BRIDGING ELECTROPHYSIOLOGY AND HEMODYNAMICS: ADAPTING EEG FOUNDATION MODELS FOR EEG-FNIRS REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) has long served as a fundamental modality for decoding brain activity. Existing EEG foundation models learn transferable neural representations from large-scale electrophysiological recordings, but their unimodal design prevents them from leveraging the complementary hemodynamic information measured by functional near-infrared spectroscopy (fNIRS). Extending these models to multimodal EEG-fNIRS data is non-trivial, as the signals differ sharply in sensor layouts, temporal resolution, and physiological origins. To overcome these challenges, we propose EF-GO, a modular framework that adapts pretrained EEG foundation models for multimodal EEG-fNIRS representation learning without retraining the original EEG backbone. EF-GO first learns fNIRS-specific representations via self-supervised masked reconstruction, then employs Region-prior-guided Graph Aggregation to map heterogeneous EEG and fNIRS channel features into a shared, region-aware graph space. To align their distinct temporal dynamics, EF-GO incorporates a time-aware Neural Ordinary Differential Equation (Neural ODE) objective to capture lagged correspondences between electrophysiological and hemodynamic representations. Experiments across six downstream tasks with two pretrained EEG backbones show that EF-GO outperformed unimodal EEG baselines across most metrics and delivered performance competitive with task-specific EEG-fNIRS models. These results demonstrate the potential of adapting pretrained EEG models for multimodal neurophysiological representation learning.

1 INTRODUCTION

Foundation models learn transferable representations through large-scale pre-training and have achieved substantial advances in natural language processing, computer vision, and multimodal learning (Devlin et al. (2019); Dosovitskiy et al. (2021); Radford et al. (2021)). Inspired by this paradigm, the brain signal analysis community has recently developed EEG foundation models that learn generalizable neural representations from large-scale unlabeled recordings. Representative models, including LaBraM (Jiang et al. (2024)), EEGPT (Wang et al. (2024)), and CBraMod (Wang et al. (2025)), have demonstrated promising transferability across brain-computer interfaces (BCIs), neurological disorder diagnosis, emotion recognition, and cognitive-state analysis. Nevertheless, these models are primarily developed for EEG or other electrophysiological signals, limiting their representations to a single physiological modality.

Brain activity, however, is inherently multimodal. Neuronal electrical activity is accompanied by cerebral hemodynamic responses through neurovascular coupling (Yeung & Chu (2022); Chiarelli et al. (2017)). Electroencephalography (EEG) records fast electrophysiological activity with millisecond-level temporal resolution, whereas functional near-infrared spectroscopy (fNIRS) measures cortical changes in oxygenated and deoxygenated hemoglobin, and provides more spatially localized measurements of hemodynamic activity (Chiarelli et al. (2017)). Because EEG and fNIRS involve electrical and optical sensing mechanisms with negligible electro-optical interference, they can be recorded concurrently to provide complementary observations of neuronal and vascular activity (Chiarelli et al. (2017); Shin et al. (2018)). Previous hybrid EEG-fNIRS studies have shown that combining the two modalities can improve decoding performance over unimodal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

approaches in applications such as motor imagery and cognitive-state recognition (Fazli et al. (2012); Liu et al. (2021); Premchand et al. (2025)).

These observations motivate the development of **multimodal brain representation models** capable of jointly modeling heterogeneous neurophysiological signals. However, data from complementary neuroimaging modalities (e.g. paired EEG-fNIRS dataset) are often limited in scale, making large-scale pretraining difficult. A practical alternative is to adapt pretrained EEG foundation models to these modalities while preserving transferable EEG knowledge.

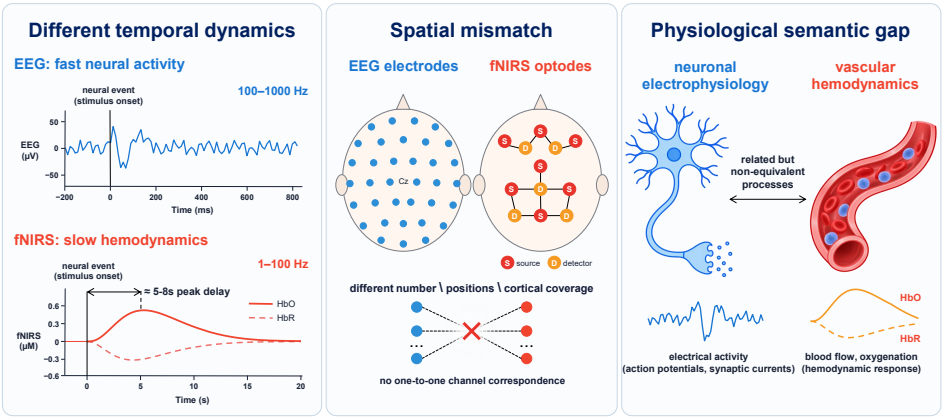


Figure 1: Cross-modal heterogeneity between EEG and fNIRS, including different temporal dynamics, spatial mismatch, and physiological semantic gap between neuronal electrophysiology and vascular hemodynamics.

*Nevertheless, adapting EEG foundation models to paired EEG-fNIRS recordings is challenging because of the **substantial heterogeneity** between the two modalities.* First, EEG and fNIRS exhibit **different temporal dynamics**. EEG records rapid neuronal electrical activity, commonly at sampling rates above 100 Hz, whereas fNIRS captures much slower hemodynamic fluctuations and is typically acquired at lower sampling rates (Chiarelli et al. (2017); Shin et al. (2018)). Moreover, the hemodynamic response follows neuronal activation with a delay of several seconds, meaning that simple sample-level synchronization does not guarantee physiological correspondence between the modalities (Cinciute (2019); Fazli et al. (2012)). Second, **spatial mismatch**. EEG electrodes and fNIRS optodes differ in number, placement, and cortical coverage across acquisition systems and datasets, preventing direct channel-wise correspondence (Shin et al. (2018); Liu et al. (2021)). Finally, EEG and fNIRS characterize **different physiological processes** (Yeung & Chu (2022), i.e., neuronal electrophysiology and vascular hemodynamics. This also makes a cross-modal semantic gap in addition to the spatial and temporal discrepancies).

To address these challenges, we propose EF-GO (EEG-fNIRS Graph-ODE), a modular framework for adapting pretrained EEG foundation models to multimodal EEG-fNIRS representation learning. Instead of retraining the EEG backbone from scratch, EF-GO retains its pretrained representations while introducing a dedicated fNIRS encoding branch. To overcome spatial heterogeneity, we introduce a Region-prior-guided Graph Aggregation module that maps heterogeneous EEG and fNIRS channel layouts into a shared brain-region-aware graph space. To model temporal heterogeneity, we further develop a Time-aware ODE based module to model the neurovascular dynamics (Chen et al. (2018)), which learns the continuous transformation from electrophysiological activity to delayed hemodynamic representations. A time-aware cross-modal alignment objective then aligns the predicted trajectories with fNIRS representations across multiple temporal offsets. We evaluate EF-GO using two pretrained EEG foundation-model backbones across six subject-independent EEG-fNIRS classification tasks. The results show that EF-GO improves the corresponding EEG-only backbones on a majority of the reported evaluation metrics and achieves competitive performance against task-specific multimodal baselines. These findings demonstrate the potential of adapting pretrained EEG models for multimodal neurophysiological representation learning. The code for EF-GO is available at `https://anonymous.4open.science/r/EF-GO-77F0/`.

Our contributions are summarized as follows:

- We propose EF-GO, a unified framework that upgrades pre-trained EEG Foundation Models into EEG-fNIRS Foundation Models. EF-GO leverages the transferable representations learned by existing EEG Foundation Models and extends them to jointly model EEG and fNIRS signals.
- We develop a cross-modal spatial-temporal alignment module that combines region-prior-guided graph aggregation with neural ODE-based neurovascular dynamics modeling. The former constructs modality-specific brain-region representations, while the latter captures the dynamic coupling between electrophysiological and hemodynamic responses, thereby bridging the spatial, temporal, and physiological heterogeneity between EEG and fNIRS.
- Extensive experiments across multiple downstream tasks demonstrate the effectiveness and generalizability of EF-GO. The results show that EF-GO outperforms existing approaches in most scenarios, and can be integrated with different EEG Foundation Models.

