (Collazos-Huertas et al., 2020), mental disorders (Sadatnejad et al., 2019), and sleep apnea (Sadatnejad et al., 2019). Of these, only the latter has relied on deep learning. Meanwhile, for multimodal language alignment, Miech et al. (2020) made significant progress by extending the NCE loss to a MIL setting and casting possible text captions as instances.

3 METHODS

3.1 EXPERIMENTAL SETUP

3.1.1 EEG-LANGUAGE PRETRAINING

Here we detail the setup for pretraining ELMs. Whereas vision-language models are typically trained by aligning a 2D image with a short caption (Radford et al., 2021; Zhang et al., 2022a), EEG-language modeling is confronted with long EEG time series and multi-paragraph medical reports. To overcome this, we employ text segmentation and time series cropping to create multiple non-overlapping samples per modality and subject. Next, we propose sub-unit alignment by pretraining on these cropped samples. In addition to considerably increasing sample size, this enables the extension of successful approaches in vision-language models. We initially describe two strategies for sub-unit alignment. First, EEG and text representations may be projected using neural networks to a new, shared latent space prior to alignment (as in CLIP; Radford et al. (2021); Zhang et al. (2022a)), denoted henceforth as $\text{ELM}_{e,l}$. Alternatively, the EEG embeddings may be projected into the output space of the language model (as in M-FLAG by Liu et al. (2023a)), denoted as ELM_{l}. This approach was found to reduce latent collapse in smaller data settings (Liu et al., 2023a). Following a description of these models, we will introduce an extension based on MIL.

For EEG-language pretraining we assume the paired input $(\mathbf{x}_{e,i}, \mathbf{x}_{l,i})$. Here $\mathbf{x}_{e,i} \in \mathbb{R}^{c \times s}$ denotes one or a batch of crops of EEG signal with c channels and s time samples belonging to EEG recording i. Meanwhile, neural signals of recording i as well as patient information is described in $\mathbf{x}_{l,i}$, which represents a natural language text report. The main goal is to train the EEG encoder function f_e, which projects a crop of EEG signal into a vector of lower dimensionality. Following pretraining, this encoder function f_e can be used for downstream applications such as pathology detection.

Dropping the recording subscript i for brevity, each pair $(\mathbf{x}_e, \mathbf{x}_l)$ is projected into the vectors $\mathbf{e} \in \mathbb{R}^d$ and $\mathbf{l} \in \mathbb{R}^d$ respectively. For every $\mathbf{x}_e$, text of the associated report is sampled according to $\tilde{\mathbf{x}}_l = z_l(\mathbf{x}_l)$, where z_l represents the language sampling function detailed below. First, both the EEG crop $\mathbf{x}_e$ and text $\tilde{\mathbf{x}}_l$ are encoded into vectors $\mathbf{h}_e$ and $\mathbf{h}_l$. For $\text{ELM}_{e,l}$, we use projectors g_e and g_l to

yield vectors $\mathbf{e}$ and $\mathbf{l}$, whereas for ELM$_l$ the text embeddings are not projected:

$$\mathbf{e} = g_e\left(f_e\left(\mathbf{x}_e\right)\right) \tag{1}$$

$$\mathbf{l} = \begin{cases} g_l\left(f_l\left(\tilde{\mathbf{x}}_l\right)\right) & \text{if ELM}_{e,l} \\ f_l\left(\tilde{\mathbf{x}}_l\right) & \text{if ELM}_l \end{cases} \tag{2}$$

To enable multimodal pretraining, the projectors g_e and g_l map $\mathbf{e}$ and $\mathbf{l}$ to a shared latent space with identical dimensionality d. For ELM$_l$, this is achieved by having g_e project to the native dimensionality of the text encoder f_l.

As paired medical EEG data and clinical reports are scarce, training the text encoder function f_l from scratch is unlikely to be successful. Furthermore, employing an existing language model and finetuning the model during multimodal pretraining can lead to training instability and collapse of the latent space (Jing et al., 2021; Liu et al., 2023a). To prevent resulting information loss, we follow the recommendations by Liu et al. (2023a) to use a pretrained language model for f_l and freeze its weights during training. For ELM$_l$, we adopt their proposed composite loss to learn f_e and g_e:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \mathcal{L}_{orth} \tag{3}$$

$$\mathcal{L}_{align} = \|\mathbf{e} - \mathbf{l}\|_2^2 = 2 - 2\mathbf{e}^\top\mathbf{l} \tag{4}$$

$$\mathcal{L}_{orth} = \sum_{j=1}\left(1 - \left(\mathbf{h}_e^\top \cdot \mathbf{h}_e\right)_{jj}\right)^2 + \sum_{j\neq k}\left(\mathbf{h}_e^\top \cdot \mathbf{h}_e\right)_{jk}^2, \tag{5}$$

where $\{j,k\} \in \{1,...,\dim\left(\mathbf{h}_e\right)\}^2$ and $\mathbf{h}_e$ denotes a batch of EEG embeddings. Whereas $\mathcal{L}_{align}$ minimizes the difference between $\mathbf{e}$ and $\mathbf{l}$, $\mathcal{L}_{orth}$ promotes independence between latent dimensions of $\mathbf{h}_e$. More specifically, the latter is achieved by manipulating the empirical correlation matrix, where the diagonal and off-diagonal elements are pushed to 1 and 0 respectively (Liu et al., 2023a).

Meanwhile, ELM$_{e,l}$ relies on the cosine similarities between normalized EEG and text embeddings, $s_{j,j}^{e2l} = \hat{\mathbf{e}}_j^\top\hat{\mathbf{l}}_j$, and between text and EEG, $s_{j,j}^{l2e} = \hat{\mathbf{l}}_j^\top\hat{\mathbf{e}}_j$, with $j = 1,2,3,...,B$ for batch size B (Radford et al., 2021). The multimodal contrastive InfoNCE loss uses a temperature hyperparameter τ (set to 0.3 using a holdout set; Appendix B.4) and is formulated as:

$$\mathcal{L}_{j,k}^{e2l} = -\log\frac{\exp\left(s_{j,k}^{e2l}/\tau\right)}{\sum_{m=1}^{B}\exp\left(s_{j,m}^{e2l}/\tau\right)} \tag{6}$$

$$\mathcal{L}_{j,k}^{l2e} = -\log\frac{\exp\left(s_{j,k}^{l2e}/\tau\right)}{\sum_{m=1}^{B}\exp\left(s_{j,m}^{l2e}/\tau\right)} \tag{7}$$

$$\mathcal{L}_{align} = \frac{1}{2B}\sum_{j=1}^{B}\sum_{k=1}^{B}\left(\mathcal{L}_{j,k}^{e2l} + \mathcal{L}_{j,k}^{l2e}\right) \tag{8}$$

Multiple instance learning. While previous approaches aim to align text and EEG crops uniformly, certain text segments likely describe specific EEG sections more accurately than others. Therefore, we introduce a MIL alignment strategy that builds on ELM$_{e,l}$ and accommodates multiple positive samples, allowing for more nuanced multimodal relationships. Whereas MIL approaches often rely on operations such as max-pooling to focus on single positive samples, we rely on insights from the video-text alignment approach (MIL-NCE) by Miech et al. (2020). For a given text sample $\mathbf{x}_l$, we sample multiple positive EEG crops $\mathbf{x}_e$ from the paired recording to approximate the $P(e|l)$ distribution, while for an EEG crop, multiple text segments are sampled to model the $P(l|e)$ distribution. We combine these and sample positives for each EEG crop and text paragraph respectively to approximate $P(e,l)$ via bidirectional alignment. This approach effectively relaxes the assumption of strong alignment for each individual $(\mathbf{x}_e, \mathbf{x}_l)$ pair, instead assuming that, on average, positive samples should have higher similarity scores than negative samples. To this end, we extend

the InfoNCE loss to multiple instances:

$$\mathcal{L}^{e|l} = -\frac{1}{B_l}\sum_{k=1}^{B_l}\log\frac{\frac{1}{|P_k|}\sum_{j\in P_k}\exp(s^{e2l}_{j,k}/\tau)}{\sum_{j=1}^{B_e}\exp(s^{e2l}_{j,k}/\tau)} \quad \text{where } |P_k| \leq N \tag{9}$$

$$\mathcal{L}^{l|e} = -\frac{1}{B_e}\sum_{k=1}^{B_e}\log\frac{\frac{1}{|Q_k|}\sum_{j\in Q_k}\exp(s^{l2e}_{j,k}/\tau)}{\sum_{j=1}^{B_l}\exp(s^{l2e}_{j,k}/\tau)} \quad \text{where } |Q_k| \leq M \tag{10}$$

$$\mathcal{L}^{e,l} = \frac{1}{2}\left(\mathcal{L}^{e|l} + \mathcal{L}^{l|e}\right) \tag{11}$$

where P_k and Q_k are sets of positive EEG crops and text paragraphs respectively. B_e and B_l are the batch sizes for EEG and text respectively, for which we sample up to N EEG crops and M text paragraphs for $\frac{B}{N}$ subjects. We set $N = 32$ and $M = 8$ as this covers all samples for a majority of subjects. We normalize using $|Q_k|$ or $|P_k|$ to account for the varying number of crops across subjects.

Language encoder. For f_l we use a transformer model which was pretrained in a contrastive manner on PubMed search logs (MedCPT; Jin et al. (2023)). See Appendix B.3 for a comparison of language models. ELM$_l$ adopts the language model's native hidden dimensionality (768), while for ELM$_{e,l}$ and ELM-MIL we project to a dimensionality of 256.

