

ELASTIQ: EEG–LANGUAGE ALIGNMENT WITH SEMANTIC TASK INSTRUCTION AND QUERYING

Muyun Jiang^{1*}, Shuailei Zhang^{1*}, Zhenjie Yang^{2*}, Mengjun Wu¹,
Weibang Jiang², Zhiwei Guo¹, Wei Zhang¹, Rui Liu¹, Shangen Zhang³,
Yong Li^{1,4}, Yi Ding^{1,†}, & Cuntai Guan^{1,†}

¹Nanyang Technological University ²Shanghai Jiao Tong University

³University of Science and Technology Beijing, Beijing ⁴Southeast University
james.jiang@ntu.edu.sg, ctguan@ntu.edu.sg

ABSTRACT

Recent advances in electroencephalography (EEG) foundation models, which capture transferable EEG representations, have greatly accelerated the development of brain–computer interfaces (BCI). However, existing approaches still struggle to incorporate language instructions as prior constraints for EEG representation learning, limiting their ability to leverage the semantic knowledge inherent in language to unify different labels and tasks. To address this challenge, we present **ELASTIQ**, a foundation model for **EEG–Language Alignment with Semantic Task Instruction and Querying**. ELASTIQ integrates task-aware semantic guidance to produce structured and linguistically aligned EEG embeddings, thereby enhancing decoding robustness and transferability. In the pretraining stage, we introduce a joint **Spectral–Temporal Reconstruction (STR) module**, which combines frequency masking as a global spectral perturbation with two complementary temporal objectives: random masking to capture contextual dependencies and causal masking to model sequential dynamics. In the instruction tuning stage, we propose the **Instruction-conditioned Q-Former (IQF)**, a query-based cross-attention transformer that injects instruction embeddings into EEG tokens and aligns them with textual label embeddings through learnable queries. We evaluate ELASTIQ on 20 datasets spanning motor imagery, emotion recognition, steady-state visual evoked potentials, covert speech, and healthcare tasks. ELASTIQ achieves state-of-the-art performance on 14 of the 20 datasets and obtains the best average results across all five task categories. Importantly, our analyses reveal for the first time that explicit task instructions serve as semantic priors guiding EEG embeddings into coherent and linguistically grounded spaces. The code and pre-trained weights will be released.

1 INTRODUCTION

Electroencephalography (EEG) provides non-invasive brain dynamics measurement with millisecond-level temporal resolution, making it particularly suitable for applications such as motor imagery (MI) decoding, emotion recognition, and steady-state visual evoked potential (SSVEP) classification. In addition to its high temporal precision, EEG offers the advantages of portability, relatively low cost, and suitability for long-term monitoring. However, EEG suffers from low signal-to-noise ratio, nonstationarity, and large variability across subjects, datasets, and tasks, which has historically limited its generalizability (Edelman et al., 2024). These shortcomings motivate the development

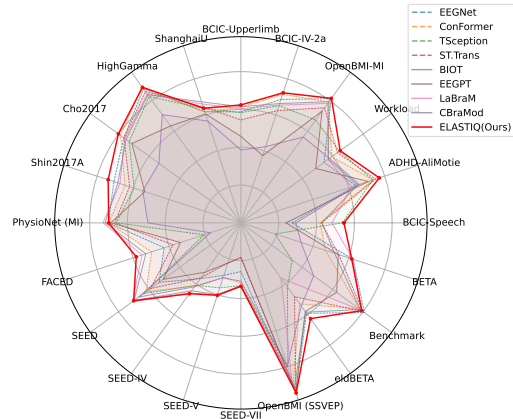


Figure 1: Comparison of ELASTIQ and baseline models on 20 different datasets.

of EEG foundation models (EEG-FMs), which aim to leverage large-scale pretraining to learn transferable representations that can overcome variability and improve downstream task performance. Typically, EEGPT (Wang et al., 2024a) applies transformer-based pretraining to capture temporal dependencies. LaBraM (Jiang et al., 2024c) leverages masked autoencoding on large EEG corpora to learn generalizable embeddings. CBraMod (Wang et al., 2024b) focuses on cross-brain modeling to facilitate cross-subject transfer. These EEG-FMs approaches constrained model training with discrete task labels (e.g., 0/1 instead of happy/angry), thereby discarding task-relevant semantic information. This absence of EEG-language coupling constraints may partly result in limited generalization. More recently, Large Language Models (LLMs) have been introduced to further enhance EEG-FMs due to their tremendous success in natural language processing (Touvron et al., 2023) and multimodal understanding (Radford et al., 2021). As a pioneering work, NeuroLM (Jiang et al., 2024b) aligns EEG and language embeddings by training a text-aligned neural tokenizer. Specifically, EEG signals are discretized into tokens and adversarially forced into the same embedding space as text. These EEG tokens are then added to the LLM vocabulary and jointly modeled with text through multi-channel autoregression and instruction tuning. While promising, current EEG-FMs and EEG-language FMs still face two major limitations: First, temporal and frequency information are reconstructed in separate branches, which may limit the ability to jointly capture cross-domain dependencies that are inherent in EEG signals. Second, instructions are only simply concatenated to the EEG sequence rather than being semantically integrated into EEG features, which restricts the model’s capacity to leverage language for guiding EEG representations.

To address these two problems, we propose **ELASTIQ**, a foundation model for EEG–Language Alignment with Semantic Task Instruction and Querying. This approach first introduces a joint spectral–temporal reconstruction framework that unifies frequency and temporal modeling. By combining global spectral perturbation with complementary temporal masking strategies, the model learns frequency-aware and contextually rich EEG representations, laying a stronger foundation for downstream tasks. To further bridge EEG signals with semantic information, we propose an Instruction-conditioned Q-Former (IQF) that aligns EEG representations with natural language. Specifically, EEG embeddings are modulated by the task-level instructions (e.g., “This is an MI task”, “Decode emotion from EEG”) and label semantics (e.g., “Left”, “Happy”), thereby guiding representation learning toward task-relevant dimensions. The modulated EEG features are then refined through a cross-attention mechanism with learnable queries, enabling instruction-driven alignment between EEG and language representations.

Our main contributions are as follows:

- We propose a novel **EEG–Language Foundation Model (ELASTIQ)** for EEG decoding across various BCI applications. ELASTIQ unifies spectral–temporal modeling with semantic task guidance to enrich EEG representations and enhance transferability across diverse downstream tasks.
- We design two key components to realize this framework: a **Spectral–Temporal Reconstruction (STR)** module that jointly captures frequency and temporal dynamics, and an **Instruction-conditioned Q-Former (IQF)** that integrates task instructions and label semantics into EEG features through query-based cross-modal alignment.
- We conduct a comprehensive evaluation on 20 EEG datasets spanning motor imagery, emotion, SSVEP, covert speech, and healthcare tasks. ELASTIQ achieves state-of-the-art (SOTA) average performances across all 5 tasks and demonstrates strong generalization across datasets.
- For the first time, we demonstrate that explicit instructions act as semantic priors that restructure EEG feature spaces for better separability, and that stronger text encoders supply richer semantics, leading to faster convergence, higher accuracy, and improved generalization.

2 METHOD

In this section, we introduce the design of **ELASTIQ**, our proposed EEG–Language foundation model. ELASTIQ is trained in two stages: a pretraining stage, where a joint spectral–temporal objective encourages frequency-aware and temporally predictive EEG representations, and an instruction tuning stage, where EEG embeddings are conditioned on task instructions and aligned with textual targets to improve decoding performance across diverse tasks. The architecture design of ELASTIQ can be found at Figure 2.

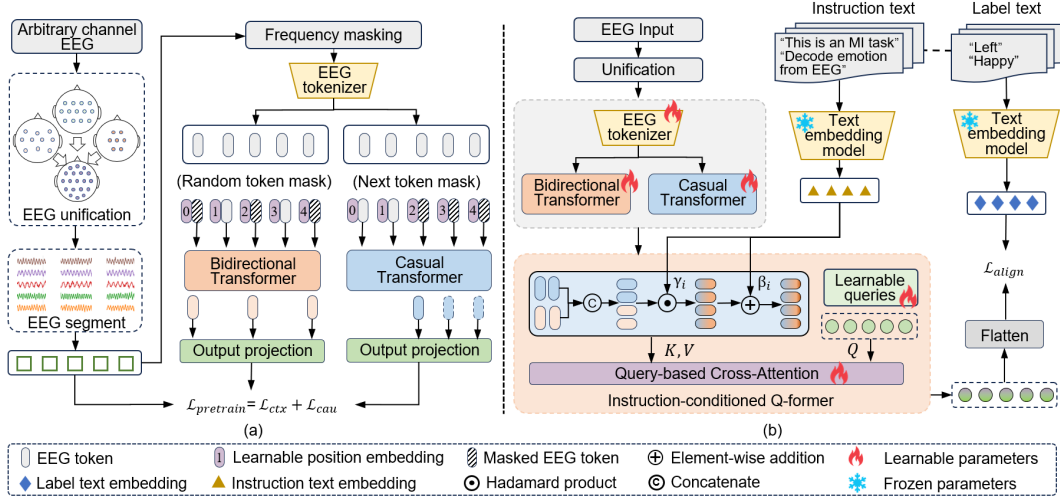


Figure 2: The architecture design of ELASTIQ. **(a)** joint Spectral-Temporal Reconstruction module (STR) for self-supervised EEG pretraining, combining frequency masking, global context modeling, and temporal sequence learning. **(b)** During instruction tuning, an Instruction-conditioned Q-Former (IQF) aligns EEG signals with language by injecting instruction embeddings and leveraging query-based cross-attention.

