TRANSFORMER MODELS FOR ROBUST EEG SOURCE LOCALIZATION

By

Simon Polishchuk

May, 2026

Director of Thesis: David Hart, PhD

Major Department: Computer Science

ABSTRACT

Electroencephalography (EEG) measures brain activity using electrodes placed on the scalp. Estimating the underlying source locations from these measurements is an ill-posed inverse problem, and traditional methods often struggle in the presence of noise and complex source configurations. Recent approaches have applied neural networks to this task, but many rely on assumptions that limit their ability to generalize across different data settings.

This work presents a transformer-based model for EEG source localization that supports both single-source and multi-source prediction. The model is designed to handle challenges commonly found in EEG data, including measurement noise, varying source configurations, and missing electrodes. Training is performed on large-scale simulated EEG data generated from known source locations.

A structured evaluation framework is introduced to assess performance under different conditions, including varying noise levels, source region densities, and electrode dropout scenarios. Across these settings, the transformer model consistently achieves lower localization error compared to both classical methods and existing neural network approaches. In addition, a set of ablation studies is conducted to analyze the impact of architectural and training design choices. The code and data used in this work will be made publicly available.

TRANSFORMER MODELS FOR ROBUST EEG SOURCE LOCALIZATION

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Data Science

By

Simon Polishchuk

May, 2026

Director of Thesis: David Hart, PhD

Thesis Committee Members:

Moritz Dannhauer, PhD

Kebin Peng, PhD

Eric Lang, PhD

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors, Dr. David Hart and Dr. Moritz Dannhauer, for their guidance, support, and invaluable contributions to this work. Their expertise and feedback played a critical role in shaping both the direction and execution of this research.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The brain generates electrical activity as neurons fire during cognitive processes. When large groups of neurons are active at the same time, this activity can be measured using electroencephalography (EEG). EEG is a widely used technique in neuroscience that records electrical potentials at the scalp using a set of electrodes. These measurements provide a sparse view of brain activity, as signals are only observed at a limited number of locations. Estimating the underlying source of this activity from EEG measurements is known as the inverse problem, which is ill-posed [10].

Traditional approaches to EEG source localization rely on numerical methods and physical models of the head to estimate the source locations. While these methods are well studied, they can be computationally expensive and often struggle in the presence of noise or when multiple sources are active [1]. More recently, machine learning methods have been applied to this problem by training models on large sets of simulated EEG data. Prior work has explored a range of architectures, including convolutional networks [8], fully-connected networks [9], and graph-based models [13].

Transformer models have achieved strong performance in areas such as natural language processing and computer vision, but their use in EEG source localization remains limited. Existing transformer-based approaches [30] typically follow the same fixed input and output structure as earlier models, which does not fully take advantage of the flexibility offered by attention mechanisms. In particular, transformers can naturally operate on a variable number of input tokens and do not require a fixed ordering of inputs. This is especially useful

for EEG data, where electrode measurements may be missing or unreliable. Additionally, the number of output tokens can be controlled through learned embeddings, allowing for direct multi-source prediction.

In this work, a transformer-based model is developed for EEG source localization that leverages these properties. Each electrode is treated as an individual input token, with spatial location included directly in the input representation. Learned embedding tokens are used to predict one or more source locations. The model is evaluated on both single-source and multi-source localization tasks and is tested under varying levels of noise and electrode dropout.

In addition to general performance improvements, the proposed model demonstrates strong gains in several challenging scenarios. In single-source localization on the 10 mm regions, the transformer achieves a mean localization error of 0.877 mm at 50% noise, compared to 1.359 mm for the MLP and 7.924 mm for LCMV. At higher noise levels (75%), the transformer maintains an error of 4.647 mm, while classical methods exceed 14 mm. In multi-source experiments, the transformer continues to perform reliably, achieving 0.564 accuracy at 50% noise, whereas other approaches degrade significantly. The model also remains robust under electrode dropout, with only minor performance degradation when five electrodes are removed. These results highlight the ability of the transformer to handle noisy, incomplete, and multi-source EEG data more effectively than existing methods.

To support training, simulated EEG data are generated using a realistic head model with varying levels of measurement noise. The brain surface is discretized into regions using a Poisson-disc-like sampling approach, which allows control over the spacing between regions. Two levels of discretization are considered, with 10 mm and 5 mm spacing, enabling evaluation across different levels of spatial resolution.

The proposed model is evaluated across a wide range of scenarios, including different noise levels, source configurations, and electrode dropout conditions. In addition to the main model, several ablation studies are conducted to examine the impact of architectural

choices and training strategies. These include experiments on iterative training, unconstrained location prediction, and alternative network configurations.

In summary, the main contributions of this work are:

- A transformer-based model for EEG source localization that supports a variable number of inputs and outputs.
- A region clustering approach that allows direct control over spatial resolution in the forward model.
- A comprehensive evaluation demonstrating improved performance over classical and existing machine learning methods, especially under noisy and incomplete data conditions.
- An ablation study analyzing the effect of network design and training strategies on model performance.

Chapter 2

Related Works and Background

2.1 EEG Source Localization

Over the past few decades, significant progress has been made in estimating neuronal sources from EEG data. These methods rely on a forward model that describes how electrical activity in the brain propagates through head tissues to the electrodes. Simple spherical models provide analytical approximations, while finite element models (FEM) offer higher accuracy by capturing the complex geometry and conductivity of the head. In these models, dipolar current sources are used to simulate the electrical activity that produces the observed EEG signals [5]. Dipoles are commonly used because they provide a flexible way to represent different types of brain activity.

Source activity can vary in both space and time, and may appear as a single source, multiple correlated sources, or distributed activity across regions of the brain. Different source localization methods aim to estimate these underlying sources from the measured EEG signals.

Since the inverse problem is ill-posed, additional assumptions are required to obtain a stable solution. The problem can be either over-determined or under-determined depending on the number of sources relative to the number of measurements. In the over-determined case, a small number of discrete sources are estimated (e.g., dipole fitting or beamforming), which tends to be more robust to noise but may miss complex activity patterns. In the under-determined case, a large number of dipoles are distributed across the brain surface,

often constrained to be perpendicular to the cortex. In this setting, regularization methods such as minimum norm estimates (MNE) are used to reduce ambiguity and obtain a stable solution [1]. Additional constraints, such as temporal or statistical priors, can further improve robustness.

A widely used distributed method is eLORETA [16], which is known for its stability and practical performance [12]. Similar to sLORETA [17], it minimizes the L2 error between predicted and measured EEG signals while applying Tikhonov regularization. Although these methods suffer from depth bias, different weighting strategies can help reduce this effect. However, they remain sensitive to noise and often require preprocessing or averaging.

In this work, both sLORETA and eLORETA are evaluated on simulated data and are found to produce identical results. Therefore, only eLORETA is reported. In addition, a discrete source localization method, the linearly constrained minimum variance (LCMV) beamformer [25], is included for comparison. A broader overview of classical EEG source localization methods can be found in [31].

2.2 Machine Learning

The use of machine learning for EEG source localization is a relatively recent development. ConvDip [8] introduced one of the first convolutional neural network approaches for this task, where the brain is discretized into voxels and the network predicts the activity level in each voxel. Since then, several approaches have extended this idea using different architectures. For example, [9] proposed a recurrent neural network with fully connected layers for spatio-temporal prediction, while [13] introduced a graph neural network for similar tasks. Other methods include learning linear combinations of basis functions [28], using neural mass models [23], combining CNNs with finite element modeling [2], and applying LSTM-based or autoencoder-based architectures [20, 11, 14]. A comprehensive overview of supervised learning approaches is provided in [19].

