CoLaP: Coupled Latent Prediction for a Synchronized EEG-fNIRS Foundation Model

Anonymous authors
Paper under double-blind review

Abstract

Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) capture the neuronal electrical activity and cortical hemodynamic changes of the same underlying brain activity, respectively. Existing brain signal foundation models rely almost exclusively on EEG as input, while synchronous EEG-fNIRS modeling methods are mostly designed for a single task on a fixed dataset, with the model architecture bound to the signal channel configuration and therefore difficult to adapt to datasets with different electrode and optode layouts. To address this problem, this paper proposes the Coupled Latent Prediction (CoLaP) foundation model. The pretraining of CoLaP consists of two objectives, intra-modality masked latent prediction and cross-modality temporal latent prediction. The former infers the latent representation of masked positions from visible context within each modality, and the latter predicts the latent representation of adjacent time windows on the cross-modality fused representation, so that the learned EEG-fNIRS representations preserve the structure of each modality while encoding the coupling between the two modalities. CoLaP is pretrained on 96,305 ten-second synchronized windows from 16 public datasets and evaluated on 7 tasks across 4 downstream datasets. The results show that CoLaP ranks among the top three on 6 of the 7 downstream tasks and outperforms all baselines on mental arithmetic, valence, and arousal recognition. Adaptation protocol experiments show that pretraining yields an average improvement of 7.74% over training from scratch. Modality ablation shows that the coupling information between EEG and fNIRS learned during pretraining transfers to different downstream tasks. For reproducibility, the code and pretrained weights are submitted anonymously with the manuscript.

1 Introduction

Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) have become the most widely used brain physiological signals owing to their non-invasiveness, portability, and controllable cost. EEG measures the electrical activity generated by the synchronous firing of neuronal populations with millisecond temporal resolution, whereas fNIRS measures the relative changes in cortical blood oxygen concentration on a time scale of seconds (Nazeer et al., 2025). The neurovascular coupling mechanism indicates that EEG and fNIRS can characterize the same neural activity from two complementary perspectives in terms of temporal scale and physiological mechanism. This complementarity has sustained the interest in joint EEG-fNIRS modeling across tasks such as emotion recognition, cognitive load assessment, motor imagery, and brain-computer interfaces.

Existing EEG-fNIRS modeling treats cross-modality fusion as the central research problem and designs networks for specific tasks, exemplified by EF-Net (Arif et al., 2024), E-FNet (Yu et al., 2025), TAFN (Bunterngchit et al., 2024), TSMMF (Si et al., 2025), and STA-Net (Liu et al., 2025). The architectures of these models are bound to their tasks, so changing the task or the subject usually requires retraining, and the resulting task-level representations are difficult to transfer. Different from these models, brain signal foundation models follow a paradigm of first pretraining a general encoder on large-scale unlabeled data and then transferring it to multiple downstream tasks. The pretraining objectives of existing models are dominated by masked reconstruction of raw waveforms or spectral details (He et al., 2022), as in LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025). In addition, some work builds on the joint-embedding predictive architecture (JEPA)

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.

to learn representations targeted at high-level semantics by predicting the latent representation of a momentum encoder, among which EEG-JEPA (Li et al., 2026b) is a representative example. The two pretraining families impose different constraints on the supervisory signal. Masked reconstruction requires recovering all details of the signal, a considerable portion of which originates from individual differences, noise, and artifacts that are irrelevant to the task, whereas JEPA only needs to approximate the latent representation of the target encoder, thereby discarding irrelevant noise and concentrating capacity on structures that remain stable across samples and tasks. Besides the pretraining objective, the input modality is also limited, as all the aforementioned models take EEG as input (Jiang et al., 2024; Wang et al., 2025; Yang et al., 2023; Döner et al., 2025; Li et al., 2026b) and lack any modeling of the coupling between EEG and fNIRS.

Taken together, large-scale pretrained models based on masked reconstruction and latent prediction have emerged for the EEG modality and can learn general representations from unlabeled corpora (Jiang et al., 2024; Wang et al., 2025; Li et al., 2026b), whereas EEG-fNIRS research remains at the level of task-specific, single-dataset, and supervised modeling, with no foundation model capable of exploiting both modalities. This observation leads to the core question of this paper: can we pretrain a synchronous EEG-fNIRS foundation model that learns general representations preserving the modality-specific structure while encoding the inter-modality coupling, and that transfers across datasets and tasks?