2 PRELIMINARIES

Brain Signals. Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) provide complementary observations of brain activity from electrophysiological and hemodynamic perspectives, respectively. EEG records neuronal electrical activity from the cerebral cortex with millisecond-level temporal resolution, whereas fNIRS measures changes in hemoglobin concentration to characterize the neurovascular coupling process, offering substantially higher spatial resolution. Since the two modalities can be acquired simultaneously without mutual interference, EEG-fNIRS multimodal analysis has attracted increasing attention in recent years for applications such as brain-computer interfaces (BCIs), affective computing, and neurological disorder diagnosis. Formally, for a synchronized EEG-fNIRS recording, the EEG samples and fNIRS samples can be defined as:

$$\mathbf{X}^e \in \mathbb{R}^{C_e \times T_e}, \quad \mathbf{X}^f \in \mathbb{R}^{C_f \times M \times T_f}, \tag{1}$$

where C_e and C_f denote the channel counts for EEG and fNIRS. $T_e = nf_e$ and $T_f = nf_f$ represent the temporal length, with n indicating the recording duration (in seconds) and f_e and f_f the EEG and fNIRS sampling frequency. Each fNIRS channel contains M chromophore concentration sequences. In our setting, $M = 2$, corresponding to oxyhemoglobin (HbO) and deoxyhemoglobin (HbR). Since EEG and fNIRS are synchronously acquired, they share the same recording duration n. However, owing to their different sampling frequencies, the temporal resolutions satisfy $T_e \gg T_f$ in most practical scenarios.

Pre-trained EEG Foundation Model. Recent advances in large-scale self-supervised pre-training have driven the rapid development of EEG Foundation Models. Given an EEG sample $\mathbf{X}^e$, a pre-trained EEG Foundation Model can be formulated as

$$\mathbf{Z}^e = F_\theta(\mathbf{X}^e), \tag{2}$$

where F_θ denotes the pre-trained EEG encoder and $\mathbf{Z}^e$ represents the learned high-level semantic representation. Benefiting from large-scale EEG pre-training, such models learn transferable EEG representations with strong generalization ability and can be adapted to a wide variety of downstream brain signal analysis tasks.

Existing EEG foundation models are pretrained primarily on electrophysiological recordings, and their input tokenization and latent representation spaces are therefore not designed to accommodate hemodynamic measurements. Consequently, they cannot be directly applied to fNIRS data without modality-specific adaptation. We investigate how to adapt pretrained EEG backbones to paired EEG-fNIRS recordings while retaining and leveraging their transferable EEG knowledge. Specifically, our objective is to learn an fNIRS encoder and align the two modalities in a shared brain-region-aware representation space that accounts for heterogeneous sensor layouts and lagged temporal dynamics. This formulation enables pretrained EEG foundation models to serve as transferable backbones for multimodal EEG-fNIRS representation learning.

3 METHODS

We propose EEG-fNIRS Graph-ODE (EF-GO), a multimodal framework for learning joint representations of neural electrophysiology and hemodynamics. As Figure 2 shows, EF-GO comprises three

sequential stages: (1) self-supervised fNIRS pre-training to acquire robust hemodynamic representations; (2) EEG-fNIRS spatial-temporal alignment, which reconciles heterogeneous sensor layouts through region-prior-guided dynamic graph aggregation and models cross-modal dynamics with a time-aware Neural ODE; and (3) downstream fine-tuning for task-specific multimodal prediction.

Instead of enforcing channel-wise correspondence between EEG and fNIRS, EF-GO maps modality-specific channel tokens into a common brain-region-aware graph space. This design accommodates heterogeneous montages across acquisition systems and datasets, while the Neural ODE explicitly captures the continuous transformation from fast electrophysiological activity to delayed hemodynamic responses.

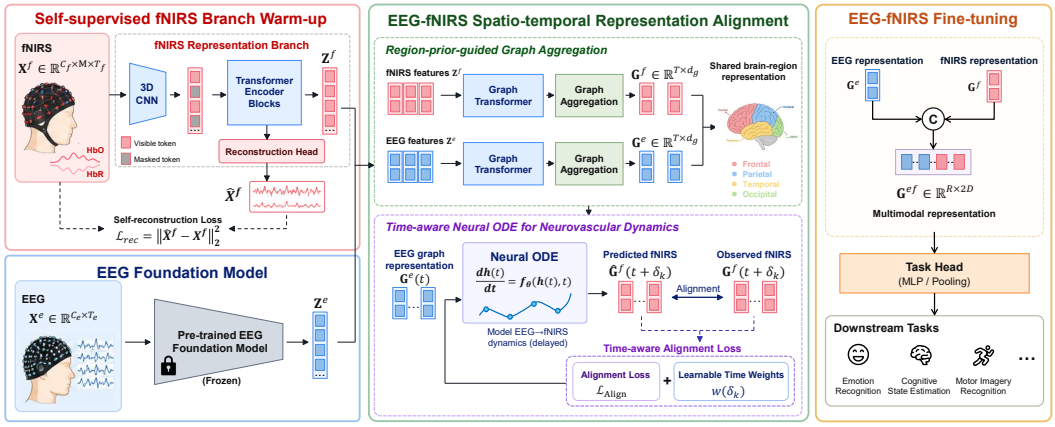


Figure 2: Overview of EF-GO. The framework consists of three stages: self-supervised fNIRS warm-up through masked reconstruction, EEG-fNIRS spatial-temporal alignment through Region-prior-guided Graph Aggregation and time-aware Neural ODE transformation, and downstream fine-tuning using the fused multimodal representations. The Neural ODE is used only during cross-modal post-training and introduces no additional ODE-solving overhead during downstream inference.

3.1 Self-supervised fNIRS Branch Warm-up

EEG and fNIRS differ substantially in sampling rates, signal-generation mechanisms, and statistical properties. Directly aligning a randomly initialized fNIRS encoder with a pre-trained EEG model is therefore prone to unstable optimization. We first pre-train the fNIRS branch on unlabeled fNIRS recordings to learn modality-specific spatial-temporal structure.

Given a n-second fNIRS sample $\mathbf{X}^f \in \mathbb{R}^{C_f \times M \times T_f}$, we first segment its temporal axis into n consecutive one-second patches. A channel-independent temporal convolutional block is first applied to each patch separately to extract local hemodynamic patterns:

$$\mathbf{F}^f = \phi_{\text{conv}}(\text{Patching}(\mathbf{X}^f)), \tag{3}$$

where $\phi_{\text{conv}}(\cdot)$ comprises 3D and 2D convolutions that jointly capture hemoglobin-specific and short-range temporal dynamics.

Then the resulting channel-window features are tokenized, and approximately half of the tokens are randomly masked. The masked token sequence is augmented with learnable spatial and temporal embeddings and processed by an fNIRS Transformer encoder:

$$\mathbf{Z}^f = E_f\left(\widetilde{\mathbf{F}}^f + \mathbf{P}^f + \mathbf{T}^f\right), \tag{4}$$

where $\widetilde{\mathbf{F}}^f$ denotes the masked tokens, and $\mathbf{P}^f$ and $\mathbf{T}^f$ are spatial and temporal embeddings, respectively.

We adopt symmetric masked reconstruction to prevent the encoder from relying solely on local visible context. Specifically, reconstruction is performed under a mask and its complement, yielding

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{comp}}, \tag{5}$$

where $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{comp}}$ are mean squared reconstruction errors over the masked and complementary token sets, respectively. This objective enables the fNIRS encoder to capture intrinsic spatial-temporal hemodynamic structure without downstream labels, providing a stable initialization for subsequent cross-modal alignment.

3.2 EEG-fNIRS SPATIAL-TEMPORAL REPRESENTATION ALIGNMENT

EEG-fNIRS representation learning is challenged by both spatial and temporal heterogeneity. Spatially, EEG electrodes and fNIRS optodes vary in number, placement, and cortical coverage across acquisition systems and datasets. Temporally, EEG reflects rapid neural activity, whereas fNIRS captures delayed hemodynamic responses induced by neurovascular coupling. EF-GO addresses these challenges using region-prior-guided graph aggregation and time-aware neural ODE for modeling neurovascular dynamics.

3.2.1 REGION-PRIOR-GUIDED GRAPH AGGREGATION

To accommodate heterogeneous EEG electrode and fNIRS optode configurations, we introduce a multi-head region fusion mechanism that maps modality-specific channel features into a fixed-dimensional whole-brain representation. Without explicit correspondence between EEG and fNIRS channels, the module organizes channels according to their anatomical regions and aggregates them through region-specific attention heads.

