# Neural decoding from stereotactic EEG: accounting for electrode variability across subjects

Georgios Mentzelopoulos[1, *], Evangelos Chatzipantazis[1], Ashwin G. Ramayya[2], Michelle J. Hedlund[2], Vivek P. Buch[2], Kostas Daniilidis[1, 3], Konrad P. Kording[1], and Flavia Vitale[1, *]

[1]University of Pennsylvania, [2]Stanford University, [3]Archimedes, Athena RC

## Abstract

Deep learning based neural decoding from stereotactic electroencephalography (sEEG) would likely benefit from scaling up both dataset and model size. To achieve this, combining data across multiple subjects is crucial. However, in sEEG cohorts, each subject has a variable number of electrodes placed at distinct locations in their brain, solely based on clinical needs. Such heterogeneity in electrode number/placement poses a significant challenge for data integration, since there is no clear correspondence of the neural activity recorded at distinct sites between individuals. Here we introduce seegnificant: a training framework and architecture that can be used to decode behavior across subjects using sEEG data. We tokenize the neural activity within electrodes using convolutions and extract long-term temporal dependencies between tokens using self-attention in the time dimension. The 3D location of each electrode is then mixed with the tokens, followed by another self-attention in the electrode dimension to extract effective spatiotemporal neural representations. Subject-specific heads are then used for downstream decoding tasks. Using this approach, we construct a multi-subject model trained on the combined data from 21 subjects performing a behavioral task. We demonstrate that our model is able to decode the trial-wise response time of the subjects during the behavioral task solely from neural data. We also show that the neural representations learned by pretraining our model across individuals can be transferred in a few-shot manner to new subjects. This work introduces a scalable approach towards sEEG data integration for multi-subject model training, paving the way for cross-subject generalization for sEEG decoding.

## 1 Introduction

Deep learning has revolutionized many fields ranging from natural language processing [Vaswani et al., 2017], to computer vision [Dosovitskiy et al., 2020], and neural decoding [Azabou et al., 2023]. Advances in these fields have shown that performance and generalization benefit from scaling up both datasets and model size, underscoring the importance of integrating data across multiple individuals. Recent work by Azabou et al. [2023] showed that such approaches can achieve excellent generalization in the motor decoding domain, by combining invasive microelectrode array recordings from multiple non-human primates. Despite their promise, however, such approaches have not been used to build scalable decoders for human brain recordings collected with stereotactic electroenchephalography (sEEG). Compared to other approaches (e.g., ECoG, microelectrodes), sEEG is minimally invasive and is currently the gold-standard clinical tool for invasive electrophysiology in humans. Therefore, building strong neural decoding models using sEEG promises to be medically relevant.

[*]Contact: gment@upenn.edu, vitalef@pennmedicine.upenn.edu
Project page: https://gmentz.github.io/seegnificant

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

There are several challenges associated with decoding from sEEG data. A major barrier is the heterogeneity of the sEEG cohorts. SEEG subjects are most often patients suffering from pharmacoresistant epilepsy that are implanted with sEEG electrodes as part of their medical care. Each subject gets implanted with a variable number of electrodes in specific locations in their brain, solely based on clinical needs [Kovac et al., 2017]. This heterogeneity makes it hard to identify any correspondence between the neural activity recorded across different subjects. This lack of correspondence poses a significant challenge to combining data across subjects in a meaningful way, resulting in most research groups resorting to within-subject models [Wu et al., 2024, Angrick et al., 2021, Petrosyan et al., 2022, Meng et al., 2021]. While somewhat effective, single-subject approaches neither scale nor generalize, suggesting a need to effectively integrate multi-subject data.

Here we introduce seegnificant: a scalable framework that can be used to decode behavior from sEEG recordings using multi-session, multi-subject data. Our approach uses a unified feature extraction backbone, designed to extract global behaviorally relevant neural representations that are shared across subjects. It uses personalized task-heads, tailored to the idiosyncrasies of individual participants, that can be used for downstream decoding tasks. To account for the heterogeneous electrode placement across individuals, we use a convolutional tokenizer that operates within electrodes individually. The tokenized latents are then processed by self-attention within the time dimension to capture long-term temporal dependencies within electrodes. Using a novel positional encoding scheme, we then imprint the latents of each electrode with information about their 3D location in the brain, using their MNI coordinates, a standarized 3D coordinate system used to localize brain regions [Talairach and Tournoux, 1988, Collins et al., 1994]. The spatially aware latents are then processed by another self-attention that operates within the electrode dimension to capture long-range dependencies across electrodes. The resulting latents, are then compressed into neural representations that capture the spatiotemporal dependencies within and between electrodes that are used for behavioral decoding.

To evaluate our approach, we combine data from 21 subjects, across 29 recording sessions, that performed a behavioral reaction time task, while their sEEG was simultaneously recorded, totaling more that 3600 behavioral trials and 100 electrode-hours (number of electrodes $\times$ number of recording hours) of sEEG recording. We demonstrate that using our approach, we are able to extract global neural representations, capable of decoding the response time of the participants during the behavioral task from neural data. Using a leave-one-out cross validation scheme, we also show that by pretraining our model on large amounts of data, we can transfer the learned neural representations to new subjects with minimal training examples. Our work introduces a novel framework designed to train models on multi-session, multi-subject sEEG data for behavioral decoding applications in a scalable way.

Our contributions can be summarized as:

- *A framework for multi-subject training based on sEEG.* We present a novel approach to training transformer based models for neural decoding based on the combined data of multi-subject, multi-session sEEG datasets.
- *Pre-trained models for behavioral decoding based on sEEG.* We trained a multi-session, multi-subject model for response time decoding based on sEEG that can be finetuned to new subjects. We will make the model and code publicly available as a resource for the community.

## 2 Related Work

### 2.1 Neural decoding using sEEG

Over the past decade, significant progress has been made in the field of neural decoding using sEEG. Angrick et al. [2021] demonstrated effective real-time speech synthesis based on sEEG using a linear discriminant analysis approach, and Petrosyan et al. [2022] showed effective speech decoding using a convolutional architecture. Meng et al. [2021] identified discriminative features for decoding speech perception and overt and imagined speech production from sEEG using a non-deep learning model. Wu et al. [2022] showed that grasp force can be successfully decoded from sEEG signals using a CNN+RNN decoder and subsequently performed a comparative study where they explored the efficacy of using a variety of convolutional architectures to classify different movement based on the sEEG neural activity [Wu et al., 2023, 2024]. Unfortunately, all aforementioned works conduct experiments on datasets containing only a few subjects (12 or less, except Wu et al. [2023]) and build

within-subject models, with no clear vision towards generalization across subjects. Diverging from within-subject approaches, in this work we combine data from many subjects to build models capable of extracting global neural representations that generalize across subjects. Contrary to previous works which used CNNs and/or RNNs, we build our models using the transformer architecture, which has been shown to excel in capturing long-range dependencies when trained on large and diverse datasets.

### 2.2 Transformer architectures for neural data

Transformers have attracted a lot of attention in the analysis of non-invasive neural signals. In the context of non-invasive electroencephalography (EEG), multiple works use the attention mechanism for various decoding tasks such as motor-imagery, visual decoding, and emotional recognition [Tang et al., 2024, Kan et al., 2023, Li et al., 2023, Liu et al., 2022, Lan et al., 2020]. Song et al. [2023] achieved SOTA performance in a visual decoding task using a convolutional tokenizer that operates within both time and electrode dimensions and is then processed using self-attention. Using the same tokenization approach, Cui et al. [2023] built neuro-GPT, a foundational model for non-invasive EEG based on transformers. However, non-invasive EEG uses a predefined number of electrodes placed at consistent locations of the scalp across subjects. This approach provides a well-defined correspondence across the neural activity recorded between subjects. In sEEG, the number and placement of electrodes across subjects is highly variable and inconsistent between subjects, complicating efforts to translate approaches from EEG to sEEG data.

