NEUROTTT: BRIDGING PRETRAIN–DOWNSTREAM TASK MISALIGNMENT IN EEG FOUNDATION MODELS VIA TEST-TIME TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale foundation models for EEG signals offer a promising path to generalizable brain–computer interface (BCI) applications, but they often suffer from misalignment between pretraining objectives and downstream tasks, as well as significant cross-subject distribution shifts. This paper addresses these challenges by introducing a two-stage alignment strategy that bridges the gap between generic pretraining and specific EEG decoding tasks. First, we propose NeuroTTT: a domain-specific self-supervised fine-tuning paradigm that augments the foundation model with task-relevant self-supervised objectives, aligning latent representations to important spectral, spatial, and temporal EEG features without requiring additional labeled data. Second, we incorporate test-time training (TTT) at inference, we perform (i) self-supervised test-time training on individual unlabeled test samples and (ii) prediction entropy minimization (Tent), which updates only normalization statistics to continually calibrate the model to each new input on the fly. Our approach, which, to our knowledge, is the first to unify domain-tuned self-supervision with test-time training in large-scale EEG foundation models, yields substantially improved robustness and accuracy across diverse BCI tasks (imagined speech, stress detection, motor imagery). Using CBraMod and LaBraM as backbones, our method pushes their performance to a markedly higher level. Results on three diverse tasks demonstrate that the proposed alignment strategy achieves state-of-the-art performance, outperforming conventional fine-tuning and adaptation methods. Our code is available at `https://anonymous.4open.science/r/NeuroTTT/`.

1 INTRODUCTION

Electroencephalography (EEG) is a non-invasive technique for measuring brain electrical activity and underpins a variety of brain-computer interface (BCI) applications including but not limited to imagined speech decoding (Proix et al., 2022), mental stress detection (Badr et al., 2024), emotion recognition (Li et al., 2022) and motor imagery classification (Altaheri et al., 2023). Early EEG decoding approaches heavily relied on handcrafted features and traditional machine learning (Ramoser et al., 2000; Ang et al., 2008). In recent years, deep learning models have achieved superior performance by learning directly from raw EEG signals in an end-to-end fashion (Craik et al., 2019; Al-Saegh et al., 2021). However, most of deep learning models employ supervised learning methods tailored for specific tasks or datasets, and lack generalization ability. Inspired by the success of foundation models in natural language processing (NLP) and computer vision (CV) (Devlin et al., 2019; Liu et al., 2024), researchers begin developing large-scale EEG foundation models – pretrained foundation models intended to serve as general feature extractors for diverse EEG tasks (Lai et al., 2025). Most of these works leverage massive EEG datasets in generative-based self-supervised pretraining and finetuned on specific downstream datasets (Yang et al., 2023; Duan et al., 2023b). These models have shown promising performance in capturing generic EEG representations and addressing issues like limited data and heterogeneous channel configurations.

However, fine-tuning strategies for EEG foundation models remain relatively underexplored. A fundamental yet often overlooked issue is the misalignment between pre-training objectives/data and downstream EEG task requirements: generic self-supervised objectives (e.g., masked signal recon-

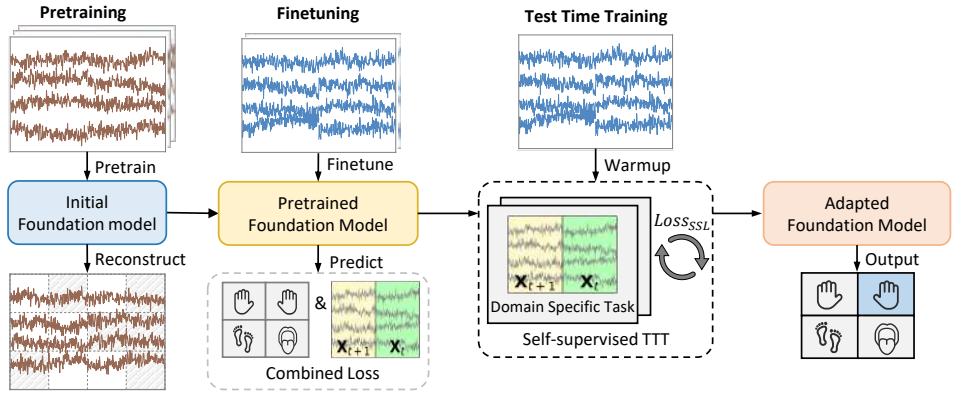

Figure 1: NeuroTTT pipeline. A pre-trained EEG foundation model (e.g., CBraMod or LaBraM) is adapted in two stages. Stage 1 (fine-tuning): the backbone feeds a supervised task head and lightweight self-supervised (SSL) heads; jointly minimize the main task loss and weighted SSL losses to align the representation to task-relevant EEG features, producing an adapted model. Stage 2 (inference): test-time adaptation is applied via one-step SSL (or BatchNorm statistics) updates on each test sample.

struction or neural code prediction (He et al., 2022; Wang et al., 2024b; Jiang et al., 2024)) may fail to emphasize the task-critical spectral, spatial, or temporal features needed for downstream domain specific applications, leaving downstream decoders under-served without targeted alignment during adapting. Compounding this, domain shift is pervasive in EEG: inter- and intra-subject variability (Saha & Baumert, 2020), session effects, and device/protocol differences routinely degrade cross-subject generalization (Melnik et al., 2017). Aligning the distribution of a target subject with that of the source subjects has long been a thorny challenge, motivating transfer learning and domain adaptation techniques for BCIs (Zhao et al., 2020; Lan et al., 2018; Duan et al., 2023a). Addressing both issues—**pretrain–downstream misalignment and distribution shift**—is therefore essential for robust EEG decoding. Simple adaptation strategies like training a shallow classifier (e.g. a linear or MLP head) on top of the pretrained foundation model often yield suboptimal results. In practice, we observe that the features from these foundation models are not sufficiently aligned to task-specific patterns, making downstream adaptation challenging.

We hypothesize that introducing domain-specific self-supervised objectives during fine-tuning can help bridging the gap between pretraining and target tasks. By augmenting the foundation model with lightweight, task-relevant self-supervised learning (SSL) tasks, we can encourage the model to learn features that are more congruent with the downstream task's discriminative information. Crucially, these SSL tasks are plug-and-play and do not require additional labeled data, we exploit unlabeled structure in EEG signals (such as frequency content or sensor topology) to shape the representation. This prepares a better feature space for final classification and yields improved performance with minimal overhead. Our SSL fine-tuning strategy is flexible: new self-supervised tasks can be designed as needed for different EEG contexts, leveraging known neuroscientific priors (e.g. known frequency bands or brain region asymmetries).