Let the channel-level token representation of modality $m \in \{e, f\}$ at time window t be

$$\mathbf{Z}_t^m = \left[\mathbf{z}_{t,1}^m, \ldots, \mathbf{z}_{t,C_m}^m\right] \in \mathbb{R}^{C_m \times d_m}, \tag{6}$$

where C_m and d_m denote the number of channels and token dimension, respectively. First, a Graph Transformer block exchanges information adaptively across all channels, producing context-aware node representations.

$$\mathbf{Z}_t^{'m} = \text{LayerNorm}\left(\text{Dropout}\left[\text{Self-attention}^m\left(\mathbf{Z}_t^m\right)\right] + \mathbf{W}^m \mathbf{Z}_t^m\right), \tag{7}$$

Then, region-prior-guided multi-head attention pooling aggregates the resulting channel nodes according to their anatomical assignments. Each pooling head is associated with a predefined brain region, and the region-specific outputs are concatenated to form a unified whole-brain representation. The pooling weight for channel i is

$$\alpha_{t,i}^{m,k} = \text{softmax}_i\left(\frac{(\mathbf{u}^k)^\top \widetilde{\mathbf{k}}_{t,i}^{m,k}}{\sqrt{d_k}} + \gamma R_{k,i}^m\right), \tag{8}$$

where $\mathbf{u}^k$ is a learnable global query, $\widetilde{\mathbf{k}}_{t,i}^{m,k}$ is the key projection of the graph-updated node feature, and γ controls the contribution of the anatomical prior. We set the number of heads to 5, corresponding to the frontal, parietal, left & right temporal, and occipital lobes. The graph representation at time window t is then computed as

$$\mathbf{G}_t^m = \text{Concat}_{k=1}^K\left(\sum_{i=1}^{C_m} \alpha_{t,i}^{m,k} \widetilde{\mathbf{v}}_{t,i}^{m,k}\right), \tag{9}$$

where $\widetilde{\mathbf{v}}_{t,i}^{m,k}$ denotes the corresponding value projection. Applying this operation to all temporal windows yields $\mathbf{G}^e, \mathbf{G}^f \in \mathbb{R}^{T \times d_g}$, where T is the number of temporal windows, and d_g denotes the common graph embedding dimension, respectively.

3.2.2 TIME-AWARE NEURAL ODE FOR NEUROVASCULAR DYNAMICS

Although graph aggregation maps EEG and fNIRS into a spatially compatible representation space, their temporal dynamics remain fundamentally different. EEG captures rapid electrophysiological activity, whereas fNIRS reflects the slower hemodynamic response induced by neurovascular coupling. Consequently, an fNIRS state is more closely related to the preceding evolution of neural activity than to the EEG state observed at exactly the same time point.

We model the latent dynamics associated with neurovascular coupling as a continuous-time initial-value problem, for which Neural ODEs are well suited. Hemodynamic responses evolve smoothly and continuously following neural activation. Meanwhile, a shared dynamics function can generate latent states at multiple evolution horizons, allowing the model to capture variable cross-modal delays without imposing a fixed hemodynamic response function.

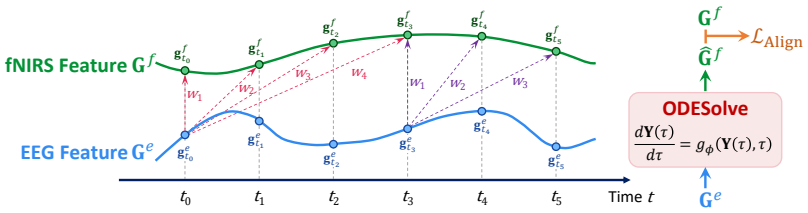


Figure 3: ODE-based temporal alignment between EEG and fNIRS representations.

As Figure 3 shows, let a Transformer-based dynamics function $g_\phi(\cdot)$ learn the continuous gradient flow of the neurovascular coupling mechanism between EEG and fNIRS, and treat the EEG representation as the initial state of the Neural ODE:

$$\frac{d\mathbf{Y}(\tau)}{d\tau} = g_\phi\left(\mathbf{Y}(\tau), \tau\right), \qquad \mathbf{Y}(0) = \mathbf{G}^e, \tag{10}$$

In this way, the ODE injects continuous-time embeddings into the current state and models dependencies across temporal windows using self-attention. Numerical integration produces a sequence of fNIRS prediction trajectories:

$$\widehat{\mathbf{G}}^{f,(\ell)} = \text{ODESolve}\left(g_\phi, \mathbf{G}^e, \ell\right), \qquad \ell = 0, \ldots, T-1. \tag{11}$$

We supervise the predicted trajectories using a time-aware alignment objective. The graph embedding is partitioned into K sub-embeddings based on brain regions. For the ℓ-th ODE step and k-th region, the trajectory matching error is

$$\mathcal{L}_{\ell,k} = \mathcal{L}\left(\widehat{\mathbf{G}}^{f,(\ell)}_{k,1:T-\ell}, \mathbf{G}^f_{k,\ell+1:T}\right), \tag{12}$$

To adaptively emphasize informative temporal trajectories and feature subspaces, EF-GO learns a separate time-aware weight $\omega_{\ell,k} = \sigma(\alpha_{\ell,k})$ for each brain region, where $\sigma(\cdot)$ is the Sigmoid function. All learnable temporal parameters $\alpha_{\ell,k}$ are initialized to 1.0. The final alignment loss is

$$\mathcal{L}_{\text{align}} = \frac{2}{T(T+1)} \sum_{\ell=0}^{T-1} \sum_{k=1}^{K} \sigma(\alpha_{\ell,k}) \mathcal{L}_{\ell,k}. \tag{13}$$

Unlike approaches that impose a fixed hemodynamic delay, this objective adaptively reweights ODE trajectories across temporal steps and feature subspaces, allowing the model to focus on the most informative EEG-fNIRS correspondences.

3.3 EEG-fNIRS Downstream Fine-tuning

After fNIRS pre-training and cross-modal post-training, EF-GO is fine-tuned for downstream EEG-fNIRS classification. The EEG encoder, fNIRS encoder, and Region-prior-guided Graph Aggregation modules are jointly optimized. The Neural ODE transformation and its alignment loss are excluded from the downstream inference path. Consequently, EF-GO retains the benefits of cross-modal dynamic alignment while avoiding the numerical integration overhead of ODE solving during inference.

Specifically, the EEG and fNIRS graph sequences are flattened over the temporal and feature dimensions and concatenated:

$$\mathbf{H}^{ef} = \text{Concat}\left[\text{vec}(\mathbf{G}^e), \text{vec}(\mathbf{G}^f)\right] \in \mathbb{R}^{2 \cdot T \cdot d_g}. \tag{14}$$

The fused representation is passed to a task-specific linear prediction head $D_\psi(\cdot)$:

$$\hat{\mathbf{y}} = D_\psi\left(\mathbf{H}^{ef}\right). \tag{15}$$

4 EXPERIMENTS

Training. The training is divided into two steps: fNIRS branch warm-up and EEG-fNIRS representation alignment post-training. We collected publicly accessible EEG-fNIRS datasets for training. These data consist of recordings from more than one hundred individuals, including various EEG and fNIRS configurations with different channel layouts and sampling frequencies.

Data Preprocessing. All EEG-fNIRS data follow the same data preprocessing procedures. For EEG signals, a 0.5-50 Hz bandpass filter and a 50 Hz notch filter are applied to eliminate offsets and filter out noise. Then the signals are re-referenced using an average reference and uniformly resampled to 200 Hz. For fNIRS signals, all signals are converted to oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) concentrations using the Beer-Lambert Law, and a 0.01-0.2 Hz bandpass filter is applied to eliminate offsets and filter out noise.

Downstream Tasks. To comprehensively validate the performance and stability of our model, we selected datasets from 6 different downstream tasks across 3 major categories, covering brain-computer interfaces (BCI) Shin et al. (2016), cognitive tasks Shin et al. (2018), and emotion recognition Ning et al. (2026). Table 1 presents the basic information of all downstream tasks. All datasets are evaluated under a subject-independent scheme. In each dataset, data from 3 subjects are used as the validation set, data from 3 subjects as the test set, and data from the remaining subjects are used as the training set.

Table 1: Overview of downstream tasks and datasets.

Tasks	Dataset	EEG Rate	EEG Channels	fNIRS Rate	fNIRS Channels	Duration	Samples	Label
Cognitive Task	WG	200 Hz	28	10 Hz	36	10s	1,560	2-class
Cognitive Task	n-Back	200 Hz	28	10 Hz	36	2s	14,040	3-class
Emotion Recognition	Valence	1000 Hz	64	47.62 Hz	51	1s	49,120	3-class
Emotion Recognition	Arousal	1000 Hz	64	47.62 Hz	51	1s	49,120	3-class
Mental Arithmetic	hBCI-MA	200 Hz	30	10 Hz	36	3s	17,400	2-class
Motor Imagery	hBCI-MI	200 Hz	30	10 Hz	36	3s	17,400	2-class

Baselines & Metrics. We compared EF-GO with classic small-scale EEG-fNIRS models and EEG foundation models. Classic small-scale EEG-fNIRS models, including FGANet (Kwak et al. (2022)), M2NN (He et al. (2022)), STA-Net (Liu et al. (2025)), comprehensively consider the differences between the EEG and fNIRS signal modalities. On the other hand, EEG foundation models represent the most well-developed branch in brain signal representation, including LaBraM (Jiang et al. (2024)), CBraMod (Wang et al. (2025)). We adopted the common metrics currently used for EEG foundation models to measure model performance. For binary classification tasks, we use Balanced Accuracy, AUC-PR, and AUC-ROC as evaluation metrics; for multi-class classification tasks, we use Balanced Accuracy, Weighted F1, and Cohen's Kappa as evaluation metrics.