In the context of invasive microelectrode recordings, prior works have attempted to use transformers for neural decoding with success. Ye and Pandarinath [2021] introduced the Neural Data Transformers (NDT) to model neural population activity as an alternative to RNN models [Glaser et al., 2020, Sussillo et al., 2016]. Le and Shlizerman [2022] extended the NDT model to process data in both the electrode and time dimensions. However, all aforementioned works trained single-subjects models. Azabou et al. [2023] diverged from single subject approaches by combining data from multiple non-human primates and achieved excellent decoding performance in a variety of different motor decoding tasks. However microelectrode recordings have a well-defined tokenization scheme based on extracellular neuronal spikes, which are not detectable by mm-scale sEEG electrodes. Thus, it is unclear how spike-based tokenization approaches can be meaningfully translated to sEEG.

### 2.3 Shared trunk architectures

Shared trunk architectures are a popular choice for multi task learning applications. They follow a simple outline: a global feature extractor, whose parameters are shared across tasks, followed by task-specific branches, with parameters specific to each task [Crawshaw, 2020]. They have been successfully used in multiple domains including facial recognition [Zhang et al., 2014], instance-aware semantic segmentation [Dai et al., 2015], image retrieval [Zhao et al., 2018] and classification [Liu et al., 2018]. Their effectiveness stems from their ability to extract shared representations that are common to all tasks. Combining data from multiple different tasks also helps mitigate the large scale data requirements for effective training of deep networks. In this work, while we are not dealing with multiple different tasks, we use a shared trunk architecture to model inter-person differences between subjects in our cohort. Our network is composed of a backbone, common to all subjects, that extracts global representations and subject-specific heads that tailor the model's output to the unique statistical profile of each subject.

## 3 Methodology

SEEG is a neural recording modality capable of recording local field potentials (LFPs) that reflect the coordinated activation of hundreds of thousands of neurons in the vicinity of an electrode, providing a mesoscale measurement of brain activity [Saez and Gu, 2023]. Due to the distributed nature of sEEG, subjects typically get implanted with tens to hundreds of electrodes in widespread locations across their brain. The resulting data are multi-variate time series that are typically epoched around a behavioral stimulus of interest (e.g. a visual stimulus presented during each trial of a behavioral task).

Neural activity recorded via sEEG across the electrodes of an individual are, of course, not independent. The captured recordings represent a conversation between thousands of neurons in distributed brain networks. The challenge is to find a way to interpret those recordings in the context of this

broader conversation, instead of interpreting them in isolation [Azabou et al., 2023]. This challenge is exacerbated when trying to integrate neural recordings across different subjects, each of which has a variable number of electrodes distributed across different brain regions. This is equivalent to monitoring the neural conversation with several microphones placed in different brain locations unique to each subject. For each monitored neural population, we do not know the identity or the functional tuning (what stimuli they respond to).

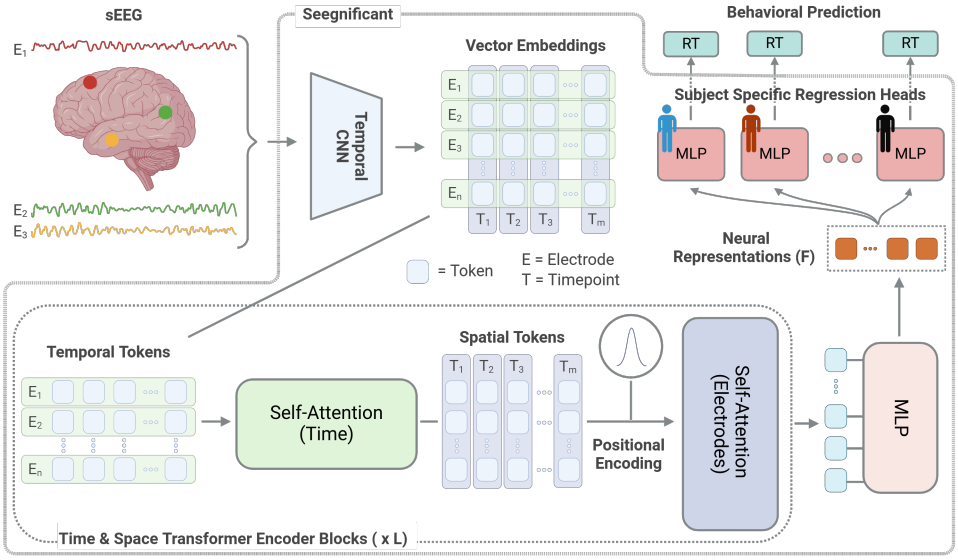


Figure 1: *Outline of our network architecture*. The sEEG signals are converted to vector embeddings through temporal convolutions and then processed by sequential self-attention operations in the time and electrode dimensions in an alternating fashion. The latents are compressed and projected through subject-specific task heads to obtain behavioral predictions.

### 3.1 Signal processing

Motivated by the lack of correspondence between the neural activity recorded between subjects, we sought to identify the subset of electrodes across subjects that respond to a behavioral stimulus of interest (i.e., share the same functional tuning in the neural conversation) while subjects perform a behavioral task. This can be achieved using the methodology introduced by Coon and Schalk [2016] and refined by Paraskevopoulou et al. [2021]. This methodology identifies electrodes whose high-$\gamma$ band activity (narrowband neural activity with frequency content between 70-150 Hz) is modulated by the behavioral stimulus of interest. We note that high-$\gamma$ band activity is used because it has been associated with the instantaneous firing of neuronal populations in the vicinity of an electrode [Miller et al., 2009] and, thus, it likely represents neural computation that is relevant to the behavioral stimulus of interest.

As described by Paraskevopoulou et al. [2021], separately for each electrode, the neural signals are filtered to the high-$\gamma$ band and the envelope is extracted using the Hilbert transform. The high-$\gamma$ envelopes are then epoched around the behavioral stimulus of interest and separated into a baseline period (short time window prior to stimulus) and task period (short time window after the stimulus). The median high-$\gamma$ envelope of the task ($Median_{task}$) and baseline ($Median_{baseline}$) periods are computed across the trials of the behavioral task. Then for each electrode, the signal-to-noise ratio (SNR) is calculated as

$$SNR = \frac{\sigma^2(Median_{task})}{\sigma^2(Median_{baseline})}, \tag{1}$$

where $\sigma^2(\cdot)$ refers to the variance of $(\cdot)$. Electrodes whose high-$\gamma$ band activity is significantly modulated by the behavioral stimulus can then be identified using a bootstrap randomization test and subsequently be used for decoding applications.

The implementation details of this analysis can be found in Appendix A.1.

### 3.2 Network architecture

#### 3.2.1 Tokenization

To combat the inhomogeneity of the number/placement of electrodes across different individuals, we designed a convolutional tokenizer that operates on the voltage traces of each electrode separately. Instead of performing a two-dimensional convolution across both the time and electrode dimension, which would blend information between the dissimilar electrodes that are randomly placed across individuals, we perform a one-dimensional convolution across the temporal dimension only. This way, each electrode gets assigned a $K$-dimensional learnable embedding for chunks of time, defined by the length of the convolutional kernel.

Specifically, let $x_e \in R^{T_{trial}}$ denote the LFP recorded by electrode $e$ for a given trial of length $T_{trial}$. Let $k_i$ denote the $i^{th}$ convolutional kernel, from a collection of $K$ kernels. Our approach returns learnable vector embedding $z_e \in R^{T_{trial} \times K}$ by performing convolutions of the form,

$$z_{e,i} = x_e * k_i, \text{ for } i \in 0, 1, ..., K, \tag{2}$$

on $x_e$ (convolutional kernel weights are shared across all the electrodes of all subjects). After temporal convolutions, batch normalization is applied to the $K$ features extracted by the convolution followed by average pooling along the time dimension (which reduces the time dimension from $T_{trial}$ to $T$), returning $z_e \in R^{T \times K}$.