EEG encoder. For the EEG encoder f_e we use a randomly initialized residual convolutional neural network, with an identical backbone architecture across all comparisons. We use nonlinear MLPs with a single-hidden layer for g_e and g_l, as well as for the projector head in EEG-only self-supervised learning. More details are provided in Appendix B.2.

3.1.2 EEG-ONLY SELF-SUPERVISED LEARNING

We compare the representations learned by EEG-language pretraining to those obtained via EEG-only pretraining. First, we employ multiple methods that train for invariance to data augmentations. This is achieved by sampling data augmentations for each EEG crop $\mathbf{x}_e$, resulting in two differing data views $\{\mathbf{x}'_e, \mathbf{x}''_e\}$. It is important the data augmentations do not destroy the semantic information in $\mathbf{x}_e$. Training to align the embeddings of these views while preventing collapse has been shown to yield data representations useful for downstream tasks (Chen et al., 2020; Mohsenvand et al., 2020; Yang et al., 2021). We implement the following methods (Appendix B.5): Bootstrap-Your-Own-Latent (BYOL; Grill et al. (2020)), Variance-Invariance-Covariance Regularization (VICReg; (Bardes et al., 2021)), and Contrast with the World Representation (ContraWR; Yang et al. (2021)). Additionally, we compare against methods using the temporal ordering of EEG crops: Relative Positioning (RP; Banville et al. (2021)), Temporal Shuffling (TS; Banville et al. (2021)), Contrastive Predictive Coding (CPC; Banville et al. (2021)).

3.2 DATASETS AND PREPROCESSING

• **TUEG.** The Temple University Hospital (TUH) EEG Corpus is the largest available corpus of hospital EEG data with varying montages, channel counts, and sampling frequencies (n=26846 (Obeid & Picone, 2016)). For each patient, one or more EEG sessions are provided, each of which contains one or more recordings. For most of the dataset, no labels are available beyond patient age and sex. However, many EEG sessions are associated with a natural-language clinical report.

• **TUAB.** The TUH Abnormal EEG corpus is a subset of TUEG which was manually labeled by clinicians indicating whether the EEG displays pathological abnormalities (Lopez et al., 2015). This enables the binary classification task of predicting the status of {normal, abnormal}. As training (n=2717) and evaluation (n=276) sets are provided, we use the latter as a hold-out test set.

3.2.1 TEXT PROCESSING

In order to categorize the textual content in the clinical reports, we employed regular expressions matching for commonly-occurring headings (an overview is provided in Appendix F.3). These enabled the segmentation of individual reports into their respective headings with associated text paragraphs, providing insight into which information in physician reports is encoded in the EEG. We cluster headings into four categories. First, the *clinical history* cluster of headings contains

demographic information in terms of patient age and sex, as well as a brief description of relevant current and/or past pathology and symptoms. The *record description* cluster includes the physician's observations of the EEG traces, which describes both normal and abnormal features, often in terms of oscillatory brain activity. The *medication* cluster contains the patient's current medication information. Finally, the *interpretation* cluster summarizes a physician's thoughts, often including the impression of whether the EEG is normal or pathologically abnormal, as well as a clinical correlation. To investigate whether EEG-language models can learn richer representations by being exposed to a larger variety of text, we also train models by sampling text from all four aforementioned clusters.

Due to the heterogeneity of the clinical reports, we further test the utility of summarizing the pathological status indicated by the clinical report using a large language model (LLM). Due to the sensitive nature of the clinical reports, we use the Llama-3 8B model (AI, 2024) locally and instruct for the production of a single-sentence summary of a report, which should include whether the EEG was deemed abnormal and for which reasons (Appendix F.3).

Language encoding. Given a sampled paragraph from a clinical report or the LLM-generated summary, we encode this text by relying on the embedding of a special $[cls]$ token which aggregates the representations across all tokens. As such, given a clinical report $\mathbf{x}_l$, the transformation function z_l corresponds to text segmentation or summarization yielding $\tilde{\mathbf{x}}_l$. Following tokenization, we embed into the $[cls]$ token using f_l. The resulting text embedding $\mathbf{h}_l$ may be used for multimodal pretraining.

3.2.2 EEG PROCESSING

From the EEG dataset, recordings longer than 2.5 hours were omitted to filter out a small subset of very long, potentially overnight recordings. For training efficiency, only the first 45 minutes of a recording were used. Any recording files shorter than 70 seconds were also omitted.

EEG preprocessing. EEG data received minimal preprocessing (using MNE (Gramfort et al., 2013)). First, the initial 10 seconds were removed to reduce the impact of set-up artefacts. Afterwards, a bandpass filter of 0.1-49 Hz was applied and all recordings were resampled to 100 Hz. To reduce the impact of signal artefacts, all EEG signals had their amplitude clipped to $\pm$ 800 μV. As a large majority of recordings used an average-reference (AR) or linked-ear reference (LE), we only used these recordings and standardized them via transformation to the 20-channel Temporal Central Parasagittal (TCP) montage. To enable fair comparisons between methods, the optimal crop-length out of $\{$5,10,20,30,60$\}$ seconds was determined without data-leakage through training and evaluation on subsets of the training data only (Appendix B.2). Based on these results, we used 20 and 60 second crops for EEG-only and EEG-language modeling respectively.

3.2.3 DATA SUBSAMPLING

TUEG contains considerably more abnormal than normal EEGs. As vision-language models have been shown to be sensitive to imbalanced classes (Wang et al., 2024), we subsample the data to create approximately equal class representation. To do so, we rely on the LLM summaries of reports, which facilitated report classification based on regular expressions due to reoccurring phrasing. If any data of a subject was present in the retrieval or TUAB test set, all their data was excluded from the pretrain subset to avoid data leakage. Further details are provided in Appendix C and resulting sample sizes are shown in Table 1.

Table 1: Dataset sample sizes.

Data subset	EEG files	Clinical reports	Crops (60s)	Crops (20s)
Pretrain	15144	11785	270K	813K
TUAB train	2712	Not used	56579	170K
TUAB test	276	Not used	5783	17349
Retrieval test	437	437	8887	26661

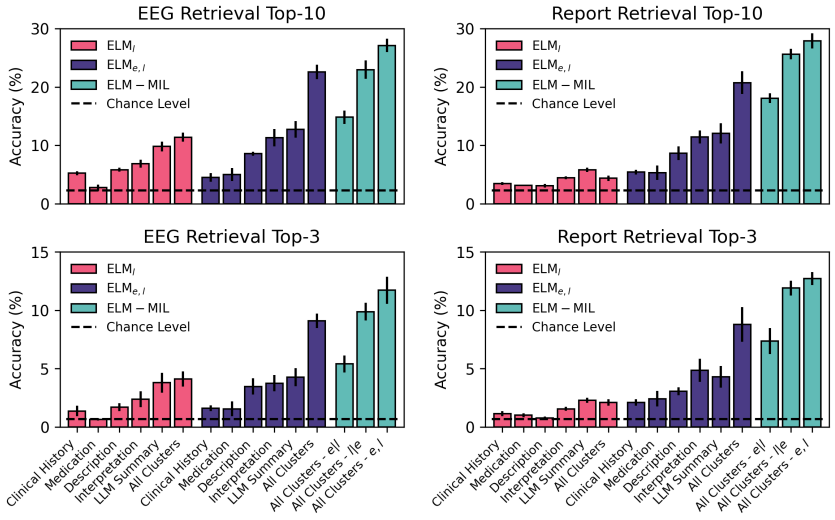

Figure 2: EEG-Language models are evaluated on their retrieval ability using top-k accuracy out of 437 patients. Either EEG is retrieved based on a queried clinical report, or vice versa. Error bars indicate standard deviations over five model training runs.

4 Experimental Results

4.1 Retrieval

To investigate the information represented in the learned embeddings resulting from EEG-language training, we perform retrieval analyses. Given a medical report describing the patient and their EEG recording, we probe the ability to recover the patient's EEG by rank-ordering candidate EEG based on embedding similarity. This analysis is also performed in opposite direction, by retrieving the associated report given an EEG recording. As reports refer to an entire recording, EEG embeddings of single crops are averaged. Given that reports consist of multiple paragraphs, embeddings of single paragraphs are also averaged. This procedure yields one EEG and report embedding per patient recording, which we use for rank-ordering based on cosine similarity.

The top-K retrieval accuracy, which scores whether the patient's EEG or report has a rank equal or better than K, is plotted in Figure 2. Many models perform considerably above chance level, indicating the successful generalization of learned multimodal EEG-language information. The text sampling markedly impacts the report retrieval. The clinical history and medication clusters, which contain no direct description of the observed EEG recording, score lowest. While including such information (description cluster) helps considerably, retrieval is particularly effective when a pathology-relevant context is provided (interpretation cluster and LLM summary). This indicates that pathology is a significant source of between-subject variation. Further clear improvements are seen when text from all clusters is sampled, indicating that these clusters contain unique information and that EEG-language modeling can capture multiple dimensions of patient information.