4.1 PERFORMANCE ON DOWNSTREAM TASKS

Table 2 compares EF-GO with other baseline models across six downstream datasets. Overall, EF-GO with LaBraM (i.e., EF-GO(L)) or EF-GO with CBraMod (i.e., EF-GO(C)) could achieve the best performance on most evaluation metrics, demonstrating its effectiveness across cognitive, affective, and BCI tasks.

The improvements are most pronounced on cognitive and emotion recognition tasks. On Cognitive-WG, EF-GO(L) improves balanced accuracy and AUC-ROC over the strongest baselines by 3.11 and 2.61 percentage points, respectively. It also achieves the best results across all three metrics on Cognitive-nBack. On REFED-Valence, EF-GO(C) provides particularly substantial improvements, increasing balanced accuracy from 0.4241 to 0.4871 and Cohen's Kappa from 0.1374 to 0.1901. On REFED-Arousal, both EF-GO(L) and EF-GO(C) achieve better balanced accuracy and Cohen's Kappa. It is worth noting that the two EEG backbones seem to excel in different tasks. EF-GO(C) is notably better at affective tasks, while EF-GO(L) excels at cognitive tasks. In BCI tasks, EF-GO(L) achieved the best performance on the hBCI-MI dataset and the best balanced accuracy on hBCI-MA, while EF-GO(C) was able to achieve the second-best results on hBCI-MA. Meanwhile, the specialized EEG-fNIRS models also exhibited competitive performance on these two datasets. Overall, the EF-GO framework can enhance the performance of unimodal foundation models in most scenarios, enabling them to achieve broad performance levels comparable to specifically designed EEG-fNIRS models.

Table 2: Comparison of EEG-fNIRS downstream task performance.

	Cognitive-WG (2-class)			Cognitive-nBack (3-class)		
Method	**Balanced Accuracy**	**AUC-PR**	**AUC-ROC**	**Balanced Accuracy**	**Weighted F1**	**Cohen's Kappa**
FGANet	0.6633 ± 0.0250	0.7542 ± 0.0167	0.7283 ± 0.0170	0.3586 ± 0.0070	0.3334 ± 0.0085	0.0459 ± 0.0123
M2NN	0.7078 ± 0.0310	0.7569 ± 0.0658	0.7581 ± 0.0431	0.3944 ± 0.0069	0.4198 ± 0.0056	0.1058 ± 0.0095
STA-Net	0.6300 ± 0.0370	**0.8032 ± 0.0405**	0.7744 ± 0.0314	0.3445 ± 0.0158	0.3255 ± 0.0199	0.0159 ± 0.0234
LaBraM	0.6278 ± 0.0242	0.6802 ± 0.0297	0.7123 ± 0.0244	0.3971 ± 0.0086	0.4161 ± 0.0165	0.1117 ± 0.0176
CBraMod	0.6389 ± 0.0252	0.6777 ± 0.0284	0.6996 ± 0.0271	0.4122 ± 0.0128	0.4280 ± 0.0219	0.1408 ± 0.0247
EF-GO(L)	**0.7389 ± 0.0100**	0.7826 ± 0.0231	**0.8167 ± 0.0163**	**0.4137 ± 0.0113**	**0.4314 ± 0.0128**	**0.1458 ± 0.0201**
EF-GO(C)	0.6955 ± 0.0203	0.7880 ± 0.0106	0.7770 ± 0.0114	0.3948 ± 0.0119	0.4152 ± 0.0195	0.1134 ± 0.0177
	REFED-Valence (3-class)			**REFED-Arousal (3-class)**		
Method	**Balanced Accuracy**	**Weighted F1**	**Cohen's Kappa**	**Balanced Accuracy**	**Weighted F1**	**Cohen's Kappa**
FGANet	0.4241 ± 0.0420	0.4743 ± 0.0334	0.1374 ± 0.0714	0.4133 ± 0.0122	0.4260 ± 0.0157	0.1324 ± 0.0187
M2NN	0.4079 ± 0.0310	0.4453 ± 0.0340	0.1117 ± 0.0564	0.4165 ± 0.0178	0.4059 ± 0.0249	0.1385 ± 0.0297
STA-Net	0.4130 ± 0.0181	0.4699 ± 0.0234	0.1130 ± 0.0297	0.4170 ± 0.0329	0.4181 ± 0.0363	0.1404 ± 0.0543
LaBraM	0.3421 ± 0.0078	0.3776 ± 0.0191	0.0085 ± 0.0162	0.3448 ± 0.0117	0.3462 ± 0.0112	0.0224 ± 0.0173
CBraMod	0.3197 ± 0.0165	0.3689 ± 0.0189	-0.0188 ± 0.0274	0.4090 ± 0.0180	0.3662 ± 0.0265	0.1170 ± 0.0265
EF-GO(L)	0.4211 ± 0.0172	0.4451 ± 0.0163	0.1170 ± 0.0202	0.4342 ± 0.0205	**0.4431 ± 0.0207**	0.1559 ± 0.0331
EF-GO(C)	**0.4871 ± 0.0105**	**0.4830 ± 0.0343**	**0.1901 ± 0.0217**	**0.4623 ± 0.0125**	0.4206 ± 0.0230	**0.1770 ± 0.0174**
	hBCI-MA (2-class)			**hBCI-MI (2-class)**		
Method	**Balanced Accuracy**	**AUC-PR**	**AUC-ROC**	**Balanced Accuracy**	**AUC-PR**	**AUC-ROC**
FGANet	0.6412 ± 0.0174	0.6902 ± 0.0348	0.7062 ± 0.0225	0.6111 ± 0.0249	0.6509 ± 0.0327	0.6594 ± 0.0314
M2NN	0.6431 ± 0.0135	0.7131 ± 0.0287	0.7043 ± 0.0259	0.5233 ± 0.0118	0.5316 ± 0.0099	0.5395 ± 0.0171
STA-Net	0.6920 ± 0.0259	**0.7924 ± 0.0300**	**0.7891 ± 0.0309**	0.6533 ± 0.0091	0.6697 ± 0.0151	0.6979 ± 0.0123
LaBraM	0.6703 ± 0.0038	0.7396 ± 0.0072	0.7421 ± 0.0061	0.6441 ± 0.0082	0.7023 ± 0.0123	0.6977 ± 0.0124
CBraMod	0.6431 ± 0.0106	0.7527 ± 0.0167	0.7579 ± 0.0187	0.6377 ± 0.0086	0.7219 ± 0.0105	0.7124 ± 0.0109
EF-GO(L)	**0.6969 ± 0.0056**	0.7472 ± 0.0129	0.7638 ± 0.0060	**0.6557 ± 0.0223**	**0.7255 ± 0.0313**	**0.7130 ± 0.0312**
EF-GO(C)	0.6958 ± 0.0123	0.7784 ± 0.0104	0.7668 ± 0.0098	0.5922 ± 0.0210	0.6273 ± 0.0339	0.6236 ± 0.0280

EF-GO(L): EF-GO with LaBraM EEG backbone, EF-GO(C): EF-GO with CBraMod EEG backbone.

4.2 Ablation Experiments

To verify the effectiveness of our proposed framework, we conducted ablation experiments on multimodal fusion, graph aggregation, ODE alignment, and the major training components of EF-GO.

Table 3: Ablation experiment results on EF-GO (L).