Stacking the results of these operations across electrodes, we obtain the latent $z \in R^{E \times T \times K}$, where $E$ denotes the number of electrodes, $T$ is the number of time samples, and $K$ is the number of features extracted by the convolutions. By representing our data in this way, our network is capable of handling inputs with varying numbers of electrodes and time lengths. We do note, however, that the time resolution (i.e., sampling rate) needs to be fixed for the convolutions to be meaningful across subjects.

#### 3.2.2 Capturing long-range spatiotemporal dependencies

The latent $z \in R^{E \times T \times K}$ semantically represents the cumulative neural activity of all electrodes of a subject, for a given trial. It can also be conceptualized as a separable dual tokenization across the time and electrode dimensions, with electrode-latents $z_e \in R^{T \times K}$ and time-latents $z_t \in R^{E \times K}$. This separability, enables us to process the vector embeddings separately in the time and electrode dimension, giving us a significant computational advantage to subsequent self-attention layers, whose computational complexity scales quadratically with sequence length.

**Attention in the time dimension.** We first interrogate our data for long-term temporal dependencies within electrodes. To do so, for each electrode separately, we arrange the latents $z_e \in R^{T \times K}$ into a sequence of $T$ tokens, where each token has dimensionality $K$. The tokens are then projected into equally shaped queries: $Q = W_q z_e$, keys: $K = W_k z_e$, and values: $V = W_v z_e$, and processed using self-attention,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{3}$$

For this operation we used the standard transformer block [Vaswani et al., 2017], preceding the attention operation with normalization layers and following with a feed-forward network. This operation is parallelized across electrodes, and we stack the results to obtain latents of the form $z_{int} \in R^{E \times T \times K}$.

**Spatial positional encoding.** To account for the inconsistency of electrode locations across individuals, there is a need to inject information about the spatial location of each electrode into our model. To do so, we used radial basis functions to capture the information about each electrode location in the brain based on its MNI coordinates [MacDonald et al., 2000, Talairach and Tournoux, 1988]. We discretize the space of each coordinate, into $n$ bins, with midpoints $\mu_0, \mu_1, \ldots, \mu_{n-1}$. For each midpoint, we centered $m$ univariate gaussians of the form,

$$f_{i,j}(s) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(s-\mu_i)^2}{2\sigma_j^2}\right), \tag{4}$$

with variances $\sigma_j^2$, for $i = 0, 1, \ldots, n-1$ and $j = 0, 1, \ldots, m-1$. Separately for the $x$, $y$, and $z$ coordinates, we compute $f_{i,j}(x)$, $f_{i,j}(y)$, and $f_{i,j}(z)\ \forall\ i, j$, which we then concatenate into a vector $p \in R^{3 \cdot m \cdot n}$, which represents the positional encoding of an electrode. The positional encodings for each electrode are then projected to $K$ dimensions (using a linear layer) and added to the latents $z_{int} \in R^{E \times T \times K}$.

**Attention in the space dimension.** Having captured temporal dependencies within electrodes in our data, and having "stamped" the latents with their spatial positional encodings, we then use self-attention again, this time in the electrode dimension to capture long-range dependencies between the neural activity across electrodes in our data. The latents $z_{int} \in R^{E \times T \times K}$ are arranged, separately for each timepoint, into $z_t \in R^{E \times K}$ as a sequence of $E$ tokens of dimension $K$, which are projected into equally shaped queries: $Q = W_q z_t$, keys: $K = W_k z_t$, and values: $V = W_v z_t$. The tokens are processed with self-attention using equation 3. Again, the standard transformer block is used, with appropriate pre-normalization followed by a feed-forward network. This time, the operation is parallelized across timepoints, and the results are stacked to obtain latents of the form $z_l \in R^{E \times T \times K}$.

Multiple layers of self-attention in time and self-attention in space can be stacked back-to-back to enhance the fitting ability of the network, if necessary. Leveraging the versatility of transformers to accept inputs of varying lengths, the attention operation in the both time and space, as described above, does not constrain our model to accept inputs of fixed length in terms of the number of electrodes, or timepoints. In the case where inputs do have unequal number of electrodes or timepoints, computations can be efficiently parallelized by masking tokens that correspond to padded electrodes and timepoints during the self-attention operations.

**Extracting global features**. In behavioral neuroscience, experimenters are usually interested in decoding few behavioral variables, such as a response time or hit rate to a given stimulus [Posner and Petersen, 1990, Galvao-Carmona et al., 2014]. With this in mind, the information contained in the latent $z_l \in R^{E \times T \times K}$, will ultimately need to be compressed into a single value. Therefore, following the self-attention operations, we unroll the latents $z_l$ and, using a feed-forward network, we project them to a low dimensional neural representation $F \in R^d$, where $d$ denotes the final number of the extracted features, with $d \ll E \cdot T \cdot K$, that will be used for the downstream decoding tasks.

#### 3.2.3 Personalizing to individual subjects

Our architecture is built around the idea of extracting neural representations that are common across sEEG subjects. We also, however, need to take into account that performance to behavioral tasks is intrinsically different between individuals [Fozard et al., 1994, Davranche et al., 2006, Green and Bavelier, 2003, Der and Deary, 2006]. Therefore, we designed a separate task head for each individual, composed of a shallow feed-forward network, that maps the extracted neural representations $F$ to the behavioral outcome of a given trial.

## 4 Experiments

In this section, we test and validate our approach for multi-session, across-subject behavioral decoding using sEEG on a large and diverse cohort of sEEG subjects performing a behavioral task.

### 4.1 Experimental setup

**Dataset.** The core design constraint of our approach is to enable the unified training of our architecture (see Fig. 1) on a diverse, multi-session, multi-subject sEEG dataset despite the inhomogeneity of electrode number/placement in each subject. Towards that goal, we built a large and diverse dataset of subjects that performed a repetitive reaction time task, while their sEEG was simultaneously recorded. Contrary to most studies, which collected a few hundred behavioral trials from few subjects, *we collected more than 3600 behavioral trials and more than 100 electrode-hours of recording from 21 subjects over 29 recording sessions*. Our dataset is composed of 13 females and 8 males, with ages ranging from 16 to 57 years old, with unique electrode number/placement for each subject, solely based on clinical needs. Across subjects electrodes span white and grey matter; cortical, subcortical, and deep structures (see Fig. 2).

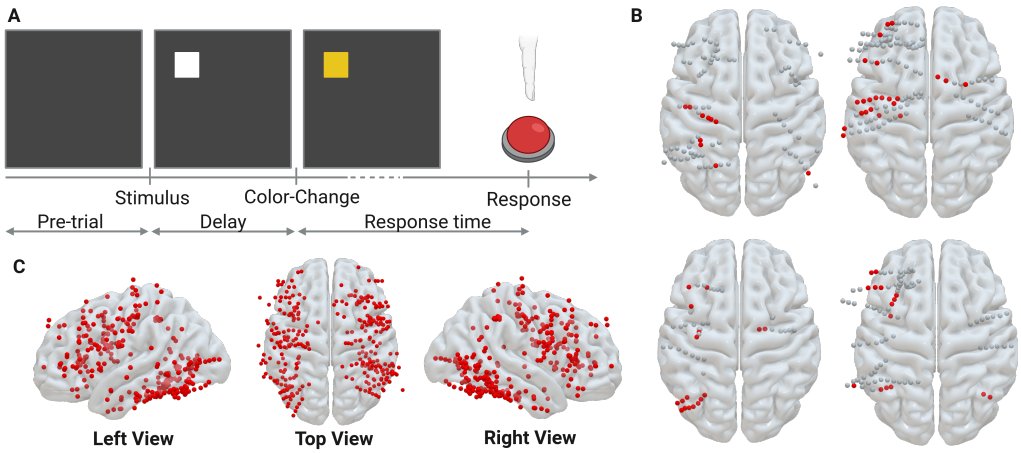


Figure 2: *Overview of behavioral experiment.* **A**. Schematic of the color change detection task. **B**. Electrode placement projected onto the MNI brain template for four example subjects in our cohort. Red dots show electrodes used for model training; grey dots show electrodes excluded from model training (see section 3.1). **C**. Electrodes used for model training across all subjects in our cohort projected on an MNI brain template.