To address this question, this paper proposes the Coupled Latent Prediction (CoLaP) foundation model, which processes EEG and fNIRS with two modality-specific encoders and is pretrained in a self-supervised manner on synchronized unlabeled data to learn general EEG-fNIRS representations. The contributions of this paper are as follows.

- We propose CoLaP, which to the best of our knowledge is the first foundation model pretrained on multiple heterogeneous synchronized EEG-fNIRS datasets. CoLaP consists of modality-specific encoders, a mask predictor, and a temporal predictor, where the encoders and predictors are all agnostic to the channel count, so the same set of weights can be applied directly to datasets with different electrode and optode layouts.
- We design two self-supervised objectives, intra-modality masked latent prediction and cross-modality temporal latent prediction. The former infers the latent representation of masked positions from visible context within each modality, learning the modality-specific information. The latter predicts the latent representation of adjacent time windows on the fused representation produced by joint self-attention, learning the coupling information between the two modalities.
- We collect 16 public synchronous EEG-fNIRS datasets as the pretraining corpus, totaling 8,676.8 EEG channel-hours and 17,990 fNIRS channel-hours, and evaluate the model on 7 tasks across 4 downstream datasets.

2 METHOD

2.1 OVERVIEW

The pretraining framework of CoLaP is illustrated in Figure 1. Given synchronized EEG and fNIRS signals, the two streams are first divided into fixed-length time windows, where the k-th time window is temporally adjacent to the $(k+1)$-th window. Each window is then divided into P patches, and a positional embedding and a modality embedding are added to each patch to form the input to the encoder. The pretraining consists of two self-supervised objectives. The intra-modality masked latent prediction first masks a subset of patches in window k, encodes the masked input into latent representations through the modality encoder, and then predicts the latent representations of the masked positions from the visible context through the mask predictor. The cross-modality temporal latent prediction feeds the patches of the two modalities in window k into their corresponding modality encoders, applies joint self-attention to the encoded results for cross-modality interaction, and then predicts the latent representation of each modality in window $k+1$ through the temporal predictor.

Patching. Different datasets vary in sampling rate and channel count, so we divide the signal within each window into a fixed number of patches; the input length is then set by the patch count rather

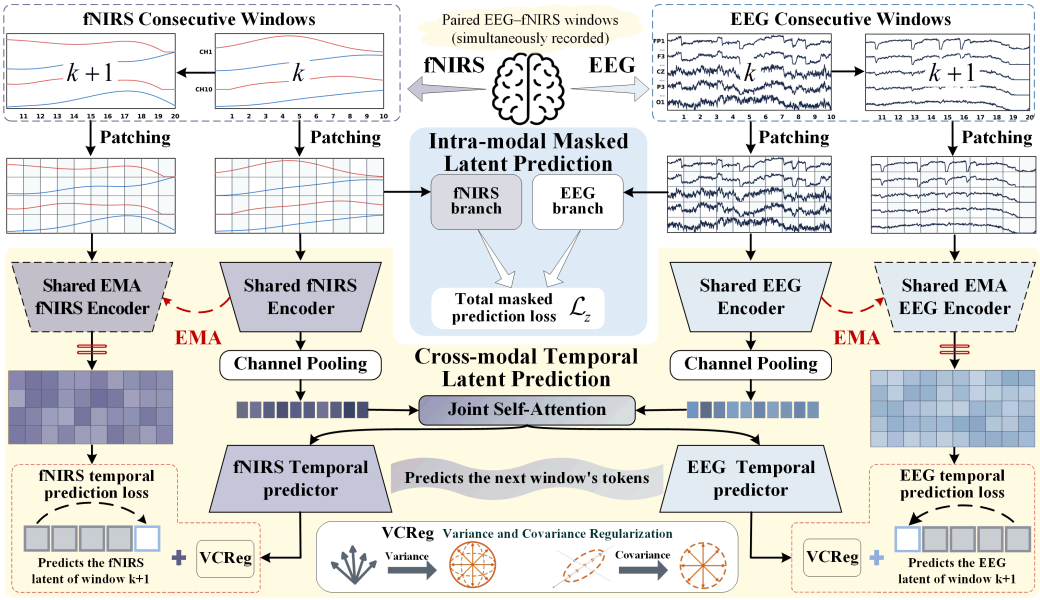


Figure 1: The pretraining framework of CoLaP.