Beyond enhancing the feature space alignment, we also address model adaptation at inference time. EEG signals are notoriously variable across recording sessions and subjects(Corsi-Cabrera et al., 2007); thus even a well-finetuned model may face distribution shifts when deployed on new subjects or conditions. To tackle this, we employ Test-Time Training (TTT) techniques that adjust the model on-the-fly for each test sample or batch (Sun et al., 2019; 2020). This test-time training with self-supervision effectively calibrates the model to the incoming data, improving robustness to shifts without requiring any labeled data or extensive retraining. Importantly, the adaptation is lightweight and privacy-friendly, all computations occur on the test sample itself with no need to share or store data from the source domain. We further incorporate a complementary test-time strategy: entropy minimization of the model's predictions (exploiting the Tent method for fully test-time training) (Wang et al., 2020). By minimizing the prediction entropy on unlabeled test inputs, we encourage the model's outputs to be more confident and cluster test features toward the learned

decision boundaries. This approach updates only a minimal subset of parameters (e.g. batch normalization statistics) during inference, providing a stable yet effective adjustment. Notably, we find that entropy-based TTA can outperform full-parameter self-supervised TTT in challenging cross-subject scenarios, likely because it avoids overfitting by tuning only normalization layers while leaving the backbone intact.

In summary, we propose a novel two-stage alignment approach for EEG foundation models, our main contributions can be summarized as follows:

1. We propose NeuroTTT, a two-stage fine-tuning paradigm for EEG foundation models that first uses domain-specific self-supervised tasks to relieve distribution shift and pre-train–downstream misalignment. This approach is lightweight and general, adding no additional deployment cost while significantly aligning the representation to task-specific features.
2. We propose to apply test-time self-supervised training for brain signal models, enabling on-the-fly calibration to individual subjects and trials without any labeled data or source model modifications.
3. We explore test-time entropy minimization in EEG contexts and demonstrate its potential to improve generalization while offering a stable and privacy-preserving adaptation mechanism.
4. We evaluate the performance of NeuroTTT on 3 different downstream tasks and demonstrate state-of-the-art results on representative EEG foundation models, suggesting that our approach is a practical step toward more task-aligned, generalizable brain foundation models.

2 Related Work

Large Brain Models: The concept of foundation models in EEG is nascent, inspired by the breakthroughs of large language models and vision transformers. Early efforts like BENDR and EEG2Vec explored self-supervised pre-training on EEG (Kostas et al., 2021; Bethge et al., 2022), but were limited in scale and generality. Recent advances have produced truly large EEG models trained on extensive and diverse datasets. CBraMod (Criss-Cross Brain Model) is a transformer-based foundation model that separately models spatial and temporal dependencies of EEG through a criss-cross attention mechanism (Wang et al., 2024b). Pre-trained via masked EEG patch reconstruction on a massive corpus, CBraMod achieved state-of-the-art results across numerous BCI tasks, demonstrating the potential of a single model to generalize widely. LaBraM (Large Brain Model) introduced a unified EEG representation learned from about 2,500 hours of data from public datasets (Jiang et al., 2024). It uses a vector-quantized spectral tokenizer and masked token prediction to learn a compact but rich encoding of EEG signals. LaBraM likewise showed superior performance on varied tasks (abnormal EEG detection, event classification, emotion recognition, etc.), highlighting the benefits of cross-dataset pre-training. BIOT (Biosignal Transformer) further extended foundation modeling to multimodal biosignals (Yang et al., 2023), enabling cross-dataset learning by tokenizing each channel into a "sentence" of patch tokens; it demonstrated the feasibility of training one model on mixed EEG, ECG, and motion data. Despite these successes, a limitation of current EEG foundation models is their reliance on generic self-supervised objectives (e.g. masked autoencoding, sequence prediction) that are agnostic to any specific task. While these objectives ensure broad coverage of EEG variability, they might underemphasize fine-grained features crucial for particular tasks (for instance, the difference between EEG frequency bands is critical for motor imagery but a vanilla autoencoder may not prioritize that).

3 Method

NeuroTTT comprises three interconnected components that together ensure a better alignment of the foundation model with each target EEG task: Stage I: Domain-Specific Self-Supervised fine-tuning, Stage II: Self-Supervised Test-Time Training, or Test-Time Entropy Minimization at inference. An overview of the proposed alignment strategy is shown in Figure 2.

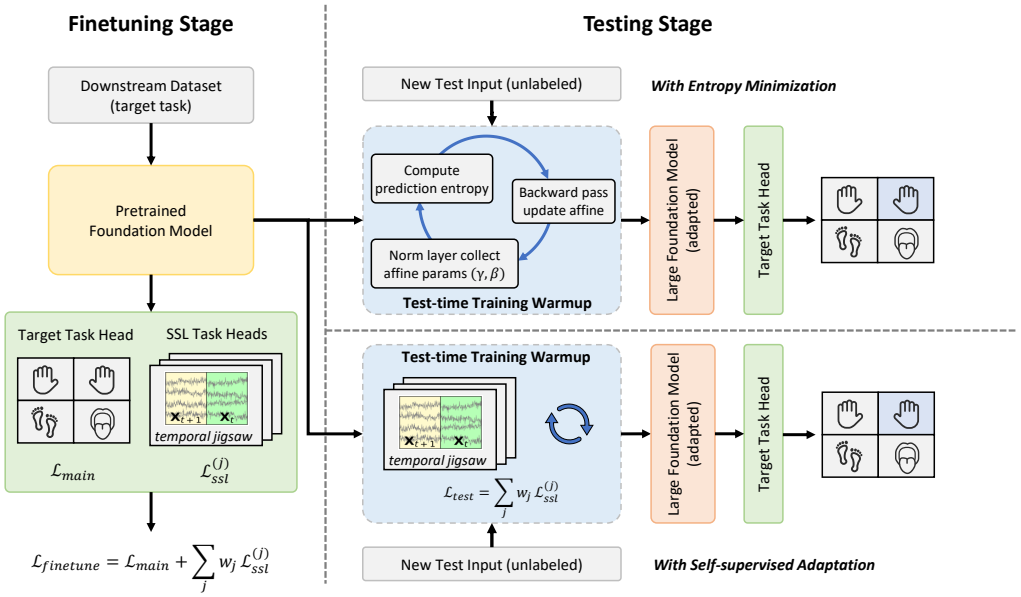

Figure 2: Overview of our NeuroTTT structure, illustrated on the Motor Imagery task. During fine-tuning, we augment the supervised head with two domain-specific self-supervised objectives: a Stopped-Band Prediction task plus a Temporal Jigsaw task. At inference, we perform self-supervised test-time training on individual samples or Tent-based entropy minimization to continually calibrate the model to each new input.

3.1 Stage I — Domain-Specific Self-Supervised Fine-tuning

In the preliminary stage, we adapt the pre-trained EEG foundation model to a specific downstream task by self-supervised fine-tuning. The pre-trained model (e.g. CBraMod or LaBraM) used as a shared feature extractor, feeding into different downstream branches: (a) the main task branch, which is a classifier for the labeled downstream task, and (b) two self-supervised task branches, each consisting of a small task-specific head dedicated to a particular unlabeled pretext task. All branches share the same foundation model backbone, but have separated objective functions. During fine-tuning, we jointly optimize the supervised loss on the main task and the losses of the self-supervised tasks. By doing so, we explicitly encourage the backbone's representations to capture not only generic EEG features (from pre-training) but also task-relevant features that the self-supervised objectives focus on. This effectively aligns the feature space with the downstream task, preparing a better representation for the main-task.