**Behavioral task.** Subjects performed the color-change detection task (see Fig. 2A & Appendix A.1). In each trial, a visual stimulus was presented. After a variable foreperiod delay, the stimulus changed color. At that time, the participant responded by pressing a button as fast as they could. *Our goal was to decode the trial-wise response time of the subjects using their sEEG.*

**Design choices**. Throughout all experiments, we use the neural data from a window of [0, 1500] msec after the stimulus color change to train models and make behavioral predictions. All models were trained with $N_t = 105$ temporal tokens within each electrode and $N_e \in [3, 28]$ electrode tokens within each timepoint, which varied based on the number of electrodes for each participant. The dimensionality of each token was fixed to $K = 2$. All training details are provided in Appendix A.2.

### 4.2 Training within-subject models

To test whether our modeling approach would be able to decode the trial-wise response time from sEEG data, we began with a within-subject approach and tested whether our architecture was capable of decoding the response times for each subject. We trained a separate model for each subject in our cohort, combining data across sessions for participants that had multiple behavioral sessions. The experimental setup and the preprocessing of the sEEG data was kept the same for all subject. *Across the 21 single-subject models, the average test set $R^2$ was 0.30 $\pm$ 0.05 (mean $\pm$ sem).*

### 4.3 Training a multi-session, multi-subject model

To investigate whether training on more data, despite the heterogeneity (see Fig. 2B, C), would improve response time decoding performance, we trained a unified model on the combined neural data of all the behavioral trials across all participants in our dataset. This effectively increased the number of training samples 21-fold compared to the data available for the single-subject models.

This model achieved a test set $R^2$ of 0.54 $\pm$ 0.01 (mean $\pm$ sem) on the combined data of all participants. The root mean square error (RMSE) for all predicted response times in the test set, across subjects, was RMSE = 82 $\pm$ 2 (mean $\pm$ sem) msec. For reference, the mean response time in the test set, across subjects, is 410 $\pm$ 5 (mean $\pm$ sem) msec, which is an order of magnitude greater than the RMSE. Since the response time profiles of each subject are distinct, we also evaluated the model performance within-subjects. *Across subjects, the average per-subject test set $R^2$ was 0.39 $\pm$ 0.05 (mean $\pm$ sem).* Compared to training on single-subjects, the multi-subject training approach boosted decoding performance for our proposed architecture by $\Delta R^2 = 0.09$, on average. We then compared the performance of the single-subject models to that of the multi-subject model,

head to head (see Fig. 3A). For the majority of subjects, there was a clear performance boost from multi-subject training.

We then explored whether finetuning the multi-session, multi-subject model to individual subjects would further boost the performance gains, compare to the single-subject models. *The finetuned models for each subject achieved an average test set $R^2$ of 0.41 $\pm$ 0.05, across subjects*, indicating that finetuning the multi-session, multi-subject model to single subjects can further boost the performance gains. Importantly, finetuning the unified model boosted the performance gains of the multi-subject model compared to the single-subject models by $\Delta R^2 = 0.11$, on average (see Fig. 3B for per-subject head to head comparisons). These results suggest that sEEG-based neural decoding benefits from multi-subject joint model training, despite the heterogeneity of electrode number/placement across subjects, demonstrating the power of multi-subject approaches compared to single-subject ones.

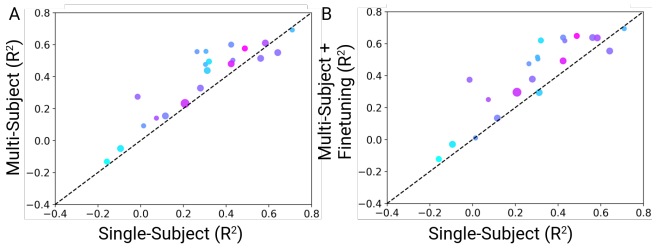


Figure 3: *Comparing decoding performance between the single-subject and multi-subject models for each subject.* **A**. Single-subject vs multi-subject model. **B**. Single-subject vs finetuned, multi-subject model. Circle size denotes the number of trials and color the number of electrodes (cyan to magenta represents ascending order)

### 4.4 Transferring to new subjects

Having shown that training our model across diverse, multi-subject data boosts decoding performance, we were interested in identifying whether the representations learned by pretraining our models on multiple subjects can efficiently be transferred to new subjects, unseen from the model during training. This is important in any real-world clinical scenario, where only a few number of behavioral trials can be collected from a subject. To test this, we employed a leave-one-out cross validation approach. We trained 21 models, each of which was trained on the combined data of all subjects but one. The average test set $R^2$ of the multi-subject models was 0.48 $\pm$ 0.006 (mean $\pm$ sem) across all trials of all subject. This indicates that our training framework is robust to data variations and that our model did not depend on the data of any specific subject to train effectively.

The weights of the pretrained models were then used as the basis for finetuning on the data of each left-out participant, unseen from the models during training. *Across subjects, the models trained on other subjects and transferred to a new one achieved an average test set $R^2$ of 0.38 $\pm$ 0.05 (mean $\pm$ sem).* This is a clear improvement to training single-subject models from scratch and comparable to training multi-session, multi-subject models. Specifically, the transferred single-subject models showed a per-subject decoding performance increase $\Delta R^2 = 0.08$ compared to our single-subject models trained from scratch (see Fig. 4A) and a per-subject decoding performance decrease $\Delta R^2 = -0.01$, compared to the multi-session, multi-subject model (Fig. 4B)).

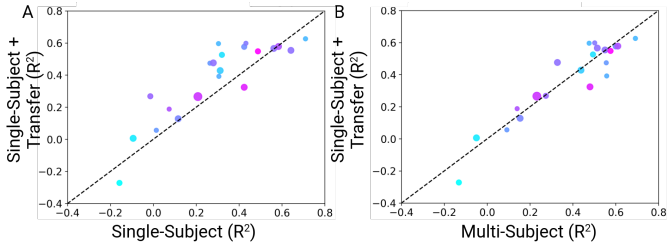


Figure 4: *Comparing decoding performance between* **(A)** Transferred single-subject finetuned models vs single-subject models trained from scratch, and **(B)** Transferred single-subject finetuned models vs the multi-session, multi-subject model.

This suggests that pretraining models on diverse, multi-subject data and using their weights as a finetune to transfer the learned representations to new subjects is almost as effective as training on

the combined data of all subjects. Given the significant computational benefit of the transferring approach, its implementation in real world medical scenarios might be more likely.

### 4.5 Comparison with baselines

We were also interested in quantifying the performance benefit of our modeling approach compared to other traditional and state-of-the-art neural decoding approaches. To do so, we trained single-subject baseline models (described in detail in A.3), and compared those with seegnificant. We observed that seegnificant outperformed the baseline models when trained on single subject, multiple subjects, and when transferred (see Fig. 5). Specifically, our single-subject models (section 4.2) outperformed all baselines by an average per-subject test set $\Delta R^2 \geq 0.03$. Our multi-subject models (section 4.3) outperformed the baselines by a per-subject test set $\Delta R^2 \geq 0.12$, which could be further increased to $\Delta R^2 \geq 0.14$ by finetuning to single subjects. Importantly, our transfer-learned single-subject models (section 4.4) also outperformed all baseline models by a per-subject test set $\Delta R^2 \geq 0.11$, on average.

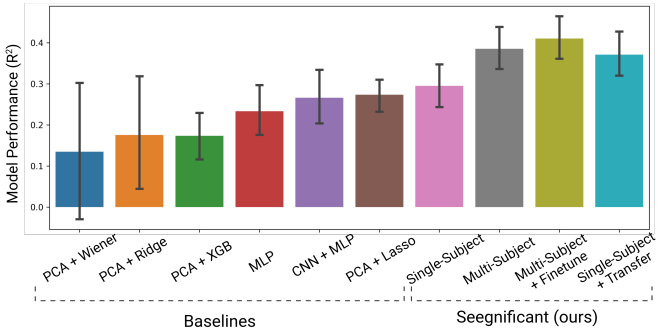


Figure 5: *Decoding performance (mean ± sem) for various baselines and our proposed models.*

### 4.6 Ablations

Last, we sought to identify the contribution of each building block of our proposed architecture to the decoding performance of our models. To investigate this, we performed an ablation study on our proposed architecture when trained on multiple subject (same as in section 4.3). To ensure a fair comparison, all training hyperparameters were kept identical across all training runs.