Despite this progress, most prior work focuses on CNNs, RNNs, or related architectures,

and transformer-based models have not been widely explored for EEG source localization. Transformers have shown strong performance in EEG classification tasks, with models such as EEGFormer [27], EEG-ViT [29], EEG-Conformer [22], and EEG-Deformer [3], but these approaches have not been directly applied to the localization problem.

To the best of current knowledge, only one prior work [30] has proposed a transformer-based approach for EEG source localization. However, that method relies on fixed input representations and distribution-based outputs similar to earlier approaches, and does not explore variations in network design. In contrast, the transformer model used in this work uses learnable embedding tokens to separate single-source from multi-source localization and includes spatial location directly in the input. It also can be extended to unconstrained location prediction. In addition, different aspects of the model are explored, including the effect of noise levels during training and variations in network configuration.

Chapter 3

Materials and Methods

For EEG source localization, brain activity must be inferred from electrical potentials measured at the scalp. These sources generate current that flows through the head tissues and produces the observed EEG signals. This process can be modeled using a finite element model, where the geometry and properties of the head tissues are obtained from magnetic resonance imaging (MRI).

In this work, the Ernie head model [24] is used to generate simulated EEG data for source localization. Both classical methods (e.g., eLORETA) and machine learning approaches are evaluated using this setup. Applying machine learning to this inverse problem introduces several design challenges. These challenges, along with the transformer model used in this work, are described in the following subsections.

3.1 Region Clustering and Forward Simulation

Although source activity can occur anywhere on the brain surface, machine learning models require a discrete set of outputs for training and evaluation. Previous approaches often discretize the brain into a 3D voxel grid [8] or use icosahedral divisions, but these methods do not fully capture the folded structure of the cortical surface.

In this work, region clusters are generated by placing points approximately equidistant across the brain surface using a specified geodesic spacing (either 5 mm or 10 mm). Cluster centers are selected from the brain surface mesh nodes based on their geodesic distances,

which are computed using Dijkstra's algorithm. Nodes that fall within the chosen geodesic threshold of a selected center are removed from further consideration. This process is repeated in a manner similar to Poisson-disc sampling until all regions are defined.

Using a 10 mm spacing results in 1803 regions, each represented by a dipole oriented perpendicular to the brain surface. A finer discretization with 5 mm spacing produces 6790 regions, as illustrated in Figure 3.1.

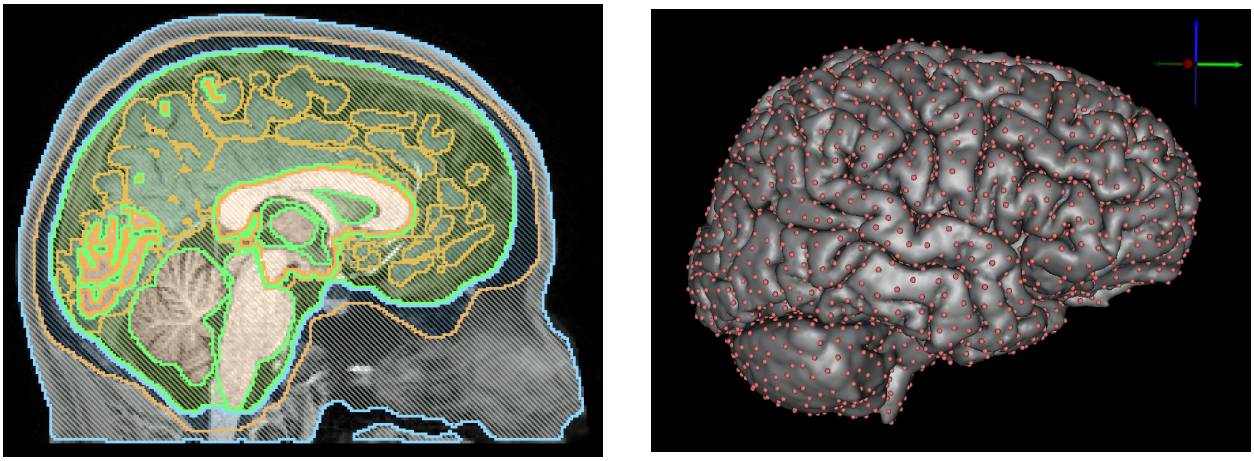

Figure 3.1: The modeled Ernie head is made up by many intricate tissues and folded brain regions as smooth surface as tetrahedral volume element discretiized from tissue-label voxels. We distributed source locations (shown as red dots) across the brain surfaces equally within 10mm (shown here) or 5mm geodesic distance between them using an in-hourse Possion disc-like approach.

The Ernie head model (from the SimNIBS software [24]) is used, which includes electrode positions based on the standard 10-10 placement system. This model contains approximately 662,000 nodes and 3.7 million tetrahedral elements. Using standard tissue conductivities [24], the EIDORS-3D package [18] is used to compute the finite element stiffness matrix $\mathbf{A}$, while in-house software [6] is used to compute the right-hand side vector $\mathbf{b}$. The linear system $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ is then solved using SCIRun 4.7 [15].

The resulting potential values at the 76 electrode locations are used to form the leadfield matrix, with each column corresponding to a dipole source. Two leadfield matrices are generated: one for the 10 mm source space (dimensions 76×1803) and one for the 5 mm source space (76×6790). These matrices are used for both classical methods (e.g., eLORETA

and LCMV) implemented in MATLAB, as well as for the machine learning models.

For the classical methods, the regularization parameters are selected from logarithmically spaced values between 10^{-5} and 10^{15} (MATLAB command: $\alpha = \text{logspace}(-5, 15, 10)$). For eLORETA, Tikhonov regularization is applied to the minimum norm estimate using these α values [16]. For LCMV, both the solution and the noise covariance matrix are regularized, resulting in 100 combinations of (α, β) values, where $\alpha = \beta$.

3.2 Synthetic Noise, Training Setup, and Dropout

To reflect the challenges of the inverse problem, synthetic measurement noise is added to the simulated EEG signals. Instead of using a single noise level, models are trained and evaluated across a range of noise strengths. The noise level is defined as a percentage of the signal amplitude, varying from 1% to 99%. For each sample and noise level, noise vectors are randomly generated and added to the EEG measurements during both training and testing. The standard deviation of the noise is defined as

$$\sigma = n\frac{l_{max}}{2} \tag{3.1}$$

where n represents the noise level and l_{max} is the maximum absolute value of the leadfield for the given sample.

During training, new noise vectors are generated for each sample. For evaluation, performance is computed using 50 noise vectors per sample for validation and 100 noise vectors per sample for testing. These noise vectors are generated once and stored to ensure consistency across all experiments.

In addition to synthetic noise, electrode dropout is considered to better reflect real EEG recording conditions. In practice, some electrodes may produce unreliable measurements due to poor contact, sensor faults, or noise artifacts. To simulate this, a small number of electrode measurements are randomly removed during training and testing. This setup

allows evaluation of how well each method performs when parts of the input data are missing.