Formally, we denote a pre-trained EEG foundation backbone f_θ with task head g_ϕ. We jointly optimize a supervised loss and lightweight, EEG-aware SSL heads that share the backbone:

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{main}}\big(g_\phi(f_\theta(x)), y\big) + \sum_j w_j \, \mathcal{L}_{\text{ssl}}^{(j)}\Big(h^{(j)}\big(f_\theta(\tilde{x}^{(j)})\big)\Big). \tag{1}$$

where w_j are weighting coefficients that balance the importance of each SSL task. $\tilde{x}^{(j)}$ is a self-supervised view that encodes EEG priors (e.g., **Stopped-Band Prediction**, **A–P Flip**, **Temporal Jigsaw**); heads $h^{(j)}$ are tiny classifiers. In practice, we choose w_j to ensure the self-supervised losses are on a comparable scale to the main loss; the additional tasks are intended to guide the representation without overwhelming the primary task objective.

We design domain-specific SSL tasks for each downstream application, informed by neuroscientific knowledge of task specific EEG signatures. One common self-supervision task across the three domains, is the Stopped Band Prediction task, with the frequency band divisions vary by downstream tasks. The second task is also specifically tailored to each domain. With an Amplitude Scaling Pre-

diction for Imagined Speech Classification, an Anterior-Posterior Flip Detection for Mental Stress Detection and a Temporal Jigsaw Task for Motor Imagery.

Mental Stress Detection – Anterior-Posterior Flip Detection: Mental stress is known to induce characteristic EEG changes, often with strong asymmetry between the frontal and parietal regions of the scalp. Under stress or cognitive load, frontal lobe activity tends to increase relative to rear regions, engaging the frontoparietal network. To leverage this domain knowledge, we design an Anterior-Posterior (A–P) Flip self-supervised task. We define pairs of EEG channels that are roughly symmetric across the anterior-posterior axis of the head. We then randomly choose whether to swap these pairs (i.e. exchange frontal channel signals with their posterior counterparts) in a given EEG sample. The model's SSL head must predict whether the input has been "flipped" or is in its original orientation (a binary classification). This pretext forces the model to learn the normal topographic distribution of activity for the task – specifically, to recognize the difference between a normal vs. front-back inverted EEG. By doing so, the encoder learns to encode the asymmetric topographic signature of stress-related brain activity and the coherence of the frontoparietal network without any explicit labels. This should yield a more transferable "stress representation" that is sensitive to spatial patterns known to correlate with mental workload or stress.

Details of the other SSL tasks can be found in Appendix B.1. After adding these tasks, the fine-tuning process optimizes the combined loss $\mathcal{L}_{finetune}$ over the training set of the target task. The outcome is a model that still leverages the general knowledge from pre-training but is specialized with domain-specific feature awareness. Because the SSL tasks are designed to be lightweight (predicting simple transformations), they do not require many extra parameters (each SSL head is a small classification layer) and do not need ground-truth labels from the dataset. They can be seen as a form of feature guiding for the model during fine-tuning. Once training is complete, we can either keep these SSL heads for potential use at test time (as we describe next) or discard them if only the main task prediction is needed during inference.

3.2 Stage II — Test-Time Training with Self-Supervision

While the preliminary fine-tuning aligns the model to the general characteristics of the task domain, there can still be residual distribution shifts at test time. For example, differences in EEG noise level, electrode impedance, or individual brain anatomy can cause a drop in performance. To mitigate this, we incorporate Self-Supervised Test-Time Training (TTT) as an online adaptation mechanism (Xiao & Snoek, 2024).

The key idea of TTT is to leverage a domain-specific self-supervised objective on each unlabeled test sample to briefly adapt the model before making the final prediction. Given a new test input $\mathbf{x}_{test}$ with no ground-truth label, our TTT procedure is:

- **Maintain Self-Supervised Head:** We retain the self-supervised tasks heads (learned during fine-tuning) for use during inference. This head produces a prediction for the self-supervised pretext task (e.g. predicting a rejected band or a Anterior-Posterior flipped indicator) on the test sample.
- **TTT Warm-Up (Adaptation Step):** When a new unlabeled test sample (or a small batch of samples) is received, we first perform a forward pass on $\mathbf{x}_{test}$ to compute the self-supervised loss $L_{SSL}(\mathbf{x}_{test}; \theta)$, since the main task loss cannot be computed without a label. For example, L_{SSL} could be a cross-entropy loss for correctly predicting an augmented transformation of $\mathbf{x}_{test}$. We then update the model parameters by taking a single gradient descent step to minimize this loss, yielding adapted parameters θ':

$$\theta' = \theta - \alpha \nabla_\theta L_{SSL}(\mathbf{x}_{test}; \theta). \tag{2}$$

  where α is a small learning rate. This adaptation step personalizes the model to the test sample by tuning it to current sample's idiosyncrasies (such as its noise characteristics or signal scale).
- **Inference after Adaptation:** After this brief adaptation on the unlabeled sample, we perform a forward pass of the adapted model on the same input (this time through the main task head) to get the final prediction (e.g. predicted class of imagined speech, stress level, or motor imagery label). We then reset the model to its original state before the next sample, or in an online scenario, we carry the updated state forward.

This procedure effectively personalizes the model to each test input on the fly. Intuitively, by solving a self-supervised puzzle on the test EEG, the model "tunes" itself to that EEG's idiosyncrasies (noise distribution, amplitude scaling, etc.) in a way that also aligns with the main task features. Our approach builds on the method proposed by Sun et al. (2020) for image classification, but here applied to time-series brain data. We emphasize that no human labels are used in this adaptation, the model is optimizing a purely self-generated objective.

3.3 Stage II — Test-Time Entropy Minimization

Alongside self-supervised TTT, we incorporate Test-Time Entropy Minimization, often referred to as Tent adaptation. This is a complementary approach that does not rely on a separate SSL head, but instead uses the model's main output entropy as a guidance for adaptation. The principle, introduced by Wang et al. (2020), is that for an unlabeled test example, a confident prediction (low entropy of the output softmax) is generally more likely to be correct, assuming the model's learned decision boundaries are relatively well-aligned. Therefore, one can adapt the model on test data by directly minimizing the entropy of its predictions.