Dataset	EEG FM	fNIRS branch	Graph Agg.	ODE Alignment	**Balanced Accuracy**	**AUC-PR**	**AUC-ROC**
Cognitive-WG	✓	×	×	×	0.6278 ± 0.0242	0.6802 ± 0.0297	0.7123 ± 0.0244
	×	✓	✓	×	0.5067 ± 0.0054	0.5038 ± 0.0033	0.5055 ± 0.0102
	✓	✓	×	×	0.6767 ± 0.0166	0.7490 ± 0.0304	0.7594 ± 0.0289
	✓	✓	✓	×	0.6934 ± 0.0206	0.7417 ± 0.0190	0.7716 ± 0.0104
	✓	w/o warm-up	✓	✓	0.7200 ± 0.0411	0.7781 ± 0.0218	0.8082 ± 0.0235
	✓	✓	w/o region prior	✓	0.7133 ± 0.0103	0.7402 ± 0.0270	0.7891 ± 0.0139
	✓	✓	✓	w/o time-aware	0.7256 ± 0.0283	0.7688 ± 0.0129	0.8076 ± 0.0171
	✓	✓	✓	✓	**0.7389 ± 0.0100**	**0.7826 ± 0.0231**	**0.8167 ± 0.0163**
hBCI-MA	✓	×	×	×	0.6703 ± 0.0038	0.7396 ± 0.0072	0.7421 ± 0.0061
	×	✓	✓	×	0.5689 ± 0.0189	0.5963 ± 0.0209	0.6047 ± 0.0216
	✓	✓	×	×	0.6770 ± 0.0159	0.7320 ± 0.0263	0.7467 ± 0.0185
	✓	✓	✓	×	0.6928 ± 0.0053	0.7365 ± 0.0069	0.7527 ± 0.0048
	✓	w/o warm-up	✓	✓	0.6868 ± 0.0146	0.7363 ± 0.0091	0.7503 ± 0.0131
	✓	✓	w/o region prior	✓	0.6910 ± 0.0134	0.7447 ± 0.0216	0.7583 ± 0.0132
	✓	✓	✓	w/o time-aware	0.6895 ± 0.0073	0.7457 ± 0.0081	0.7616 ± 0.0059
	✓	✓	✓	✓	**0.6969 ± 0.0056**	**0.7472 ± 0.0129**	**0.7638 ± 0.0060**

As shown in Table 3, the EEG-only backbone consistently outperforms the fNIRS-only branch, confirming the importance of electrophysiological information. Nevertheless, naive EEG-fNIRS fusion improves balanced accuracy over the EEG-only model, demonstrating that fNIRS provides complementary hemodynamic information, particularly for the cognitive task. Adding brain-region-aware graph aggregation further improves balanced accuracy over naive fusion. Incorporating ODE-based alignment also yields an additional gain. These results indicate that graph aggregation effectively reconciles heterogeneous sensor layouts, while the temporal transformation learned by the Neural ODE is especially beneficial for tasks with extended neurovascular dynamics.

Removing the warm-up of fNIRS reduces performance, indicating that warm-up is indeed more conducive to learning stable fNIRS representations. Besides, the brain-region prior and time-aware

weighting also contribute. Lastly, EF-GO achieved the best results across all metrics, demonstrating that the various modules can work together to achieve better outcomes.

4.3 VISUAL ANALYSIS.

To demonstrate the effectiveness of the spatial-temporal alignment in our proposed EF-GO framework, we conducted some visual analyses as Figure 4 shows.

As shown in Figure 4(a), the post-training loss continuously decreases, eventually approaching zero. The overall loss curve exhibits a steady downward trend and converges after approximately 5,000 training steps. Meanwhile, the minimum learnable temporal parameter, $\min(\alpha)$, remained above 0.99 without exhibiting any significant decline. This indicates that the ODE-based cross-modal learning objective can be effectively optimized.

Figure 4(b) illustrates the learned temporal weights $w_{k,\ell}$ across different ODE steps. All five curves display a consistent rise–peak–decline pattern: relatively small weights are assigned to the initial trajectory steps, the weights increase rapidly from $\ell = 2$ to $\ell = 4$ second, and the maximum values occur around $\ell = 5$ second. This phenomenon is very similar to the hemodynamic response function, i.e., the response does not take effect immediately but instead reaches its peak after 4-8 seconds (Stute et al. (2025)). Meanwhile, curves show similar temporal trends, suggesting a shared temporal preference with potential region-specific variation.

Figure 4(c) shows the t-SNE distribution plots for EEG representations, original fNIRS representations, and ODE-predicted fNIRS representations. The EEG features form a compact cluster that is clearly separated from the original fNIRS features, reflecting the substantial modality gap between electrophysiological and hemodynamic features. After ODE transformation, the centroid of the predicted fNIRS features moves from the EEG domain toward the centroid of the original fNIRS features, as indicated by the arrow. Although the predicted and original distributions do not completely overlap, the markedly reduced centroid separation provides qualitative evidence that the ODE module transforms EEG features toward the fNIRS representation space. These results support the role of the ODE module in reducing cross-modal discrepancy while preserving a structured latent representation.

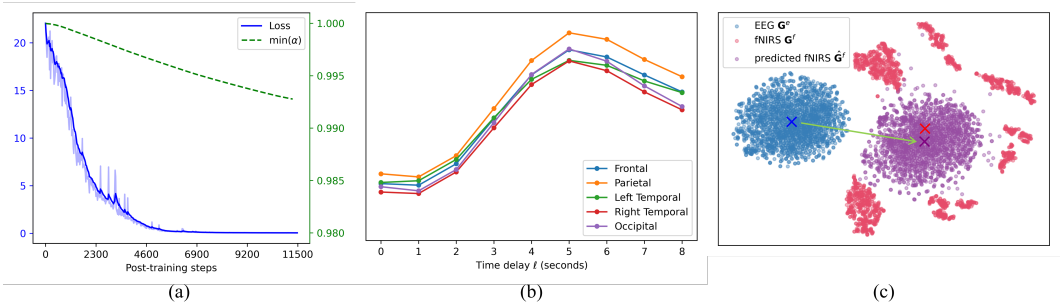


Figure 4: Visualization for EF-GO post-training. (a) Post-training loss curve and minimum time weight curve. (b) Learned temporal weighting profiles across ODE trajectory steps (ℓ). (c) t-SNE distributions of EEG and fNIRS features before and after alignment.

5 CONCLUSION

In this paper, we presented EF-GO, a unified framework that upgrades pre-trained EEG Foundation Models into EEG-fNIRS Foundation Models through cross-modal spatial-temporal alignment. By introducing region-prior-guided brain graph aggregation and Neural ODE-based neurovascular temporal dynamics modeling, EF-GO effectively bridges the spatial and temporal heterogeneity between EEG and fNIRS while preserving the transferable knowledge of existing EEG Foundation Models. Extensive experiments across multiple downstream tasks demonstrate that EF-GO can improve multimodal representation learning and downstream task performance with modest additional training cost. We hope this work provides a practical paradigm for extending single-modality Brain Foundation Models toward more general Multimodal Brain Foundation Models, facilitating the integration of complementary neuroimaging modalities in future brain signal foundation models.

AI USE STATEMENT

In this work, we used generative AI tools for <assist with translation, support qualitative and thematic data analysis, interpret results>. We have not used generative AI tools for <clean and reformat dataset>, and <formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, generate synthetic data sets>are not applicable to this work. Additionally, we used generative AI tools for <sourcing/searching for information, edit a research paper to improve readability>. We have reviewed all AI-assisted work. All AI-assisted content was reviewed by the authors. Translations were checked against the original text; qualitative and thematic analyses and interpretations were verified using the original data; retrieved information and references were confirmed through primary sources; and language edits were reviewed to ensure that they did not alter the intended meaning, findings, or arguments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will remain committed to open-source and reproducibility efforts. The EF-GO architecture, training objectives, and individual model components are described in Section 3 and Appendix B, with additional analyses and ablation results provided in Section 4 Appendix C and Appendix D. The datasets, task settings, sampling rates, channel configurations, and subject-independent data splits are documented in Section 4 and Appendix B. To further support reproducibility, we provide anonymized source code, at `https://anonymous.4open.science/r/EF-GO-77F0/`. The code will be made publicly available upon publication later.

REFERENCES

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Antonio M. Chiarelli, Filippo Zappasodi, Francesco Di Pompeo, and Arcangelo Merla. Simultaneous functional near-infrared spectroscopy and electroencephalography for monitoring of human brain activity and oxygenation: A review. *Neurophotonics*, 4(4):041411, 2017. doi: 10.1117/1.NPh.4.4.041411.

Sigita Cinciute. Translating the hemodynamic response: Why focused interdisciplinary integration should matter for the future of functional neuroimaging. *PeerJ*, 7:e6621, 2019. doi: 10.7717/peerj.6621.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Siamac Fazli, Jan Mehnert, Jens Steinbrink, Gabriel Curio, Arno Villringer, Klaus-Robert Müller, and Benjamin Blankertz. Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*, 59(1):519–529, 2012. doi: 10.1016/j.neuroimage.2011.07.084.

Qun He, Lufeng Feng, Guoqian Jiang, and Ping Xie. Multimodal multitask neural network for motor imagery classification with EEG and fNIRS signals. *IEEE Sensors Journal*, 22(21):20695–20706, 2022. doi: 10.1109/JSEN.2022.3205956.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Youngchul Kwak, Woo-Jin Song, and Seong-Eun Kim. FGANet: fNIRS-guided attention network for hybrid EEG-fNIRS brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:329–339, 2022. doi: 10.1109/TNSRE.2022.3149899.

S Lee, HR Jung, IN Wang, MK Jung, H Kim, and DJ Kim. Multimodal eeg and fnirs biosignal acquisition during motor imagery tasks in patients with orthopedic impairment, 2022.