The simulation framework is further extended to include varying source strengths in multi-source scenarios. Rather than assuming that both dipoles contribute equally, a weighting parameter α is introduced to control the relative contribution of each source. For each two-source example, the EEG signal is generated as a weighted combination of the corresponding leadfield vectors with additive noise:

$$\mathbf{x} = \alpha\, L(c_1) + (1 - \alpha)\, L(c_2) + \epsilon \tag{3.2}$$

where $L(c_i)$ denotes the leadfield vector associated with source location c_i, and ϵ represents additive noise.

To study the effect of source strength imbalance, experiments are conducted using different ranges of α, including relatively balanced and more imbalanced settings. This formulation introduces variability in source dominance and allows evaluation under more realistic multi-source EEG conditions.

3.3 Neural Network Architecture

A transformer-based architecture is used to solve the inverse problem. The model takes the 76 electrode measurements from the leadfield matrix, along with their spatial locations, as input and predicts the index of the source region. To provide a clear comparison, a baseline fully-connected network is also implemented following assumptions of other techniques. This baseline is primarily used to evaluate the effectiveness of the transformer model.

3.3.1 Baseline: Multi-Layer Perceptron

For comparison, a simple fully-connected network, or Multi-Layer Perceptron (MLP), is used as a baseline. The network consists of three fully-connected layers, similar to the setup described in [9]. The hidden layer sizes are 100 and 1000, and each layer is followed by a

ReLU activation function. The model assumes a fixed input size based on the training data. For the dropout experiments, missing input values are handled by setting the corresponding channels to zero.

3.3.2 Transformer Network

The transformer architecture follows the encoder structure introduced in [26]. Instead of using a single fixed input vector, each electrode is treated as an individual token. Each token consists of the electrode value concatenated with its spatial location. This allows the model to learn spatial relationships between electrodes, which is important for EEG data.

Each token is projected into a 512-dimensional embedding space using a linear layer. Following the design of Vision Transformers [4], a learned embedding token is added to the input sequence, resulting in a total of 77 tokens. For multi-source prediction, additional embedding tokens are included. These tokens are then processed through a stack of transformer encoder layers with self-attention.

After the encoder, the embedding tokens are passed through a linear projection to produce predictions over the set of region clusters. The region with the highest score is selected as the predicted source location, and the output is trained using a cross-entropy loss. In the multi-source setting, predictions are matched to the closest ground-truth sources, and the corresponding losses are summed. A visualization of this architecture is shown in Figure 3.2.

In the final model, the embedding dimension is set to 512, with 8 attention heads and 3 transformer layers. A dropout rate of 0.1 is used within the network to reduce overfitting. The use of self-attention enables the model to capture both local interactions between nearby electrodes and broader relationships across the entire sensor layout.

A key advantage of this formula is its flexibility with respect to input size. Since each electrode is treated as a separate token, the model can handle missing electrodes by simply omitting those tokens during inference. In addition, the order of the input tokens does not affect the output, which reduces the risk of errors due to input ordering.

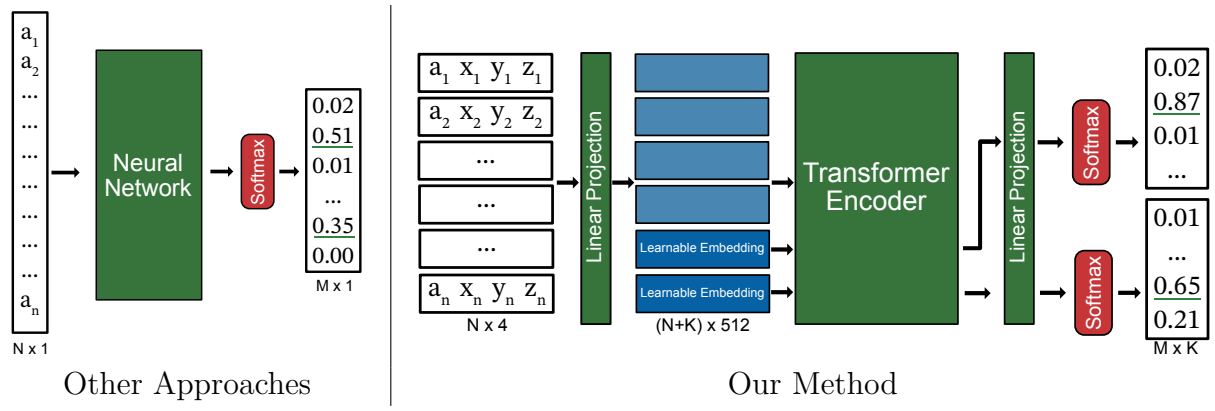

Figure 3.2: Network diagram for the Transformer model compared to other approaches. In most other methods [8, 9, 30], the neural network has a single input vector for the N electrodes and single output predictor vector for the M source regions, with multi source prediction handled by taking the K top values from the output confidences. In comparison, our approach can directly appends location information to the electrode values to treat each as individual token, and can integrate K learned embedding tokens [4], so that each region source is predicted directly. In our setup, N=76, K=1 for single-source or K=2 for multi-source, and M=1803 for 10mm or M=6790 for 5mm.

To further extend the model, an auxiliary regression head can be added to predict the relative source strength parameter α. This head operates on a pooled representation of the learned embedding tokens and is trained using an L1 loss. The overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \lambda \mathcal{L}_{\alpha} \tag{3.3}$$

where λ is a weighting parameter that controls the contribution of the source strength prediction task.

Chapter 4

Results

After training, the proposed method is evaluated on the test set as described in section 3.2 for single-source, multi-source, and dropout tests. Two metrics are reported to assess performance: the accuracy of predicting the correct source region (Acc.) and the average Euclidean distance between the predicted region center and the true region center, which corresponds to the mean localization error (MLE).

4.1 Single-Source Region Prediction

The ability of the transformer network to correctly predict the source region is evaluated and compared with the baseline multilayer perceptron (MLP) described in Section 3.3.1. The transformer network was trained using a learning rate of 1e−4 with the OneCycle learning rate scheduler [21] for 250 steps on the 10 mm source space and 300 steps on the 5 mm source space. The MLP network was trained with a learning rate of 1e−4 for 150 steps and with a learning rate of 1e−5 for an additional 50 steps. Both networks were observed to converge after these training durations.

Additionally, comparisons are performed with eLORETA [16] and the LCMV beamformer [25], two classical source localization methods that rely on optimal spatial filter design. As described in Section 3.2, 100 noise vectors for each brain region are saved for testing at each noise level. These noise vectors are added to the data before being passed to the classical approaches in order to ensure fair comparisons.

Table 4.1: Comparison of the transformer network to the classical eLORETA method, LCMV, ConvDip, and baseline MLP approach on the **10mm** regions. Average Accuracy (Acc.) and Euclidean Distance (MLE) in millimeters are reported for a test set of 100 different noise vectors per possible region. The transformer network performs better on every metric compared to the other techniques.