For both EEG and report retrieval, ELM$_{e,l}$ models tend to outperform ELM$_l$ models. However, this discrepancy in performance is particularly prevalent for report retrieval. This is likely due to omission of a text projection head in ELM$_l$, which may therefore lack the flexibility to appropriately separate the EEG reports in latent space. Due to the benefit of pretraining using all text clusters, we pretrain our ELM-MIL models in this manner only and observe that these can further improve retrieval performance. Interestingly, sampling multiple positive EEG crops (i.e., $e|l$) performs considerably better than the inverse ($l|e$), yet bidirectional alignment and sampling multiple positives jointly (e, l) scores highest. These results indicate for the additional flexibility of this approach to aid in multimodal alignment, supporting the hypothesis that not all EEG and text pairs are equally informative.

Table 2: Pathology detection via linear probing at 1%, 10%, and 100% labeled data of the TUAB training set. The (second) best scores are printed (underlined) bold. Standard deviations over five model training runs are included.

		Balanced Accuracy			AUROC		
Method	Text Sampling	1%	10%	100%	1%	10%	100%
SV	-	71.36±1.10	81.06±0.30	84.13±0.29	79.87±1.30	89.23±0.51	91.83±0.32
BYOL	-	72.69±0.57	79.03±1.16	79.94±2.14	78.85±0.81	86.75±0.76	88.82±0.70
VICReg	-	71.76±0.81	79.6±1.07	82.46±0.96	78.7±1.11	86.04±0.80	88.78±1.04
ContraWR	-	73.30±1.44	80.72±1.69	82.44±1.22	80.30±1.91	86.67±1.32	88.44±1.20
RP	-	74.52±1.06	82.16±0.38	83.36±0.42	82.63±0.87	89.78±0.43	91.43±0.34
TS	-	74.99±0.86	82.16±0.64	84.10±0.66	82.51±0.91	89.58±0.55	91.50±0.32
CPC	-	73.20±0.79	78.44±1.00	79.95±1.49	81.48±1.02	86.44±1.07	87.92±1.14
ELM_l	Clinical History	76.36±0.54	79.88±1.32	82.61±1.43	84.48±1.07	87.87±1.05	89.40±0.80
ELM_l	Medication	75.71±1.14	80.41±0.77	83.20±1.17	84.27±0.92	88.10±0.86	89.79±0.76
ELM_l	Description	79.61±0.69	81.87±0.69	83.88±0.89	87.88±0.80	89.73±0.63	90.67±0.42
ELM_l	Interpretation	80.57±0.62	81.98±1.41	83.08±1.01	88.86±0.60	89.82±0.78	90.53±0.62
ELM_l	LLM Summary	82.36±0.56	83.70±0.54	84.37±0.51	90.35±0.34	90.97±0.35	91.58±0.22
ELM_l	All Clusters	79.07±0.87	81.07±0.75	83.18±0.60	87.12±0.48	88.61±0.36	89.78±0.23
$\text{ELM}_{e,l}$	Clinical History	79.86±0.00	82.71±0.00	84.13±0.00	87.61±0.00	90.72±0.00	91.81±0.00
$\text{ELM}_{e,l}$	Medication	79.86±0.00	82.58±0.00	82.31±0.00	88.41±0.00	90.57±0.00	91.81±0.00
$\text{ELM}_{e,l}$	Description	81.47±0.29	83.64±0.54	84.84±0.91	89.14±0.53	91.70±0.19	92.71±0.14
$\text{ELM}_{e,l}$	Interpretation	82.83±0.35	84.09±0.52	84.51±0.58	90.92±0.35	92.48±0.31	93.13±0.27
$\text{ELM}_{e,l}$	LLM Summary	82.18±0.83	83.16±1.04	83.24±0.44	90.35±0.37	91.57±0.53	92.27±0.42
$\text{ELM}_{e,l}$	All Clusters	82.64±0.24	84.13±0.35	85.39±0.45	90.98±0.29	92.53±0.21	93.26±0.24
ELM-MIL $l\|e$	All Clusters	82.53±1.80	**86.38**±0.77	**87.62**±0.43	89.88±1.47	92.92±0.54	93.52±0.34
ELM-MIL $e\|l$	All Clusters	**83.71**±0.59	<u>84.37</u>±0.97	85.65±0.97	**92.37**±0.43	**93.25**±0.27	<u>93.65</u>±0.16
ELM-MIL e, l	All Clusters	<u>83.10</u>±0.56	84.21±0.82	<u>87.11</u>±0.76	<u>91.54</u>±0.44	<u>93.14</u>±0.24	**93.91**±0.17

4.2 Pathology classification

4.2.1 Linear probing

Next, we study the learned representations in their relevance to clinical pathology. To do so, we first train linear probes to detect pathology on the representations of pretrained models on TUAB under varying amounts of labels (Table 2; Appendix D). Models are trained on single EEG crops, across which we average predictions to obtain a recording-level prediction. We find EEG-language pretraining yields large improvements for pathology detection over EEG-only pretraining, with multimodal models being particularly effective at small sample sizes: at 1% of exposed labels, performance increases reach 8.7% balanced accuracy and 9.7% AUROC. Our ELM-MIL models are found to score highest on this task too, albeit with variability between the variants.

We evaluate models out-of-distribution on the NMT EEG Dataset (Khan et al. (2022); Section A.1) and investigate two additional tasks for clinical event detection (TUEV and TUSZ: A.2). We observe strong performance for ELM-MIL across evaluations.

Given the broad outperforming of ELMs compared to EEG-only models, we investigated whether the strategy of sub-unit multimodal modeling provides inherent benefits. We provide this additional set of analyses in appendix A.3, which indicates that our sub-unit alignment strategy promotes the encoding of between-subject information even in the absence of semantically relevant text. This allows $\text{ELM}_{e,l}$ to nearly match the best EEG-only pretraining strategy for pathology detection when reports are randomly shuffled.

4.2.2 Zero-shot pathology detection.

Next, we investigate the unique ability of multimodal language modeling to leverage the language modality to perform 'zero-shot' classification. Without any explicit labels for downstream training, EEG may be classified by computing its similarity in latent space to text prompts representing the

Table 3: Pathology detection via zero-shot classification. The (second) best scores are printed (underlined) bold. Standard deviation over five model training runs are included.

Method	Text Sampling	Balanced Accuracy (%)	AUROC (%)	F1-Score (%)
ELMl	Clinical History	50.00±0.00	34.34±1.66	0.00±0.00
ELMl	Medication	50.00±0.00	74.99±1.80	0.00±0.00
ELMl	Description	50.00±0.00	43.70±1.99	0.00±0.00
ELMl	Interpretation	50.00±0.00	89.23±0.31	0.00±0.00
ELMl	LLM Summary	71.98±0.62	90.92±0.35	61.77±1.24
ELMl	All Clusters	50.00±0.00	84.52±0.57	0.00±0.00
ELMe, l	Clinical History	62.50±8.74	67.57±10.41	60.96±6.13
ELMe, l	Medication	51.09±17.79	52.21±23.54	50.67±12.53
ELMe, l	Description	64.03±3.95	71.40±4.32	63.56±3.81
ELMe, l	Interpretation	82.34±1.42	91.80±0.47	80.10±1.63
ELMe, l	LLM Summary	58.87±15.48	67.98±20.88	64.32±9.06
ELMe, l	All Clusters	<u>83.16</u>±1.15	**91.91**±0.67	<u>81.25</u>±1.35
ELM-MIL $l\|e$	All Clusters	68.86±7.89	75.23±9.28	68.07±6.82
ELM-MIL $e\|l$	All Clusters	79.10±2.93	87.26±3.19	77.60±2.52
ELM-MIL e, l	All Clusters	**84.31**±0.57	<u>91.56</u>±1.31	**82.13**±0.64

candidate classes. As suggested by Radford et al. (2021), we create a prompt ensemble over 21 variations of the phrasing "The EEG is {normal, abnormal}" (Appendix D). Results in table 3 indicate that, despite a small dataset, EEG-language models can reach high levels of zero-shot pathology detection. The best models outperform nearly all linear probes at 1% labels and even match EEG-only models at 100% labels. The clinical history, medication, and description models perform poorly, which follows from these models not being exposed to the explicit phrasing indicating the EEG status as normal or abnormal per se. Their performance can likely be improved by designing appropriate prompts.

Notably, while the ELM$_{e,l}$ models trained on either the interpretation cluster or all clusters both perform well with high consistency, training on LLM summaries resulted in highly variable scores. As the LLM-generated text was considerably more uniform with repetitive phrasing across reports, the lack of variability in combination with limited data may have lead to unstable language representations of the text projector. Meanwhile, we observe the opposite pattern for ELM$_l$, where LLM summaries enabled the only consistently above-chance zero-shot classifier. As with the report retrieval analysis, the fixed text representations of a language model which is not finetuned for EEG is likely inadequate to reliably separate between diverse descriptions of pathological and normal EEG. Meanwhile, the rigid LLM-generated text may have aided in this scenario by consistently yielding divergent text representations with which normal and abnormal EEG may be aligned. In line with previous evaluations, our ELM-MIL approach further improves performance, but requires the bidirectional approach (e, l).

4.2.3 EEG-Language modeling with MIL-InfoNCE

Whereas for InfoNCE the temperature parameter sets the relative focus across negative samples (Wang & Liu, 2021), for MIL-InfoNCE it does so too for positive samples. We therefore test the sensitivity of our methods to the parameter (Figure 3). We find that MIL-InfoNCE is more robust to changes of τ for pathology detection, while retrieval performance can be further improved by lowering τ. This may be explained as retrieval being subject-based rather than class-based (see Appendix 12). Moreover, performance increases from $\tau < 1$ indicates the utility of this additional hyperparameter of InfoNCE, which is absent in NCE.