Rihui Li, Dalin Yang, Feng Fang, Keum-Shik Hong, Allan L. Reiss, and Yingchun Zhang. Concurrent fNIRS and EEG for brain function investigation: A systematic, methodology-focused review. *Sensors*, 22(15):5865, 2022. doi: 10.3390/s22155865.

Yang Li, Xin Zhang, and Dong Ming. Early-stage fusion of EEG and fNIRS improves classification of motor imagery. *Frontiers in Neuroscience*, 16:1062889, 2023. doi: 10.3389/fnins.2022.1062889.

Mutian Liu, Banghua Yang, Lin Meng, Yonghuai Zhang, Shouwei Gao, Peng Zan, and Xinxing Xia. STA-Net: Spatial–temporal alignment network for hybrid EEG-fNIRS decoding. *Information Fusion*, 119:103023, 2025. doi: 10.1016/j.inffus.2025.103023.

Ziming Liu, Jeremy Shore, Miao Wang, Fengpei Yuan, Aaron Buss, and Xiaopeng Zhao. A systematic review on hybrid EEG/fNIRS in brain–computer interface. *Biomedical Signal Processing and Control*, 68:102595, 2021. doi: 10.1016/j.bspc.2021.102595.

Alexandra Medvedeva, Nikolay Syrov, Lev Yakovlev, Yana Alieva, Artemiy Berkmush-Antipova, Galina Ivanova, Natalia Shusharina, and Alexander Kaplan. Multisession fnirs-eeg data of post-stroke motor recovery. recordings during intact and paretic hand movements. *Scientific Data*, 13 (1):448, 2026.

Manuel Muñoz-Caracuel, Vanesa Muñoz, Francisco J Ruíz-Martínez, Dalila Di Domenico, Sabrina Brigadoi, and Carlos M Gómez. Multivariate analysis of the systemic response to auditory stimulation: An integrative approach. *Experimental physiology*, 106(4):1072–1098, 2021.

Xiaojun Ning, Jing Wang, Zhiyang Feng, Tianzuo Xin, Shuo Zhang, Shaoqi Zhang, Zheng Lian, Yi Ding, Youfang Lin, and Ziyu Jia. Refed: A subject real-time dynamic labeled eeg-fnirs synchronized recorded emotion dataset. *Advances in Neural Information Processing Systems*, 38, 2026.

Pablo Ortega, Tong Zhao, and A Aldo Faisal. Hygrip: full-stack characterization of neurobehavioral signals (fnirs, eeg, emg, force, and breathing) during a bimanual grip force control task. *Frontiers in Neuroscience*, 14:919, 2020.

Tustanah Phukhachee, Thanate Angsuwatanakul, Keiji Iramina, and Boonserm Kaewkamnerdpong. A simultaneous eeg-fnirs dataset of the visual cognitive motivation study in healthy adults. *Data in Brief*, 53:110260, 2024.

Brian Premchand, Zhuo Zhang, Kai Keng Ang, Juanhong Yu, Isaac Okumura Tan, Josephine Pei Wen Lam, Anna Xin Yi Choo, Ananda Sidarta, Patrick Wai Hang Kwong, and Lau Ha Chloe Chung. A personalized multimodal bci–soft robotics system for rehabilitating upper limb function in chronic stroke patients. *Biomimetics*, 10(2):94, 2025.

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.

Milan Rybář, Riccardo Poli, and Ian Daly. Simultaneous eeg and fnirs recordings for semantic decoding of imagined animals and tools. *Scientific Data*, 12(1):613, 2025.

Jian Shi, Danyang Chen, Xingwei Zhao, Zhixian Zhao, Shengjie Li, Yeguang Xu, Tao Ding, Zheng Zhu, Peng Zhang, Qing Ye, et al. Hefmi-ich: a hybrid eeg-fnirs motor imagery dataset for brain-computer interface in intracerebral hemorrhage. *Scientific Data*, 12(1):1816, 2025.

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.

Jaeyoung Shin, Alexander von Lühmann, Do-Won Kim, Jan Mehnert, Han-Jeong Hwang, and Klaus-Robert Müller. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, 5:180003, 2018. doi: 10.1038/sdata.2018.3.

Katharina Stute, Louisa K Gossé, Samuel Montero-Hernandez, Guy A Perkins, Meryem A Yücel, Simone Cutini, Turgut Durduran, Ann-Christine Ehlis, Marco Ferrari, Judit Gervain, et al. The fnirs glossary project: a consensus-based resource for functional near-infrared spectroscopy terminology. *Neurophotonics*, 12(2):027801–027801, 2025.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024. doi: 10.52202/079017-1239.

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. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023. doi: 10.52202/075280-3420.

Yuxuan Yao, Hongbo Wang, Li Chen, Yiheng Peng, and Jingjing Luo. Foundation models for EEG decoding: Current progress and prospective research. *Journal of Neural Engineering*, 22(6), 2025. doi: 10.1088/1741-2552/ae17e9.

Michael K. Yeung and Vivian W. Chu. Viewing neurovascular coupling through the lens of combined EEG–fNIRS: A systematic review of current methods. *Psychophysiology*, 59(6):e14054, 2022. doi: 10.1111/psyp.14054.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26304–26321, 2023.

A Related Work

A.1 Brain Foundation Models

Large-scale self-supervised pre-training has recently emerged as a promising paradigm for learning transferable representations from large-scale recordings. Based on this paradigm, unlike conventional EEG models trained for a specific dataset or task, brain foundation models aim to acquire reusable representations that can be adapted to diverse clinical, cognitive, and brain-computer interface applications (Yao et al. (2025)). Existing approaches mainly differ in their signal tokenization strategies, pre-training objectives, and mechanisms for accommodating heterogeneous channel configurations.

Early efforts investigated general-purpose representation learning across physiological or intracranial signals. BIOT (Yang et al. (2023)) tokenizes individual signal channels into a unified sequence representation, enabling cross-dataset learning under mismatched channels, variable recording lengths, and missing values. Brant (Zhang et al. (2023)) scales self-supervised pre-training to large intracranial neural recordings and jointly models temporal and frequency-domain information. Although these models demonstrate the potential of large-scale neural-signal pre-training, BIOT is designed for heterogeneous biosignals rather than paired neuroimaging modalities, whereas Brant primarily targets intracranial recordings. LaBraM (Jiang et al. (2024)) introduces a neural tokenizer and masked EEG modeling objective to learn generic representations from approximately 2,500 hours of EEG recordings. EEGPT (Wang et al. (2024)) employs dual self-supervised objectives and hierarchical spatial-temporal representation alignment to improve transferability across EEG tasks. CBraMod (Wang et al. (2025)) introduces criss-cross attention to separately model spatial and temporal dependencies while accommodating different EEG channel configurations. More recently, REVE (El Ouahidi et al. (2025)) uses geometry-aware positional encoding and large-scale masked pre-training to process EEG recordings with arbitrary electrode layouts and sequence lengths.

Despite their strong transferability, these foundation models predominantly operate on a single electrophysiological modality. Even models trained across different types of biosignals generally encode each signal independently and do not explicitly model the physiological coupling between simultaneously recorded electrophysiological and hemodynamic activity. In contrast, EF-GO investigates how the transferable knowledge of pretrained EEG foundation models can be extended to paired EEG-fNIRS recordings through modality-specific pre-training and cross-modal spatial-temporal alignment.

A.2 EEG-fNIRS Multimodal Representation Learning

EEG and fNIRS provide complementary measurements of neuronal electrical activity and cerebral hemodynamics, respectively. Earlier hybrid EEG-fNIRS systems primarily relied on handcrafted features followed by feature-level or decision-level fusion. For example, simultaneous EEG and fNIRS measurements were shown to improve motor-imagery decoding compared with unimodal systems (Fazli et al. (2012)). Subsequent reviews have summarized the benefits and methodological challenges of concurrent EEG-fNIRS analysis, including differences in sampling rate, sensor configuration, physiological origin, and hemodynamic delay (Liu et al. (2021); Li et al. (2022)).

With the development of deep learning, recent studies have moved from handcrafted fusion toward end-to-end multimodal representation learning. Li et al. (Li et al. (2023)) systematically compared early-, intermediate-, and late-stage fusion and found that early fusion can improve motor-imagery classification. FGANet (Kwak et al. (2022)) uses fNIRS-derived spatial attention to guide EEG feature extraction and combines modality-specific predictions to reduce performance degradation associated with delayed hemodynamic responses. M2NN (He et al. (2022)) introduces a multimodal multitask architecture that jointly optimizes EEG, fNIRS, and fused prediction branches for motor-imagery decoding. STA-Net (Liu et al. (2025)) uses fNIRS-guided spatial attention and EEG-guided temporal cross-attention to dynamically align EEG and fNIRS signals.