than by the sampling rate, and all datasets share a common input format. Let the multi-channel signal of modality m in the k-th time window be $X_k^m \in \mathbb{R}^{(C_m \times T_m)}$, where $m \in \{E, F\}$ denotes EEG and fNIRS and C_m is the channel count. The signal is divided along the time axis into P patches of equal length $L_m = T_m/P$, and the p-th patch of the c-th channel is denoted as $x^m(c,p) \in \mathbb{R}^{(L_m)}$. Both modalities share the window length τ and the patch count P, so the p-th patch covers the same time segment in the two modalities.

Tokenization. The tokenizer determines the information that each patch carries into the encoder. The local characteristics of EEG manifest in both the waveform and the distribution of band energy, so an EEG patch is formed by the raw segment together with a band-energy vector; those of fNIRS manifest in the rate of change of the signal, so an fNIRS patch is formed by the raw segment together with its first-order and second-order differences. Both are projected to d dimensions by linear layers that are shared across channels, so the tokenizer contains no parameter that depends on the channel count, and its size depends only on the patch length, the number of frequency bands, and the latent dimension. The explicit forms of the two patch representations are given in Appendix K.

During pretraining, the patch representation at a masked position is replaced by a learnable mask vector while the remaining positions retain the tokenizer output. With the mask matrix $M^m \in \{0,1\}^{(C_m \times P)}$ and the mask vector $\omega^m \in \mathbb{R}^d$, the masked representation is $\tilde{z}^m(c,p) = (1 - M^m(c,p)) \cdot z^m(c,p) + M^m(c,p) \cdot \omega^m$.

Embeddings. To enable the model to distinguish different temporal positions and modalities, a positional embedding and a modality embedding are added to each patch, yielding $h^m(c,p) = \tilde{z}^m(c,p) + p_p + v^m$, where $p_p \in \mathbb{R}^d$ is the positional embedding of the p-th patch and $v^m \in \mathbb{R}^d$ is the modality embedding. The input to the encoder is denoted as $H^m \in \mathbb{R}^{(C_m \times P \times d)}$.

Encoder. The encoder of CoLaP is built on the criss-cross attention proposed by CCNet (Huang et al., 2019) and adopted in CBraMod (Wang et al., 2025). In CBraMod, this attention is combined with an asymmetric conditional position encoding (ACPE) and channel embeddings. We deliberately omit both components here, since they rely on a fixed electrode layout and would bind the encoder to a specific channel configuration. Since the two modalities differ in amplitude range, noise structure, and time scale, EEG and fNIRS use separate encoder weights.

Different channels at the same patch position exhibit spatial dependencies, and different patches on the same channel exhibit temporal dependencies. As the two kinds of dependencies differ in nature, the input to the l-th layer $H^{(l)} \in \mathbb{R}^{(C \times P \times d)}$ is first normalized and then split into two halves along the feature dimension, which are used for channel-axis attention and time-axis attention respectively. Both paths apply the same single-head scaled dot-product attention and differ only in the axis over

which it is computed, so we give their joint form:

$$A_* = \mathrm{softmax}\Big(Q_* K_*^\top / \sqrt{d_k}\Big), \qquad H_* = A_* V_* \tag{1}$$

where $* \in \{s, t\}$ denotes the channel-axis and the time-axis path, Q_*, K_*, and V_* are obtained by linear projection from the corresponding half, $A_s \in \mathbb{R}^{(C\times C)}$ and $A_t \in \mathbb{R}^{(P\times P)}$ are the attention weight matrices between channels and between patches, and d_k is the dimension of a single attention head. The outputs of the two paths are concatenated along the feature dimension, added to the input, and then passed through the feed-forward network:

$$H' = H^{(l)} + (H_s\,;\, H_t), \qquad H^{(l+1)} = H' + \mathrm{MLP}(\mathrm{LN}_2(H')) \tag{2}$$

After stacking multiple layers, the encoder outputs $E^m \in \mathbb{R}^{(C_m\times P\times d)}$. The channel-axis attention does not use the positional index of the channel, the encoder contains no parameters related to the channel count, and the patch representations at the input side are generated channel by channel, so the same set of weights can be applied to datasets with different electrode and optode layouts.