#### 4.6.1 Model components

In this subsection, we trained our proposed model while ablating its components one by one. The model is composed of: 1. convolutional tokenizer (CT), 2. attention in time dimension (AT), 3. spatial positional encoding (PE), 4. attention in space dimension (AS), 5. multi-layer perceptron (MLP), 6. subject-specific regression heads (RH) (see also Fig. 1). Since there is no straightforward way to process the raw sEEG data with AT without first extracting features, we did not ablate CT. Additionally, since there is no straightforward way to project the output of the AS to the regression heads without a linear layer, we also did not ablate the MLP.

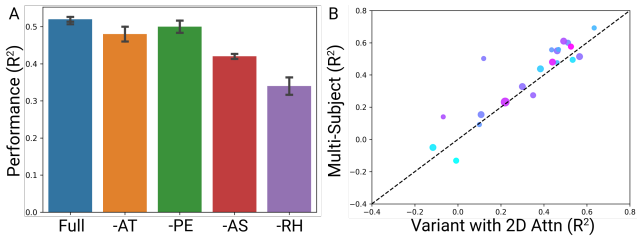


Figure 6: *Summary of ablation results.* (A) Decoding performance (mean ± sem) across different model variants. (B) Head-to-head comparison of the decoding performance of our multi-session, multi-subject model vs its 2D attention variant.

The ablation study results are summarized in Fig. 6A. Inspection of the results suggests that the subject-specific regression heads are key to the model's decoding performance ($\Delta R^2 = 18$), likely because the response time profiles across subjects are quite diverse. Ablating the attention mechanism

in the space dimension also substantially degrades decoding performance ($\Delta R^2 = 0.10$) suggesting that it might be the main mechanism by which the model is capable of dealing with the heterogenous number/placement of electrodes across subjects.

#### 4.6.2 Model variants

We also compared our architecture with a close variant. Our architecture is designed with two separate attention mechanisms over the time and space dimensions of our data (see section 3.2.2), mainly to reduce computational burden since the computational complexity of attention scales quadratically with sequence length. To ensure that this choice did not negatively affect the model's decoding performance, we compared our architecture against a variant where the two separate attention mechanisms are replaced by a single 2D attention mechanism over the combined time and space dimensions of the data (to enable this, the positional encoding was ablated). *We identified that our proposed architecture outperformed its 2D attention variant by a per-subject, test set* $\Delta R^2 = 0.06$*, on average* (see Table 1 and Fig. 6B). We also identified that our proposed architecture trained $\sim 5.5\times$ faster.

Table 1: *Decoding performance (mean* $\pm$ *sem) for the different models*.

| Model | Per-Subject $R^2$ | Train time (mins) |
|---|---|---|
| Variant | $0.33 \pm 0.05$ | $141.6 \pm 3.23$ |
| Ours | $0.39 \pm 0.05$ | $25.7 \pm 0.03$ |

## 5 Discussion

In this work, we introduce seegnificant: a novel framework and architecture that can be used for multi-session, across subject decoding based on sEEG. We show that training on diverse and large cohorts of sEEG data, despite inter-subject differences in terms of electrode number/placement, is not only possible but also leads to better decoding performance compared to single-subject approaches. We also show that the neural representations learned by pretrained models can be efficiently transferred to new subjects, despite the small number of available trials within-subjects (which is the case in any real-world clinical setting). In fact, this few-shot, transfer learning approach gives a clear benefit to training from scratch and can reach decoding performance levels that are very close to those of multi-subject pretrained models, with a significant computational benefit.

Our results suggest that training on diverse multi-subject sEEG datasets boosts decoding performance compared to single-subject approaches. While our dataset is large compared to other works [Angrick et al., 2021, Petrosyan et al., 2022, Meng et al., 2021, Wu et al., 2022, 2024], scaling up further will inevitably require combining data collected while subjects perform a variety of different behavioral tasks. Advances in other fields have shown that training on multi-task data can improve performance and generalization [Azabou et al., 2023, Ruder et al., 2017]. Therefore, investigating whether those performance/generalization gains will hold on sEEG decoding is worth exploring. Training on multi-task datasets would also give insight as to which aspects of neural computations are shared across tasks and which are not. We leave the investigation of multi-task model training as future work.

Given the clinical circumstances under which sEEG is recorded, performance to downstream decoding tasks would likely benefit from pretraining using self-supervised objectives. When patients are admitted in the epilepsy monitoring unit, their sEEG recordings are collected continuously over 5-10 days. In contrast, sEEG study participants perform only a few minutes of a behavioral task. Evidently, the amount of unsupervised data collected is orders of magnitude more than the supervised data. Strategies such as autoencoding, masked modeling, or next token prediction could enable training on the vasts amounts of data that are collected but unused by supervised learning approaches.

Overall, this work advances clinical and human neuroscience research by introducing a framework that enables multi-subject model training based on sEEG. We show that models pretrained on large and diverse sEEG datasets can be easily transferred to new subjects and therefore can be used to accelerate progress in the neural decoding domain. Given that sEEG is currently the gold standard invasive neural recording modality used in humans, our work advances progress towards sEEG-based neural decoding therapeutic interventions, ultimately bringing them closer to clinical translation.

## Acknowledgments and Disclosure of Funding

We would like to thank the following funding sources for their generous support: National Institutes of Health grant R01NS121219 (FV); Office of Naval Research grant N00014-22-1-2677 (KD); National Institutes of Health grant 6T32NS091006 (AGR); Onassis Foundation graduate student scholarship and A. G. Leventis Foundation graduate student scholarship (GM). Authors declare no competing interests.

## References

M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications Biology*, 4(1), Sept. 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02578-0. URL http://dx.doi.org/10.1038/s42003-021-02578-0.

M. Azabou, V. Arora, V. Ganesh, X. Mao, S. Nachimuthu, M. J. Mendelson, B. Richards, M. G. Perich, G. Lajoie, and E. L. Dyer. A unified, scalable framework for neural population decoding, 2023. URL https://arxiv.org/abs/2310.16046.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, Jan. 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

V. P. Buch, J. M. Bernabei, G. Ng, A. G. Richardson, A. Ramayya, C. Brandon, J. Stiso, D. S. Bassett, and T. H. Lucas. "primed to perform:" dynamic white matter graph communicability may drive metastable network representations of enhanced preparatory cognitive control. Sept. 2022. doi: 10.1101/2022.09.25.509351. URL http://dx.doi.org/10.1101/2022.09.25.509351.

J. F. Burke, K. A. Zaghloul, J. Jacobs, R. B. Williams, M. R. Sperling, A. D. Sharan, and M. J. Kahana. Synchronous and asynchronous theta and gamma activity during episodic memory formation. *The Journal of Neuroscience*, 33(1):292–304, Jan. 2013. ISSN 1529-2401. doi: 10.1523/jneurosci.2057-12.2013. URL http://dx.doi.org/10.1523/JNEUROSCI.2057-12.2013.

D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *Journal of computer assisted tomography*, 18(2): 192–205, 1994.

W. Coon and G. Schalk. A method to establish the spatiotemporal evolution of task-related cortical activity from electrocorticographic signals in single trials. *Journal of Neuroscience Methods*, 271:76–85, Sept. 2016. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2016.06.024. URL http://dx.doi.org/10.1016/j.jneumeth.2016.06.024.

M. Crawshaw. Multi-task learning with deep neural networks: A survey. *ArXiv*, abs/2009.09796, 2020. URL https://api.semanticscholar.org/CorpusID:221819295.