Although these methods demonstrate the benefits of multimodal interaction and adaptive alignment, they are predominantly task-specific architectures trained from scratch on individual EEG-fNIRS datasets. Consequently, they cannot directly exploit the transferable representations learned by large-scale EEG foundation models. Moreover, most existing approaches model neurovascular de-

lays through attention, temporal weighting, or sliding-window fusion, without explicitly learning continuous cross-modal dynamics from electrophysiological activity to hemodynamic responses.

B ARCHITECTURE OF THE FNIRS BRANCH

Table 4 summarizes the fNIRS branch used in the EF-GO framework. Let B denote the batch size, C_f the number of fNIRS channels, and n denote the number of temporal windows. Each input window contains $M = 2$ hemoglobin signals (HbO and HbR) and 20 temporal sampling points. For each fNIRS sample, the final output of the fNIRS Branch is $\mathbf{Z}^f \in \mathbb{R}^{(C_f * n) \times 120}$. The total trainable parameters of the fNIRS Branch are 1.43 M.

Table 4: Layer-wise architecture of the shared fNIRS encoder.

Stage	**Layer / operation**	**Configuration**	**Output shape**
Input	fNIRS signals	C_f channels; n windows; 2 modals (HbO and HbR); 20 points per window	$B \times C_f \times n \times 2 \times 20$
Token-axis formation	Rearrangement and singleton feature-axis insertion	Merge channel and window axes, where $P = C_f \cdot n$	$B \times 1 \times P \times 2 \times 20$
Cross-modal Conv	3-D Conv, GroupNorm, and GELU	$1 \to 8$ channels; kernel $1 \times 2 \times 7$; stride $1 \times 1 \times 4$; padding $0 \times 1 \times 3$	$B \times 8 \times P \times 3 \times 5$
Feature reshape	Merge the modal-feature and temporal-feature axes	$3 \times 5 = 15$ features	$B \times 8 \times P \times 15$
Local refinement 1	2-D Conv, GroupNorm, and GELU	$8 \to 8$ channels; kernel 1×3; stride 1; padding 0×1	$B \times 8 \times P \times 15$
Local refinement 2	2-D Conv, GroupNorm, and GELU	$8 \to 8$ channels; kernel 1×3; stride 1; padding 0×1	$B \times 8 \times P \times 15$
Token embedding	Rearrangement and flattening	Each token has $8 \times 15 = 120$ features	$B \times P \times 120$
Position encoding	Channel-position and temporal embeddings	Replicate the selected channel embeddings across n windows and add the corresponding temporal embeddings	$B \times P \times 120$
Transformer encoder	8 pre-normalized Transformer blocks	10-head self-attention; Q/K normalization; MLP $120 \to 480 \to 120$; GELU; residual connections; layer scale; stochastic depth	$B \times P \times 120$
Output normalization	LayerNorm	$\epsilon = 10^{-6}$	$B \times P \times 120$

C DATASETS FOR POST-TRAINING

Table 5 summarizes the seven simultaneous EEG–fNIRS datasets used for cross-modal alignment during post-training. These datasets cover a diverse range of experimental paradigms, including auditory processing, semantic cognition, motor execution, stroke rehabilitation, visual cognition, orthopedic rehabilitation, and hybrid motor imagery. Considerable heterogeneity exists across datasets: EEG sampling rates range from 256 to 4,000 Hz with 8–64 channels, whereas fNIRS sampling rates range from 3.67 to 13.57 Hz with 12–140 channels. Trial durations also vary from 1 to 9 s. This diversity provides a challenging yet representative setting for learning modality-invariant representations and aligning EEG and fNIRS signals despite differences in spatial coverage, temporal resolution, and acquisition protocols.

Table 5: Overview of post-training tasks and datasets.

Dataset	EEG Rate	EEG Channels	fNIRS Rate	fNIRS Channels	Duration
Auditory (Muñoz-Caracuel et al. (2021))	1000 Hz	11	3.67 Hz	76	4s
ToolAnimal (Rybář et al. (2025))	2048 Hz	64	7.81 Hz	14	1s
HYGRIP (Ortega et al. (2020))	4000 Hz	24	12.5 Hz	24	2s
Stroke (Medvedeva et al. (2026))	500 Hz	8	13.57 Hz	140	4s
VisualCognitive (Phukhachee et al. (2024))	500 Hz	31	10 Hz	48	3s
Orthopedic (Lee et al. (2022))	1000 Hz	18	7.18 Hz	12	5s
HEFMI (Shi et al. (2025))	256 Hz	32	11 Hz	90	9s

D Experimental Configurations

Table 6 and Table 7 present the model hyperparameters during the representation learning and downstream task adaptation stages, respectively. For the downstream tasks, all datasets were split using a subject-independent protocol. Specifically, all data from three subjects in each dataset were used as the test set, data from another three subjects as the validation set, and data from the remaining subjects as the training set. The optimal model is selected based on the results from the validation set for testing on the test set. Meanwhile, all experiments were repeated using five fixed random seeds (0, 1, 2, 3, 4), and the mean and standard deviation were reported.

Table 6: Hyperparameters for fNIRS warm-up and EEG-fNIRS representation alignment.

Hyperparameters	**Values**
Batch size	64
Peak learning rate	5e-6
Minimal learning rate	1e-7
Learning rate scheduler	Cosine
Optimizer	AdamW
Adam β	(0.9,0.98)
Weight decay	0.05
Total epochs	50
Warmup epochs	5
EEG patch window	200
fNIRS patch window	20
Gradient clipping	3.0
$\mathcal{L}_{\text{Align}}$	MMD distance
Warm-up mask ratio	0.5

Table 7: Hyperparameters for downstream fine-tuning.

Hyperparameters	**EF-GO(L)**	**EF-GO(C)**
MLP size	200	[200×T, 200]
Batch size	{64, 128, 192, 256}	
Peak learning rate	{5e-5, 1e-4, 5e-4}	
Minimal learning rate	1e-6	
Learning rate scheduler	Cosine	
Optimizer	AdamW	
Adam β	(0.9,0.98)	
Weight decay	0.05	
Total epochs	50	
Warmup epochs	5	
Gradient clipping	3.0	
Label smoothing	0.1 (multi-class classification)	

E MORE RESULTS

Table 8 shows the average performance among six downstream tasks. EF-GO consistently outperforms the baselines across the six downstream tasks. EF-GO(L) achieves the best balanced accuracy, AUC-PR, AUC-ROC, and weighted F1, while EF-GO(C) obtains the highest Cohen's kappa. It is worth noting that task-specific EEG-fNIRS models performed poorly in terms of average performance. These results demonstrate the effectiveness and generalizability of EF-GO across different classification settings and EEG backbones.

Table 8: Average performance among six downstream tasks.

	All tasks	**2-class classification tasks**		**3-class classification tasks**	
Method	**Balanced Accuracy**	**AUC-PR**	**AUC-ROC**	**Weighted F1**	**Cohen's Kappa**
FGANet	0.5186 ± 0.1226	0.7535 ± 0.1022	0.7382 ± 0.0740	0.4112 ± 0.0585	0.1052 ± 0.0420
M2NN	0.5155 ± 0.1220	0.7356 ± 0.1455	0.7243 ± 0.1273	0.4237 ± 0.0163	0.1187 ± 0.0142
STA-Net	0.5250 ± 0.1367	0.7751 ± 0.0628	0.7474 ± 0.0364	0.4045 ± 0.0597	0.0898 ± 0.0534
LaBraM	0.5094 ± 0.1507	0.7815 ± 0.1040	0.7777 ± 0.0837	0.3800 ± 0.0286	0.0475 ± 0.0457
CBraMod	0.5101 ± 0.1333	0.7684 ± 0.0922	0.7586 ± 0.0648	0.3877 ± 0.0285	0.0797 ± 0.0703
EF-GO(L)	**0.5601 ± 0.1393**	**0.8003 ± 0.0864**	**0.8019 ± 0.0744**	**0.4399 ± 0.0060**	0.1396 ± 0.0165
EF-GO(C)	0.5546 ± 0.1153	0.7785 ± 0.1037	0.7542 ± 0.0818	0.4396 ± 0.0308	**0.1602 ± 0.0335**

EF-GO(L): EF-GO with LaBraM EEG backbone, EF-GO(C): EF-GO with CBraMod EEG backbone.

Table 9 quantitatively evaluates EEG–fNIRS alignment at different stages of post-training. After 5 epochs, the predicted fNIRS representations exhibit only a marginal improvement in cosine similarity over the EEG representations (0.0068 vs. -0.0132), while producing higher MSE and MMD, indicating that the cross-modal mapping has not yet converged. After 20 epochs, alignment improves substantially across all three metrics. The cosine similarity between the predicted and observed fNIRS representations reaches 0.8195, compared with 0.5804 for the direct EEG–fNIRS comparison. Meanwhile, the MSE decreases from 0.828 to 0.331, and the MMD distance decreases from 160.64 to 59.97, corresponding to reductions of approximately 60.0% and 62.7%, respectively. These results demonstrate that sufficient post-training enables EF-GO to transform EEG representations toward the fNIRS feature space, producing representations that more closely match the observed hemodynamic features.