The same encoder performs two forward passes on window k. One pass takes as input a partially masked set of patches and produces the output for masked latent prediction, and the other takes the complete patches and produces the output for cross-modal temporal prediction. The two passes share the same parameters but backpropagate their gradients separately, and no intermediate result is passed between them, so the unmasked input does not leak into the prediction of the masked positions.

EMA Encoder. We introduce an exponential moving average (EMA) encoder for each modality. This EMA encoder shares the structure of the modality-specific encoder but is not trained directly. Instead, its weights are updated as the exponential moving average of the trainable encoder weights, $\bar{\theta} \leftarrow \beta \cdot \bar{\theta} + (1-\beta)\cdot\theta$, where θ denotes the parameters of the trainable encoder, $\bar{\theta}$ the parameters of the EMA encoder, and β the momentum coefficient. Because the EMA weights are updated slowly, the EMA encoder lags behind the trainable encoder, so both the masked latent prediction and the cross-window prediction are always supervised by relatively stable representations. The two EMA encoders receive complete, unmasked windows and do not participate in gradient computation. Their output for window k serves as the supervision of the masked positions, and their output for window $k+1$ serves as the supervision of the cross-window prediction.

2.2 Intra-modal masked latent prediction

The intra-modality masked latent prediction infers the latent representations of masked positions from the visible context without any labels. It uses only the information within a single modality and therefore provides an independent pretraining signal for each modality.

EEG and fNIRS Branches. The masked latent prediction is performed within a single modality, and EEG and fNIRS each have a branch of the same structure. As shown in Figure 2, each branch consists of three parts. The shared encoder encodes the masked window k into $\tilde{E}^m$, the mask predictor produces predictions at the masked positions, and the shared EMA encoder observes the unmasked signal of the same time window and provides the supervision target.

The masked latent prediction operates on every channel and patch position of the encoder output. The mask predictor is applied channel by channel, mapping the encoding of the c-th channel over its P patches, $\tilde{E}^m(c,\cdot)$, into P d-dimensional vectors, one for each patch position:

$$\hat{z}^m(c,\cdot) = g^m\Big(\tilde{E}^m(c,\cdot)\Big), \qquad \hat{z}^m(c,p) \in \mathbb{R}^d \tag{3}$$

where g^m is the mask predictor, composed of two self-attention layers and a linear output layer.

Masked Prediction Objective. Let r_m denote the masking ratio, which differs between the two modalities. EEG uses block masking, which divides the channel axis and the time axis into blocks and randomly masks some blocks in their entirety according to the ratio, while fNIRS uses random patch masking. The supervision signal is the output of the shared EMA encoder on the same time window. Let $\bar{E}^m(c,p) \in \mathbb{R}^d$ denote this output and $\hat{z}^m(c,p)$ is the output of the mask predictor at a

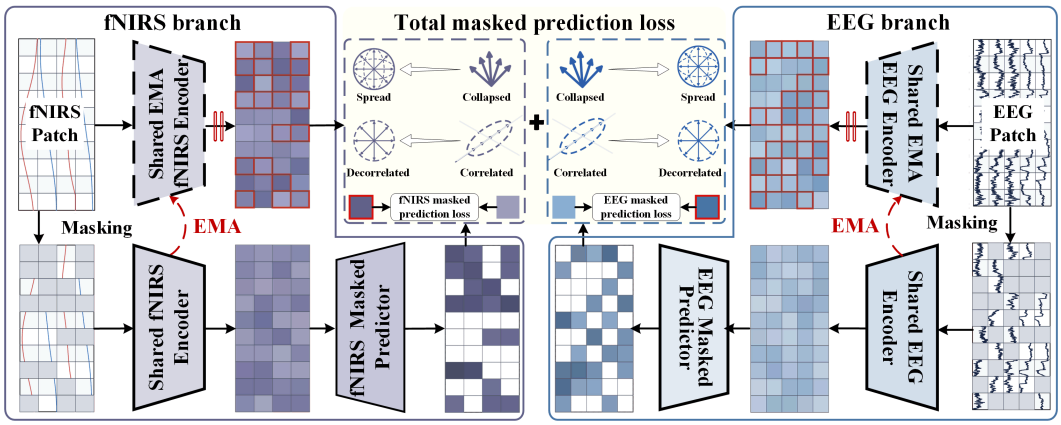


Figure 2: The architecture of the EEG and fNIRS branches.