2.1 PRETRAINING STAGE

The pretraining of ELASTIQ learns EEG features that are frequency-aware, structurally consistent, and temporally predictive through a joint Spectral-Temporal Reconstruction framework, which serves as a conditional bottleneck enforcing spectral invariance and contextual recovery. Let $X^{C \times T} \in \mathbb{R}$ denote EEG trial with C means EEG channels and T time points. To manage long recordings and improve training stability, X is segmented into a sequence of non-overlapping windows of fixed length t , yielding segments $x_i^{C \times t} \in \mathbb{R}$ for $i = 1, \dots, \lfloor T/t \rfloor$. Each segment captures synchronized activity across all channels within the temporal window.

Spectral Masking To realize a random frequency band cutoff, we first suppress a randomly chosen frequency band in each segment before tokenization. For x_i , compute its spectrum $X_{f,i} = \text{FFT}(x_i)$ and randomly select a band $[f_{\min}, f_{\max}]$ of width f_{band} to remove, producing a masked spectrum $\tilde{X}_{f,i} = \mathcal{M}_{[f_{\min}, f_{\max}]}(X_{f,i})$. Then we conduct an inverse transform to get the perturbed signal via $\tilde{x}_i = \text{iFFT}(\tilde{X}_{f,i})$. This encourages invariance to the loss of localized spectral components and complements the dual spectral-temporal objectives.

Convolutional Tokenizer We adopt a lightweight tokenizer consisting of a temporal convolution, a spatial convolution, batch normalization, and pooling:

$$\tilde{\mathbf{z}}_i = \text{Tokenizer}(\tilde{x}_i) = \text{Pool}(\text{BatchNorm}(\text{Conv}_S(\text{Conv}_T(\tilde{x}_i)))) \quad (1)$$

The resulting token embeddings $\tilde{\mathbf{z}}_i$ lie in $\mathbb{R}^{N \times d}$, where N is the number of tokens and d the embedding dimension.

Mask token modeling branch A bidirectional transformer is trained with a random masking strategy, where a subset of token positions $\mathcal{M}$ is replaced by mask tokens and the model reconstructs the corresponding input token $\tilde{\mathbf{z}}_i$ from the unmasked context. Each reconstructed token is then mapped back to the input space through a two-layer MLP decoder:

$$g(\tilde{\mathbf{z}}_i) = W_2 \sigma(W_1 \tilde{\mathbf{z}}_i + b_1) + b_2, \quad (2)$$

where $\sigma(\cdot)$ denotes a non-linear activation, W_1 and W_2 are learnable weight matrices, and b_1 and b_2 are the corresponding bias terms. The reconstruction loss is computed against the original input

segment:

$$\mathcal{L}_{ctx} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|g(\tilde{\mathbf{z}}_i) - x_i\|_2^2. \quad (3)$$

Causal token modeling branch A causal transformer is optimized with a future mask, restricting each token at position i to attend only to $\{1, \dots, i\}$, thus preventing information leakage from the future. This imposes an autoregressive task in which the model predicts the next-token representation $\tilde{\mathbf{z}}_{i+1}$. The prediction is decoded through the same two-layer MLP, yielding $g(\tilde{\mathbf{z}}_{i+1})$, and the next-token loss is defined as

$$\mathcal{L}_{cau} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|g(\tilde{\mathbf{z}}_{i+1}) - x_{i+1}\|_2^2. \quad (4)$$

Joint objective The overall pretraining objective combines structural and temporal components:

$$\mathcal{L}_{pretrain} = \lambda_{ctx} \mathcal{L}_{ctx} + \lambda_{cau} \mathcal{L}_{cau}, \quad (5)$$

where λ_{ctx} and λ_{cau} are balancing coefficients (set to 1 by default). This design enforces that latent tokens must be decodable through $g(\cdot)$ back into the input domain, ensuring that the learned representations remain both contextually and temporally consistent with the original signals.

2.2 INSTRUCTION TUNING STAGE

The goal of instruction tuning in ELASTIQ is to close the modality gap between EEG signals and natural language semantics by learning conditional representations that align neural dynamics with textual meaning. To incorporate semantic guidance, each trial x_i is paired with a natural language instruction s_{ins} that specifies the task context (e.g., “Decode motor imagery”, “Decode emotional states”), together with a textual target s_{tgt} corresponding to the class label (e.g., “Left Hand”, “Happy”). The objective of instruction tuning is then defined as

$$\min_{\theta} \mathcal{L}(f_{\theta}(x_i, s_{ins}), s_{tgt}), \quad (6)$$

where f_{θ} denotes the model parameterized by θ , which conditions EEG embeddings on the instruction and aligns them with the textual target.

Instruction as a Conditioning Prior Let $\mathbf{m} \in \mathbb{R}^{2N \times d}$ denote the sequence of tokenized EEG embeddings obtained from the pretraining encoder, where $2N$ is the number of concatenated tokens output by both the transformer and d is the embedding dimension. Given the instruction text s_{ins} , we obtain its embedding $\mathbf{e}_{ins} \in \mathbb{R}^k$ using a frozen pretrained language encoder such as BERT (Devlin et al., 2019) or SBERT (Reimers & Gurevych, 2019).

$$\mathbf{e}_{ins} = f_{\text{text}}(s_{ins}), \quad (7)$$

where $f_{\text{text}}(\cdot)$ maps the instruction sentence into a normalized semantic vector. For **BERT**, we use the hidden state of the [CLS] token as the sentence embedding. For **SBERT**, the model directly outputs a sentence embedding via mean pooling over token representations. In both cases, the resulting vector is ℓ_2 -normalized:

$$\mathbf{e}_{ins} = \frac{\mathbf{e}_{ins}}{\|\mathbf{e}_{ins}\|_2}. \quad (8)$$

This normalized vector serves as the high-level semantic prior for guiding EEG representations toward the language semantic space.

To fuse this conditioning prior with the EEG embedding space, we employ a Feature-wise Linear Modulation (FiLM) operator (Perez et al., 2017), which parametrizes a family of affine transformations. Specifically, the modulation parameters $(\gamma, \beta) \in \mathbb{R}^d$ are derived from the instruction embedding via a nonlinear projection:

$$(\gamma, \beta) = \tanh(W_{\gamma\beta} \mathbf{e}_{ins} + \mathbf{b}_{\gamma\beta}), \quad (9)$$

where $W_{\gamma\beta} \in \mathbb{R}^{2N \times d}$ and $\mathbf{b}_{\gamma\beta} \in \mathbb{R}^{2N}$ jointly generate scaling and shifting coefficients. The conditioned EEG representation is then given by:

$$\tilde{\mathbf{m}} = \gamma \odot \mathbf{m} + \beta, \quad (10)$$

where $\odot$ denotes element-wise multiplication. This formulation ensures that $\mathbf{m}$ is shaped by instruction semantics rather than generic alignment. The instruction embedding $\mathbf{e}_{\text{ins}}$ biases the EEG latent space toward task-relevant features, producing representations that are both aligned with textual targets $\mathbf{e}_{\text{tgt}}$ and regularized on an instruction-informed manifold for improved semantic fidelity and generalization.

Query-based Cross-Attention To align between neural representations and instruction semantics, we introduce a set of N_q learnable query vectors $\mathbf{Q}_0 \in \mathbb{R}^{N_q \times d}$, which function as compact latent probes. Rather than directly inheriting the full complexity of the EEG embedding space, these queries serve as bottlenecks through which information must be filtered. The proposed Instruction-conditioned Q-Former (IQF) employs cross-attention to couple $\mathbf{Q}_0$ with the instruction-modulated EEG embeddings $\tilde{\mathbf{z}}$, yielding

$$\mathbf{Q} = \text{QFormer}(\mathbf{Q}_0, \tilde{\mathbf{m}}) = \text{softmax}\left(\frac{\mathbf{Q}_0 W_Q (\tilde{\mathbf{m}} W_K)^\top}{\sqrt{d}}\right) \tilde{\mathbf{m}} W_V, \quad (11)$$

where W_Q, W_K, W_V denote the query, key, and value projections, and d is the key dimension.

Conceptually, this operation projects EEG embeddings onto a lower-rank query subspace regularized by the instruction prior. The learnable queries act as semantic filters, retaining task-relevant features while suppressing irrelevant variance. This constrained information flow can be viewed as conditional information maximization, yielding embeddings that are semantically consistent and generalizable. The refined EEG representation $\mathbf{h} \in \mathbb{R}^k$ is obtained by aggregating the query outputs $\mathbf{Q}$ through a multilayer perceptron, serving as a task-adapted summary of the EEG signal.

Semantic Alignment with Textual Targets Given the ground-truth label $y \in \mathcal{C}$, we obtain its textual prototype embedding $\mathbf{e}_{\text{tgt}} \in \mathbb{R}^k$ by encoding the corresponding class name with the same frozen language model used for instructions:

$$\mathbf{e}_{\text{tgt}} = f_{\text{text}}(s_{\text{tgt}}), \quad (12)$$

Using the same encoder for both instructions and labels ensures that they are represented in a shared semantic space. To align $\mathbf{h}$ with its semantic prototype, we minimize a cosine similarity loss between them:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(1 - \cos(\mathbf{h}, \mathbf{e}_{\text{tgt}}^c)\right) \mathbb{I}[y = c], \quad (13)$$

where $\mathbb{I}[y = c]$ is the indicator function for the ground-truth class. This objective enforces $\mathbf{h}$ to be maximally aligned with its corresponding prototype $\mathbf{e}_{\text{tgt}}$ while remaining orthogonal to irrelevant ones. As a result, EEG embeddings and language prototypes cohabit a shared latent space $\mathcal{Z}$, where distances reflect cross-modal semantic consistency and class-level discriminability.