Formally, for a given test sample $\mathbf{x}test$, we define the entropy of the model's predictive distribution $p\theta(y \mid \mathbf{x}_{test})$ as:

$$L_{ent}(\theta; \mathbf{x}_{test}) = -\sum_{c=1}^{C} p_\theta(y = c \mid \mathbf{x}_{test}) \log p_\theta(y = c \mid \mathbf{x}_{test}), \tag{3}$$

where C is the number of output classes. In our implementation, we apply Tent as follows: when a batch of test data is observed, we feed it through the model's main task head to obtain predictions (probability distribution over classes). We compute the entropy of these predictions and take a gradient step to reduce L_{ent} for adjusting the model parameters to minimize this entropy. Crucially, to preserve stability, we follow the established practice of updating only the Batch Normalization (BN) parameters during this process, while keeping the other weights fixed:

$$\theta_{BN} \leftarrow \theta_{BN} - \alpha \nabla_{\theta_{BN}} L_{ent}(\theta; \mathbf{x}_{test}). \tag{4}$$

Here θ_{BN} denotes the collection of affine scale and shift parameters (and, optionally, the running mean and variance) of all BatchNorm layers. By adjusting only these normalization parameters, the model recalibrates its internal feature statistics to better match the current test batch, thereby lowering its prediction uncertainty.

The Tent adaptation is extremely lightweight – it adds negligible computational cost (just one forward-backward pass per batch) and only modifies a tiny fraction of the model's parameters. This makes it well-suited for real-time adaptation scenarios and ensures that the risk of overfitting is low (since BN layers have limited capacity compared to the entire network). We found that using entropy minimization at test time yields even better results than self-supervised TTT in some cases, especially for cross-subject transfer. This somewhat counter-intuitive outcome is likely because the full-model TTT (as above) adapts many parameters on a single sample, which, if the sample is noisy or unrepresentative, could momentarily destabilize the model. In contrast, Tent's BN-only, entropy-driven update provides a gentler nudging of the model towards the target domain's feature distribution, maintaining overall stability of the learned representations. Additionally, Tent naturally works in an online continual adaptation setting (updating BN for each batch) without needing to reset the model state between samples.

4 Experiments

We conduct experiments to evaluate the effectiveness of our proposed alignment strategies on multiple EEG datasets and foundation models. We first describe the experimental setup, including the pre-trained models used, the downstream tasks and datasets, and implementation details such as preprocessing. We then present quantitative results comparing various fine-tuning and adaptation

methods, followed by ablation studies that dissect the contribution of each component of our approach.

4.1 Experimental Setup

Pre-trained Foundation Models: We evaluate our methods using two representative state-of-the-art large-scale EEG foundation models, each with distinct architectures and pre-training strategies, which provides a robust test of our alignment methods. Detailed descriptions of both models are provided in Appendix C.3. For all experiments, we use these foundation models as feature extractors and apply various adaptation strategies, as outlined in the subsequent sections. Both models internally downsample input signals to 200 Hz, thereby capturing frequency components up to 100 Hz in accordance with the Nyquist theorem.

Downstream BCI Tasks and Datasets: We evaluate on three diverse EEG decoding tasks to demonstrate generality: Imagined Speech Classification (Jeong et al., 2022), Mental Stress Detection (Goldberger et al., 2000; Zyma et al., 2019) and Motor Imagery Classification (Brunner et al., 2008). All the downstream EEG tasks with the corresponding datasets and preprocessing are presented in Appendix C.1.

Performance Metrics: For the binary Mental Stress Detection task, we use Balanced Accuracy, Area Under the Precision-Recall Curve (AUC-PR), and Area Under the Receiver Operating Characteristic Curve (AUROC) as evaluation metrics, with AUROC designated as the monitoring score. For the multi-class tasks—Imagined Speech Classification and Motor Imagery Classification—we evaluate performance using Balanced Accuracy, Cohen's Kappa, and Weighted F1 Score, with Cohen's Kappa selected as the monitoring metric.

Finetuning and Adaptation Strategies: In all of our finetuning and adaptation settings, the foundation model remains fully trainable. This design choice is crucial for better aligning the pre-trained representations with the specific demands of each downstream EEG task. By allowing gradient updates throughout the model, we enable deeper integration of task-specific information, improving both accuracy and generalization. We compare the following strategies on each task:

- **Full Supervised Finetuning:** Train the entire foundation model and a small task head (1–3 linear layers) end-to-end on labeled data. This is the de facto standard approach.
- **Lora Finetuning:** A parameter-efficient fine-tuning method that only introduces low-rank updates to a pre-trained model's weights (Hu et al., 2022), popular in LLMs finetuning.
- **SHOT (Source Free Domain Adaptation):** A source-free domain adaptation setting for downstream tasks (Liang et al., 2020). In essence, SHOT uses the source-trained classifier's "hypothesis" and adjusts the feature extractor to better fit target data without source data.
- **NeuroTTT:** Our full method. We fine-tune the model on the training set using our multi-task loss (with the appropriate domain-specific SSL tasks for each dataset) as described in the Method 3 section. At inference, we apply test-time training with two variants:

  (a) TTT with SSL, which performs a single gradient step on the SSL task for each test sample; (b) TTT with Tent, which applies entropy minimization to test batches by updating batch normalization statistics.

All models are optimized using the Adam optimizer (Kingma & Ba, 2014). For full-parameter finetuning—including any LoRA layers or SSL heads—we use a learning rate of 1e-4. We train each model until validation performance converges. For Mental Stress Detection and Motor Imagery Classification, convergence typically occurs within 20 epochs. In contrast, Imagined Speech Classification requires longer training due to slower convergence under domain-specific SSL supervision; we train for 50 epochs under supervised-only settings and extend to 150 epochs when incorporating domain-specific SSL. More hyperparameter setting can be found in Appendix C.4.

4.2 Results

Table 1 summarizes the results for Imagined Speech Classification, tested on BCI Competition 2020-III dataset. Table 2 reports performance in Mental Stress Detection task, tested on MentalArithmetic Stress dataset. Table 3 shows the cross-subject results in Motor Imagery tasks, tested on BCI Competition IV-2a dataset.

Table 1: Results on BCIC2020-III (5-class).

Aligning Strategies	CBraMod (Wang et al., 2024b)			LaBraM-Base (Jiang et al., 2024)		
	Balanced-Accuracy	Cohen's Kappa	Weighted F1	Balanced-Accuracy	Cohen's Kappa	Weighted F1
Linear	0.3976 ± 0.0113	0.2470 ± 0.0141	0.3968 ± 0.0120	0.2529 ± 0.0260	0.0661 ± 0.0325	0.2290 ± 0.0565
Shallow MLP	0.4733 ± 0.0098	0.3417 ± 0.0122	0.4735 ± 0.0098	0.2493 ± 0.0315	0.0517 ± 0.0394	0.2304 ± 0.0442
Lora	0.3531 ± 0.0158	0.1913 ± 0.0198	0.3525 ± 0.0159	0.2484 ± 0.0372	0.0606 ± 0.0215	0.2374 ± 0.0410
SHOT	0.5008 ± 0.0010	0.3760 ± 0.0013	0.5004 ± 0.0011	0.2493 ± 0.0283	0.0717 ± 0.0229	0.2433 ± 0.0274
TTT with SSL	0.5887 ± 0.0170	0.4858 ± 0.0212	0.5886 ± 0.0164	0.2720 ± 0.0083	0.0900 ± 0.0104	0.2649 ± 0.0137
TTT with Tent	**0.5898 ± 0.0148**	**0.4872 ± 0.0185**	**0.5895 ± 0.0142**	**0.2742 ± 0.0047**	**0.0928 ± 0.0159**	**0.2713 ± 0.0126**

Table 2: Results on MentalArithmetic.