	Acc. ↑						**MLE ↓**					
Method \Noise	1%	10%	25%	50%	75%	99%	1%	10%	25%	50%	75%	99%
LCMV	1.000	0.999	0.965	0.594	0.345	0.238	0.000	0.002	0.517	7.924	14.29	17.79
eLORETA	1.000	0.999	0.986	0.869	0.527	0.307	0.000	0.002	0.203	4.161	14.04	25.20
ConvDip	1.000	0.999	0.994	0.944	0.814	0.639	0.000	0.003	0.089	1.278	5.837	14.35
MLP	1.000	0.999	0.992	0.940	0.819	0.651	0.000	0.003	0.125	1.359	5.611	13.73
Transformer	1.000	**1.000**	**0.998**	**0.960**	**0.848**	**0.686**	0.000	**0.000**	**0.034**	**0.877**	**4.647**	**12.40**

Table 4.2: Comparison of the transformer network to other approaches on on the **5mm** regions. The transformer network performs better on both metrics in most cases, especially at high noise levels.

	Acc. ↑						**MLE ↓**					
Method \Noise	1%	10%	25%	50%	75%	99%	1%	10%	25%	50%	75%	99%
LCMV	1.000	0.999	0.856	0.336	0.173	0.114	0.000	**0.006**	1.748	11.29	16.46	19.15
eLORETA	1.000	0.998	0.942	0.607	0.295	0.141	0.000	0.013	0.685	7.478	18.04	28.03
ConvDip	1.000	0.999	0.983	0.851	0.623	0.424	0.000	0.008	0.212	2.841	10.07	19.74
MLP	1.000	0.998	0.983	0.858	0.645	0.450	0.000	0.013	**0.209**	2.674	9.319	18.57
Transformer	1.000	**0.999**	**0.986**	**0.878**	**0.669**	**0.471**	0.000	0.055	0.341	**2.246**	**8.602**	**17.79**

Average performance across all brain regions and 100 noise realizations per region is reported. Results for the 10 mm source spacing are shown in Table 4.1 and the left side of Figure 4.1, while results for the 5 mm spacing are shown in Table 4.2 and the right side of Figure 4.1. At low noise levels, all evaluated methods achieve comparable accuracy. As noise levels increase, however, the learned models maintain substantially higher accuracy, with the transformer consistently outperforming the MLP by approximately 2%–3%. Beyond improved localization accuracy, the proposed method also exhibits a significantly faster forward-pass runtime compared to the classical approaches, indicating its potential suitability for real-time or near–real-time analysis in future applications.

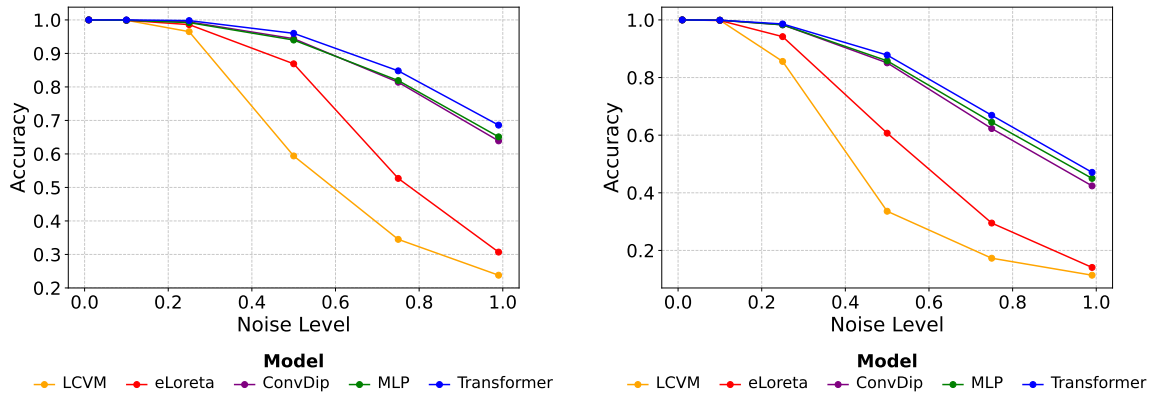

Figure 4.1: A graph of the accuracy of each method on the 10mm (left) and 5mm (right) regions. The transformer model does better than the classical methods and other deep learned approaches.

4.2 Multi-Source Region Prediction

Multi-source localization highlights a key advantage of the transformer architecture over traditional techniques. Whereas many existing methods rely on the assumption that the highest-confidence outputs correspond directly to source locations, the transformer supports explicit multi-source prediction by increasing the number of learned embedding tokens within the attention mechanism.

In these experiments, two sources of comparable strength are generated using the forward model, with the inter-source distance constrained to 10 cm. During both training and evaluation, the model is required to predict the locations of both dipoles. A permutation-invariant loss is used to account for the lack of ordering in the predicted source locations. Predicted and ground-truth source locations are matched by minimizing the global pairwise distance between them, which additionally prevents degenerate solutions in which the model predicts the same location for both sources. Results for all evaluated methods are reported in Table 4.3 and illustrated in Figure 4.2.

Table 4.3: Comparison of the transformer network to other approaches on on the 10mm regions when predicting two sources of similar strength. While the other methods break down as noise level increases in a multi-source setup, the transformer maintains reasonable performance up to 50% noise.

	Acc. ↑						**MLE ↓**					
Method \Noise	1%	10%	25%	50%	75%	99%	1%	10%	25%	50%	75%	99%
LCMV	0.215	0.220	0.220	0.199	0.165	**0.155**	21.22	21.23	22.04	23.28	25.38	**26.60**
eLORETA	0.019	0.011	0.012	0.001	0.005	0.004	27.69	28.44	29.04	28.98	29.27	29.75
ConvDip	0.991	0.915	0.663	0.268	0.082	0.024	0.149	0.377	6.731	21.35	33.27	41.54
MLP	0.933	0.784	0.495	0.195	0.047	0.024	1.189	3.599	10.94	23.51	34.85	43.44
Transformer	**0.995**	**0.986**	**0.863**	**0.564**	**0.297**	0.138	**0.128**	**0.203**	**2.100**	**9.955**	**20.57**	30.87

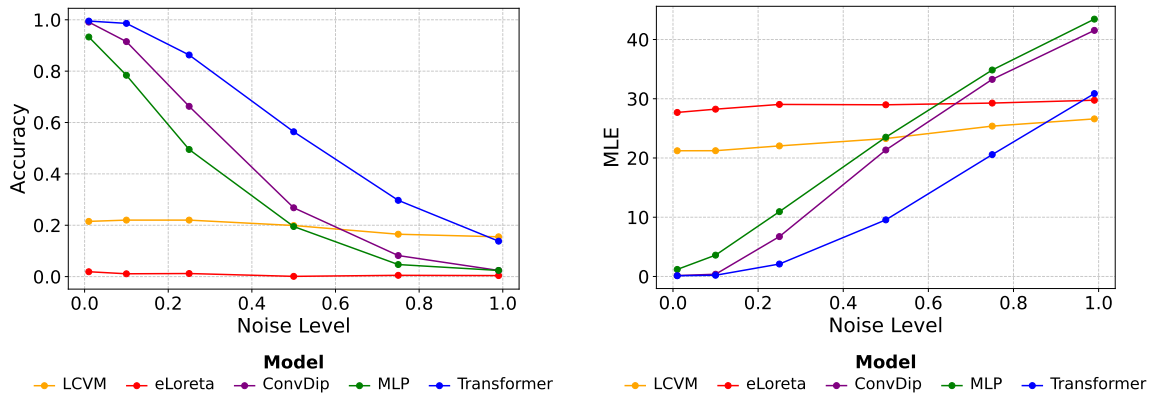

Figure 4.2: A graph of the accuracy (left) and mean localization error (right) of each method on a two-source setup. The transformer model does better than the classical methods and other deep learned approaches on multi-source prediction.