3 EXPERIMENTS

3.1 DATASET

Pretraining dataset We use 9 datasets, namely Stieger2021 (Stieger et al., 2021), SEED-FRA (Liu et al., 2022b), SEED-GER (Liu et al., 2022b), SEED-SD (Li et al., 2025), SEED-Neg, ChineseEEG (Mou et al., 2024), Chisco (Zhang et al., 2024), LargeSpanish, ThinkOutLoud (Nieto et al., 2022) as the pretraining datasets. The total duration of these datasets is around 1153 hours. More details about the pretraining datasets can be found in Appendix F.

Downstream Dataset We systematically evaluate our ELASTIQ on the five different BCI tasks with 20 datasets in total. **Motor Imagery:** OpenBMI-MI (Lee et al., 2019), BCIC-IV-2a (Tangermann et al., 2012), BCIC-Upperlimb (Jeong et al., 2022), SHU-MI-MI (Yang et al., 2025), High-Gamma (Schirmermeister et al., 2017), Cho2017 (Cho et al., 2017), Shin2017A (Shin et al., 2016), PhysioNet-MI (Schalk et al., 2004). **Emotion:** SEED (Duan et al., 2013), SEED-IV (Zheng et al., 2018), SEED-V (Liu et al., 2021), SEED-VII (Jiang et al., 2024a), FACED (Chen et al., 2023). **SSVEP:** OpenBMI-SSVEP (Lee et al., 2019), BETA (Liu et al., 2020), eldBETA (Liu et al., 2022a), Benchmark (Wang et al., 2016). **Covert speech:** BCIC2020-3 (Jeong et al., 2022). **Healthcare:** ADHD-AliMotie (Nasrabadi et al., 2020), Mental Workload (Zyma et al., 2019). More details about the downstream datasets can be found in Appendix G.

3.2 EXPERIMENTAL SETUP

Baselines & Metrics In this paper, we selected both the state-of-the-art traditional models and the EEG-FMs as baselines. For the traditional models, we selected EEGNet (Lawhern et al., 2018), TSception (Ding et al., 2022), ST-Transformer (Song et al., 2021) and Conformer (Song et al., 2022). For the EEG foundation model, we selected BIOT (Yang et al., 2023), EEGPT (Wang et al., 2024a), LaBraM (Jiang et al., 2024c), CBraMod (Wang et al., 2024b). To provide a reliable evaluation across imbalanced datasets, we adopted **balanced accuracy** and **Cohen’s Kappa** as performance metrics. Balanced accuracy accounts for class imbalance by averaging recall across classes, while Cohen’s Kappa measures the agreement between predicted and true labels beyond chance level, providing a more robust assessment of model performance.

EEG Preprocessing and Unification EEG recordings from different studies typically use diverse electrode montages. ELASTIQ performs channel unification by interpolating all signals onto the standardized 10–10 electrode layout, with 65 channels in Appendix K. For datasets recorded with fewer than 65 channels, we perform spatial interpolation to enforce a consistent topological structure across inputs. Then, signals are downsampled to 200 Hz. The MI datasets are band-pass filtered to 0.3–40 Hz, while all other datasets are filtered to 0.3–70 Hz. For segmentation, we distinguish between pre-segmented and continuous datasets. For pre-segmented datasets such as MI, SSVEP, and Covert Speech, we directly use the original trial-based divisions provided by the dataset. For continuous datasets, we follow dataset-specific conventions: FACED recordings are divided into 10-second windows, the SEED series datasets are split into 4-second segments, and the Workload dataset is segmented into 5-second windows. More details about preprocessing and evaluation are provided in Appendix H.

Evaluation settings To ensure fair and meaningful comparisons, we designed our evaluation protocol around the availability of subjects in each dataset. In line with conventional statistical practice (Snedecor & Cochran, 1989), we adopt **30** subjects as a practical threshold to distinguish between small- and large-scale datasets. For datasets with more than **30** participants, we adopted a **cross-subject transfer setting**: the training and validation set includes EEG data from a subset of subjects, while the test set is drawn from entirely unseen individuals. This split directly assesses the model’s ability to generalize across subjects. For datasets with fewer than 30 participants, we employed a **multi-subject adaptation setting**: for each subject, a portion of their trials is allocated to training and validation, while the remaining trials are held out for testing. This strategy balances two considerations: (i) it guarantees enough training samples per individual to stabilize learning, and (ii) it still evaluates robustness to intra-subject variability across sessions or trials. Our evaluation settings were chosen to reflect realistic deployment scenarios: large datasets test generalization across people, while small datasets emphasize consistency within individuals under varying conditions.

Implementation Details Pre-training and instruction tuning are both conducted in an end-to-end manner. Detailed model hyperparameters and training parameters are provided in Appendix D. Total trainable parameters are around 26.42 M. Our model is trained using a 4xH100 cluster with PyTorch.

3.3 EXPERIMENTAL RESULTS

We compare ELASTIQ with baselines on 20 downstream datasets in Table 1. **Motor Imagery:** Results show that our proposed method achieves the state-of-the-art performance on five MI datasets. Specifically, our method achieves the best performance improvement in BCIC-IV-2a (72.34%), OpenBMI-MI (81.44%), BCIC-Upperlimb (62.39%), Cho2017 (80.11%), and Shin2017A (73.84%). **Emotion Recognition:** Results show that our proposed method achieves the state-of-the-art performance on four emotion datasets. Specifically, our method provides the best performance improvement in FACED (58.19%), SEED-IV (46.3%), SEED-V (40.26%), and SEED-VII (33.56%). **SSVEP:** Results show that our proposed method achieves the state-of-the-art performance on two SSVEP datasets. Specifically, our method provides great performance improvement in OpenBMI-SSVEP (94.62%), eldBETA (62.62%). **Covert Speech:** Results show that our proposed method achieves the state-of-the-art performance (54.53%) on this covert speech dataset. **Healthcare:** Results show that our proposed method achieves the state-of-the-art performance on two datasets. Specifically, our method provides great performance improvement in ADHD-AliMotie

Table 1: Results across datasets with traditional and foundation EEG models. To facilitate comparison, the best three results per dataset are highlighted, where darker shading corresponds to higher performance.