We perform additional ablations to investigate crucial aspects of the ELM-MIL e, l model. First, we find that additional positive EEG and text samples improve downstream performance (Table 4). We additionally ablate the aggregation method for positive samples and find MIL-InfoNCE to outperform considered alternatives. We compare to aligning only the most similar positive sample (denoted Max+InfoNCE), using attention to create a weighted average across positive samples based on similarity values (Attn+InfoNCE; Ilse et al. (2018)), as well as taking the sum instead of mean

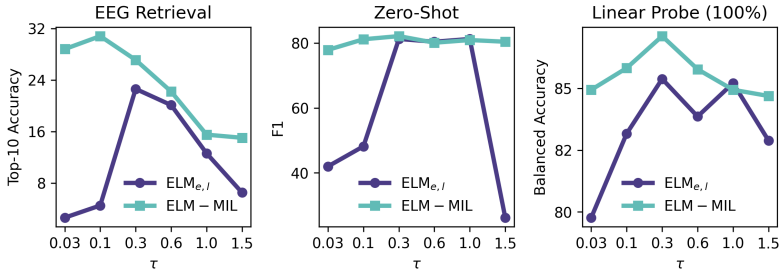

Figure 3: Model comparisons across EEG retrieval and pathology detection under different values of the temperature parameter τ.

across log-probabilities (Sum+InfoNCE). The latter does not account for the varying amount of text and EEG crops across subjects.

Table 4: Ablation studies (Means over five training runs). Ret: EEG Retrieval (Top-10 accuracy), LP: Linear Probe (Balanced accuracy at 100%), ZS: Zero-shot classification (F1).

(a) Aggregation

Method	Ret.	LP	ZS
Max+InfoNCE	3.9	77.5	43.2
Attn+InfoNCE	8.3	84.9	17.5
Sum+InfoNCE	24.7	86.0	78.8
MIL-InfoNCE	**27.1**	**87.1**	**82.1**

(b) Positive EEG Samples

N	Ret.	LP	ZS
2	19.8	85.9	78.8
4	21.8	85.9	79.2
8	25.3	86.5	80.0
16	26.5	86.9	78.5
32	**27.1**	**87.1**	**82.1**

(c) Positive Text Samples

M	Ret.	LP	ZS
2	**28.1**	85.8	80.4
4	27.1	86.0	80.4
8	27.1	**87.1**	**82.1**

5 Discussion

The current work presents a first application of multimodal pretraining using natural language and functional brain data in a medical context. Our findings indicate that ELMs provide better representations than EEG-only SSL. To enable this, we perform sub-unit alignment following timeseries cropping and text segmentation. We further improve downstream performance via MIL-InfoNCE to address misalignment. The most useful representations were obtained via a combination of our ELM-MIL models and exposure to a variety of textual information. Such multimodal models were also found to be capable of zero-shot pathology detection. Using linear probing, sizable performance improvements over EEG-only SSL were observed, with the largest gains in contexts with few annotated samples. We additionally show strong performance of ELMs via external validation and clinical event detection tasks.

Some considerations of the current study deserve mention. No additional paired EEG-report datasets are currently publicly available, which for now prevents assessing the generalizability of our results across datasets. Although great care was taken to prevent data leakage and no model development involved any of the evaluation data, future work is required to properly study generalizability and scaling behavior as investigated using large transformer models in (Yang et al., 2024; Jiang et al., 2024). While the retrieval analyses suggest that certain models learn richer data representations, a lack of annotations hindered a more detailed assessment of their utility for downstream tasks. Future research could benefit from annotations for specific pathologies, enabling more precise model comparisons. Additionally, we observed lower pathology detection scores for EEG-only SSL than a previous study (Mohsenvand et al., 2020), despite using the same data augmentations. Their work pretrained larger models on multiple datasets to output sequential representations. However, many such adaptations to the EEG encoder or its training could also be applied to EEG-language modeling. Finally, although several models displayed accurate zero-shot pathology detection, the variability in results may be due to the challenges of language modeling with limited data. Further research is needed to explore additional inductive biases or regularization of the text projector to address this issue.

Reproducibility Statement

Section 3.2.2 contains details about the data preprocessing, while Appendix C provides details about data subsampling. Information on model architecture, hyperparameters, and optimization is provided in Section 3.1 and Appendix B. We provide the code of our methods as supplementary material, which we will additionally host publicly upon manuscript publication.

Ethics Statement

This study uses an already existing repository of EEG data, which was collected following ethical guidelines, including participant consent and anonymization. We have ensured that data handling complies with privacy and security standards. As part of this, the manuscript and code release have been carefully reviewed and stripped of any instances of clinical reports.

References

Meta AI. Llama 3. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`, 2024.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Colin D Binnie and Hermann Stefan. Modern electroencephalography: its role in epilepsy management. *Clinical Neurophysiology*, 110(10):1671–1697, 1999.

Julian Caicedo-Acosta, David Cárdenas-Peña, D Collazos-Huertas, Jorge I Padilla-Buritica, G Castano-Duque, and Germán Castellanos-Dominguez. Multiple-instance lasso regularization via embedded instance selection for emotion recognition. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 244–251. Springer, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Yinda Chen, Che Liu, Wei Huang, Sibo Cheng, Rossella Arcucci, and Zhiwei Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Diego Collazos-Huertas, Julian Caicedo-Acosta, German A Castaño-Duque, and Carlos D Acosta-Medina. Enhanced multiple instance representation using time-frequency atoms in motor imagery classification. *Frontiers in neuroscience*, 14:155, 2020.

Mohammad-Javad Darvishi-Bayazi, Mohammad Sajjad Ghaemi, Timothee Lesort, Md Rifat Arefin, Jocelyn Faubert, and Irina Rish. Amplifying pathological detection in eeg signaling pathways through cross-dataset transfer learning. *Computers in Biology and Medicine*, 169:107893, 2024.

Denis A Engemann, Federico Raimondo, Jean-Rémi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, et al. Robust eeg-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11): 3179–3192, 2018.

Lukas AW Gemein, Robin T Schirrmeister, Patryk Chrabaszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of eeg pathology. *NeuroImage*, 220:117021, 2020.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu12. Large brain model for learning generic rep-resentations with tremendous eeg data in bci. *ICLR*, 2024.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Jin Jing, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J Halford, Alice Lam, Douglas Maus, Fonda Chan, Marjan Dolatshahi, Carlos F Muniz, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. *JAMA neurology*, 77(1): 49–57, 2020.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The nmt scalp eeg dataset: an open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

Ann-Kathrin Kiessner, Robin T Schirrmeister, Lukas AW Gemein, Joschka Boedecker, and Tonio Ball. An extended clinical eeg dataset with 15,300 automatically labelled recordings for pathology decoding. *NeuroImage: Clinical*, 39:103482, 2023.

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.

Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 637–647. Springer, 2023a.

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*, 2023b.

Sebas Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult eegs. In *2015 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2015.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

Raman K Malhotra and Alon Y Avidan. Sleep stages and scoring technique. *Atlas of sleep medicine*, pp. 77–99, 2013.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pp. 238–253. PMLR, 2020.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:195498, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19 (6):066020, 2022.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: A deep recurrent neural network for abnormal eeg identification. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*, pp. 47–56. Springer, 2019.

Khadijeh Sadatnejad, Mohammad Rahmati, Reza Rostami, Reza Kazemi, Saeed S Ghidary, Andreas Müller, and Fatemeh Alimardani. Eeg representation using multi-instance framework on the manifold of symmetric positive definite matrices. *Journal of Neural Engineering*, 16(3):036016, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

Samuel L Smith, Andrew Brock, Leonard Berrada, and Soham De. Convnets match vision transformers at scale. *arXiv preprint arXiv:2310.16764*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

L Veloso, J McHugh, E von Weltin, S Lopez, I Obeid, and J Picone. Big data resources for eegs: Enabling deep learning research. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–3. IEEE, 2017.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022.

Xingyi Wang, Yuliang Ma, Jared Cammon, Feng Fang, Yunyuan Gao, and Yingchun Zhang. Self-supervised eeg emotion recognition models based on cnn. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1952–1962, 2023.

Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, 2024.

David Western, Timothy Weber, Rohan Kandasamy, Felix May, Samantha Taylor, Yixuan Zhu, and Luke Canham. Automatic report-based labelling of clinical eegs for classifier training. In *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6. IEEE, 2021.

Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022a.

Zhi Zhang, Sheng-hua Zhong, and Yan Liu. Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2022b.

A Additional Results

A.1 External Validation on the NMT Scalp EEG Dataset

We leverage the NMT Scalp EEG Dataset (Khan et al., 2022) in order to validate our results out-of-distribution. Models are trained only on TUEG data and their representations are subsequently evaluated using linear probes for abnormality classification without any finetuning (Table 5). The NMT dataset deviates considerably from TUEG. Data was recorded from a South Asian population at the Pak-Emirates Military Hospital, Rawalpindi, Pakistan, using a different EEG recording setup. Furthermore, the NMT participants are considerably younger, feature more males (66.6%), and their EEG recordings are labeled predominantly normal (83.8% in the training set, while the test set is balanced). This enables a challenging and imbalanced external validation for representation learning methods. We apply the same preprocessing as for TUEG and use the provided train/test split, yielding n=2216 and n=183 respectively. We observe that our ELM-MIL approach scores highest in all-but-one setting, indicating significant transferability of the learned representations. Without finetuning, the model even matches the best performance reported by Khan et al. (2022), who used supervised learning to train on TUH and finetune on NMT (Accuracy=82%, AUC=87%). Compared to TUAB, the notably lower 1% performance likely results from the heavily imbalanced dataset, meaning only 4 abnormal recordings are shown, compared to the 13-14 for TUAB 1%. The comparatively overall lower scores on the NMT dataset are in line with previous findings (Khan et al., 2022).