masked position. The masked latent prediction objective is computed only at the masked positions.

$$L_z^m = \frac{\sum_{(c,p)} M^m(c,p)\, \ell\big(\hat{z}^m(c,p),\, \bar{E}^m(c,p)\big)}{\sum_{(c,p)} M^m(c,p)} \tag{4}$$

where $\ell(\cdot,\cdot)$ is the smooth L1 loss (Huber, 1964), M^m is the mask matrix, the denominator is the number of masked positions, and $\bar{E}^m$ does not participate in gradient computation.

The variance and covariance regularization terms (VCReg; Bardes et al., 2022) constrain the distribution of the representations during training:

$$L_{vc} = \lambda_v L_v + \lambda_c L_c \tag{5}$$

where L_v is the variance term and L_c is the covariance term. The variance term penalizes only the dimensions whose standard deviation falls below a given lower bound. Let $\hat{z}$ denote the set of prediction outputs at the masked positions, and compute the standard deviation of each dimension along the batch dimension:

$$L_v = \frac{1}{d} \sum_i \max(0,\, \gamma_v - \mathrm{std}_i(\hat{z})) \tag{6}$$

where γ_v is the prescribed lower bound of variance. The covariance term suppresses the correlation between different dimensions. Let $R \in \mathbb{R}^{(N \times d)}$ denote the flattened matrix of the representations participating in the regularization at any position, where each row corresponds to a token and N is the number of tokens at that position. Normalizing each dimension yields $\bar{R}$, and the correlation matrix is computed as $\mathrm{Cov} = \bar{R}^\top \bar{R}/N$:

$$L_c = \frac{\sum_{i \neq j} \mathrm{Cov}_{(i,j)}^2}{d(d-1)} \tag{7}$$

The variance term is applied to the prediction outputs at the masked positions, and the covariance term is applied to the representations participating in the prediction.

The intra-modality masked prediction objective consists of the masked prediction losses of the two modalities together with the variance and covariance terms:

$$L_z = \lambda_E L_z^E + \lambda_F L_z^F + L_{vc} \tag{8}$$

where λ_E and λ_F are the weights of the masked prediction losses of the two modalities, and λ_v and λ_c are the weights of the variance and covariance terms, respectively.

2.3 Cross-modal temporal latent prediction

The cross-modality temporal latent prediction lets the two modalities exchange information through joint self-attention and then predicts the latent representation of the adjacent window. Unlike the masked latent prediction, no masking is applied here.

Joint Self-Attention and Temporal Predictor. Cross-modality computation must be independent of the channel configuration, while the encoder output still retains the channel axis. The C_m representations at each patch are therefore aggregated into a d-dimensional token along the channel axis:

$$e^m(p) = W_p^m \left[\operatorname{mean}_c E^m(c,p) \,;\, \max_c E^m(c,p) \,;\, \operatorname{std}_c E^m(c,p) \right] \in \mathbb{R}^d \tag{9}$$

where mean_c, $\max_c$, and std_c are taken along the channel axis and $W_p^m \in \mathbb{R}^{(d\times 3d)}$ is a learnable projection. Each modality then yields P tokens, $S^m = [e^m(1), \ldots, e^m(P)] \in \mathbb{R}^{(P\times d)}$.

The two token sequences are concatenated in temporal order into $Z = [S^E; S^F] \in \mathbb{R}^{(2P\times d)}$ and fed into the joint self-attention (Vaswani et al., 2017), whose query, key, and value are linear projections of Z:

$$A = \operatorname{softmax}\Big(QK^\top/\sqrt{d_k}\Big) \in \mathbb{R}^{2P\times 2P}, \qquad O = AV \in \mathbb{R}^{2P\times d} \tag{10}$$

where d_k is the dimension of a single attention head. The tokens of the two modalities reside in the same matrix, so the cross-modality interaction is completed in a single computation, and the upper-right and lower-left blocks of A give the fNIRS-to-EEG and EEG-to-fNIRS weights. Let the first and the last P rows of O be o^E and o^F, which are fused with the respective token sequences through residual connections:

$$u^E = \operatorname{LN}\big(S^E + F_E\big(\big[\, S^E \,;\, o^E \,\big]\big)\big), \qquad u^F = \operatorname{LN}\big(S^F + F_F\big(\big[\, S^F \,;\, o^F \,\big]\big)\big) \tag{11}$$

where F_E and F_F are fusion networks of identical structure but independent parameters, each mapping a $2d$-dimensional input to d dimensions through two linear layers with a Gaussian error linear unit (GELU) (Hendrycks & Gimpel, 2016) activation between them. The two fused representations then pass through one shared temporal self-attention block Φ, which makes the different temporal positions visible to one another, and the temporal predictors produce the prediction of window $k+1$ position by position:

$$\hat{e}^E = g^E\big(\Phi(u^E)\big) \in \mathbb{R}^{P\times d}, \qquad \hat{e}^F = g^F\big(\Phi(u^F)\big) \in \mathbb{R}^{P\times d} \tag{12}$$

where each predictor g^m is composed of a position-wise feed-forward network and a temporal self-attention layer and keeps the sequence length unchanged.

Temporal Prediction Objective. The target of the cross-window prediction is the latent representation of window $k+1$ produced by the shared EMA encoder, denoted $\bar{e}^m(p) \in \mathbb{R}^d$. After the target is normalized, the difference is measured by the cosine distance and averaged over the P positions:

$$L_t^m = \frac{1}{P}\sum_p \left[\, 1 - \cos(\hat{e}^m(p),\, \operatorname{LN}(\bar{e}^m(p)))\right] \tag{13}$$

where $\bar{e}^m$ does not participate in gradient computation. The total objective adds three covariance terms to the cross-window prediction loss, and they act on the output of the temporal predictors, the output of the joint self-attention, and the fused representations, respectively:

$$L_t = \lambda_t \left(L_t^E + \gamma L_t^F\right) + \lambda_t\lambda_p L_c^e + \lambda_c \left(L_c^o + L_c^u\right) \tag{14}$$

where λ_t is the weight of the cross-window prediction loss, γ the ratio of the fNIRS loss to the EEG loss, and λ_p the ratio of the covariance term of the prediction outputs to the cross-window loss. L_c^e, L_c^o, and L_c^u are the covariance terms of the two predictors' outputs, the joint self-attention output, and the fused representations, computed as in the intra-modality objective, and the latter two share its coefficient λ_c.

2.4 Objective and Optimization

The total objective consists of the intra-modality masked prediction objective and the cross-modal temporal prediction objective:

$$L = L_z + \alpha L_t \tag{15}$$

where α is a scalar coefficient dynamically adjusted during training to control the proportion of the cross-modal temporal prediction objective in the total objective. It does not participate in gradient updates. At each training step, the proportion is computed from the current loss values as $\rho =$

Table 1: Results on the four downstream tasks of EEG-NIRS BCI and EEG-NIRS cognitive tasks (%). The first six rows are hybrid EEG-fNIRS models, the next four rows are EEG-only models, and the last row is a general biosignal model.