Dataset	Metrics	Traditional models				EEG Foundation models				
		EEGNet	Conformer	TScep.	STTran.	BIOT	EEGPT	LaBraM	CBraMod	ELASTIQ
BCIC-IV-2a	B-Acc	0.6866	0.7151	0.6526	0.6243	0.4244	0.3743	0.6548	0.6637	0.7234
	Kappa	0.5818	0.6266	0.5367	0.4990	0.2326	0.1657	0.5420	0.5514	0.6311
OpenBMI-MI	B-Acc	0.8128	0.7904	0.6729	0.7527	0.5613	0.7306	0.7812	0.7925	0.8144
	Kappa	0.6234	0.5709	0.3458	0.5054	0.1225	0.4613	0.5850	0.5720	0.6287
BCIC-Upperlimb	B-Acc	0.5871	0.6123	0.5830	0.5448	0.3854	0.4577	0.6039	0.6143	0.6239
	Kappa	0.3851	0.4351	0.3775	0.3192	0.0788	0.1919	0.4145	0.4230	0.4391
SHU-MI	B-Acc	0.6183	0.6163	0.6365	0.6456	0.5664	0.6061	0.6612	0.6735	0.6374
	Kappa	0.2364	0.2326	0.2720	0.2913	0.1329	0.2121	0.3385	0.3468	0.2747
HighGamma	B-Acc	0.8770	0.8917	0.8263	0.8588	0.7076	0.6337	0.8425	0.8378	0.8861
	Kappa	0.8155	0.8475	0.7395	0.7882	0.5615	0.4506	0.7640	0.7567	0.8291
Cho2017	B-Acc	0.7661	0.7872	0.7311	0.7639	0.5361	0.7133	0.7598	0.7511	0.8011
	Kappa	0.5322	0.5744	0.4622	0.5278	0.0722	0.4267	0.5210	0.5022	0.6022
Shin2017A	B-Acc	0.7067	0.6444	0.5939	0.6189	0.5460	0.5342	0.6725	0.6844	0.7384
	Kappa	0.4529	0.2892	0.1874	0.2379	0.0915	0.0682	0.3597	0.3680	0.7364
PhysioNet-MI	B-Acc	0.7005	0.6986	0.6594	0.6678	0.4918	0.6881	0.7299	0.7182	0.6992
	Kappa	0.4010	0.3972	0.3189	0.3356	0.0163	0.3763	0.4520	0.4363	0.3983
FACED	B-Acc	0.4267	0.4968	0.2053	0.3773	0.1696	0.3360	0.5495	0.5290	0.5819
	Kappa	0.3509	0.4284	0.1086	0.2985	0.0642	0.2496	0.4843	0.4643	0.5243
SEED	B-Acc	0.5351	0.6203	0.6367	0.5911	0.6673	0.5082	0.7102	0.6988	0.7011
	Kappa	0.3151	0.4306	0.4602	0.3892	0.5033	0.2674	0.5628	0.5513	0.5543
SEED-IV	B-Acc	0.3648	0.4129	0.4089	0.3591	0.4181	0.3233	0.4591	0.4444	0.4630
	Kappa	0.1577	0.2139	0.1888	0.1405	0.1956	0.0869	0.2720	0.2577	0.2754
SEED-V	B-Acc	0.2921	0.3041	0.3616	0.2238	0.3058	0.2237	0.4015	0.3996	0.4026
	Kappa	0.1132	0.1327	0.1972	0.0343	0.1312	0.0333	0.2561	0.2554	0.2575
SEED-VII	B-Acc	0.2580	0.3177	0.3317	0.1850	0.3084	0.1823	0.3328	0.3319	0.3356
	Kappa	0.1409	0.2061	0.2209	0.0511	0.1932	0.0480	0.2248	0.2239	0.2275
OpenBMI-SSVEP	B-Acc	0.9419	0.8960	0.8044	0.9392	0.7963	0.9383	0.8713	0.9165	0.9462
	Kappa	0.9242	0.8614	0.7392	0.9189	0.7283	0.9178	0.8283	0.8886	0.9283
BETA	B-Acc	0.6244	0.4700	0.1909	0.5450	0.3350	0.5420	0.6236	0.5209	0.6163
	Kappa	0.6147	0.4564	0.1702	0.5333	0.3179	0.5233	0.6060	0.5087	0.5902
eldBETA	B-Acc	0.5913	0.5294	0.4190	0.4825	0.5437	0.5794	0.3619	0.5841	0.6262
	Kappa	0.5402	0.4705	0.3464	0.4179	0.4866	0.5268	0.2821	0.5321	0.5801
Benchmark	B-Acc	0.8077	0.7762	0.3357	0.7958	0.4762	0.6262	0.7879	0.7804	0.7902
	Kappa	0.8028	0.7705	0.3187	0.7906	0.4628	0.6166	0.7749	0.7747	0.7900
BCIC-Speech	B-Acc	0.2707	0.4187	0.5360	0.4267	0.2907	0.2387	0.4822	0.4288	0.5453
	Kappa	0.0883	0.2733	0.4200	0.2833	0.1133	0.0483	0.3865	0.2865	0.4317
ADHD-AliMotie	B-Acc	0.6392	0.7352	0.7341	0.7493	0.6581	0.7084	0.6158	0.6489	0.7699
	Kappa	0.2808	0.4704	0.4757	0.5032	0.3187	0.4193	0.2272	0.3148	0.5282
Mental Workload	B-Acc	0.5444	0.5944	0.6458	0.6319	0.5799	0.4896	0.5694	0.5722	0.6493
	Kappa	0.0976	0.1789	0.3024	0.2676	0.1465	0.0222	0.1656	0.1688	0.3134
Average	Macro-Acc	0.6012	0.6164	0.5483	0.5892	0.4884	0.5217	0.6236	0.6309	0.6678
	Kappa	0.4384	0.4572	0.3624	0.4140	0.2539	0.3205	0.4673	0.4766	0.5391

(76.99%) and Mental Workload (64.93%). In terms of overall performance, ELASTIQ attains an average macro-accuracy of 66.78% and an average Kappa of 53.91%, substantially surpassing the other baselines, thus demonstrating superior generalization across diverse EEG decoding scenarios. Interestingly, when averaged across all tasks, our proposed ELASTIQ, LaBraM, and CBraMod consistently rank as the top three performers.

3.4 EFFECT OF INSTRUCTION DETAIL DURING INFERENCE

To assess whether the model truly exploits instructions to modulate EEG representations, we measure its instruction sensitivity in a direct inference setting without fine-tuning. We report results under three instruction levels: **No Instructions**: Input *Default* or *None*; **Informed Task**: Provide an instruction specifying the EEG task; **Informed Task & Targets**: Provide both the task type and target classes (e.g., This is an MI task; decode *Left* vs. *Right*). The instruction-tuned model is directly applied to the held-out test sets without adaptation. The quantitative results in Table 2 corroborate these visual patterns. Compared to the “No Instruction” condition, providing task-level instructions yields an average gain of 1.77% in balanced accuracy, while the most detailed condition (task & targets) further improves performance to 2.12%, respectively.

Table 2: Performance comparison of selected datasets with different instruction conditions. Results are reported in B-Acc/Kappa. Best results are highlighted in bold.

Dataset	# Class	No Instructions	Informed Task	Informed Task & Targets
OpenBMI-MI	2	0.6271 / 0.2542	0.6606 / 0.3213	0.6660 / 0.3321
HighGamma	2	0.5533 / 0.3300	0.5882 / 0.3823	0.5898 / 0.3846
Cho2017	2	0.6483 / 0.2967	0.6656 / 0.3311	0.6806 / 0.3611
Shin2017A	2	0.6020 / 0.2051	0.6289 / 0.2588	0.6192 / 0.2395
SEED	3	0.5565 / 0.3417	0.5684 / 0.3592	0.5768 / 0.3716
SEED-IV	4	0.3695 / 0.1568	0.3874 / 0.1859	0.4035 / 0.2005
SEED-V	5	0.2942 / 0.1177	0.3036 / 0.1340	0.3060 / 0.1340
SEED-VII	7	0.2474 / 0.1222	0.2508 / 0.1262	0.2519 / 0.1273
FACED	9	0.3258 / 0.2374	0.3257 / 0.2374	0.3254 / 0.2373
OpenBMI-SSVEP	4	0.7692 / 0.6922	0.7929 / 0.7239	0.7935 / 0.7247
Average	—	0.4789 / 0.2128	0.4966 / 0.2422	0.5001 / 0.2473

3.5 VISUALIZATION OF LEARNED FEATURE SPACES UNDER DIFFERENT INSTRUCTION CONDITIONS

To better understand how instruction conditioning reshapes the latent representation space, we provide a qualitative visualization of the learned features in Figure 3. We first project the high-dimensional EEG embeddings into two dimensions using UMAP (McInnes et al., 2018), and then estimate their probability densities with kernel density estimation (KDE), where darker regions indicate higher sample concentration. Class labels are represented by fixed prototype points, obtained by encoding the corresponding label text and projecting them through the same model and UMAP mapping. Our analysis proves that without instructions, the feature embeddings exhibit weak separation, with substantial overlap between categories. By contrast, when instructions are provided, the feature distributions become more structured: clusters corresponding to different classes are pushed farther apart.

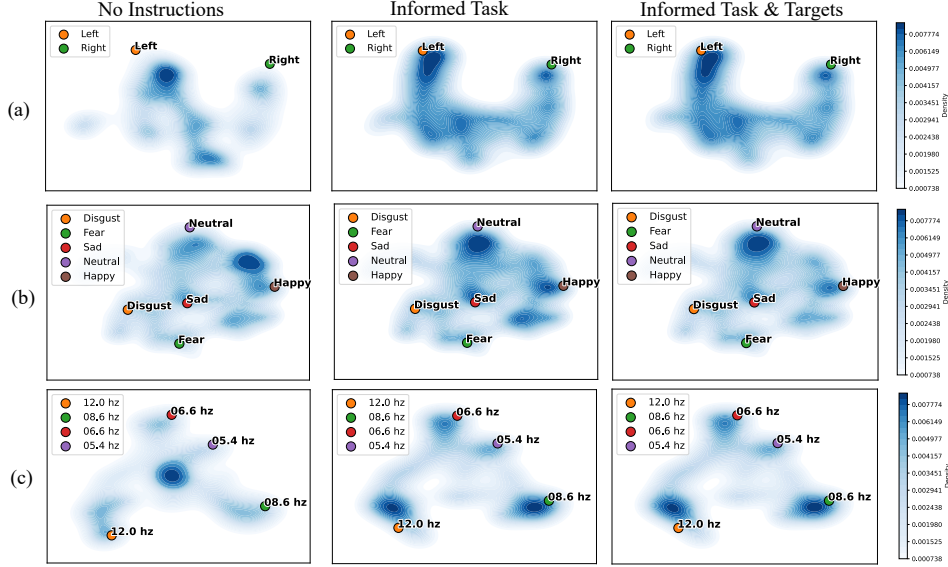


Figure 3: KDE visualization of EEG embeddings under different instruction settings (a) OpenBMI-MI, (b) SEED-V, (c) OpenBMI-SSVEP

3.6 EFFECT OF TARGET TEXT EMBEDDINGS AND INFERENCE REGIME ON ELASTIQ

To assess the role of target text embeddings, we replace them with label IDs and train the model using cross-entropy classification. Comparisons are conducted under both direct inference and fully fine-tuned settings. Results in Table 3 show that across datasets and settings, ELASTIQ consistently outperforms the baseline that relies solely on label IDs. We further evaluate variants using BERT-

and SBERT-based encoders, with the SBERT variant yielding superior performance, suggesting stronger EEG–text alignment from richer semantic representations.

Table 3: Comparison of ELASTIQ with and without target text embeddings under fully fine-tuned and direct inference settings. Best results are highlighted in bold.