Table 5: Linear probing for abnormality classification on the NMT dataset using 1%, 10%, and 100% labeled training data. The (second) best SSL scores are printed (underlined) bold. Standard deviations over runs are included.

	Balanced Accuracy			AUROC		
Method	1%	10%	100%	1%	10%	100%
BYOL	57.94±1.07	68.97±1.05	71.30±2.20	63.78±1.70	76.48±2.10	80.65±2.50
ContraWR	58.25±1.16	66.43±0.93	67.76±0.38	<u>65.72</u>±1.01	72.47±0.95	75.42±1.01
VICReg	55.32±0.91	67.59±0.37	71.58±1.29	61.57±1.49	74.19±0.63	78.50±1.60
TS	57.64±1.07	**72.00**±1.50	<u>77.70</u>±2.29	64.90±0.70	<u>81.36</u>±1.53	<u>87.08</u>±1.02
RP	57.76±0.48	<u>71.45</u>±1.23	77.54±2.37	64.92±0.81	80.42±1.83	86.50±2.17
CPC	<u>58.89</u>±1.82	69.50±0.83	71.87±1.24	65.24±2.06	77.84±1.12	79.98±1.60
ELM-MIL e, l	**60.60**±0.54	68.57±0.90	**81.00**±1.18	**69.49**±2.26	**81.42**±1.15	**89.77**±0.21

A.2 EEG Event Detection

To further evaluate learned representations, we use the TUH EEG Seizure Corpus (TUSZ; Shah et al. (2018)) and TUH EEG Events Corpus (TUEV; Obeid & Picone (2016)), which are subsections of TUEG. Rather than recording-level predictions, these tasks require classification of single, short EEG crops. We pretrain models using 5-second EEG crops, drop subjects which feature in either TUSZ or TUEV yielding a pretraining sample size of n=14480 recordings, and train with lower learning rates for better stability (base learning rate of 0.02 for ELM-MIL and 0.1 otherwise). For ELM-MIL we increase the amount of positive EEG crops N to 120.

- **TUSZ**: This corpus has sections of recordings labeled to contain either seizure or background activity. We crop the recordings into 5-second segments and perform binary classification using 5-fold cross validation on the provided *train* and *dev* sets (n=6491), while testing on the *eval* set (n=865; Table 6).
  We find considerable performance differences between models, with BYOL and Temporal Shuffling performing well as EEG-only pretraining methods, while ELM-MIL scores highest across most settings.
- **TUEV**: A corpus containing annotated EEG with six classes, of which three are clinical (spike and slow wave (SPSW), generalized periodic epileptiform discharge (GPED), periodic lateralized epileptiform discharge (PLED)) as well as eye movements, artifacts, and background activity. We only use the provided *train* set (n=359) as the *test* set does not include the TUEG subject

Table 6: Linear probing for seizure classification on the TUSZ dataset using 1%, 10%, and 100% labeled training data. The (second) best SSL scores are printed (underlined) bold. Standard deviations over runs are included.

	Balanced Accuracy			AUROC		
Method	1%	10%	100%	1%	10%	100%
SV	56.98±1.68	65.97±2.30	77.35±1.30	84.00±0.90	87.33±0.93	90.38±0.39
BYOL	**70.95**±2.55	79.43±0.64	81.39±0.90	77.45±3.14	86.60±0.79	88.89±1.02
ContraWR	64.63±2.80	75.07±1.53	77.56±0.59	68.86±3.13	82.75±1.66	85.68±0.74
VICReg	63.30±1.27	74.36±0.99	78.02±1.18	69.61±1.82	82.01±0.93	86.17±0.98
TS	68.66±1.90	<u>80.11</u>±0.58	<u>82.92</u>±0.91	<u>78.60</u>±1.90	<u>87.63</u>±0.58	<u>89.77</u>±0.72
RP	59.04±1.51	69.48±1.14	72.27±1.30	64.41±2.50	76.72±1.37	79.52±1.26
CPC	64.72±1.49	74.34±1.84	78.25±1.99	72.01±1.00	81.77±2.24	85.91±2.00
ELM-MIL	<u>70.04</u>±4.07	**81.02**±0.95	**83.68**±0.46	**78.98**±5.18	**88.98**±0.86	**91.51**±0.33

Table 7: Linear probing for event classification (6-classes) on the TUEV dataset using 100% labeled training data. The (second) best SSL scores are printed (underlined) bold. Standard deviations over runs are included.

Method	Balanced Acc.	AUROC
SV	40.98±2.38	86.28±0.78
BYOL	46.19±2.39	82.40±1.60
ContraWR	**48.84**±1.19	84.11±1.64
VICReg	46.75±1.15	83.26±1.83
TS	45.00±1.66	<u>84.86</u>±1.32
RP	38.93±1.09	78.95±1.59
CPC	41.83±2.08	81.94±1.75
ELM-MIL	**48.84**±2.80	**87.69**±1.01

identifiers, which would have prevented the exclusion of these subjects from the pretraining data. By performing 5-fold cross validation while splitting on the subject level using the *train* set, we can guarantee to avoid data leakage. For each 1-second event, we include two seconds of context before and after, yielding 5-second crops.

In terms of overall performance (Table 7), ELM-MIL scores well across both metrics. Next, we investigated per-class performance as TUEV includes distinctly different event categories (Figure 4). We observe that ELM-MIL scores well across the three clinical events (SPSW, GPED, PLED) with over 3.5% better average scores. However, the model underperformed on artifact and eye movement detection, which may indicate models may lose sensitivity to events not described in the text. Interestingly, a portion of reports include sections on such technical problems, but these were segmented out for the current study. Follow-up research is needed to further investigate the effects of including such text.

A.3 Language-independent effects of sub-unit alignment.

Language-independent effects of sub-unit alignment. Given the broad outperforming of ELMs compared to EEG-only models, especially for $\text{ELM}_{e,l}$, we further investigate whether the general setup of multimodal pretraining provides inherent benefits. EEG recordings are split into multiple crops, which in turn are all aligned to the same clinical report during pretraining. It follows that EEG crops of a single recording are indirectly aligned to one another to some extent (Figure 1C). We investigated this hypothesis by shuffling reports between patients prior to pretraining. We find that while embeddings of single EEG crops of an untrained encoder are only minimally more similar within-subject than between-subject (ratio of $\sim$1.1x), this effect is much more pronounced after pretraining $\text{ELM}_{e,l}$ on correctly paired reports ($\sim$6.3x), and even more so after pretraining on shuffled reports ($\sim$15.7x; figure 5). Linear probing reveals that training $\text{ELM}_{e,l}$ on shuffled reports clearly boosts pathology detection over using an untrained encoder and manages to almost match EEG-only

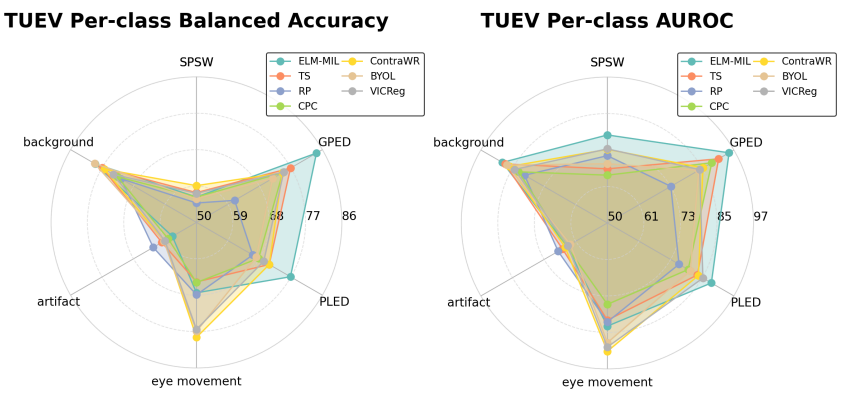

Figure 4: Per-class scores for TUEV show that ELM-MIL outperforms for the clinical events. SPSW: Spike and sharp wave, GPED: generalized periodic epileptiform discharges, PLED: periodic lateralized epileptiform discharges.