Aligning Strategies	CBraMod (Wang et al., 2024b)			LaBraM-Base (Jiang et al., 2024)		
	Balanced-Accuracy	AUC-PR	AUROC	Accuracy	AUC-PR	AUROC
Linear	0.6389 ± 0.0092	0.4880 ± 0.0262	0.7788 ± 0.0408	0.6123 ± 0.0254	0.4826 ± 0.1124	0.6865 ± 0.0523
Shallow MLP	0.6486 ± 0.0310	0.4459 ± 0.0527	0.7301 ± 0.0379	0.6125 ± 0.0249	0.2537 ± 0.0575	0.7329 ± 0.0191
Lora	0.6001 ± 0.0144	0.5080 ± 0.0297	0.7327 ± 0.0239	0.5944 ± 0.0136	0.2475 ± 0.0245	0.7362 ± 0.0085
SHOT	0.6308 ± 0.0297	0.6063 ± 0.0613	0.7696 ± 0.0067	0.5827 ± 0.0274	0.4783 ± 0.0343	0.5826 ± 0.0267
TTT with SSL	0.7196 ± 0.0553	0.5720 ± 0.0798	0.8081 ± 0.0652	0.7161 ± 0.0231	0.6494 ± 0.0337	0.8180 ± 0.0081
TTT with Tent	**0.7280 ± 0.0509**	**0.5925 ± 0.0615**	**0.8256 ± 0.0409**	**0.7269 ± 0.0106**	**0.6401 ± 0.0402**	**0.7916 ± 0.0475**

As shown in the tables, our method pushes Large Brain Models to a markedly higher performance level. Notably, LoRA shows the worse performance among all methods. This partial adaptation can struggle to align pre-trained EEG features to a new task because the frozen backbone's representations remain largely unchanged. If the pre-trained features are not already well-suited for the target EEG task, a small low-rank tweak may be insufficient to bridge the gap. In EEG applications, tasks and datasets often differ substantially (e.g. different cognitive tasks, pathologies, or recording setups), so the foundation model's features may not cleanly separate the new classes without significant reorientation. Since LoRA can only adjust the model via a limited subspace of parameters, it may fail to reshape the feature space enough for the downstream classification, leaving misalignments between what the backbone encodes and what the new task requires. Recent work on EEG foundation models has indeed observed that parameter-efficient "additive" tuning methods (like adapters or LoRA) yield inconsistent performance across tasks and pre-trained models. These methods often heavily depend on how closely the pre-training data matches the downstream domain. In other words, if the EEG foundation model wasn't pre-trained on data very similar to the target task, a LoRA-based adaptation might not adequately realign the representation to the new task's patterns. This is one reason LoRA underperforms: it cannot fully compensate when the pretrained feature space and the downstream data distribution are misaligned. Addressing this misalignment is one of the central problems our method tackles.

Unlike the large pretrain–downstream gap, the distribution shift between subjects is typically less severe. In an already well-learned feature space, applying full-parameter SSL fine-tuning can actually mislead the representation, especially when the incoming EEG is noisy or unrepresentative. By contrast, Tent's BN-only, entropy-driven updates provide a gentler, more controlled push toward the target domain's distribution, preserving the overall stability of the learned features. This explains why Tent often outperforms test-time SSL in our evaluations.

Table 3: Results on BCIC-IV-2a.

Aligning Strategies	CBraMod (Wang et al., 2024b)			LaBraM-Base (Jiang et al., 2024)		
	Balanced-Accuracy	Cohen's Kappa	Weighted F1	Balanced-Accuracy	Cohen's Kappa	Weighted F1
Linear	0.5028 ± 0.0230	0.3370 ± 0.0307	0.4809 ± 0.0293	0.41111 ± 0.00285	0.21482 ± 0.00382	0.34972 ± 0.03412
Shallow MLP	0.4578 ± 0.0317	0.2771 ± 0.0423	0.4535 ± 0.0315	0.34479 ± 0.02853	0.12639 ± 0.03721	0.22937 ± 0.01984
Lora	0.4164 ± 0.0090	0.2218 ± 0.0122	0.3952 ± 0.0227	0.34375 ± 0.03313	0.12500 ± 0.04472	0.27821 ± 0.04114
SHOT	0.5354 ± 0.0267	0.3766 ± 0.0365	0.5228 ± 0.0346	0.36310 ± 0.02880	0.36310 ± 0.02880	0.15080 ± 0.03840
TTT with SSL	0.5435 ± 0.0112	0.3914 ± 0.0149	0.5249 ± 0.0161	0.43835 ± 0.01856	0.25120 ± 0.01146	0.37380 ± 0.05552
TTT with Tent	**0.5674 ± 0.0062**	**0.4232 ± 0.0082**	**0.5537 ± 0.0068**	**0.45110 ± 0.01950**	**0.26819 ± 0.01257**	**0.41559 ± 0.05358**

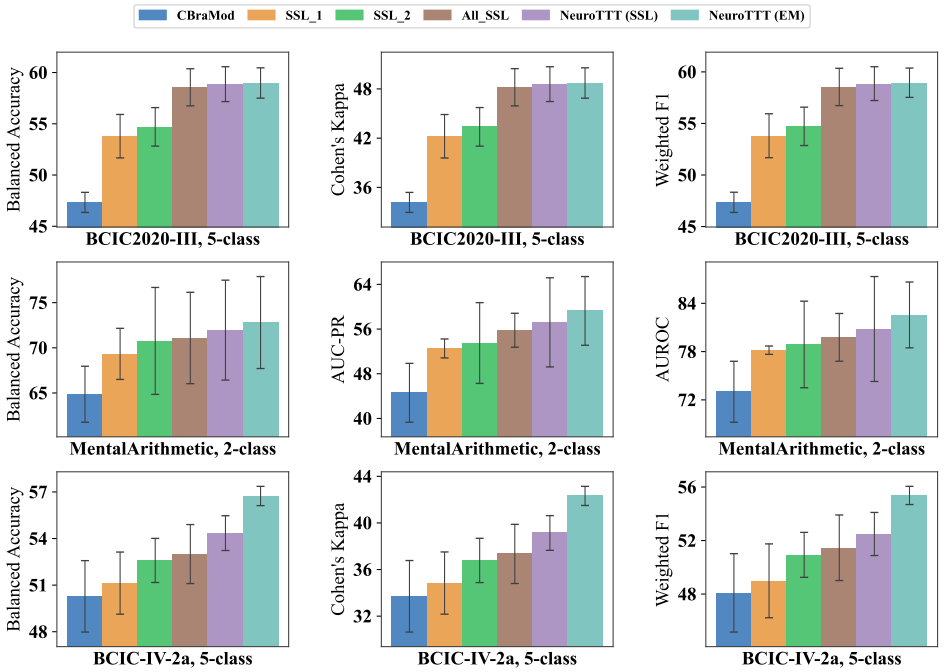

Figure 3: Ablation Study of NeuroTTT on CBraMod.