Method	OpenBMI-MI		SEED	
	Fully fine-tune	Direct inference	Fully fine-tune	Direct inference
Without target text embedding	0.7912 / 0.5987	0.6515 / 0.3183	0.6328 / 0.4581	0.5426 / 0.3294
ELASTIQ (BERT base-uncased)	0.8077 / 0.6154	0.6617 / 0.3267	0.6448 / 0.4729	0.5493 / 0.3391
ELASTIQ (SBERT mpnet-base-v2)	0.8144 / 0.6287	0.6660 / 0.3321	0.7011 / 0.5543	0.5768 / 0.3716

3.7 EFFECT OF JOINT SPECTRAL-TEMPORAL RECONSTRUCTION MODULE

To evaluate the contribution of each strategy during STR pretraining, we performed an ablation study in which individual masking components were selectively removed. Results in Table 4 indicate that across OpenBMI-MI, BCI-IV-2a, and SEED, retaining all three masking strategies yields the best B-Acc/Kappa. These results indicate that all three components contribute to the model’s performance.

Table 4: Ablation study of masking strategies. Best results are highlighted in bold.

Frequency Mask	Random Mask	Causal Mask	OpenBMI-MI	BCI-IV-2a	SEED	SEED-IV
✓	✓		0.8078 / 0.6195	0.7093 / 0.6181	0.6882 / 0.5487	0.4689 / 0.2812
✓		✓	0.7896 / 0.6017	0.7015 / 0.6094	0.6771 / 0.5298	0.4462 / 0.2591
	✓	✓	0.8033 / 0.6152	0.7168 / 0.6265	0.6957 / 0.5499	0.4598 / 0.2721
✓	✓	✓	0.8144 / 0.6287	0.7234 / 0.6311	0.7011 / 0.5543	0.4630 / 0.2754

3.8 COMPARISON WITH NEUROLM

NeuroLM (Jiang et al., 2024b) is among the earliest studies exploring EEG-language alignment. To ensure a fair evaluation, ELASTIQ was re-implemented under the same settings as NeuroLM. Table 5 shows that ELASTIQ (26.4M params) outperforms much larger NeuroLM variants in SEED and Workload datasets.

Table 5: Comparison of ELASTIQ (ours) and NeuroLM. Best results are highlighted in bold.

Methods	#Params	SEED			Workload		
		Balanced Acc.	Cohen’s Kappa	Weighted F1	Balanced Acc.	AUC-PR	AUROC
NeuroLM-B	254 (M)	0.5554±0.0075	0.3393±0.0117	0.5599±0.0068	0.6172±0.0113	0.5824±0.0080	0.6253±0.0160
NeuroLM-L	500 (M)	0.6006±0.0047	0.4067±0.0063	0.6048±0.0050	0.6311±0.0250	0.5869±0.0155	0.6247±0.0339
NeuroLM-XL	1696 (M)	0.6034±0.0010	0.4082±0.0036	0.6063±0.0030	0.6345±0.0442	0.5889±0.0423	0.6130±0.0764
ELASTIQ (ours)	26.4 (M)	0.7029±0.0061	0.5543±0.0093	0.6904±0.0078	0.6423±0.0125	0.6021±0.0104	0.6321±0.0137

4 CONCLUSION

We introduced ELASTIQ, a foundation model for EEG–Language Alignment with Semantic Task Instruction and Querying. By combining a joint Spectral-Temporal Reconstruction module and an Instruction-conditioned Q-Former, ELASTIQ learns language-guided EEG representations that transfer effectively across tasks. Extensive evaluations on 20 datasets covering MI, emotion recognition, SSVEP, covert speech, and healthcare applications show that ELASTIQ achieves, on average, state-of-the-art (SOTA) performance, highlighting the value of instruction-informed alignment for generalizable EEG decoding. More importantly, our work establishes natural language as both an interpretable anchor and a transferable supervision signal, highlighting its central role in shaping future EEG foundation models and BCI systems.

REFERENCES

- Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.
- Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. Eeg datasets for motor imagery brain–computer interface. *GigaScience*, 6(7):gix034, 2017.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- Yi Ding, Neethu Robinson, Su Zhang, Qiuhaio Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2022.
- Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 81–84. IEEE, 2013.
- Bradley J Edelman, Shuailei Zhang, Gerwin Schalk, Peter Brunner, Gernot Müller-Putz, Cuntai Guan, and Bin He. Non-invasive brain-computer interfaces: state of the art and trends. *IEEE reviews in biomedical engineering*, 2024.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.
- Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.
- Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 2024a.
- Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neuroilm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024b.
- Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024c.
- Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025.
- Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.
- Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019.

- Ziyi Li, Le-Yan Tao, Rui-Xiao Ma, Wei-Long Zheng, and Bao-Liang Lu. Investigating the effects of sleep conditions on emotion responses with eeg signals and eye movements. *IEEE Transactions on Affective Computing*, 2025.
- Bingchuan Liu, Xiaoshan Huang, Yijun Wang, Xiaogang Chen, and Xiaorong Gao. Beta: A large benchmark database toward ssvep-bci application. *Frontiers in neuroscience*, 14:627, 2020.
- Bingchuan Liu, Yijun Wang, Xiaorong Gao, and Xiaogang Chen. eldbeta: a large eldercare-oriented benchmark database of ssvep-bci for the aging population. *Scientific data*, 9(1):252, 2022a.
- Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang 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, 2021.
- Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people. *Journal of Neural Engineering*, 19(2):026012, 2022b.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- Xinyu Mou, Cuilin He, Liwei Tan, Junjie Yu, Huadong Liang, Jianyu Zhang, Yan Tian, Yu-Fang Yang, Ting Xu, Qing Wang, et al. Chineseeeg: A chinese linguistic corpora eeg dataset for semantic alignment and neural decoding. *Scientific Data*, 11(1):550, 2024.
- Ali Motie Nasrabadi, Armin Allahverdy, Mehdi Samavati, and Mohammad Reza Mohammadi. EEG data for ADHD / Control children. IEEE Dataport, June 10 2020. URL <https://dx.doi.org/10.21227/rzfz-zn36>.
- Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific data*, 9(1):52, 2022.
- Ethan Perez, Harm De Vries, Florian Strub, Vincent Dumoulin, and Aaron Courville. Learning visual reasoning without strong priors. *arXiv preprint arXiv:1707.03017*, 2017.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.
- Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.
- Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Do-Won Kim, Jichai Jeong, Han-Jeong Hwang, and Klaus-Robert Müller. Open access dataset for eeg+ nirs single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2016.
- George W Snedecor and Witliam G Cochran. Statistical methods, 8thedn. *Ames: Iowa State Univ. Press Iowa*, 54:71–82, 1989.

- Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.
- Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.
- James R Stieger, Stephen Engel, Haiteng Jiang, Christopher C Cline, Mary Jo Kreitzer, and Bin He. Mindfulness improves brain–computer interface performance by increasing control over neural activity in the alpha band. *Cerebral Cortex*, 31(1):426–438, 2021.
- Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.
- Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024b.
- Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.
- Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. A multi-day and high-quality eeg dataset for motor imagery brain-computer interface. *Scientific Data*, 12(1):488, 2025.
- Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.
- Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*, 11(1):1265, 2024.
- Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3):1110–1122, 2018.
- Igor Zyma, Sergii Tukaev, Ivan Seleznev, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A RELATED WORK

Self-supervised Pretaining. Self-supervised pretraining has emerged as a powerful paradigm in representation learning, reducing the reliance on large amounts of annotated data while leveraging abundant unlabeled signals. Self-supervised methods design pretext tasks that encourage models to learn meaningful feature representations from the inherent structure of data. Early successes in natural language processing, such as BERT (Devlin et al., 2019) and GPT series (Radford et al., 2019), demonstrated that masked language modeling and next-word prediction can yield representations transferable to diverse downstream tasks. Similarly, in computer vision, contrastive learning approaches like SimCLR (Chen et al., 2020), MoCo (He et al., 2020) and masked image modeling like MAE (He et al., 2022) showed that pretraining on large-scale unlabeled images leads to robust and generalizable visual features.

EEG Foundation model. The concept of foundation models has recently expanded into the EEG domain, aiming to build large-scale pre-trained backbones that generalize across datasets, tasks, and clinical conditions. Several pioneering efforts have been proposed. BIOT (Yang et al., 2023) explored scalable transformer-based architectures for biomedical signals, positioning EEG as a central modality. EEGPT (Wang et al., 2024a), inspired by advances in language modeling, leveraged transformer pretraining strategies such as masked prediction and contrastive learning to enhance generalization across heterogeneous EEG datasets. LaBraM (Jiang et al., 2024c) introduced a large-brain-model framework, emphasizing cross-dataset pretraining to capture universal EEG representations. CBraMod (Wang et al., 2024b) extended this idea by focusing on cross-brain modularity, enabling adaptation across diverse cognitive and motor tasks. Beyond EEG-specific approaches, NeuroLM (Jiang et al., 2024b) proposed a broader neural language model for neuroscience data, while PhysioOmni (Jiang et al., 2025) further expanded the scope to multi-physiological modalities, integrating EEG with signals such as ECG and EMG to learn cross-modal representations. Collectively, these efforts highlight the emerging trajectory of EEG-FMs: moving from task-specific networks toward unified, pre-trained architectures capable of powering downstream applications with minimal fine-tuning, and paving the way for general-purpose brain decoding systems.

B EFFECT OF INCORRECT INSTRUCTIONS

To further examine the role of language guidance, we analyze cases where the model is deliberately given misleading instructions that do not match the underlying EEG dataset. Figure 4 shows examples on OpenBMI-MI and SEED-V datasets.