4.3 Localization under Varying Source Strengths

To evaluate how differences in source contribution affect localization performance, experiments are conducted using two simultaneously active sources with unequal strengths. In realistic EEG scenarios, multiple brain regions may be active at the same time, but their contributions to the measured signal are rarely equal. This setup allows analysis of how well the model can identify both dominant and weaker sources.

Performance is evaluated using three complementary metrics:

- **Overall accuracy**, measuring correct localization of both sources
- **Strong source accuracy**, measuring correct prediction of the dominant source

Table 4.4: Localization performance under varying source strength ranges (α) and noise levels. Results are reported for overall accuracy, dominant source accuracy, and weak source accuracy.

Noise	$\alpha \in [0.4, 0.6]$			$\alpha \in [0.25, 0.75]$			$\alpha \in [0.1, 0.9]$		
	Acc	Strong	Weak	Acc	Strong	Weak	Acc	Strong	Weak
0.01	0.989	0.994	0.994	0.976	0.990	0.985	0.970	0.992	0.977
0.10	0.967	0.985	0.980	0.937	0.981	0.954	0.866	0.981	0.880
0.25	0.825	0.917	0.891	0.742	0.920	0.797	0.507	0.904	0.560
0.50	0.488	0.721	0.640	0.376	0.738	0.485	0.264	0.786	0.337

- **Weak source accuracy**, measuring correct prediction of the less dominant source

This evaluation separates the model's ability to identify the most prominent signal from its ability to recover weaker sources that contribute less to the overall EEG measurement.

The results in Table 4.4 show that the model consistently performs better on the dominant source across all noise levels and strength ranges. As the imbalance between sources increases, performance on the weaker source decreases, which is expected due to its reduced contribution to the overall signal. This effect becomes more pronounced at higher noise levels.

Despite this, the model maintains reasonable performance across all tested conditions, indicating robustness to moderate variations in source strength. These results demonstrate that the proposed approach can effectively handle multi-source EEG data where source contributions are not equal, while also highlighting the inherent difficulty of detecting weaker sources in noisy environments.

4.4 Handling Dropout of Electrodes

In real EEG data collection, it is common for a small number of electrodes to produce unreliable measurements. Electrodes may be faulty, poorly calibrated, or partially disconnected from the scalp, resulting in inaccurate signals. When such "bad" electrodes are spatially clustered, accurate source localization becomes especially challenging, since the dipolar structure of the underlying brain activity may be disrupted. Many existing source localization methods

Table 4.5: Comparison of the transformer network to the classical eLORETA method, LCMV, ConvDip, and baseline MLP approach on the 10mm regions when dropping 5 electrodes randomly. The transformer network shows minimal performance drop off when presented with missing information.

	Acc. ↑						**MLE ↓**					
Method \Noise	1%	10%	25%	50%	75%	99%	1%	10%	25%	50%	75%	99%
LCMV	1.000	0.999	0.961	0.581	0.326	0.223	0.000	0.002	0.575	8.320	14.95	18.35
eLORETA	1.000	0.999	0.982	0.796	0.495	0.282	0.000	0.003	0.255	4.837	15.44	26.91
ConvDip	1.000	0.999	0.991	0.929	0.780	0.600	0.001	0.018	0.149	1.699	7.215	16.54
MLP	0.999	0.997	0.974	0.894	0.755	0.592	0.022	0.064	0.511	2.775	8.245	16.79
Transformer	1.000	**1.000**	**0.996**	**0.949**	**0.820**	**0.649**	0.001	**0.001**	**0.059**	**1.190**	**5.814**	**14.34**

struggle under these conditions, whereas the transformer model can naturally accommodate them due to its ability to operate on a variable number of input channels.

To evaluate the model's ability to generalize in the presence of missing data, these conditions are simulated by randomly dropping five electrode measurements from each training and testing sample. This experiment is performed for both single-source and multi-source settings, requiring the model to make predictions despite incomplete input data.

For single-source localization, the results are reported in Table 4.5. The transformer maintains high accuracy and low localization error even when electrodes are missing. At a noise level of 99%, the mean localization error increases from 12.40 mm to 14.34 mm when dropout is introduced, indicating only a modest reduction in performance. Compared to the classical and baseline neural network methods, the transformer remains the most robust across all tested noise levels.

Multi-source localization under electrode dropout is summarized in Table 4.6. As expected, performance decreases as noise increases, but the model remains stable at low and moderate noise levels. Accuracy remains above 0.96 at 10% noise and above 0.83 at 25% noise, while the mean localization error stays below 3 mm up to 25% noise. At 50% noise, performance degrades more noticeably, reflecting the added difficulty of recovering multiple active sources when both noise and missing channels are present.

Overall, these results show that the transformer is resilient to electrode dropout in both single-source and multi-source settings. This is an important property for practical EEG

Table 4.6: Multi-source localization performance under electrode dropout.

Noise	Accuracy	MLE (mm)
0.01	0.987	0.293
0.10	0.966	0.548
0.25	0.831	2.863
0.50	0.490	11.752

applications, where imperfect sensor contact and missing channels are common during real data collection.

4.5 Additional Ablations

Beyond the primary transformer model, several alternative network configurations and encoding strategies are explored. These experiments are intended to assess the impact of architectural and training design choices. The model presented in the main results is found to provide the most effective overall configuration. Details and results for the additional ablation studies are provided here.

4.5.1 Ablation: Unconstrained Location Prediction

While the region-based clustering approach yields strong performance, it is informative to examine whether discretizing the brain into predefined regions is strictly necessary. To this end, an unconstrained location prediction model is evaluated. In this configuration, the transformer network is modified to directly predict a continuous source location rather than a region index.

Specifically, the final projection layer is adapted to output six values corresponding to an (x, y, z) location and a surface normal direction (n_x, n_y, n_z). For baseline comparison, the linear network is modified in an identical manner. The (x, y, z) coordinates are passed through a tanh activation to constrain predictions within normalized brain boundaries, while the (n_x, n_y, n_z) components are normalized using an L2 norm to enforce a unit surface normal. This network configuration is illustrated in Figure 4.3. Unlike the region-based formulation,

this approach does not require a predefined number of source regions.

For training, a loss function based on the mean squared error of the predicted location is used, scaled by an approximation of the angular distance between two points on a spherical surface. This formulation provides an approximation of the great-circle distance between the predicted and true source locations:

$$s \approx ||\mathbf{p_2} - \mathbf{p_1}|| \left((1 - (\mathbf{n}_1 \cdot \mathbf{n}_2)) \frac{\pi - 2}{4} + 1 \right) \tag{4.1}$$

This linear approximation is chosen to avoid numerical instabilities that arise in the exact great-circle distance formulation. A detailed discussion of these instabilities and the derivation of the approximation is provided in the following section.