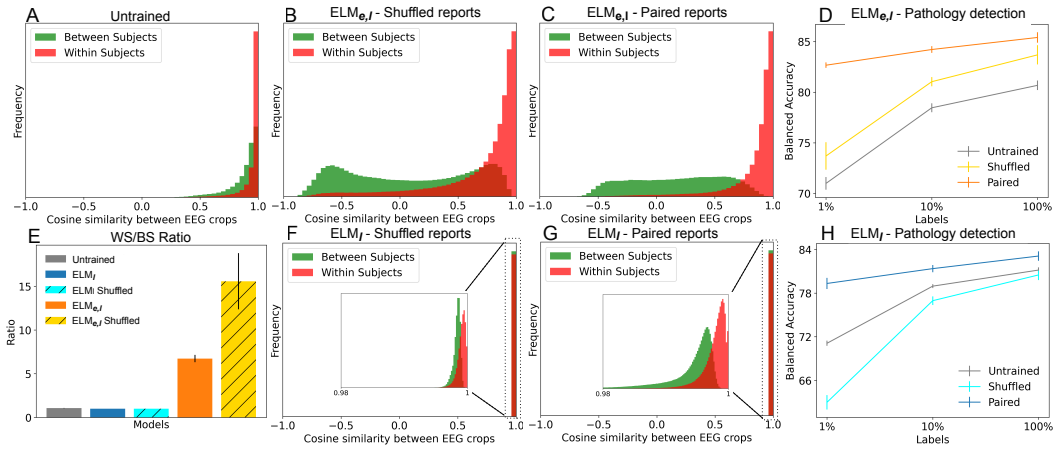

Figure 5: **A-C, E-G)** We investigate the distributions of cosine similarity values of EEG crop embeddings between- and within subjects (denoted BS and WS respectively). We plot these for an untrained encoder (one example run), as well as EEG encoders of ELMs trained with paired or shuffled reports. We find that ELM$_{e,l}$ produces dissimilar between-subject EEG embeddings, while ELM$_l$ does not. **E)** shows the ratio between WS and BS similarity values across five runs (with standard deviations). **D,H)** The downstream performance via linear probing is shown on the right, with error bars representing standard deviations across five training runs.

pretraining without the need for augmentations (mean accuracies of 73.70%, 81.04%, 83.69%). On the contrary, the ratios for ELM$_l$ are close to 1 after training using paired and shuffled reports, with the latter resulting in decreased pathology detection accuracy.

Conceptually, while shuffling reports destroys the semantic relevance of reports, it still provides a unique subject-specific reference to which the EEG embeddings are aligned to. Pretraining then reduces to promoting invariance to within-subject information, as all EEG crops of a patient are aligned to the same report. However, while for ELM$_l$ these reports occupy arbitrary positions in the latent language space due to the absence of the text projector, ELM$_{e,l}$ exhibits additional dynamics. Namely, for a given EEG crop (or text paragraph) in a batch belonging to subject i (that is, $id = i$), nearly all negative contrastive samples will belong to a different patient ($P(id = i) \ll P(id \neq i)$). The negative contrast therefore largely amounts to minimizing similarity between patients. This can be viewed as encoding between-subject information and these results imply that training with this objective is a useful pretext task for EEG timeseries. Naturally, this will depend on the downstream

tasks, but both retrieval and pathology detection require between-subject information. The advantage of retrieval and linear probing of $\text{ELM}_{e,l}$ may thus be, at least in part, due to the inherent utility of our extension of multimodal language modeling to timeseries by using sub-unit alignment, independent of language. Still, pathology detection with only few annotations is considerably better using paired reports, indicating the importance of relevant clinical language for label-efficiency.

A.4 Post-hoc investigation of data leakage

To maximize the amount of data available in this data-scarce setting, the TUAB training set was included during pretraining. We investigate whether this gave a disproportionate advantage to linear probes trained on ELM representations by repeating the "1% labels" context using unseen subjects as follows: Given only the TUAB test set, we train linear probes using 10-fold cross validation (times five random seeds), each time splitting 10-20-70% of the test set into train/validation/test. This gives the same labeled sample size as 1% of the TUAB training set without relying on samples seen during pretraining. As seen in Table 8, results are highly similar, strongly suggesting that the advantage of ELMs is not due to the inclusion of the TUAB training set in the pretraining set.

Table 8: Effect of overlap in subjects used for pretraining and linear probing. Higher standard deviations result from a smaller test set.

Method	Overlap	**Balanced Accuracy**
TS	Yes	$74.99_{\pm 0.86}$
TS	No	$74.56_{\pm 1.12}$
$\text{ELM}_{e,l}$ All Clusters	Yes	$82.64_{\pm 0.24}$
$\text{ELM}_{e,l}$ All Clusters	No	$82.28_{\pm 0.64}$

B Training Details

In this section, we provide further detailed information of the model training. Unless stated otherwise, ablation and hyperparameter analyses were performed on a data subset consisting of 5000 and 500 EEG recordings divided into a training and test set respectively. To prevent data leakage, this data had no overlap with the patients used for evaluation of the main results.

B.1 Optimization

All models are pretrained using the LARS optimizer (You et al., 2017) with a cosine decay learning rate schedule over 50 epochs, with a warm-up of 4 epochs. The base learning rate is set to 0.3 for EEG-only, 0.01 for ELMs, and 0.06 for ELM-MIL, scaled with the batch size (BaseLR $\times$ BatchSize/256; Grill et al. (2020)). We use a weight-decay parameter of 1×10^{-4}. Models were trained on either an Nvidia Geforce GTX 3090 or Tesla V100 GPU and require less than 24GB of memory. Training took approximately 9 hours for EEG-language modeling or 18 hours for EEG-only modeling due to data augmentations. We used CUDA v11.3 and PyTorch v1.12.1.

B.2 EEG Encoder

We use a CNN architecture with a residual stream as the EEG encoder for all analyses (Figure 6). The model uses parallel convolutions, involving reflection padding and 1D-convolutions with kernel sizes $\{4, 8, 16\}$ with 32 filters each. These outputs are concatenated, resulting in a 96 dimensional representation and 747K trainable parameters. We compare input lengths of EEG crops varying from 5 to 60 seconds. This presents a trade-off where longer crops result in a greater information content per crop, while reducing the total sample size. As EEG-only pretraining relies on data augmentations, this introduces an additional influence of crop length. Specifically, longer crop lengths likely make the pretraining task easier, as augmentations introduce relatively lesser distortion due to the greater information content. We therefore compare performance of different crop lengths for both EEG-language and EEG-only pretraining. As the EEG encoder progressively downsamples the signal, we adjust the pooling layers to the input length. These adjustments are shown in Table B.2. For EEG-language pretraining we evaluate zero-shot pathology detection, while for EEG-only pretraining

we are required to compare the performance of a linear probe. Results are shown in Figure 7. Due to computational resources, we only compare crop lengths for BYOL and ELM$_l$ as representations of EEG-only and EEG-language modeling. We observe that for EEG-only pretraining an intermediate crop-length of 20 seconds performs best, which matches the findings by Mohsenvand et al. (2020). Meanwhile, zero-shot pathology detection is found to be relatively insensitive to crop lengths of at least 10 seconds, with 60 second crops scoring highest, while the shortest crop length consistently leads to unstable text representations and chance-level performance.

For the EEG projector, we use a linear layer with an output dimension of 512 followed by batch normalization, exponential linear units, and a final linear layer with output size 256.

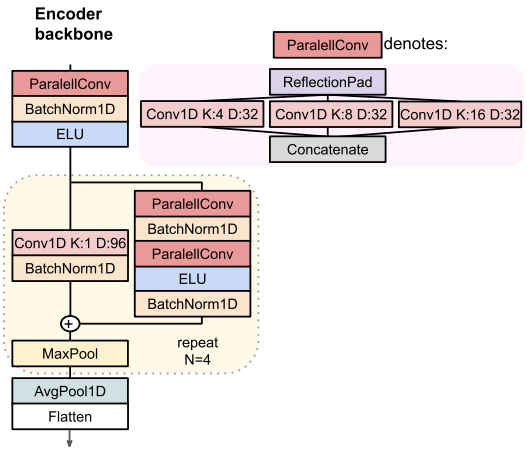

Figure 6: An identical EEG encoder architecture is used across all analyses. The size of the max pool operation depends on the input length. These are detailed in table B.2. K: Kernel size, D: Output dimensionality.

Table 9: Multiple input lengths for the cropped EEG timeseries were compared, which included adjustments to the pooling layer.

Model Setups			**Batch Size**	
Input Dim.	Max Pool Size	Intermediate Dim.	EEG+Text	EEG
500	[2,2,2,2]	[166, 55, 18, 6]	2048	2048
1000	[3,3,3,3]	[333, 111, 37, 12]	2048	2048
2000	[3,3,3,3]	[666, 222, 74, 24]	2048	1024
3000	[4,4,4,4]	[750, 187, 46, 11]	1024	800
6000	[4,4,4,4]	[1500, 375, 93, 23]	800	400

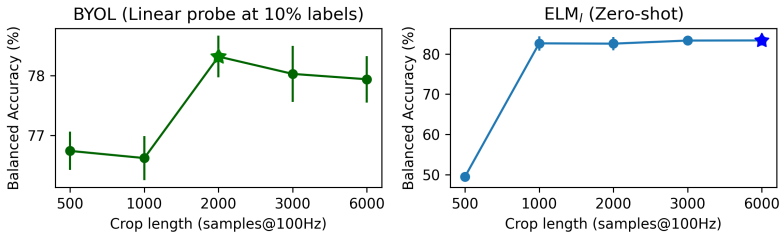

Figure 7: Comparison of pathology detection based on EEG input crop length, ranging from 5 to 60 seconds, via averaged balanced accuracy scores. Error bars indicate the standard deviation across five random seeds.

B.3 LANGUAGE ENCODER

We compare three pretrained language models in their ability to perform zero-shot pathology detection following EEG-language pretraining (Table 10). We find that MedCPT performs best (Jin et al., 2023), which is trained using contrastive learning with 255 million user click logs from PubMed.

For the text projector of $\text{ELM}_{e,l}$, we use a linear layer with output size 1024 followed by batch normalization, rectified linear units, and a final linear layer with output size 256 and batch normalization.

Table 10: Zero-shot classification comparison between language models for $\text{ELM}_{e,l}$.