When provided with correct instructions, the learned feature spaces become more structured, with compact intra-class clusters and clearer inter-class separation. However, when misleading instructions are introduced, the feature space is distorted toward the semantics of the given instruction rather than the ground-truth task. For example, MI data conditioned on emotion-related instructions form clusters resembling affective categories, and SEED-V data prompted with MI or SSVEP instructions are reorganized into motor or frequency-based groupings.

These results emphasize the strong controllability of our model through natural language. While correct instructions enhance discriminability, misleading instructions actively reshape the representation space according to the semantic prior they provide. This highlights both the power and sensitivity of instruction-conditioned alignment in EEG-FMs.

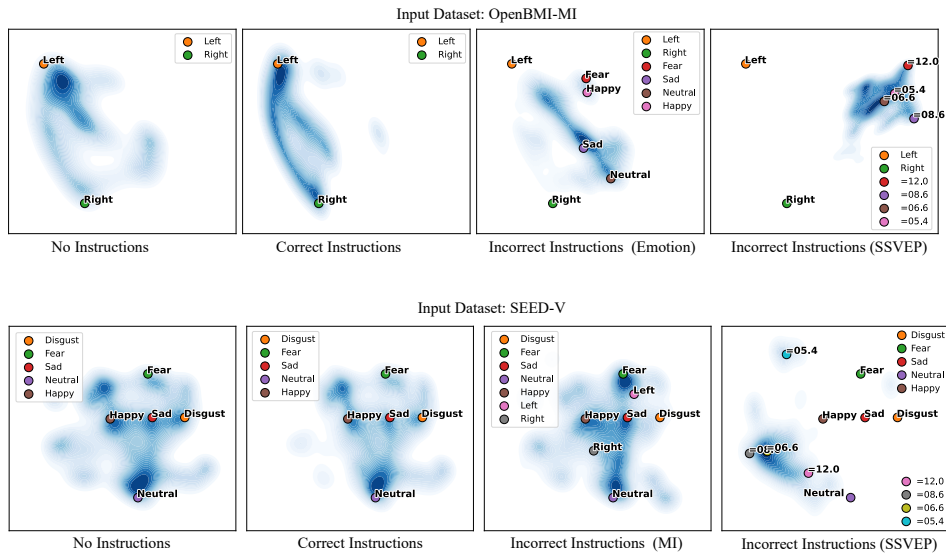


Figure 4: Comparison of KDE visualization of features between incorrect and correct instructions.

C TOKEN CLUSTERING VISUALIZATION FOR DUAL MASKING BRANCHES

We further visualize the feature learnt by the bidirectional transformer and the causal transformer. Results in Figure 5 yield clearly separable clusters, indicating that they extract distinct and complementary features from EEG signals. This validates the design of joint STR: by optimizing structural and temporal objectives separately, the model learns representations that capture different aspects of EEG dynamics, which together provide a richer and more transferable embedding.

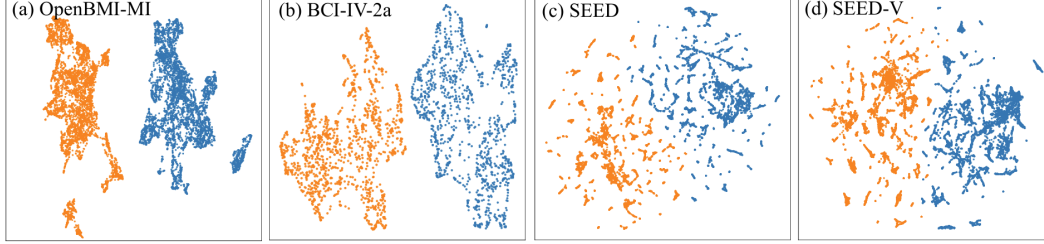


Figure 5: Token clustering after UMAP visualization. **Yellow points**: tokens from bidirectional transformer. **Blue points**: tokens from causal transformer

D PARAMETER SETTINGS

We list the hyperparameters selected in our model in Table 6, and training parameters in Table 7. Parameter size of ELASTIQ can be found at Table 8

Table 6: Hyperparameters for ELASTIQ

Hyperparameters	Value
<i>Tokenization</i>	
Sampling rate	200 hz
Segment window length	0.5s
Input channels	65
Temporal kernel size	(1, 40), padding (1, 20)
Spatial kernel size	(65, 1)
Pooling	(1, 10)
<i>Transformers</i>	
Transformer encoder layers	12
Token size	256
Feed-forward scale	4
Attention head number	8
Dropout	0.1
Mask ratio	0.5
Frequency cutoff range	1–50 Hz
Frequency cutoff Band	6 Hz continuous band
<i>Q-Former</i>	
Number of queries	8
Q-Former layers	4
Text embedding dimension	768
Text embedding model	bert-base-uncased / all-mpnet-base-v2

E INSTRUCTION AND TARGET TEXTS FOR DOWNSTREAM DATASETS

We list the instruction and target texts for each downstream dataset in Table 9.

Table 7: Training Parameters for ELASTIQ

Hyperparameters	Value
Batch size	512
Peak learning rate	1×10^{-3}
LR scale (transformer params)	0.1
LR scale (other params)	1.0
Minimal learning rate	1×10^{-4}
Learning rate scheduler	Cosine Annealing lr
Optimizer	AdamW ($\beta = 0.9, 0.999$)
Weight decay	1×10^{-3}
Precision	bf16-mixed

Table 8: Parameter Size of ELASTIQ

Hyperparameters	Value
Tokenizer	0.27 M
Dual Transformer	20.23 M
Q-Former	6.19 M
Total trainable parameters	26.42 M

F PRE-TRAINING DATASET DESCRIPTION

- SEED-FRA (Liu et al., 2022b): Eight French subjects participated in the experiments. Twenty-one film clips (positive, neutral and negative emotions) were chosen as stimuli used in the experiments.
- SEED-GER (Liu et al., 2022b): Eight German subjects participated in the experiments. Twenty film clips (positive, neutral and negative emotions) were chosen as stimuli used in the experiments.
- SEED-SD (Li et al., 2025): SEED-SD is a multimodal EEG and eye-tracking dataset collected from 40 healthy participants under three sleep-related conditions—sleep deprivation, sleep recovery, and normal sleep. In each condition, participants watched 24 video clips (six per emotion) designed to evoke four basic emotions: happiness, sadness, fear, and neutral; each clip lasts about 2.5 minutes.
- Chisco dataset (Zhang et al., 2024): The Chisco dataset is a large-scale EEG corpus collected from three subjects for imagined speech decoding, featuring over 20,000 sentences and more than 900 minutes of high-density EEG per subject. It covers 6,000+ everyday phrases across 39 semantic categories, with trials designed to include both reading and imagined speech phases.
- ThinkOutLoud (Nieto et al., 2022): This open-access EEG dataset comprises recordings from 10 participants, collected using a 136-channel system across three paradigms—inner speech, pronounced speech, and visualized condition.
- Stieger2021 (Stieger et al., 2021): This database contains EEG recordings from 62 healthy participants, each completing 7–11 sessions of BCI training to control a computer cursor in one- and two-dimensional spaces using motor imagery. Data were collected with 62 electrodes, and accompanying behavioral measures.

G MORE DETAILS FOR EXPERIMENTAL SETTING ON DOWNSTREAM DATASETS

- BCIC-IV-2a dataset (Tangemann et al., 2012) comprises recordings from nine subjects, each participating in two sessions of a four-class MI paradigm (left hand, right hand, foot, and tongue). EEG was collected using 22 scalp electrodes and three EOG channels at 250 Hz.

Table 9: Datasets, corresponding instruction formulations, and decoding targets.

Dataset	Instructions	Targets
OpenBMI-MI	Decode motor imagery; Decode (Left vs Right) hand motor imagery	Right, Left
BCIC-IV2a	Decode motor imagery; Decode (Left vs Right vs Foot vs Tongue) motor imagery	Left, Right, Foot, Tongue
BCIC-Upperlimb	Decode motor imagery; Decode (Cylindrical, Spherical, Lumbrical) hand movements	Cylin, Sphe, Lumbrical
SHU-MI	Decode motor imagery; Decode (Left vs Right) hand motor imagery	Right, Left
HighGamma	Decode motor imagery; Decode (Left vs Right vs Foot) motor imagery	Left, Right, Foot
Cho2017	Decode motor imagery; Decode (Left vs Right) hand motor imagery	Left, Right
Shin2017A	Decode motor imagery; Decode (Left vs Right) hand motor imagery	Left, Right
PhysioNet-MI	Decode motor imagery; Decode (Left vs Right) hand motor imagery	Left, Right
FACED	Decode emotional states; Decode emotional states (Anger, Fear, Disgust, Sadness, Amusement, Inspiration, Joy, Tenderness, Neutral)	Anger, Fear, Disgust, Sad, Amusement, Inspiration, Joy, Tenderness, Neutral
SEED	Decode emotional states; Decode emotional states (Positive, Neutral, Negative)	Positive, Neutral, Negative
SEED-IV	Decode emotional states; Decode emotional states (Neutral, Sad, Fear, Happy)	Neutral, Sad, Fear, Happy
SEED-V	Decode emotional states; Decode emotional states (Disgust, Fear, Sad, Neutral, Happy)	Disgust, Fear, Sad, Neutral, Happy
SEED-VII	Decode emotional states; Decode emotional states (Happy, Surprise, Neutral, Sad, Disgust, Fear, Anger)	Happy, Surprise, Neutral, Sad, Disgust, Fear, Anger
OpenBMI-SSVEP	Decode SSVEP; Decode SSVEP frequencies (5.4hz, 6.6hz, 8.6hz, 12.0hz)	12.0, 08.6, 06.6, 05.4
eldBETA	Decode SSVEP; Decode SSVEP frequencies (8.0hz, 9.5hz, 11.0hz, 8.5hz, 10.0hz, 11.5hz, 9.0hz, 10.5hz, 12.0hz)	08.0, 09.5, 11.0, 08.5, 10.0, 11.5, 09.0, 10.5, 12.0
Benchmark	Decode SSVEP; Decode SSVEP frequencies from 8.0hz to 15.8hz with 0.2hz interval, total 40 classes	40 freq. classes (8.0–15.8Hz, step 0.2Hz)
BETA	Decode SSVEP; Decode SSVEP frequencies from 8.0hz to 15.8hz with 0.2hz interval, total 40 classes	40 freq. classes (8.0–15.8Hz, step 0.2Hz)
BCIC-Speech	Decode covert speech; Decode covert speech (hello, help-me, stop, thank-you, yes)	hello, help-me, stop, thank-you, yes
ADHD-AliMotie	Decode mental disorder; Decode ADHD vs Healthy	Healthy, ADHD
Workload	Decode mental workload states; Decode mental workload states (Resting vs Workload)	Resting, Workload