Results for the unconstrained location prediction task are reported in Table 4.7. The transformer consistently outperforms the MLP in this setting, although both approaches perform worse than the region-based formulation and classical localization methods. At low noise levels, both models achieve near-perfect accuracy. Performance degrades substantially at higher noise levels, with accuracies at 50% noise dropping to 0.548 for the MLP and 0.771 for the transformer. These results indicate that, while the unconstrained approach is capable of learning from electrode data, additional refinements are likely required for robust performance under high-noise conditions. Future work may explore improvements

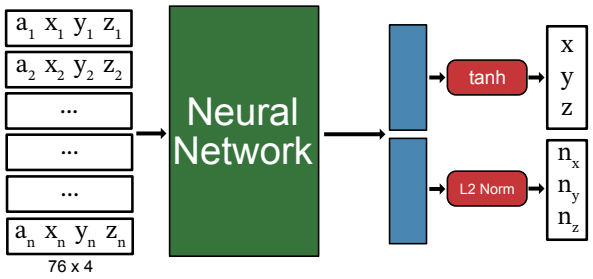

Figure 4.3: Network configuration for unconstrained location prediction. Rather than generating a vector of prediction confidences over discrete regions, the network outputs six values. Three values are passed through a tanh function to represent a location within the normalized brain boundaries $(-1, 1)$, while the remaining three are L2-normalized to represent the surface normal direction at that location.

Table 4.7: Comparison of unconstrained location prediction methods on the 10mm regions. While neither method outperformed the region-prediction based approaches, the transformer network showed significantly better performance over a standard MLP. Future developments of this approach could lead to better generalization across brain models.

	MLP		Transformer	
Noise Level	Acc. ↑	MLE ↓	Acc. ↑	MLE ↓
0.01	1.000	0.517	1.000	0.472
0.10	0.999	0.683	0.999	0.805
0.25	0.951	1.716	0.988	1.344
0.50	0.548	5.378	0.771	3.730
0.75	0.213	10.45	0.374	8.507
0.99	0.085	15.81	0.162	14.18

that enable reliable generalization without relying on a predefined set of brain regions.

4.5.2 Derivation of the Unconstrained Location Loss Function

The unconstrained location prediction loss introduced in Equation 4.1 is formulated as a linear approximation of the great-circle distance. This section outlines the derivation of that approximation and motivates its use during training.

A sphere can be uniquely defined by two points on its surface and the corresponding surface normals at those points. The distance between the two points along the surface is given by the arc length, which can be expressed as

$$s = r\theta \tag{4.2}$$

where s denotes the arc length, r is the radius of the sphere, and θ is the angle between the two points. The radius r can be obtained using the Law of Cosines, yielding

$$r = \frac{d}{\sqrt{2 - 2\cos\theta}} \tag{4.3}$$

where d represents the Euclidean distance between the two surface points. The angular separation θ can be computed from the surface normals using a dot product:

$$\theta = \cos^{-1}(\mathbf{n}_1 \cdot \mathbf{n}_2) \tag{4.4}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are the unit normals at the two surface locations. Substituting these expressions into Equation 4.2 gives

$$s = d\frac{\cos^{-1}(\mathbf{n}_1 \cdot \mathbf{n}_2)}{\sqrt{2 - 2(\mathbf{n}_1 \cdot \mathbf{n}_2)}} \tag{4.5}$$

where $d = \|\mathbf{p}_2 - \mathbf{p}_1\|$.

The formulation in Equation 4.5 introduces numerical instability during training. As the angle between the two points approaches zero, both the inverse cosine term and the square-root denominator approach zero, leading to unstable gradients. While the overall range of this expression spans from 1 to $\frac{\pi}{2}$ as the dot product varies from 1 to -1, the presence of near-zero divisions makes it unsuitable for optimization.

To address this issue, a linear function is introduced that preserves the same range and domain while avoiding these instabilities:

$$f(x) = (1 - x)\frac{\pi - 2}{4} + 1 \tag{4.6}$$

This function maps values of x from 1 to -1 onto the interval $[1, \frac{\pi}{2}]$. Substituting $x = \mathbf{n}_1 \cdot \mathbf{n}_2$ into this expression and replacing the angular component in Equation 4.5 yields the final approximation used for training:

$$s \approx \|\mathbf{p}_2 - \mathbf{p}_1\| \left((1 - (\mathbf{n}_1 \cdot \mathbf{n}_2))\frac{\pi - 2}{4} + 1 \right) \tag{4.7}$$

This approximation provides a stable and computationally efficient estimate of the great-circle distance while avoiding numerical issues during optimization.

4.5.3 Ablation: Encoder–Decoder Transformer Architecture

An additional transformer configuration based on an encoder–decoder architecture is evaluated. In the encoder-only formulation used in the main experiments, the number of output regions must be known at training time, as predictions are produced through a final linear classification layer. An encoder–decoder architecture offers an alternative formulation, in which the network can explicitly cross-attend to individual region representations when determining the most likely source location. This ablation explores the feasibility of such an approach.

Like in the encoder-only architecture, the 76 electrode measurements are concatenated with their spatial locations, projected into a 32-dimensional embedding space, and treated as source tokens. Brain region centers are similarly projected into the same 32-dimensional space and treated as target tokens. This reduced dimensionality is chosen to mitigate the computational cost associated with extensive cross-attention. The electrode tokens are first processed by a transformer encoder consisting of six layers with four attention heads. The resulting representations are then passed to a decoder, which performs cross-attention over the region tokens using a single layer with four attention heads. Each region token is subsequently scored using a linear projection to a single scalar value, and the region with the highest score is selected as the predicted source location. The resulting score vector is trained using a cross-entropy loss. A schematic of the full encoder–decoder architecture is shown in Figure 4.4.

In practice, this encoder–decoder configuration does not reliably converge, even with the reduced embedding dimensionality. When convergence does occur, training is substantially slower and performance remains inferior to the encoder-only architecture. The large number of region tokens involved in cross-attention limits overall accuracy, which peaks at approximately 78% even at a noise level of 1%. While it is possible that pretraining strategies, higher-dimensional embeddings, or additional computational resources could improve performance, this architecture is not competitive under the current experimental setup.

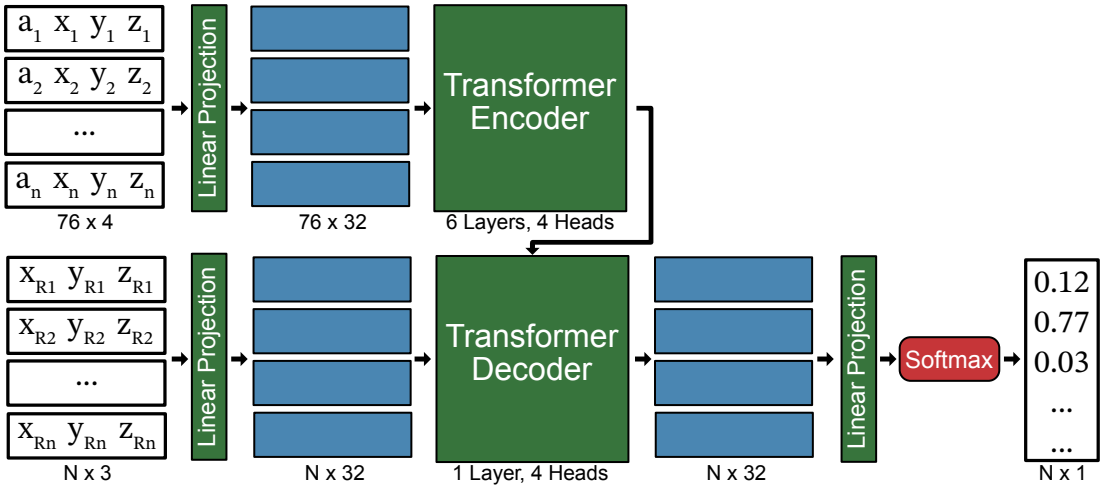

Figure 4.4: Encoder–decoder transformer architecture. In this configuration, brain region locations are projected into the same embedding space as the EEG measurements and electrode locations. The encoder processes the EEG tokens, while the decoder performs cross-attention over the region tokens. Each region token is scored via a linear projection. Due to the large number of regions, the embedding dimensionality is reduced to 32 to improve training efficiency; however, this configuration generally fails to converge.