Language Model	**Balanced Accuracy**	**AUROC**
BiomedBERT (Gu et al., 2021)	78.61±2.90	85.78±2.58
Bio-ClinicalBERT (Alsentzer et al., 2019)	80.86±1.19	87.33±0.68
MedCPT (Jin et al., 2023)	**82.58**±0.25	**88.37**±0.39

B.4 TEMPERATURE PARAMETER

For $\text{ELM}_{e,l}$, the softmax operation used in the loss computation includes a temperature hyperparameter τ. We compare zero-shot pathology detection for multiple values. We observe poor performance for low temperature values, but stable zero-shot classification for higher parameter values. We set $\tau = 0.3$ for all further analyses.

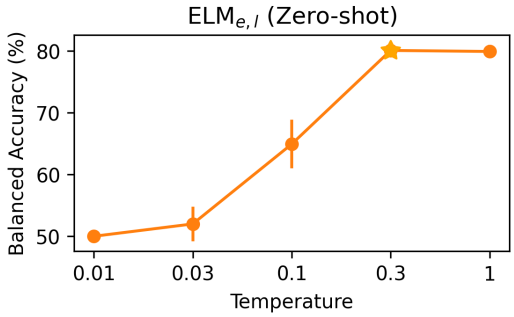

Figure 8: Comparison of temperature values for $\text{ELM}_{e,l}$ on zero-shot pathology detection. Error bars indicate the standard deviation across three random seeds.

B.5 EEG-ONLY PRETRAINING

We implement the following methods for EEG-only SSL:

Bootstrap-Your-Own-Latent. BYOL relies on two encoder models: an online and a target network (Grill et al., 2020). During pretraining, the online network is trained to predict the target model's output. Meanwhile, the weights of the target network are updated using a moving average of the weights of the online network, which has been empirically shown to prevent collapse of the latent space. For alignment, ℓ_2 normalization is applied to the EEG embeddings $\{\mathbf{h}'_e, \mathbf{h}''_e\}$ and the mean square distance is minimized. We adopt the recommended parameter value for the exponential moving average (Grill et al., 2020). The projection head is a 2-layer non-linear MLP with a hidden dimension of width 256 and an output dimension of 32.

Variance-Invariance-Covariance Regularization. VICReg allows for the use of a single encoder model and prevents collapse by applying two explicit regularization terms to each of the embedding batches $\{\mathbf{h}'_e, \mathbf{h}''_e\}$ (Bardes et al., 2021). The 'variance' term maintains the standard deviation (computed batch-wise) of every embedding dimension above a threshold, thereby avoiding a trivial solution. In addition, latent collapse is avoided through the 'covariance' term which decorrelates pairs of embedding dimensions. The method minimizes the mean square distance between $\{\mathbf{h}'_e, \mathbf{h}''_e\}$.

Hyperparameters are set to their recommended values (Bardes et al., 2021). The projection head is a 2-layer non-linear MLP with a hidden dimension of width 256 and an output dimension of 256.

Contrast with the World Representation. ContraWR was proposed to improve augmentation-based SSL for EEG (Yang et al., 2021). The method, which is contrastive in nature, maximizes similarity between $\{\mathbf{h}'_e, \mathbf{h}''_e\}$ while preventing collapse by minimizing similarity with 'negative' samples. ContraWR forms a negative representation by aggregating across all negative batch elements, aiming to compensate for the low signal-to-noise of EEG data by creating a more reliable negative contrast. It relies on a triplet loss based on Info-NCE (Gutmann & Hyvärinen, 2010). We also here set the hyperparameters to the values recommended by the authors (Yang et al., 2021). The projection head is a 2-layer non-linear MLP with a hidden dimension of width 256 and an output dimension of 32.

Relative Positioning. Pairs of EEG crops are sampled and assigned binary labels based on their temporal proximity (Banville et al., 2021). Crops close in time are labeled positive, while those far apart are labeled negative. We use the same EEG encoder as for all other methods to create representations and use the suggested contrastive module to compute the element-wise absolute difference between representations. A logistic regression model then predicts the label. The method is trained using binary logistic loss. For all methods by (Banville et al., 2021), we use the hyperparameters reported to work best on TUAB, including between-subject sampling of EEG crops.

Temporal Shuffling. An extension of Relative Positioning by sampling triplets of EEG crops. The task is to determine whether the crops are in temporal order or shuffled (Banville et al., 2021). The contrastive module concatenates absolute differences between representations. As with Relative Positioning, a logistic regression model is used for prediction, and the method is trained end-to-end using binary logistic loss.

Contrastive Predictive Coding. This method uses an autoregressive encoder to summarize a sequence of EEG crops into a context vector (Banville et al., 2021). The task is to predict which future crop actually follows the context, among negative samples. A bilinear model is used for prediction at each future step. The method is trained end-to-end using the InfoNCE loss.

B.5.1 Data augmentations

For EEG-only pretraining, we adapt the data augmentations proposed by Mohsenvand et al. (2020), which were found to perform well on the TUAB dataset. For a given EEG crop, we apply the same augmentation to each channel. Parameters are sampled independently for each EEG crop and uniformly from the ranges displayed in table 11. Augmentations are visualized for a single EEG channel in figure 9.

Table 11: Data augmentation parameter ranges; adapted from Mohsenvand et al. (2020).

Data Augmentation	Min	Max
Amplitude Scale	0.5	1.5
Time Shift in samples	-60	60
DC shift in microvolts	-10	10
Zero-Masking in samples	0	200
Additive Gaussian Noise (sigma)	0	0.2
Band-Stop Filter (5Hz width, Hz)	2.8	47

C Details on Data Subsampling

To alleviate class imbalance in the TUEG dataset, we perform data subsampling. We rely on the LLM summaries of reports, which were more consistent in their phrasing regarding the normal or abnormal status. This allowed for a more reliable classification using regular expressions. All reports for which no clear classification was made were omitted. 5015 reports in the potential training set were classified as normal, which were associated with 7526 EEG recordings. For our 'pretrain' data subset, we subsampled the abnormal EEGs to match the amount of normal EEG recordings. This resulted in 7526 abnormal EEG recordings, with 6770 reports. Although only a minor subset of these

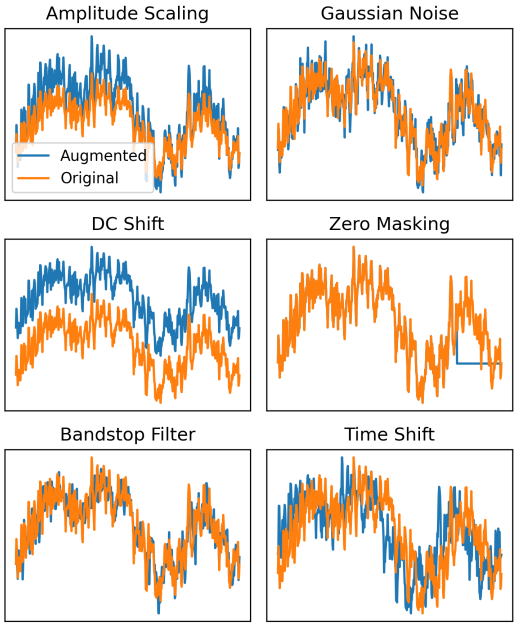

Figure 9: Data augmentations visualized for a single channel of EEG data.

preliminary classifications was manually verified, it is important to note that this process was solely to alleviate severe class imbalance and was not used for further analysis.

For EEG-language modeling, the pretrain subset was effectively smaller, as a report had to be omitted from pretraining when it did not contain at least one relevant heading. Out of the 15144 total EEG files, this resulted in pretrain sample sizes of: 14836 (clinical history cluster), 14320 (medication cluster), 14800 (description cluster), 14794 (interpretation cluster), and 14946 (all clusters).

To test for retrieval performance, we supplemented the TUAB test set with data from the TUH EEG Epilepsy Corpus (Veloso et al., 2017) in an attempt to create a larger, roughly balanced evaluation set of those with and without pathology. For this, we only selected the first recording of a subject so that no multiple files from the same subject were present. Additionally, we only included reports which had at least one heading from each text cluster to allow for a fair comparison.

D CLASSIFICATION

To study the predictive capability of learned representations after pretraining, we train linear probes and perform zero-shot classification.

Linear probe

For linear evaluation, we train logistic linear regression models using 10-fold cross validation for each pretrained model using sklearn (Pedregosa et al., 2011). We perform grid-search over 45 logarithmically-spaced values for L2 regularization between 10^{-6} and 10^{5} via a validation set.

Supervised Learning

For the supervised learning baseline, we use the identical EEG encoder backbone as used for all other analyses and use 60 second crops. We add an MLP (hidden dimensionality of 256) with dropout $p = 0.5$ and output dimensionality equal to the amount of classes. The ADAM learning rate is set to 0.001 and we use the validation set to select weight decay out of $[0.1, 0.01, 0.0001]$. We use a batch size of 256 and train using the cross entropy loss. When using 100% labels, we first train on the training set for up to 50 epochs (with early stopping after 5 epochs without improvement) and select the epoch which resulted in the best validation loss. Subsequently, we continue training on the train and validation sets together until the loss has decreased below the best validation loss.