W. Cui, W. Jeong, P. Thölke, T. Medani, K. Jerbi, A. A. Joshi, and R. M. Leahy. Neuro-gpt: Towards a foundation model for eeg, 2023. URL https://arxiv.org/abs/2311.03764.

J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3150–3158, 2015. URL https://api.semanticscholar.org/CorpusID:8510667.

K. Davranche, B. Burle, M. Audiffren, and T. Hasbroucq. Physical exercise facilitates motor processes in simple reaction time performance: An electromyographic analysis. *Neuroscience Letters*, 396(1):54–56, Mar. 2006. ISSN 0304-3940. doi: 10.1016/j.neulet.2005.11.008. URL http://dx.doi.org/10.1016/j.neulet.2005.11.008.

G. Der and I. J. Deary. Age and sex differences in reaction time in adulthood: Results from the united kingdom health and lifestyle survey. *Psychology and Aging*, 21(1):62–73, Mar. 2006. ISSN 0882-7974. doi: 10.1037/0882-7974.21.1.62. URL http://dx.doi.org/10.1037/0882-7974.21.1.62.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

J. L. Fozard, M. Vercruyssen, S. L. Reynolds, P. A. Hancock, and R. E. Quilter. Age differences and changes in reaction time: The baltimore longitudinal study of aging. *Journal of Gerontology*, 49(4):P179–P189, July 1994. ISSN 0022-1422. doi: 10.1093/geronj/49.4.p179. URL http://dx.doi.org/10.1093/geronj/49.4.p179.

A. Galvao-Carmona, J. J. González-Rosa, A. R. Hidalgo-Muñoz, D. Páramo, M. L. Benítez, G. Izquierdo, and M. Vázquez-Marrufo. Disentangling the attention network test: behavioral, event related potentials, and neural source analyses. *Frontiers in Human Neuroscience*, 8:813, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00813. URL https://www.frontiersin.org/article/10.3389/fnhum.2014.00813.

J. I. Glaser, A. S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording. Machine learning for neural decoding. *eneuro*, 7(4):ENEURO.0506–19.2020, July 2020. ISSN 2373-2822. doi: 10.1523/eneuro.0506-19.2020. URL http://dx.doi.org/10.1523/ENEURO.0506-19.2020.

C. S. Green and D. Bavelier. Action video game modifies visual selective attention. *Nature*, 423(6939):534–537, May 2003. ISSN 1476-4687. doi: 10.1038/nature01647. URL http://dx.doi.org/10.1038/nature01647.

H. Kan, J. Yu, J. Huang, Z. Liu, H. Wang, and H. Zhou. Self-supervised group meiosis contrastive learning for eeg-based emotion recognition. *Applied Intelligence*, 53(22):27207–27225, Sept. 2023. ISSN 1573-7497. doi: 10.1007/s10489-023-04971-0. URL http://dx.doi.org/10.1007/s10489-023-04971-0.

S. Kovac, V. N. Vakharia, C. Scott, and B. Diehl. Invasive epilepsy surgery evaluation. *Seizure*, 44:125–136, Jan. 2017. ISSN 1059-1311. doi: 10.1016/j.seizure.2016.10.016. URL http://dx.doi.org/10.1016/j.seizure.2016.10.016.

Y.-T. Lan, W. Liu, and B.-L. Lu. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9207625. URL http://dx.doi.org/10.1109/IJCNN48605.2020.9207625.

T. Le and E. Shlizerman. Stndt: Modeling neural population activity with a spatiotemporal transformer, 2022. URL https://arxiv.org/abs/2206.04727.

M. Leszczyński, A. Barczak, Y. Kajikawa, I. Ulbert, A. Y. Falchier, I. Tal, S. Haegens, L. Melloni, R. T. Knight, and C. E. Schroeder. Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Science Advances*, 6(33), Aug. 2020. ISSN 2375-2548. doi: 10.1126/sciadv.abb0977. URL http://dx.doi.org/10.1126/sciadv.abb0977.

C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, and X. Chen. Eeg-based emotion recognition via transformer neural architecture search. *IEEE Transactions on Industrial Informatics*, 19 (4):6016–6025, Apr. 2023. ISSN 1941-0050. doi: 10.1109/tii.2022.3170422. URL http://dx.doi.org/10.1109/TII.2022.3170422.

S. Liu, E. Johns, and A. J. Davison. End-to-end multi-task learning with attention, 2018. URL https://arxiv.org/abs/1803.10704.

W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, June 2022. ISSN 2379-8939. doi: 10.1109/tcds.2021.3071170. URL http://dx.doi.org/10.1109/TCDS.2021.3071170.

D. MacDonald, N. Kabani, D. Avis, and A. C. Evans. Automated 3-d extraction of inner and outer surfaces of cerebral cortex from mri. *NeuroImage*, 12(3):340–356, Sept. 2000. ISSN 1053-8119. doi: 10.1006/nimg.1999.0534. URL http://dx.doi.org/10.1006/nimg.1999.0534.

K. Meng, D. B. Grayden, M. J. Cook, S. Vogrin, and F. Goodarzy. Identification of discriminative features for decoding overt and imagined speech using stereotactic electroencephalography. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, Feb. 2021. doi: 10.1109/bci51272.2021.9385355. URL http://dx.doi.org/10.1109/BCI51272.2021.9385355.

K. J. Miller, L. B. Sorensen, J. G. Ojemann, and M. den Nijs. Power-law scaling in the brain surface electric potential. *PLoS Computational Biology*, 5(12):e1000609, Dec. 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000609. URL http://dx.doi.org/10.1371/journal.pcbi.1000609.

S. E. Paraskevopoulou, W. G. Coon, P. Brunner, K. J. Miller, and G. Schalk. Within-subject reaction time variability: Role of cortical networks and underlying neurophysiological mechanisms. *NeuroImage*, 237:118127, Aug. 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118127. URL http://dx.doi.org/10.1016/j.neuroimage.2021.118127.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

A. Petrosyan, A. Voskoboinikov, D. Sukhinin, A. Makarova, A. Skalnaya, N. Arkhipova, M. Sinkin, and A. Ossadtchi. Speech decoding from a small set of spatially segregated minimally invasive intracranial eeg electrodes with a compact and interpretable neural network. *Journal of Neural Engineering*, 19(6):066016, Nov. 2022. ISSN 1741-2552. doi: 10.1088/1741-2552/aca1e1. URL http://dx.doi.org/10.1088/1741-2552/aca1e1.

M. I. Posner and S. E. Petersen. The attention system of the human brain. 13(1):25–42, Mar. 1990. doi: 10.1146/annurev.ne.13.030190.000325. URL https://doi.org/10.1146/annurev.ne.13.030190.000325.

A. G. Ramayya, V. Buch, A. Richardson, T. Lucas, and J. I. Gold. Simple human response times are governed by dual anticipatory processes with distinct and distributed neural signatures. June 2022. doi: 10.1101/2022.06.13.496029. URL http://dx.doi.org/10.1101/2022.06.13.496029.

S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4324–4331, 2017. URL https://arxiv.org/abs/1705.08142.

I. Saez and X. Gu. Invasive computational psychiatry. *Biological Psychiatry*, 93(8):661–670, Apr. 2023. ISSN 0006-3223. doi: 10.1016/j.biopsych.2022.09.032. URL http://dx.doi.org/10.1016/j.biopsych.2022.09.032.

Y. Song, Q. Zheng, B. Liu, and X. Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. ISSN 1558-0210. doi: 10.1109/tnsre.2022.3230250. URL http://dx.doi.org/10.1109/TNSRE.2022.3230250.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, Feb. 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL http://dx.doi.org/10.1016/j.neucom.2023.127063.

D. Sussillo, R. Jozefowicz, L. F. Abbott, and C. Pandarinath. Lfads - latent factor analysis via dynamical systems, 2016. URL https://arxiv.org/abs/1608.06315.

J. Talairach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: An approach to cerebral imaging*. Thieme Medical Publishers, 1988.