4.3 Ablation Study:

We perform ablation studies to validate the contribution of each component of our approach: each domain-specific self-supervised finetuning tasks and the test-time training mechanisms, as shown in Figure 3. The ablation experiment is designed as follows:

Isolated SSL Task Fine-tuning. To assess the contribution of each self-supervised objective, we fine-tuned the model with each SSL task in isolation. The results show that every domain-specific SSL task injects distinct, informative signals into the representation; combining them produces additive, cumulative gains.

Effect of Test-Time Training (TTT). We compared models with vs. without TTT to evaluate its effectiveness. Notably, TTT delivers modest improvements on the BCIC dataset (within-subject, non–cross-subject), but yields substantial benefits on the two cross-subject tasks—exactly as expected given the stronger distribution shifts across subjects.

5 Conclusion

We proposed NeuroTTT: a comprehensive approach to aligning large EEG foundation models with domain-specific downstream tasks through a combination of targeted self-supervised finetuning and test-time training techniques. By injecting task-relevant inductive biases (frequency bands, amplitude scaling, topographic orientation, temporal order), we substantially improved the suitability of pre-trained representations for specific BCI tasks. Further, by leveraging test-time training and entropy minimization, our models can seamlessly adapt to new inputs on the fly, addressing one of the biggest challenges in EEG applications: distribution shift.

Our results on three diverse tasks and two state-of-the-art foundation models demonstrate that the proposed alignment strategy yields state-of-the-art performance, outperforming conventional finetuning and adaptation methods. The success of our approach suggests that bridging the gap between foundation model pre-training and target task requirements is both necessary and highly effective in the EEG domain. Rather than treating a foundation model as a monolithic feature extractor, we show that a small amount of task-specific "attention" in the form of self-supervision can unlock significant gains. Additionally, this work is a step toward more generalizable and personalized EEG AI systems: models that come with broad knowledge but can specialize themselves to each user or context safely and efficiently.

REFERENCES

Ali Al-Saegh, Shefa A Dawwd, and Jassim M Abdul-Jabbar. Deep learning for motor imagery eeg-based classification: A review. *Biomedical Signal Processing and Control*, 63:102172, 2021.

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 2390–2397. IEEE, 2008.

Yara Badr, Usman Tariq, Fares Al-Shargie, Fabio Babiloni, Fadwa Al Mughairbi, and Hasan Al-Nashash. A review on evaluating mental stress by deep learning using eeg signals. *Neural Computing and Applications*, 36(21):12629–12654, 2024.

David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Lewis L Chuang, Ozan Özdenizci, and Albrecht Schmidt. Eeg2vec: Learning affective eeg representations via variational autoencoders. In *2022 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 3150–3157. IEEE, 2022.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):34, 2008.

James F Cavanagh and Michael J Frank. Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, 18(8):414–421, 2014.

Douglas Cheyne, Sonya Bells, Paul Ferrari, William Gaetz, and Andreea C Bostan. Self-paced movements induce high-frequency gamma oscillations in primary motor cortex. *Neuroimage*, 42 (1):332–342, 2008.

María Corsi-Cabrera, L Galindo-Vilchis, Yolanda Del-Río-Portilla, Consuelo Arce, and J Ramos-Loyo. Within-subject reliability and inter-session stability of eeg power and coherent activity in women evaluated monthly over nine months. *Clinical Neurophysiology*, 118(1):9–21, 2007.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

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.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. Domain-specific denoising diffusion probabilistic models for brain dynamics. *arXiv preprint arXiv:2305.04200*, 2023a.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023b.

Andreas K Engel and Pascal Fries. Beta-band oscillations—signalling the status quo? *Current opinion in neurobiology*, 20(2):156–165, 2010.

Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517, 2012.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. *arXiv preprint arXiv:2406.04328*, 2024.

Ole Jensen and Ali Mazaheri. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in human neuroscience*, 4:186, 2010.

Ole Jensen and Claudia D Tesche. Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience*, 15(8):1395–1399, 2002.

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, 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*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Junhong Lai, Jiyu Wei, Lin Yao, and Yueming Wang. A simple review of eeg foundation models: Datasets, advancements and future perspectives. *arXiv preprint arXiv:2504.20069*, 2025.

Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R Müller-Putz. Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):85–94, 2018.

Shanglin Li, Motoaki Kawanabe, and Reinmar J Kobler. Spdim: Source-free unsupervised conditional and label shift adaptation in eeg. *arXiv preprint arXiv:2411.07249*, 2024.

Zhunan Li, Enwei Zhu, Ming Jin, Cunhang Fan, Huiguang He, Ting Cai, and Jinpeng Li. Dynamic domain adaptation for class-aware cross-subject and cross-session eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(12):5964–5973, 2022.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Andrew Melnik, Petr Legkov, Krzysztof Izdebski, Silke M Kärcher, W David Hairston, Daniel P Ferris, and Peter König. Systems, subjects, sessions: to what extent do these factors influence eeg data? *Frontiers in human neuroscience*, 11:150, 2017.

Barry S Oken, Martin C Salinsky, and Siegward-M Elsas. Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical neurophysiology*, 117(9):1885–1901, 2006.

Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251, 2012.

Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

Timothée Proix, Jaime Delgado Saa, Andy Christen, Stephanie Martin, Brian N Pasley, Robert T Knight, Xing Tian, David Poeppel, Werner K Doyle, Orrin Devinsky, et al. Imagined speech can be decoded from low-and cross-frequency intracranial eeg features. *Nature communications*, 13 (1):48, 2022.

Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training for out-of-distribution generalization. *arXiv preprint arXiv:2406.04328*, 2019.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Huiyang Wang, Hongfang Han, John Q Gan, and Haixian Wang. Lightweight source-free domain adaptation based on adaptive euclidean alignment for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 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.

Zehao Xiao and Cees GM Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Cross-data biosignal learning in the wild. *arXiv preprint arXiv:2305.10351*, 2023.

He Zhao, Qingqing Zheng, Kai Ma, Huiqi Li, and Yefeng Zheng. Deep representation-based domain adaptation for nonstationary eeg classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):535–545, 2020.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

Large Language Models (LLMs) were used in this work solely for language polishing and clarity improvement. The authors drafted all scientific content, including research ideas, methodology, experiments, and results. LLM assistance was limited to refining grammar, improving readability, and rephrasing sentences for clarity. No content was generated, altered, or fabricated regarding the scientific contributions. The authors take full responsibility for the final manuscript.