Table 9: Quantitative analysis of EEG and fNIRS representations alignment.

Comparison	Cosine sim ↑	MSE ↓	MMD distance ↓
- after 5 epochs post-training			
EEG vs. fNIRS	-0.0132	2.018	386.01
predicted fNIRS vs. fNIRS	0.0068	3.184	569.21
- after 20 epochs post-training			
EEG vs. fNIRS	0.5804	0.828	160.64
predicted fNIRS vs. fNIRS	0.8195	0.331	59.97

Figure 5 illustrates the EEG aggregation weights on a dataset lacking electrodes from the left and right temporal lobes. It can be observed that the left and right temporal lobes can adaptively generate features for the missing brain regions based on the characteristics of other brain regions.

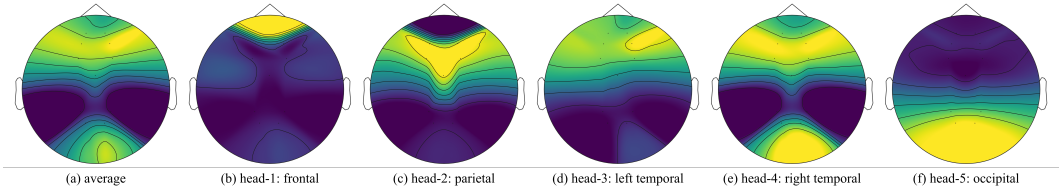


Figure 5: EEG aggregation weights on a dataset lacking the left and right temporal electrodes.

F CONVERGENCE OF ODE-BASED EEG–FNIRS ALIGNMENT

Here, to simplify notation, let $z_i^e \in \mathbb{R}^d$ denote the EEG representation of the i-th sample. Its transformation into the fNIRS representation space is modeled by

$$\frac{\mathrm{d}z_i(s)}{\mathrm{d}s} = f_\theta(z_i(s), s), \qquad z_i(0) = z_i^e, \tag{16}$$

where f_θ is a neural vector field. The solution at time t is denoted by

$$\widehat{z}_{i,t}^f = \Phi_\theta^t(z_i^e). \tag{17}$$

Let $\widehat{P}_{\theta,t}$ and $P_{f,t}$ denote the distributions of the predicted and observed fNIRS representations at time t, respectively. The learnable temporal parameter is denoted by a_t, and its effective loss weight is $w_t = \mathrm{sigmoid}(a_t) = \frac{1}{1+\exp(-a_t)}$. The ODE alignment objective is defined as

$$\mathcal{L}(\theta, \mathbf{a}) = \frac{1}{C} \sum_{t=1}^{T} w_t D_t(\theta), \tag{18}$$

where $C > 0$ is a fixed normalization constant and

$$D_t(\theta) = \mathrm{MMD}_k^2 \left(\widehat{P}_{\theta,t}, P_{f,t} \right) \geq 0. \tag{19}$$

The Gaussian kernel is given by

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right). \tag{20}$$

Assumptions. We assume that:

1. $f_\theta(s, z)$ is continuous in s and globally Lipschitz continuous in z: $\|f_\theta(s, z_1) - f_\theta(s, z_2)\| \leq L_f \|z_1 - z_2\|$.
2. The ODE parameters and the temporal parameters remain bounded. In particular, there exist finite constants $a_{\min}$ and $a_{\max}$ such that $a_{\min} \leq a_t \leq a_{\max}, \quad t = 1, \ldots, T$.
3. Dropout is disabled during ODE integration, and the derivatives required by the RK4 method are uniformly bounded.
4. The kernel bandwidth satisfies $\sigma \geq \sigma_{\min} > 0$.
5. The numerical objective has an L-Lipschitz continuous gradient with respect to all trainable parameters.

From Assumption (2), the effective temporal weights satisfy

$$0 < w_{\min} := \mathrm{sigmoid}(a_{\min}) \leq w_t \leq \mathrm{sigmoid}(a_{\max}) =: w_{\max} < 1. \tag{21}$$

Therefore, the weighted and unweighted alignment losses satisfy

$$\frac{w_{\min}}{C} \sum_{t=1}^{T} D_t(\theta) \leq \mathcal{L}(\theta, \mathbf{a}) \leq \frac{w_{\max}}{C} \sum_{t=1}^{T} D_t(\theta). \tag{22}$$

Thus, the learnable sigmoid weights only rescale the individual alignment terms and cannot eliminate them as long as the temporal parameters remain bounded.

ODE well-posedness and stability. Under the Lipschitz assumption, the Picard–Lindelöf theorem guarantees a unique ODE solution on every finite time interval. Moreover, Grönwall's inequality gives

$$\left\| \Phi_\theta^t(z_1^e) - \Phi_\theta^t(z_2^e) \right\| \leq e^{L_f t} \|z_1^e - z_2^e\|. \tag{23}$$

Hence, the ODE alignment map is stable with respect to perturbations in the EEG representation.

Convergence of the numerical alignment objective. Let $\Phi^t_{\theta,h}$ denote the RK4 approximation with step size h. Under the stated smoothness assumptions,

$$\sup_{i,t} \left\| \Phi^t_{\theta,h}(z^e_i) - \Phi^t_{\theta}(z^e_i) \right\| \le C_T h^4. \tag{24}$$

The Gaussian kernel is Lipschitz continuous in each argument, with

$$L_k = \sup_{x,y} \|\nabla_x k_\sigma(x,y)\| = \frac{e^{-1/2}}{\sigma}. \tag{25}$$

Let $\mathcal{L}_h(\theta, \mathbf{a})$ denote the alignment objective computed using the RK4 solutions. Then

$$|\mathcal{L}_h(\theta,\mathbf{a}) - \mathcal{L}(\theta,\mathbf{a})| \le \frac{4L_k C_T h^4}{C} \sum_{t=1}^{T} w_t \le \frac{4TL_k C_T}{C} h^4. \tag{26}$$

Consequently,

$$\mathcal{L}_h(\theta,\mathbf{a}) \longrightarrow \mathcal{L}(\theta,\mathbf{a}) \qquad \text{as } h \to 0. \tag{27}$$

Therefore, the absence of weight normalization changes only the constant in the error bound and does not change the fourth-order convergence of the RK4-based alignment objective.

Optimization convergence. Let

$$\boldsymbol{\vartheta} = (\theta, \mathbf{a}) \tag{28}$$

contain both the ODE parameters and the learnable temporal parameters. Consider the gradient update

$$\boldsymbol{\vartheta}_{r+1} = \boldsymbol{\vartheta}_r - \eta \nabla \mathcal{L}_h(\boldsymbol{\vartheta}_r), \qquad 0 < \eta < \frac{2}{L}. \tag{29}$$

By the descent lemma,

$$\mathcal{L}_h(\boldsymbol{\vartheta}_{r+1}) \le = \mathcal{L}_h(\boldsymbol{\vartheta}_r) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla \mathcal{L}_h(\boldsymbol{\vartheta}_r)\|^2 . \tag{30}$$

Since $\mathcal{L}_h \ge 0$, it follows that

$$\sum_{r=0}^{\infty} \|\nabla \mathcal{L}_h(\boldsymbol{\vartheta}_r)\|^2 < \infty, \tag{31}$$

and hence

$$\lim_{r\to\infty} \|\nabla \mathcal{L}_h(\boldsymbol{\vartheta}_r)\| = 0. \tag{32}$$

Furthermore, for the pure alignment objective,

$$\frac{\partial \mathcal{L}_h}{\partial a_t} = \frac{1}{C} \operatorname{sigmoid}(a_t) \left[1 - \operatorname{sigmoid}(a_t)\right] D_{h,t}(\theta). \tag{33}$$

Because a_t is bounded, the sigmoid derivative has a strictly positive lower bound on $[a_{\min}, a_{\max}]$. Therefore,

$$\left| \frac{\partial \mathcal{L}_h}{\partial a_t} \right| \longrightarrow 0 \quad \Longrightarrow \quad D_{h,t}(\theta) \longrightarrow 0. \tag{34}$$

Thus, bounded learnable temporal parameters cannot produce a trivial solution by suppressing all alignment weights. Under the stated assumptions, the ODE solution is unique and stable, the RK4-based weighted MMD objective converges to the exact objective with error $\mathcal{O}(h^4)$, and the optimization procedure converges to a first-order stationary point.

In the experiments, the raw temporal parameters decreased only from approximately 1.00 to 0.99, corresponding to effective sigmoid weights decreasing from approximately 0.7311 to 0.7291. This observation supports the bounded-parameter assumption and indicates that no temporal alignment term was suppressed during training.