J. Tang, Z. Ma, K. Gan, J. Zhang, and Z. Yin. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment. *Information Fusion*, 103:102129, Mar. 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.102129. URL http://dx.doi.org/10.1016/j.inffus.2023.102129.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

X. Wu, G. Li, S. Jiang, S. Wellington, S. Liu, Z. Wu, B. Metcalfe, L. Chen, and D. Zhang. Decoding continuous kinetic information of grasp from stereo-electroencephalographic (seeg) recordings. *Journal of Neural Engineering*, 19(2):026047, Apr. 2022. ISSN 1741-2552. doi: 10.1088/1741-2552/ac65b1. URL http://dx.doi.org/10.1088/1741-2552/ac65b1.

X. Wu, S. Jiang, G. Li, S. Liu, B. Metcalfe, L. Chen, and D. Zhang. Deep learning with convolutional neural networks for motor brain-computer interfaces based on stereo-electroencephalography (seeg). *IEEE Journal of Biomedical and Health Informatics*, 27(5):2387–2398, May 2023. ISSN 2168-2208. doi: 10.1109/jbhi.2023.3242262. URL http://dx.doi.org/10.1109/JBHI.2023.3242262.

X. Wu, G. Li, X. Gao, B. Metcalfe, and D. Zhang. Channel selection for stereo- electroencephalography (seeg)-based invasive brain-computer interfaces using deep learning methods. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:800–811, 2024. ISSN 1558-0210. doi: 10.1109/tnsre.2024.3364752. URL http://dx.doi.org/10.1109/TNSRE.2024.3364752.

J. Ye and C. Pandarinath. Representation learning for neural population activity with neural data transformers. *Neurons, Behavior, Data analysis, and Theory*, 5(3), Aug. 2021. ISSN 2690-2664. doi: 10.51628/001c.27358. URL http://dx.doi.org/10.51628/001c.27358.

Z. Zhang, P. Luo, C. C. Loy, and X. Tang. *Facial Landmark Detection by Deep Multi-task Learning*, page 94–108. Springer International Publishing, 2014. ISBN 9783319105994. doi: 10.1007/978-3-319-10599-4_7. URL http://dx.doi.org/10.1007/978-3-319-10599-4_7.

X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu. *A Modulation Module for Multi-task Learning with Applications in Image Retrieval*, page 415–432. Springer International Publishing, 2018. ISBN 9783030012465. doi: 10.1007/978-3-030-01246-5_25. URL http://dx.doi.org/10.1007/978-3-030-01246-5_25.

# A Appendix / supplemental material

## A.1 Data collection and signal processing details

**Experimental setup**. Stereotactic EEG was recorded from 23 participants diagnosed with medically refractory epilepsy who underwent surgical implantation of intracranial electrodes for seizure localization. Patients were implanted with intraparencymal depth electrodes ("stereo EEG," Ad-tech, 1.1 diameter, 4 contacts spaced 5 mm apart), except in one patient who also had subdural grid electrodes (Ad-tech, 4 mm contacts, spaced 10 mm apart). sEEG was recorded using a Natus recording system. Based on the amplifier and the discretion of the clinical team, signals were sampled at either 512 or 1024 Hz. Clinical circumstances alone determined the number and placement of the implanted electrodes. To participate in this study, participants provided written informed consent in accordance with the IRB of the University of Pennsylvania and were compensated for their time.

**Behavioral task**. Participants performed the color-change-detection-task, a stimulus-detection task with a variable foreperiod delay [Buch et al., 2022, Ramayya et al., 2022] while their sEEG activity was simultaneously recorded. Participants viewed visual stimuli on a laptop and responded by pressing a button on a game controller. The behavioral task is composed of sequences of trials described here. Each trial began with the presentation of a visual stimulus (small white box) at a randomized location on the screen as a fixation target (one of nine locations on a 3 x 3 grid). The stimulus changed color (from white to yellow) after one of two randomly selected foreperiod delays: (i) 500 ms, or (ii) 1500 ms. The response time (time between color change and button press) for each trial was recorded.

**Signal preprocessing**. The sEEG signals were first converted to a bipolar montage by taking the difference between pairs of immediately adjacent contacts on the same electrode shank. The resulting bipolar signals were treated as new virtual electrodes (henceforth, electrodes) whose virtual location was the midpoint between the adjacent contacts [Ramayya et al., 2022, Burke et al., 2013]. Within electrodes, we computed the line noise SNR, as the ratio of each electrode's spectral power in the frequency range of 58-62 Hz over the spectral power in the range 18-22 Hz. Electrodes for which the SNR was greater than one were excluded from further analysis. Additionally, disconnected electrodes, defined as electrodes whose voltage had a standard deviation of zero were also excluded from further study. For the remaining electrodes, we investigated for noise contaminated trials defined as trials where (i) the voltage was not recorded due to saturation of the amplifier, or (ii) the mean $\pm$ SD of the voltage of the trial was greater than 10 times the mean of all trials of that electrode. Trials that were determined as noise contaminated were then removed.

**Signal processing details**. To combat the heterogeneity of electrode placement between subjects, we identified the subset of electrodes across subjects that responded to the stimulus color change while participants performed the color-change-detection-task. To do so, we followed the methodology used by Paraskevopoulou et al. [2021]. Specifically, we identified electrode locations where high-$\gamma$ band activity was significantly modulated by the stimulus color change.

We first filtered the data, separately for each electrode, to the high-$\gamma$ band (70-150 Hz) using a $4^{\text{th}}$ order Butterworth filter and subsequently extracted the high-$\gamma$ envelope using the Hilbert Transform. Then, we epoched the high-$\gamma$ envelopes around the color-change ([-500, 1500] ms) and separated each trial into a baseline period ([-500, 0] ms, before the color-change) and a task period ([0, 1500] ms, after the color-change). For each electrode, we computed the median high-$\gamma$ amplitude time course of the task ($Median_{task}$) and baseline ($Median_{baseline}$) periods across trials. Then for each electrode, we calculated the signal-to-noise ratio (SNR) as

$$SNR = \frac{\sigma^2(Median_{task})}{\sigma^2(Median_{baseline})}, \tag{5}$$

where $\sigma^2(x)$ refers to the variance of $x$. To identify electrodes with statistically significant SNR values, we applied a bootstrap randomization test in which we randomly shuffled the task and baseline labels from all locations 10,000 times and computed one random SNR value for each such iteration. We then calculated p-values for each electrode location as the fraction of randomized SNR values that were larger than the computed SNR. We determined responsive electrodes as those with a p-value $< 0.05$, which corresponds to a confidence level $\alpha = 0.05$. Prior to determining significance, we adjusted the p-values calculated across all electrodes of all subjects using false discovery rate [Benjamini and Hochberg, 1995].

The *broadband* voltage traces of all responsive electrodes were then z-scored (within electrodes), and downsampled to 400 Hz. Those electrodes were then used for decoding response times by our models. We note that using this procedure, we also rejected two participants from the study for whom only zero and one electrodes were returned as significant (lowering our participant count from 23 to 21). This electrode selection procedure lowered our computational cost, and because it merely uses high-$\gamma$ to select electrodes, it avoids problems of statistical double dipping.

### A.2 Model training details

**Model complexity.** The total number of trainable parameters for the multi-session, multi-subject model is 797,095. From those, 753,394 are shared across subjects (shared trunk) and the rest 43,701 parameters are subject-specific (parameters of the regression heads with 2,081 parameters per subject).

**Training hyperparameters**. All models were implemented and trained using Pytorch 2.1.0+cu121 [Paszke et al., 2019]. AdamW was used as the optimizer [Leszczyński et al., 2020] (with $b_1 = 0.5$ and $b_2 = 0.999$). All models were trained for 1000 epochs. A step learning rate scheduler was used with an initial learning rate set to $10^{-3}$ and decayed by a factor of 0.5 every 200 epochs for single-subject models and by a factor of 0.9 every 100 epochs for multi-subject models. Batch size was fixed to 64 and 1024 for all single-subject and multi-subject models, respectively. All models were optimized using Huber loss, except for when finetuning the multi-session, multi-subject model (see section 4.3) to individual subjects, where MSE loss was used.