B DOMAIN-SPECIFIC SELF-SUPERVISED FINE-TUNING

B.1 DOMAIN-SPECIFIC SELF-SUPERVISED TASKS

In addition to Anterior-Posterior Flip Detection, the other self-supervised tasks are as follows:

Stopped Band Prediction: The band-prediction task is motivated by the fact that EEG signals are composed of multiple frequency bands (delta, theta, alpha, beta, gamma, etc.), which often correlate with cognitive states. Jayalath et al. (2024) employ band rejection during MEG pretraining and demonstrate that band prediction enables the model to learn frequency-aware representations more effectively. To leverage this, we apply a band-stop filter to the input EEG in a random frequency band (removing or attenuating one band's power) and train the model to identify which band was removed. This encourages the model to learn representations that are sensitive to frequency-specific information, effectively tuning it to the spectral features crucial for specific downstream tasks. The band-prediction task is framed as a classification problem, where each class corresponds to the index of a removed frequency band. During fine-tuning, we randomly reject one band and compute a cross-entropy loss between the band-prediction head's output and the true index of the dropped band. This task is lightweight and does not require any additional data beyond the unlabeled EEG sample itself.

Table 4: Functional frequency bands in Imagined Speech Classification.

Band	**Hz**	**Association**
Delta–Theta (δ/θ)	0.5–8	Tracks the speech envelope and syllable-rate rhythms; aligns parsing parsing at syllabic/prosodic timescales (Giraud & Poeppel, 2012).
Alpha–Beta (α/β)	8–30	α: inhibitory gating of task-irrelevant regions; β: top-down prediction and maintenance of internal models (Jensen & Mazaheri, 2010).
Low Gamma (γ_{low})	30–70	Local cortical computations supporting feature integration during speech analysis (Giraud & Poeppel, 2012).
High Gamma (γ_{high})	70–100	Auditory-cortex high-γ encodes acoustic–phonetic detail predictive of intelligibility (Pasley et al., 2012).

Table 5: Functional frequency bands in Mental Stress Detection.

Band	**Hz**	**Association**
Frontal-midline Theta (θ)	4–8	Increases with working-memory load and cognitive-control demands (Jensen & Tesche, 2002).
Alpha (α)	8–12	Decreases with task engagement/alertness relative to relaxed rest (Oken et al., 2006).
Low Beta (β_{low})	13–20	Indexes maintenance of the current sensorimotor/cognitive and top-down control (Engel & Fries, 2010).
High Beta (β_{high})	20–30	Strengthens with heightened vigilance/executive control; supports long-range top-down communication (Engel & Fries, 2010).

Stopped Band Prediction task is a shared self-supervision task across the three downstream tasks, with the frequency band divisions vary by tasks. The task-specific band divisions are summarized in Tables 4, 5, 6. Both foundation models used in our study (CBraMod and LaBraM) were pre-trained at a sampling rate of 200 Hz. According to the Nyquist theorem, this limits the representable frequency range to approximately 100 Hz. Although higher frequency bands may be relevant for certain tasks like imagined speech decoding, our design is constrained by this 100 Hz upper limit.

Table 6: Functional frequency bands in Motor Imagery.

Band	Hz	Association
Theta (θ)	3–7	Cognitive control/monitoring engaged during MI preparation (frontal-midline θ) (Cavanagh & Frank, 2014).
Mu (μ, α)	8–13	Classic MI biomarker: contralateral μ-ERD over sensorimotor cortex; post-event μ-ERS (Pfurtscheller & Da Silva, 1999).
Beta (β)	13–30	β-ERD during MI/preparation; post-movement β rebound (ERS) reflects network reset (Pfurtscheller & Da Silva, 1999).
Low Gamma (γ_{low})	30–45	Brief motor-cortex γ bursts reflecting fast local processing; EEG analyses often cap at $\sim$45 Hz for SNR (Cheyne et al., 2008).

Imagined Speech Classification – Amplitude Scaling Prediction: In imagined speech decoding (using the BCI Competition 2020-III dataset (Jeong et al., 2022)), an important feature is the amplitude modulation of EEG signals, as certain cognitive or speech imagery events can manifest as subtle amplitude changes in specific frequency ranges. We adopt an amplitude scaling pretext task, inspired by prior work on scaling transformations for speech detection Jayalath et al. (2024). We artificially scale the raw EEG signal by a random factor α (drawn from a discrete set of scaling factors, e.g. 16 possible values from -2x to 2x) and task the model with predicting the scaling factor. This is implemented as a regression or classification problem (we treat it as classification over discrete scale indices). By learning to recognize how a scaled EEG differs from a normal one, the model becomes sensitive to the amplitude dynamics of the signal, which can improve its ability to detect the presence or absence of imagined speech patterns. The amplitude of EEG signal is stretched or telescoped through 16 scale factors. where the scale-augmented signal can be expressed as $x_i^{st} = \alpha * x_i$. The objective is to predict the scaling factor α.

Motor Imagery – Temporal Jigsaw Task: Motor imagery (using the BCI Competition IV-2a dataset (Brunner et al., 2008)) entails a subject imagining limb movements, which produces EEG patterns with temporal dynamics (like event-related desynchronization in specific frequency bands following a cue). To ensure the model captures temporal order and dependencies, we introduce a temporal jigsaw task. We segment each EEG trial into a small number of consecutive time segments (for example, split into two or three chunks of equal length). We then randomly shuffle the order of these segments along the time axis and feed the jumbled sequence into the model. The model's task is to predict the correct temporal order of the segments (essentially solving a jigsaw puzzle in time). In the simplest case of two segments, this reduces to a binary classification: whether the input is in correct chronological order or reversed. For more segments, we can have the model output a permutation or rank for each segment's position. This task compels the model to learn temporal correlations and the progression of neural patterns over time – e.g. distinguishing the initial cue response period from the later motor imagery period. Mastering this pretext should enhance the model's understanding of EEG time structure, benefiting tasks like motor imagery classification where timing (such as the onset of motor-related rhythms) is crucial.

C More Details for Experimental settings

C.1 Downstream BCI Tasks and Datasets

Downstream BCI Tasks and Datasets: We evaluate on three diverse EEG decoding tasks:

Imagined Speech Classification – We use the BCI Competition 2020 Phase III (BCIC2020-3) dataset (Jeong et al., 2022), which involves subjects imagining speaking specific phrases. The task is to classify which phrase is being imagined ("hello", "help me", "stop", "thank you" and "yes"). EEG signals were recorded at 64 channels and 256 Hz sampling rate. This is a within-subject classification problem: models are trained and tested on data from the same subject (with a standard train/test split per subject provided by the competition). We adhere to the original competition data splits for consistency with prior work. We attempted a cross-subject evaluation for this dataset (training on some subjects and testing on others) but found that both CBraMod and LaBraM performed at chance level in that scenario, likely due to the extreme variability and limited data per subject in

imagined speech. Therefore, we report results in the within-subject setting, averaging performance across subjects.

Mental Stress Detection – We use public MentalArithmetic Stress dataset (Goldberger et al., 2000; Zyma et al., 2019), in which subjects perform arithmetic tasks under time pressure (stress condition) or relaxed conditions. EEG signals were recorded at 20 channels and 500 Hz sampling rate. The goal is to detect whether a given EEG trial corresponds to a high-stress condition or a low-stress condition (binary classification). We use a strict cross-subject protocol follow CBraMod: Subject 1 to 28 are set as training set, subject 29 to 32 are validation set and subiect 33 to 36 are set to test set. Evaluating is demonstrated on completely held-out subjects, to test generalization to unseen individuals' stress responses.