- OpenBMI-MI dataset (Lee et al., 2019) provides a large-scale benchmark for brain-computer interface research. Its MI subset contains data from 54 subjects, each participating in two sessions. Subjects performed left- and right-hand motor imagery tasks,

with approximately 100 trials per session, recorded using a 64-channel EEG system at 1000 Hz.

- BCIC-Upperlimb dataset (Jeong et al., 2022) is from BCI Competition 2021 – Track 4. It provides EEG recordings of subjects performing three unilateral grasp movements (cylindrical, spherical, lumbrical) across three consecutive days (train/validation/test), designed to evaluate upper-limb movement decoding and session-to-session transfer.
- SHU-MI dataset (Yang et al., 2025) includes high-quality multi-day recordings from 62 participants. Fifty-one subjects performed a two-class MI paradigm (left vs. right hand grasping), while eleven subjects performed a three-class paradigm (left hand, right hand, and foot). Each participant contributed three sessions, with both raw and preprocessed EEG data publicly available.
- High-Gamma dataset (Schirrneister et al., 2017) was collected at TU Berlin and contains 128-channel EEG recordings from 14 subjects. Participants performed four tasks (left hand, right hand, both feet, and rest). Each subject completed 13 runs, yielding approximately 1000 four-second trials.
- Cho2017 dataset (Cho et al., 2017) contains EEG recordings from 52 subjects performing four-class motor imagery tasks (left hand, right hand, foot, tongue) using a 62-channel montage at 1,000 Hz sampling rate.
- PhysioNet-MI (Schalk et al., 2004) is a publicly available dataset on PhysioNet. It comprises EEG recordings from 109 healthy subjects performing both motor execution and motor imagery tasks involving the left and right hands.
- Shin2017A (Shin et al., 2016) dataset contains EEG recordings from 30 healthy subjects (29 right-handed, 1 left-handed; average age 28.5 ± 3.7 years). Subjects performed two-class hand motor imagery tasks (left vs. right hand) using a 30-channel EEG montage at 1000 Hz. Each participant completed three sessions with 20 trials per session (10 per class).
- SEED dataset (Duan et al., 2013) contains EEG and eye movement data of 12 subjects and EEG data of another 3 subjects. Data was collected when they were watching film clips.
- SEED-IV (Zheng et al., 2018) contains data from 15 subjects, each undergoing three sessions. During each session, 24 movie clips were used to elicit four discrete emotions: happy, sad, fear, and neutral. EEG was recorded using a 62-channel NeuroScan system at 1000 Hz, along with synchronized eye-tracking signals.
- SEED-V (Liu et al., 2021) expands the categories to five (happy, sad, fear, disgust, and neutral) and includes recordings from 20 subjects, each with three sessions and 15 clips per session.
- SEED-VII (Jiang et al., 2024a) further extends to seven categories (happy, sad, fear, disgust, neutral, anger, and surprise), employing 80 video stimuli, and was recorded with EEG and Tobii Pro Fusion eye-tracking from 20 subjects.
- FACED dataset (Chen et al., 2023) includes EEG recordings from 123 healthy participants exposed to 28 film clips designed to induce nine fine-grained emotions: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutral. EEG was collected using a 32-channel cap (10–20 system) at 250 Hz. In addition to categorical labels, dimensional ratings such as valence, arousal, familiarity, and liking were provided.
- Benchmark (Wang et al., 2016) is one of the most widely adopted SSVEP corpora, consisting of 35 subjects with 64-channel EEG recordings. Participants performed a cued spelling task involving 40 visual targets driven by joint frequency and phase modulation within the 8–15.8 Hz range. This dataset has become a de facto standard for assessing algorithmic performance in high target-count speller systems.
- BETA (Liu et al., 2020) extends the Benchmark dataset by including 70 subjects under a similar 40-target spelling paradigm with 64-channel recordings. The larger subject pool provides a solid basis for evaluating cross-subject transferability and generalization of SSVEP decoding algorithms.
- eldBETA (Liu et al., 2022a) focuses on the aging population, containing EEG data from 100 elderly participants (aged 52–81). Each subject completed a 9-target SSVEP task with 64 channels. This dataset enables the investigation of age-related changes in neural responses and the development of BCI systems tailored for elderly users.

- OpenBMI–SSVEP (Lee et al., 2019) is part of the OpenBMI dataset and provides recordings from 30 healthy adults across two sessions. The paradigm consisted of 4 visual targets presented at the screen edges, with stimulation frequencies of 5.45, 6.67, 8.57, and 12 Hz. The dataset is well-suited for studying low-frequency responses, small-class classification, and cross-session robustness.
 - The BCIC2020-3 dataset (Jeong et al., 2022), released as part of the International BCI Competition 2020, contains multi-class imagined speech EEG recordings from 15 healthy subjects. Participants were instructed to imagine speaking five short phrases while 64-channel EEG signals were recorded, providing a benchmark resource for covert speech decoding research.
 - ADHD–AliMotie (Nasrabadi et al., 2020) recruited 121 children, including 61 diagnosed with ADHD and 60 healthy controls. Participants were between 7 and 12 years old, comprising both boys and girls. EEG was recorded using 19 electrodes at a sampling rate of 128 Hz.
 - Mental Workload (Zyma et al., 2019) contains 36 subjects performing serial subtraction. EEG was recorded using 19 electrodes at a sampling rate of 500 Hz.
- Please find the summary of downstream datasets in Table 10.

Table 10: Summary of downstream EEG datasets used in this study.

Dataset	Task	#Classes	#Subjects	#Channels	Sampling
BCIC-IV-2a	MI	4	9	22	250 Hz
OpenBMI-MI	MI	2	54	64	1000 Hz
BCIC-Upperlimb	MI	3	9	22	250 Hz
SHU-MI	MI	2	62	64	250 Hz
High Gamma	MI	3	14	128	1000 Hz
Cho2017	MI	2	62	64	1000 Hz
Shin2017A	MI	2	62	64	1000 Hz
PhysioNet	MI	2	62	64	160 Hz
SEED	Emotion	4	15	62	1000 Hz
SEED-IV	Emotion	4	15	62	1000 Hz
SEED-V	Emotion	5	20	62	1000 Hz
SEED-VII	Emotion	7	20	62	1000 Hz
FACED	Emotion	9	123	32	250 Hz
OpenBMI–SSVEP	SSVEP	4	30	64	1000 Hz
BETA	SSVEP	40	70	64	1000 Hz
eldBETA	SSVEP	9	100	64	1000 Hz
Benchmark	SSVEP	40	35	64	1000 Hz
BCIC2020-3	Covert speech	5	22	23	256 Hz
ADHD–AliMotie	Healthcare	2	121	64	128 Hz
Mental Workload	Healthcare	2	36	64	500 Hz

H EVALUATION SETTINGS ON DOWNSTREAM DATASETS

We report the details about the evaluation settings for all downstream datasets in Table 11.

I IMPACT OF TEXT ENCODER ON INSTRUCT TUNING LOSS AND VALIDATION ACCURACY

We further investigate the impact of text encoders by comparing two ELASTIQ variants: one with BERT base (Devlin et al., 2019) and one with SBERT mpnet v2 (Reimers & Gurevych, 2019). Figure 6 shows that the SBERT-based model converges faster and reaches a lower validation loss than BERT, while also attaining substantially higher validation accuracy. This suggests that sentence-level semantic embeddings from SBERT provide richer guidance for EEG–language alignment than token-level BERT representations.

Table 11: Train/validation/test split strategies for downstream EEG datasets.