Zero-shot classification

For zero-shot pathology detection, we perform an ensemble over 21 binary prompts, listed in Table 12. Prompt ensembling was shown to improve performance (Radford et al., 2021), but we employ it here also as the limited data is likely to lead to less stable representations, which may lead to sensitivity to phrasing. To inspect whether results are sensitive to changes to the prompt set, we perform a post-hoc analysis using the held-out test set that iteratively leaves one prompt out of the ensemble (Figure 10). We observe that results are consistent across such reduced prompt sets, except for the ELM_l model trained on the clinical history or interpretation clusters, although neither model reaches competitive performance. This set was only initially verified on the training set to enable model- and parameter-comparisons using zero-shot performance. Tuning is likely to enable further performance improvements, although the flexibility of the zero-shot approach may introduce severe risk of overfitting on the TUEG dataset.

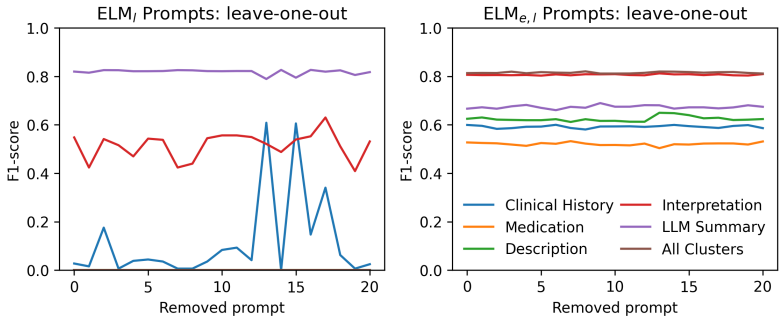

Figure 10: Analysis of the sensitivity to prompts in the ensemble used for zero-shot classification. We plot the average F1-score across five random seeds. Note that for ELM_l, multiple models have a consistent F1-score of 0 and are therefore not individually visible.

Table 12: Prompt ensemble used for zero-shot classification.

Normal EEG Prompts	**Abnormal EEG Prompts**
Normal EEG.	Abnormal EEG.
No pathology present.	Pathology present.
No abnormalities.	Abnormalities observed.
Normal routine EEG.	Markedly abnormal EEG.
Normal awake record.	Abnormal awake record.
Normal EEG record.	Abnormal EEG record.
This EEG is normal.	This EEG is abnormal.
This is a normal EEG.	This is an abnormal EEG.
This EEG is within normal limits	This EEG is mildly abnormal.
Normal awake EEG.	Abnormal awake EEG.
Normal asleep EEG.	Abnormal asleep EEG.
Normal awake and asleep EEG.	Abnormal awake and asleep EEG.
Normal EEG in wakefulness and drowsiness.	Abnormal EEG in wakefulness and drowsiness.
No pathology.	Abnormal EEG due to:
EEG shows no pathology.	Abnormal EEG for a subject of this age due to:
No abnormalities.	Abnormalities in the EEG.
No abnormalities observed.	Abnormalities observed.
EEG shows no abnormalities.	EEG shows abnormalities.
No clinical events detected.	Clinical events detected.
No indications of pathology observed.	Indications of pathology observed.
The EEG is normal.	The EEG is pathologically abnormal.

E CLINICAL IMPLICATIONS

To better understand the clinical utility of the learned EEG representations, we conducted additional experiments using 5-second EEG segments from the TUSZ and TUEV datasets. The strong performance across both recording-level classification and event detection tasks suggests that our model learns clinically relevant features at multiple temporal scales. With further progress, this capability may support various future clinical applications, from rapid screening of prolonged recordings to real-time event detection. Whereas the present study focuses on establishing an initial application to explore viability, future work may benefit from focussing on improving the interpretability of these representations through techniques such as channel-specific attribution. Additionally, the multimodal nature of our approach opens possibilities for automated report generation, which could assist in clinical documentation while maintaining human oversight. Certain clinical limitations also deserve further attention, such as a careful study of how the frequency of specific pathology and clinical events in reports impacts model performance. Finally, biases present in language models may impact multimodal pretraining, which should be investigated in future work.

F ADDITIONAL VISUALISATIONS

F.1 EEG EMBEDDINGS OF PATHOLOGY

We provide t-SNE (complexity=40, (Van der Maaten & Hinton, 2008)) visualisations of the averaged EEG embeddings per subject after pretraining. These are post-hoc plots for which we use models trained on the entire pretraining subset and display embeddings of hold-out TUAB patients. $\text{ELM}_{e,l}$ and ELM-MIL show the clearest visual separation between abnormal and normal EEGs, which is in line with the linear probing results.

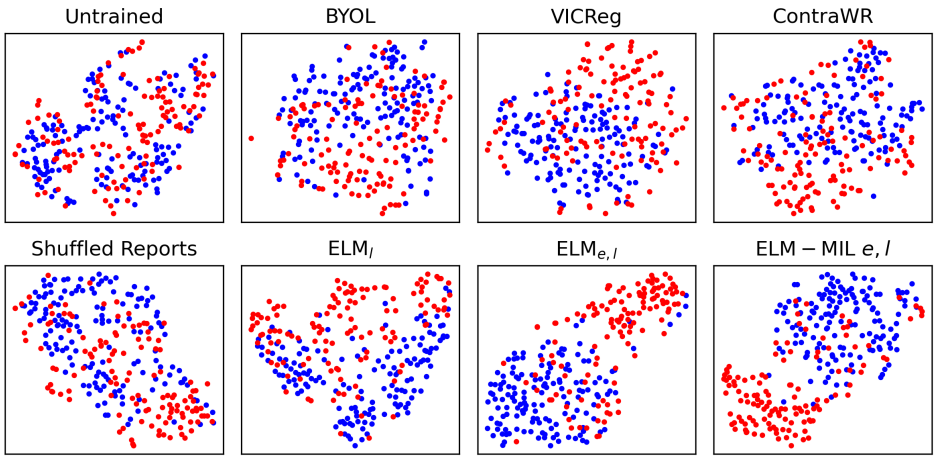

Figure 11: Example EEG embeddings averaged within-subject of pretrained models on the TUAB hold-out data (red: abnormal, blue: normal). The data is projected using t-SNE. The 'untrained' and 'shuffled reports' plots feature the same setup as the $\text{ELM}_{e,l}$ model, with the latter being trained on reports randomly shuffled between subjects.

F.2 WITHIN-SUBJECT EEG EMBEDDINGS

We provide additional visualizations of t-SNE projections of EEG crops (Figure 12). Specifically, we compare $\text{ELM}_{e,l}$ using InfoNCE and ELM-MIL using MIL-InfoNCE across three temperature parameters $\tau = [0.1, 0.3, 1.0]$. To do so, we randomly sample three normal (blue shades) and three abnormal (red shades) subjects. We observe that whereas both methods exhibit diminished subject clustering at a higher temperature ($\tau = 1.0$), at low temperatures ($\tau = 0.1$) this only occurs for InfoNCE. Meanwhile, subject clustering gets more pronounced for MIL-InfoNCE. This may explain

the observation that retrieval performance increases by reducing τ for MIL-InfoNCE, which as a task requires subject rather than class separation per se.

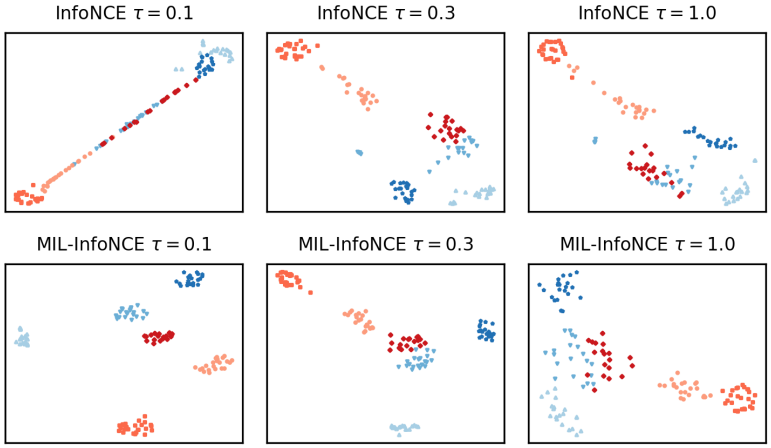

Figure 12: A comparison of subject clustering using t-SNE projections of embeddings of EEG crops. Red (blue) shades indicate three randomly sampled abnormal (normal) subjects.

F.3 Report and content segmentation

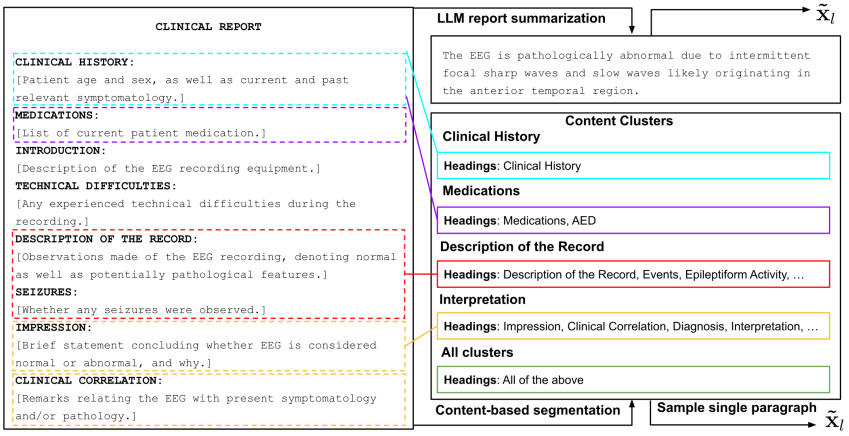

Figure 13: An example set of headings which may make up a clinical report. Paragraphs are extracted from the reports into content-based clusters or an LLM-generated summary.