4.5.4 Ablation: Iterative Training Approach

The results shown in Tables 4.1 and 4.2 are obtained by training separate models at fixed noise levels. While each model addresses the same source localization task, training is repeated independently for each noise condition. This setup motivates an investigation into whether information learned at lower noise levels can be reused when training models at higher noise levels.

To explore this idea, an iterative training strategy is considered in which a model trained at a given noise level is used to initialize training at the next, higher noise level. The goal of this approach is to reduce overall training cost by avoiding repeated training from scratch for each noise setting.

Under this strategy, models are trained sequentially from low to high noise levels, ranging from 0.01 to 0.99. Each noise level is trained for 50 epochs before advancing to the next stage, and model weights are saved after each step to support stable learning transitions. Both the MLP and transformer models are evaluated at the current noise level throughout

Table 4.8: Comparison of average accuracy of iterative models on each noise level on the 5mm regions. Both the iterative MLP and Transformer network were trained for 50 epochs per noise level using 1e-4 oneCycle learning rate schedule. The iterative approach does not outperform the standard approach, but does train for all noise levels in a much shorter time.

Noise Level	Standard MLP	Standard Transformer	Iterative MLP	Iterative Transformer
0.01	1.000	1.000	1.000	1.000
0.10	0.998	0.999	0.998	0.999
0.25	0.983	0.986	0.954	0.984
0.50	0.858	0.878	0.800	0.865
0.75	0.645	0.669	0.593	0.653
0.99	0.450	0.471	0.417	0.458

the iterative process. The resulting performance is summarized in Table 4.8.

Across noise levels, the iteratively trained transformer outperforms both the standard-trained and iteratively trained MLP, with the largest gains observed at higher noise levels. However, iterative training does not surpass the performance of standard single-noise-level training for the transformer when longer training schedules are used, such as 250 epochs at 99% noise. Moreover, increasing the number of epochs per noise level reduces the computational benefit of the iterative approach. Consequently, the standard training pipeline remains the most effective overall, while the iterative strategy may serve as a practical alternative for quickly training reasonably strong models across multiple noise conditions.

4.5.5 Ablation: Network Architecture

To validate the architectural choices used in the proposed model, an ablation study is conducted comparing multiple network designs and training configurations.

4.5.6 Positional Encodings

In many transformer-based architectures for computer vision and natural language processing, positional encodings are added after the projection layers to provide information about spatial or sequential structure. In this work, an alternative strategy is considered by incorporating electrode location information through a positional encoding rather than appending

Table 4.9: Comparison of average accuracy for the transformer model with and without positional encoding on the 10 mm region prediction task. The model using positional encoding underperforms across all noise levels. This behavior is likely related to the sparse nature of EEG measurements and the direct inclusion of electrode locations in the input representation.

Noise Level	No Pos. Enc. Acc. ↑	With Pos. Enc. Acc. ↑
0.01	1.000	1.000
0.10	1.000	0.987
0.25	0.998	0.981
0.50	0.960	0.879
0.75	0.848	0.810
0.99	0.686	0.175

locations directly to the input vectors. Results for this comparison are reported in Table 4.9.

Across all noise levels, the inclusion of positional encoding results in lower accuracy compared to directly concatenating electrode locations with the input features. This suggests that explicitly providing spatial information through the linear projection is more effective for this task. The sparse nature of EEG measurements, combined with the direct incorporation of spatial coordinates, may reduce the need for an additional positional encoding mechanism.

4.5.7 Hyperparameters

A hyperparameter study is performed to assess the sensitivity of both the MLP and transformer models to training configuration choices. The results are summarized in the following tables. Table 4.10 examines the impact of different learning rates on model performance. Table 4.11 evaluates the effect of step and OneCycle learning rate schedulers. Tables 4.12 and 4.13 explore variations in transformer embedding dimensions and the number of attention layers.

Overall, these ablations indicate that the model parameters and training configurations used in the main results provide strong and consistent performance. Two notable exceptions are observed. First, for the 10 mm source space, the MLP model trained with a learning rate of 1e−3 using the OneCycle scheduler outperforms the baseline MLP trained with a

Table 4.10: Ablation study on the learning rate for the 10mm regions. The accuracy is reported.

	MLP			Transformer		
Noise Level	1e-5	1e-4	1e-3	1e-5	1e-4	1e-3
0.01	1.000	1.000	1.000	1.000	1.000	0.999
0.10	1.000	0.999	1.000	1.000	1.000	0.996
0.25	0.986	0.992	0.996	0.997	0.998	0.980
0.50	0.860	0.940	0.950	0.956	0.960	0.902
0.75	0.617	0.819	0.827	0.836	0.848	0.745
0.99	0.400	0.651	0.661	0.672	0.686	0.579

step scheduler. However, the same configuration performs worse for the 5 mm source space, highlighting the MLP's sensitivity to training hyperparameters. Second, modest performance improvements are observed for the transformer when increasing the embedding dimension and number of layers. These gains come at the cost of substantially increased computational requirements and longer training times, making the baseline transformer configuration a more practical choice.

Table 4.11: Ablation study on the learning rate scheduler for the 10mm regions. The accuracy is reported.

	MLP			Transformer	
Noise Level	Step	Onecycle (1e-3)	Onecycle (1e-4)	Step	Onecycle
0.01	1.000	1.000	1.000	1.000	1.000
0.10	0.999	1.000	1.000	1.000	1.000
0.25	0.992	0.996	0.991	0.997	0.998
0.50	0.940	0.952	0.924	0.958	0.960
0.75	0.819	0.836	0.797	0.847	0.848
0.99	0.651	0.675	0.633	0.686	0.686

Table 4.12: The accuracy of the transformer model on 10mm regions with various sizes for the projection dimension. While the 768 dimensional transformer performs slightly better than the presented 512 method, it requires significantly more computational resources.

Noise Level	64 Transformer	512 Transformer	768 Transformer
0.50	0.958	0.960	0.960
0.75	0.839	0.848	0.850
0.99	0.676	0.686	0.690

Table 4.13: An ablation study on the accuracy of the transformer model on 10mm regions when varying number of layers. While the 4 dimensional transformer performs slightly better than the presented 3 layer method, it requires significantly more computational resources and training time.

Noise Level	2 Layer Transformer	3 Layer Transformer	4 Layer Transformer
0.50	0.957	0.960	0.962
0.75	0.839	0.848	0.850
0.99	0.675	0.686	0.691

Chapter 5

Conclusion and Future Work

A transformer-based model for EEG source localization is presented and evaluated for both single-source and multi-source settings. The results show improved performance compared to classical methods and baseline neural networks, particularly in more challenging conditions such as high noise levels and electrode dropout. In single-source experiments, the transformer consistently achieves highest accuracy and lowest mean localization error across most noise levels.