Motor Imagery Classification – We use the well-known BCI Competition IV 2a dataset (Brunner et al., 2008), which contains EEG recordings from subjects performing motor imagery of four classes (left hand, right hand, both feet, tongue). EEG signals were recorded at 22 channels and 250 Hz sampling rate. We follow the standard cross-subject evaluation: Subject 1-5, 6-7, 8-9 are used for training, validation, and test, respectively, the predefined splits from the competition.

These tasks cover a range of EEG signal characteristics (from cognitive/internal speech processes to emotional stress responses to sensorimotor rhythms) and vary in number of classes and difficulty. They provide a solid testbed for our alignment approach.

C.2 Preprocessing

We follow the preprocessing pipelines recommended by the foundation model authors to ensure compatibility with the models. For CBraMod and LaBraM, EEG signals are re-referenced and band-pass filtered (follow the setting in CBraMod) and then downsampled to 200hz follow the pre-trained sampling rate. Then the signals will be segmented into short windows (1-second segments). We do not apply any task-specific feature engineering – the raw (preprocessed) signals are fed into the models. The same preprocessing is applied consistently across all adaptation methods for fairness.

C.3 Foundation Models

We test our methods with two recent large-scale EEG foundation models as backbones:

CBraMod (Wang et al., 2024b): a Criss-Cross Brain Transformer model (Wang et al., 2025) that was pre-trained on 12 public EEG datasets using a masked patch reconstruction objective. CBraMod features a dual-stream attention architecture separating spatial and temporal attention, and it produces a 768-dimensional feature vector per EEG trial (after temporal pooling). We use the official pre-trained weights released by the authors.

LaBraM (Jiang et al., 2024): the Large Brain Model which was pre-trained on over 2,500 hours of EEG from around 20 datasets. LaBraM employs a ViT-like backbone with an EEG-specific tokenizer (based on vector-quantized neural spectral codes) and was trained to predict masked neural code tokens. We use the official pre-trained weights released by the authors.

These two models are representative of current state-of-the-art EEG foundation models; they have complementary architectures and pre-training approaches, which provides a robust test of our alignment methods. For all experiments, we treat the foundation model as a fixed feature extractor initially and then apply the different adaptation strategies on top (as described below).

C.4 Implementation Details

Finetuning and Adaptation Strategies:

- **Full Supervised Finetuning:** The entire foundation model, along with the task-specific classification head, is trained end-to-end using labeled downstream data with 1-3 linear layers. This serves as a strong baseline for comparison.
- **Lora Finetuning:** A parameter-efficient fine-tuning method that only introduces low-rank updates to a pre-trained model's weights. We insert trainable rank-r matrices in each attention projection (queries, keys, values) and in the feed-forward layers of each transformer

Table 7: Hyperparameters for NeuroTTT (Test Stage)

Hyperparameters	**Settings**
Steps	1/4
Batch size	SSL: 1; Tent: 4
Learning rate	1×10^{-5}
Optimizer	Adam
w_1 (Imagined Speech)	0.6
w_2 (Imagined Speech)	0.6
w_1 (Mental Stress)	0.2
w_2 (Mental Stress)	0.1
w_1 (Motor Imagery)	0.1
w_2 (Motor Imagery)	0.8

block, following Hu et al. (2022). We allow these LoRA layers to be trained on the labeled data while the original weights are frozen. This efficiently adapts the backbone without full fine-tuning. We tried to apply LoRA to as many layers as possible (all self-attention and intermediate dense layers in CBraMod/LaBraM) to maximize adaptability.

- **SHOT (Source Free Domain Adaptation):** We simulate a source-free domain adaptation setting for downstream tasks. We first train a classifier on the source training set (treating the foundation model as fixed feature extractor, akin to linear probing on source). Then, at test time, we adapt this classifier using the SHOT method: we freeze the classifier's weights and instead update the feature extractor (foundation model) on the target (unlabeled test) data by maximizing the mutual information of predictions and minimizing entropy, as described by Liang et al. (2020). In essence, SHOT uses the source-trained classifier's "hypothesis" and adjusts the feature extractor to better fit target data without source data. Notably, while source-free domain adaptation methods like SHOT have been actively explored in other biosignals (e.g., EMG), they remain underexplored for EEG. Only very recent efforts have applied SHOT-style techniques to shallow neural networks in EEG (Wang et al., 2024a; Li et al., 2024), and, to our knowledge, there is virtually no work integrating SHOT within EEG finetuning.
- **Domain-SSL + Test-Time Training (TTT):** This is our full method. We fine-tune the model on the training set using our multi-task loss (with the appropriate domain-specific SSL tasks for that dataset) as described in the Method 3 section. At inference, we apply test-time adaptation. In practice, we evaluate two variants:

  (a) TTT with SSL, which performs a single gradient step on the SSL task for each test sample;

  (b) TTT with Tent, which applies entropy minimization to test batches by updating batch normalization statistics.

For test-time adaptation, NeuroTTT (SSL) uses one gradient update per test sample on the self-supervised task, without online training. For Tent, we apply several updates per test batch to the batch normalization layers using momentum-based exponential moving averages. More hyperparameters for NeuroTTT are shown in Table 7.

D Limitations and Future Works

Limitations: While our results demonstrate consistent gains across tasks and backbones, several limitations merit discussion:

1. Task-Engineered SSL Dependence. Our alignment relies on manually designed domain-specific SSL tasks (e.g., band prediction, A–P flip, temporal jigsaw). These choices encode neuroscientific priors but may not be universally optimal; subpar task design can bias representations or yield negligible gains. Automating SSL design or meta-selecting pretext tasks remains future work.

2. Scope of TTT Parameter Updates. Our current TTT with SSL uses full-parameter updates during inference. While effective, this may be unnecessary or induce instability on noisy trials. Future work will evaluate restricted adaptation—e.g., updating only late blocks, normalization layers, or lightweight adapters—to better trade off stability, compute overhead, and accuracy.
3. TTT Stability vs. Benefit Trade-off. Test-time self-supervised training can be sensitive to noise and atypical trials, occasionally perturbing well-formed features. Although Tent offers a more stable BN-only alternative, its adaptation capacity is limited to normalization statistics and may under-correct severe shifts.

Future Work: We believe this framework can be extended to other neural recording modalities (MEG, sEEG, fNIRS) and other tasks not explored in this paper. Future work will explore automated ways of generating domain-specific SSL tasks (perhaps via meta-learning) and more advanced test-time adaptation schemes (e.g. continuous adaptation in lifelong settings). We also envision incorporating subject metadata or physiological priors into the adaptation to further enhance personalization. Ultimately, aligning brain foundation models to downstream tasks and users is crucial for making these models truly practical, and our work takes an important step in that direction.