**Hyperparameters of the spatial positional encoding.** For all models, the gaussian kernels used in the spatial positional encoding were centered at $\mu \in \{-90, -70, -50, \ldots, 50, 70\}$ and had variances $\sigma^2 \in \{1, 2, 4, \ldots, 64\}$.

**Controlling for random data splits**. For all models, the train/validation/test split was 70/15/15 %. To ensure that our results would not be biased in any way by the data splits, experiments performed in sections 4.2, 4.3, 4.5, 4.6.2, A.4.2, and A.4.3 were run using 5 different data splits and the results were averaged across the different splits. For experiments described in section 4.4, multi-subject models were trained for 1 data split (since training 21 multi-session, multi-subject models $\times$ 5 times for the different data split would be very computationally expensive). Finetuning to single-subjects was performed across 5 different data splits and the results across the splits were averaged. For the ablation study described in section 4.6.1, models were trained for 3 data splits, to reduce computational burden.

**Finetuning.** For finetuning multi-session, multi-subject models to single subjects in section 4.3, all model weights were updated throughout the single-subject finetuning. For transferring multi-session, multi-subject models to single subjects in section 4.4, since the regression heads of the left-out participant were completely untrained, we gradually unfroze model weights. For the first 400 training epochs, only the regression head (2,081 parameters per subject) of the left-out participant was trained (the remaining 753,394 model parameters were frozen). For the remaining 600 epochs, all model parameters were trained. All training parameters were identical to those described in paragraph "Training hyperparameters" above.

**Computational resources.** All models were trained on a machine with an AMD EPYC 7502P 32-Core Processor and 1 Nvidia A40 GPU with 44.99 GiB of memory. Single-subject models trained on average within 5 mins and multi-subject models trained within an hour. The memory requirement to train the multi-session, multi-subject model with a batch size of 1024 was $\sim$ 8 GiB.

### A.3 Baseline model details

**Non deep learning models.** The following non-deep learning models were trained on single-subjects, using scikit-learn (version 1.2.2): 1. Wiener Filter, 2. Ridge Regression, 3. Lasso Regression, 4. Gradient-Boosting (XGB). All models were used as part of a pipeline composed of: standard scaling $\rightarrow$ PCA $\rightarrow$ model. For Ridge and Lasso Regression, we optimized over the hyperparameter `alpha` $\in \{0.1, 1, 10\}$. For XGB, we optimized over the hyperparameter `n_estimators` $\in \{50, 100, 200\}$. All other hyperparameters were kept at their default values.

**Deep learning models.** The following deep learning models were trained on single-subjects, using the training procedure described in section A.2 (same as that used for our proposed architecture): 1. MLP (4 layers), 2. CNN + MLP (1 + 3 layers, respectively).

### A.4 Additional experiments

#### A.4.1 Model real-time applicability

To understand whether our multi-session, multi-subject model (see section 4.3) could be used in real time, which is essential for any real-world neural decoding system, we measured our model's inference time on two machines: a commercial laptop and a server. *The results, summarized in Table 2, show that our model runs in* $< 10$ *msec on both machines, on either CPU or GPU*. This indicates that our proposed architecture can be easily used in real time, since real-time systems run on the order of 100 msec.

Table 2: *Model inference time on different hardware. Units are in msec.*

| Machine | CPU | GPU |
|---|---|---|
| AMD EPYC 7502P + Nvidia A40 | 9.1 | 5.1 |
| Intel Core i9 + Nvidia A2000 | 4.0 | 7.9 |

#### A.4.2 Contribution of spatial positional encoding

Following our ablation study (see section 4.6.1), we sought to better understand the contribution of our spatial positional encoding to the model's performance. Contrary to other positional encoding techniques, constructed to encode the order of tokens in a sequence (see Vaswani et al. [2017] for Fourier and Su et al. [2024] for rotation based approaches), our approach encodes tokens with their 3D location in the brain, based on each electrode's MNI coordinates.

To quantify the performance gains, we trained a variant of our multi-session, multi-subject model, which was identical to that described in section 4.3 with the positional encoding ablated. We observed that *ablating the the positional encoding leads to a per-subject decoding performance decrease of* $\Delta R^2 = -0.02$*, on average*. We show a within-subject head to head comparison of the performance of the two models in Fig. 7.

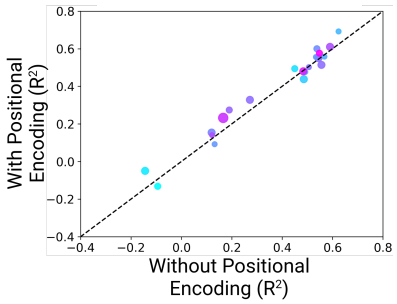


Figure 7: *Per subject decoding performance comparison between the model with and without spatial positional encoding*.

We also investigated whether the decoding performance of the model with and without positional encoding is significantly different. Specifically, we performed a Wilcoxon rank-sum test between the groups: (1) test set $R^2$ scores of all subjects obtained by training the multi-subject model with spatial positional encoding, and (2) test set $R^2$ scores of all subjects obtained by training the multi-subject model without spatial positional encoding. The test returned a test-statistic = 0.34 and a p-value = 0.73, indicating that there is no significant difference between the performance of the model with and without spatial positional encoding.

Those results, while not significant, suggests that informing electrode-tokens with their 3D location in the brain prior to self-attention in the space dimension might slightly benefit decoding performance. Intuitively, this would makes sense, since the functional computations happening within roughly the same brain regions are similar across humans.

#### A.4.3 Comparison of our proposed spatial positional encoding against other approaches.

The results of section A.4.2 suggest that the decoding performance gains introduced by the positional encoding scheme used in this work are only modest. Therefore, we explored whether other positional encoding schemes would boost decoding performance further. To do so, we trained variants of our proposed architecture on the combined data of all subjects (same training scheme as the one used in section 4.3) with the positional encoding replaced by the: 1. positional encoding scheme introduced

by Vaswani et al. [2017], 2. a variant of the positional encoding scheme used in this work where RBFs are replaced by fourier functions. While the explored schemes are far from comprehensive, we identified that our proposed positional encoding scheme performed as well as or better than other approaches (see Table 3). However, the decoding performance gains provided by all positional encodings explored in this work are modest. Therefore, it would be worth exploring other positional encoding schemes, perhaps based on whole brain MRI images and/or brain atlases other than MNI [Talairach and Tournoux, 1988].

Table 3: *Model performance using different positional encoding schemes.*

| Positinal Encoding | Per-Subject $R^2$ |
|---|---|
| Vaswani et al. [2017] | $0.16 \pm 0.04$ |
| MNI-Fourier | $0.39 \pm 0.05$ |
| MNI-RBF (Ours) | $0.39 \pm 0.05$ |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: To the best of our knowledge, all claims made in the abstract and introduction are supported by our experiments described in section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our proposed approach in sections 3 and 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results are introduced in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the methodology introduced in this work in section 3. The details to reproduce the behavioral experiment used in this work are provided in section A.1. The details of how to train the models introduced in this work are provided section A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our code is available at https://github.com/gmentz/seegnificant. The dataset used in this work contains electrophysiological data collected from human patients. To respect their privacy and comply with HIPAA regulations, we cannot make the dataset public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We refer the reader to section A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All figures with barplots include error bars and the factors of variability capture by them are included in captions of the figure. One statistical test (corrected appropriately for multiple comparisons) was used in this work, described in section A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We refer the reader to A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, this manuscript conforms in every aspect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to 1, and 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The data used in this work are not shared to protect the privacy of the human participants whose electrophysiological data was used in this study. To the best of our knowledge, the code shared with this work does not pose a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide proper citations for all the works upon which this work is built.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All details of the code are discused in section A.2. Code is publicly available at https://github.com/gmentz/seegnificant.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We do not have access to the full text of instructions given to participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Approval was obtained by the IRB of the University of Pennsylvania prior to data collection.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.