The model also extends naturally to multi-source prediction by increasing the number of learned embedding tokens and using a permutation-invariant loss. In experiments with two sources separated by 10 cm, the transformer maintains reasonable performance at moderate noise levels, while other methods degrade more quickly. In addition, the model shows robustness to missing data. When five electrodes are randomly removed, performance decreases only slightly compared to the no-dropout case. For example, at 99% noise on the 10 mm regions, accuracy decreases by 3.7% and mean localization error increases by 1.96 mm. These results suggest that the transformer can better handle noisy inputs, multiple sources, and incomplete measurements, which are common challenges in real EEG data.

One limitation of this work is the reliance on synthetically generated data for both training and evaluation. For practical use, the model should be validated on real EEG recordings, including both evoked and resting-state data, to assess its generalization across different subjects, head models, and preprocessing pipelines. Additional considerations include evaluating performance across different electrode layouts and incorporating uncertainty estimates into

the predictions.

Future work could extend this approach to include temporal information. By incorporating recurrent structures or additional attention mechanisms, time-dependent EEG signals could be modeled within a similar framework. Integration with existing EEG analysis toolboxes, such as MNE [7], would also improve usability. Further evaluation across different head models and datasets would help better understand the generalization capabilities of the transformer approach.

Finally, EEG source localization is still an area with many different methods and evaluation setups. By making the code and data from this work publicly available, the goal is to provide a consistent benchmark for future research and support further development of machine learning approaches for EEG source localization.

Bibliography

[1] Dannhauer, M., Lämmel, E., Wolters, C. H., and Knösche, T. R. Spatio-temporal regularization in linear distributed source reconstruction from eeg/meg: a critical evaluation. *Brain topography 26* (2013), 229–246.

[2] Delatolas, T., Antonakakis, M., Sakkalis, V., Wolters, C. H., and Zervakis, M. Eeg source analysis with a convolutional neural network and finite element analysis. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2023), pp. 1–4.

[3] Ding, Y., Li, Y., Sun, H., Liu, R., Tong, C., Liu, C., Zhou, X., and Guan, C. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics 29*, 3 (2025), 1909–1918.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

[5] Eichelbaum, S., Dannhauer, M., Hlawitschka, M., Brooks, D., Knösche, T. R., and Scheuermann, G. Visualizing simulated electrical fields from electroencephalography and transcranial electric brain stimulation: A comparative evaluation. *NeuroImage 101* (2014), 513–530.

[6] Gomez, L. J., Dannhauer, M., and Peterchev, A. V. Fast computational optimization of tms coil placement for individualized electric field targeting. *Neuroimage 228* (2021), 117696.

[7] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. S. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience 7*, 267 (2013), 1–13.

[8] Hecker, L., Rupprecht, R., Tebartz Van Elst, L., and Kornmeier, J. Convdip: A convolutional neural network for better eeg source imaging. *Frontiers in Neuroscience 15* (2021).

[9] Hecker, L., Rupprecht, R., van Elst, L. T., and Kornmeier, J. Long-short term memory networks for electric source imaging with distributed dipole models. *bioRxiv* (2022).

[10] Helmholtz, H. v. Ueber einige gesetze der vertheilung elektrischer ströme in körperlichen leitern, mit anwendung auf die thierisch-elektrischen versuche (schluss.). *Annalen der Physik 165*, 7 (1853), 353–377.

[11] Huang, G., Liu, K., Liang, J., Cai, C., Gu, Z. H., Qi, F., Li, Y., Yu, Z. L., and Wu, W. Electromagnetic source imaging via a data-synthesis-based convolutional encoder–decoder network. *IEEE Transactions on Neural Networks and Learning Systems 35*, 5 (2024), 6423–6437.

[12] Jatoi, M. A., Kamel, N., Malik, A. S., and Faye, I. Eeg based brain source localization comparison of sloreta and eloreta. *Australasian physical & engineering sciences in medicine 37* (2014), 713–721.

[13] Jiao, M., Wan, G., Guo, Y., Wang, D., Liu, H., Xiang, J., and Liu, F. A graph fourier transform based bidirectional long short-term memory neural network for electrophysiological source imaging. *Frontiers in Neuroscience 16* (2022).

[14] Liang, J., Yu, Z. L., Gu, Z., and Li, Y. Electromagnetic source imaging with a combination of sparse bayesian learning and deep neural network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 31* (2023), 2338–2348.

[15] Parker, S. G., and Johnson, C. R. Scirun: a scientific programming environment for computational steering. In *Proceedings of the 1995 ACM/IEEE conference on Supercomputing* (1995), pp. 52–es.

[16] Pascual-Marqui, R. D. Discrete, 3d distributed, linear imaging methods of electric neuronal activity. part 1: exact, zero error localization. *arXiv preprint arXiv:0710.3341* (2007).

[17] Pascual-Marqui, R. D., et al. Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find Exp Clin Pharmacol 24*, Suppl D (2002), 5–12.

[18] Polydorides, N., and Lionheart, W. R. A matlab toolkit for three-dimensional electrical impedance tomography: a contribution to the electrical impedance and diffuse optical reconstruction software project. *Measurement science and technology 13*, 12 (2002), 1871.

[19] Reynaud, S., Merlini, A., Ben Salem, D., and Rousseau, F. Comprehensive analysis of supervised learning methods for electrical source imaging. *Frontiers in Neuroscience 18* (2024).

[20] Sarah, R., Adrien, M., Douraied, B. S., and François, R. Eeg source imaging by supervised learning. In *2023 31st European Signal Processing Conference (EUSIPCO)* (2023), pp. 1170–1174.

[21] Smith, L. N., and Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications* (2019), vol. 11006, SPIE, pp. 369–386.

[22] Song, Y., Zheng, Q., Liu, B., and Gao, X. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 31* (2023), 710–719.

[23] Sun, R., Sohrabpour, A., Worrell, G. A., and He, B. Deep neural networks constrained by neural mass models improve electrophysiological source imaging of spatiotemporal brain dynamics. *Proceedings of the National Academy of Sciences 119*, 31 (2022), e2201128119.

[24] Thielscher, A., Antunes, A., and Saturnino, G. B. Field modeling for transcranial magnetic stimulation: A useful tool to understand the physiological effects of tms? In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (2015), IEEE, pp. 222–225.

[25] Van Veen, B. D., Van Drongelen, W., Yuchtman, M., and Suzuki, A. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on biomedical engineering 44*, 9 (1997), 867–880.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[27] Wan, Z., Li, M., Liu, S., Huang, J., Tan, H., and Duan, W. Eegformer: A transformer–based brain activity classification method using eeg signal. *Frontiers in Neuroscience Volume 17 - 2023* (2023).

[28] Wei, C., Lou, K., Wang, Z., Zhao, M., Mantini, D., and Liu, Q. Edge sparse basis network: A deep learning framework for eeg source localization. In *2021 International Joint Conference on Neural Networks (IJCNN)* (2021), pp. 1–8.

[29] Yang, R., and Modesitt, E. Vit2eeg: Leveraging hybrid pretrained vision transformers for eeg data.

[30] Zheng, T., and Guan, Z. Eeg source imaging based on a transformer encoder network. In *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)* (2023), pp. 208–212.

[31] Zorzos, I., Kakkos, I., Ventouras, E. M., and Matsopoulos, G. K. Advances in electrical source imaging: A review of the current approaches, applications and challenges. *Signals 2*, 3 (2021), 378–391.