Dataset	Split Strategy
BCIC-IV-2a	Multi-subject: For each subject, first 75% trials train/val, last 25% test; 20% of train/val for validation.
OpenBMI-MI	Cross-subject: Subjects 1–42 for train/val, 43–54 for test; 20% of train/val for validation.
BCIC-Upperlimb	Multi-subject (Same as BCIC-IV-2a)
SHU-MI	Multi-subject (Same as BCIC-IV-2a)
HighGamma	Multi-subject (Same as BCIC-IV-2a)
Cho2017	Cross-subject: Subjects (total 49) 1–40 for train/val, 41–49 for test; 20% of train/val for validation.
Shin2017A	Multi-subject (Same as BCIC-IV-2a)
PhysioNet-MI	Cross-subject: Subjects 1–80 for train/val, 81–109 for test; 20% of train/val for validation.
SEED	Multi-subject Trial-based: For each subject (15), trials 1–9 train, 10–12 val, 13–15 test.
SEED-IV	Multi-subject Trial-based: For each subject (15), trials 1–16 train, 17–20 val, 21–24 test.
SEED-V	Multi-subject Trial-based: For each subject (16), trials 1–5 train, 6–10 val, 11–15 test.
SEED-VII	Multi-subject Trial-based: For each subject (20), trials 1–10 train, 11–15 val, 16–20 test.
FACED	Cross-subject: Subjects 1–100 train/val, 101–122 test; 20% of train/val for validation.
OpenBMI-SSVEP	Cross-subject: Subjects 1–42 train/val, 43–54 test; 20% of train/val for validation.
BETA	Cross-subject: Subjects 1–46 train, 47–50 val, 51–70 test.
eldBETA	Cross-subject: Subjects 1–75 train, 76–80 val, 81–100 test.
Benchmark	Cross-subject: Subjects 1–26 train, 27–28 val, 29–35 test.
BCIC-Speech	Multi-subject: First 300 trials train, remaining trials test (validation drawn from training set).
ADHD-AliMotie	Cross-subject: 70 subjects (35 ADHD + 35 controls) train, 10 (5+5) val, 40 (20+20) test.
Mental Workload	Cross-subject: Subjects 0–31 train/val, 32–35 test; 20% of train/val for validation.

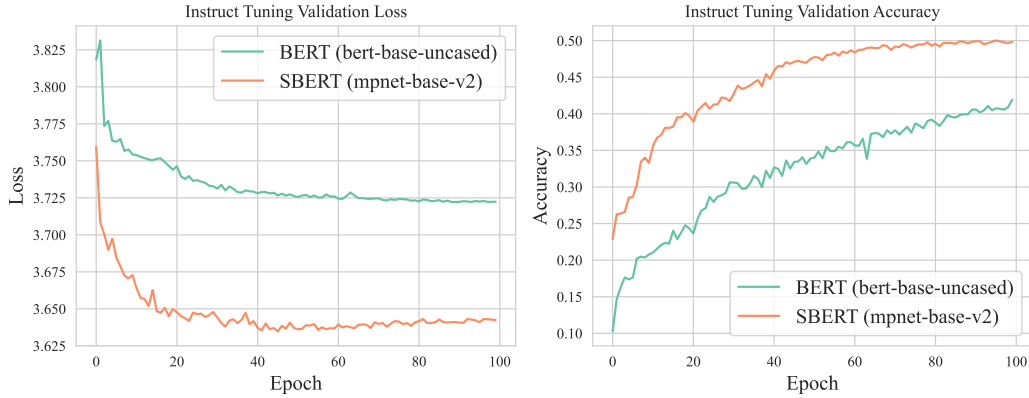


Figure 6: Instruct tuning loss and validation accuracy of ELASTIQ using different sentence embedding models: BERT (bert-base-uncased) vs. SBERT (mpnet-base-v2)

J VISUALIZATION OF PRETRAINING LOSS

Figure 7 plots the loss trajectories for the bidirectional (random masking) and causal (future masking) transformers during pretraining. Both objectives decrease monotonically, confirming effective optimization. However, the causal transformer converges more rapidly, reaching a stable minimum around epoch 25 with a lower final loss. In contrast, the bidirectional transformer converges more slowly and plateaus at a higher loss. This divergence may be explained by the fact that causal prediction imposes stronger sequential constraints that accelerate convergence, whereas bidirectional reconstruction likely requires integrating information across the entire context, making optimization more challenging.

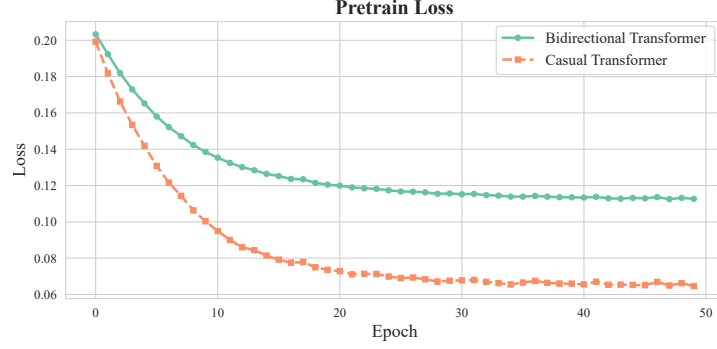


Figure 7: Pretraining loss of the bidirectional transformer with random masking and the causal transformer with next token masking

K STANDARD EEG MONTAGE USED BY ELASTIQ

In our framework, heterogeneous EEG datasets with different channel montages were spatially aligned onto a standard 10–10 electrode layout, as illustrated in Figure 8. Each electrode in the datasets was mapped to the nearest neighbor on this template, thereby preserving the spatial topology of the scalp distribution. This alignment ensures consistent channel representation across datasets and facilitates effective modeling of spatial dependencies.

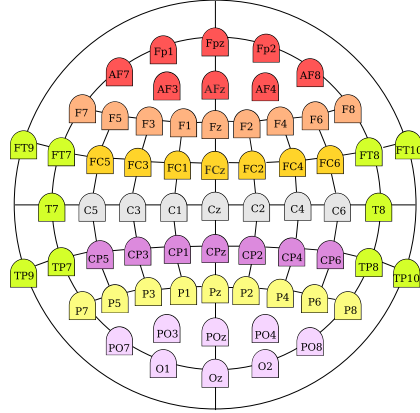


Figure 8: ELASTIQ employs the 10–10 system with 65 EEG electrodes; any input montage is interpolated to this configuration before being fed into ELASTIQ.

L TOPOGRAPHY VISUALIZATION

Figure 9 presents saliency maps derived from our model across three representative tasks, highlighting the EEG components most influential for prediction. For motor imagery (OpenBMI-MI), the model highlights contralateral motor cortex regions around C3 and C4 when distinguishing left-versus right-hand movements, consistent with established neurophysiological findings. For emotion recognition (SEED), salient activations emerge in frontal and temporal regions, reflecting neural substrates involved in affective processing. For SSVEP (Benchmark), the maps exhibit strong responses over occipital areas, in line with the visual cortex origin of steady-state responses. These results confirm that our model captures task-relevant neural patterns, thereby improving interpretability and supporting the neuroscientific plausibility of the learned representations.

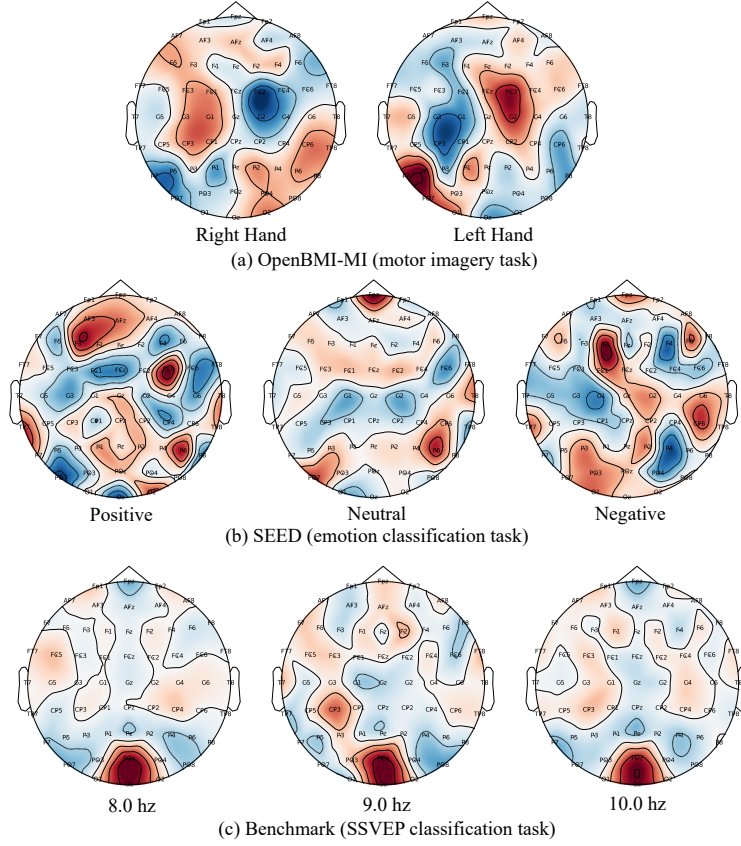


Figure 9: Topography visualization on downstream datasets

M REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. The training and evaluation pipelines are described in detail in Appendix H, including dataset preprocessing, model configurations, and evaluation protocols. All datasets used are publicly available, and the code, pretrained checkpoints, and scripts for data preprocessing and evaluation will be fully released upon publication.

N THE USE OF LARGE LANGUAGE MODELS (LLMs)

Large language models (LLMs) were exclusively used to refine the writing of this manuscript, such as improving grammar, clarity, and readability. They were not involved in generating scientific content, designing experiments, or interpreting results. All research ideas, technical contributions, and analyses presented in this paper were conceived, implemented, and validated entirely by the authors.