Method	MI		MA		NB		WG	
	BA	AUC-PR	BA	AUC-PR	BA	F1	BA	AUC-PR
EF-Net	52.78±2.39	53.63±2.54	63.54±1.21	70.74±3.79	33.02±2.24	32.37±2.18	58.61±6.86	61.54±5.55
EFRM	50.19±2.06	52.17±2.72	56.24±2.23	61.29±1.80	37.86±2.65	34.60±5.07	61.42±1.32	62.59±2.04
TSMMF	58.24±5.68	58.90±4.32	63.78±6.67	73.24±4.00	41.36±3.16	36.79±5.96	65.09±6.33	67.01±5.19
STA-Net	52.04±2.89	52.33±2.10	55.91±5.37	61.38±4.91	36.93±0.52	32.12±1.83	57.33±2.41	61.18±5.77
DeepSyncNet	49.07±1.59	49.64±1.38	65.93±4.86	70.71±3.53	40.33±2.14	38.10±3.97	58.47±2.46	57.84±4.32
TAFN	55.09±3.69	57.44±4.78	58.05±1.92	63.57±3.79	37.14±1.89	36.00±2.57	52.33±3.30	59.63±3.42
LaBraM	51.57±2.23	53.22±0.32	66.37±2.35	71.22±3.76	41.77±3.11	38.77±1.69	62.99±2.08	64.36±2.15
CBraMod	54.72±6.53	63.88±12.34	63.27±2.46	72.16±2.19	37.24±1.19	32.11±5.86	64.75±4.09	65.51±5.94
LUNA	49.81±0.47	51.82±0.29	50.56±0.79	57.52±3.33	36.93±3.01	31.02±1.43	54.95±4.31	58.48±3.69
EEG-JEPA	59.81±6.47	63.86±11.98	63.51±4.47	70.26±1.01	37.14±2.72	30.65±9.97	64.34±6.28	66.31±6.81
BIOT	50.09±0.35	50.70±1.34	54.15±2.76	57.71±4.14	37.76±3.08	31.18±6.64	55.47±3.18	58.43±2.94
CoLaP	56.76±6.19	60.03±8.72	66.93±4.19	73.81±2.72	38.27±2.15	29.89±5.91	65.00±3.51	70.62±4.48

$\alpha \cdot L_t/(L_z + \alpha \cdot L_t)$, and an exponential moving average is applied to ρ to suppress fluctuations between batches.

3 Experiments

3.1 Downstream Tasks

Datasets and Tasks. As summarized in Table 6 of Appendix D, the downstream evaluation involves 4 synchronized EEG-fNIRS datasets and 7 tasks, namely EEG-NIRS BCI (Shin et al., 2017), EEG-NIRS cognitive tasks (Shin et al., 2018b), REFED (Ning et al., 2025), and ENTER (Chen et al., 2024). The EEG-NIRS BCI dataset contains two binary classification tasks, motor imagery (MI) and mental arithmetic (MA). The EEG-NIRS cognitive tasks dataset contains two tasks, N-back (NB) and word generation (WG). The REFED dataset contains valence recognition and arousal recognition, two three-class tasks denoted as REFED-Valence and REFED-Arousal. ENTER is four-class emotion recognition.

Metrics. Binary tasks were evaluated with balanced accuracy (BA) and AUC-PR, and multi-class tasks with BA and F1; the definitions are given in Appendix E. The best epoch was selected on the validation set by AUC-PR for binary tasks and by BA for multi-class tasks, and the test results were averaged over 3 random seeds.

3.2 Comparison with EEG-fNIRS and EEG Baselines

This section compares CoLaP with two groups of baselines. The first group consists of six hybrid EEG-fNIRS models, including EF-Net (Arif et al., 2024), EFRM (Jung & An, 2025), TSMMF (Si et al., 2025), STA-Net (Liu et al., 2025), DeepSyncNet (Li, 2025), and TAFN (Bunterngchit et al., 2024). These models use the same synchronized input, the same windows, and the same subject splits as CoLaP, forming an input-matched controlled comparison. The second group consists of four EEG-only baselines and one general biosignal model, including LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), LUNA (Döner et al., 2025), EEG-JEPA (Li et al., 2026b), and BIOT (Yang et al., 2023).

EEG-NIRS BCI and Cognitive Tasks. Table 1 reports the results on the MI and MA tasks of EEG-NIRS BCI and the NB and WG tasks of EEG-NIRS cognitive tasks. On MA, CoLaP achieves the highest BA of 66.93%, higher than all hybrid and all EEG-only baselines, with LaBraM the closest at 66.37%, and the highest AUC-PR of 73.81%. On MI, CoLaP reaches a BA of 56.76%, lower only than TSMMF at 58.24% among the hybrid baselines and only than EEG-JEPA at 59.81% among the EEG-only baselines. NB is the only task on which CoLaP does not rank among the top three, with a BA of 38.27% against LaBraM at 41.77%, TSMMF at 41.36%, and DeepSyncNet at 40.33%, while the BA of all methods falls between 33% and 42%, not far from the chance level of 33.33% for three-class classification. On WG, CoLaP reaches a BA of 65.00%, higher than all EEG-only baselines, with only TSMMF slightly higher at 65.09%.

Table 2: Results on valence recognition, arousal recognition, and four-class emotion recognition in REFED and ENTER (%).

Method	Input	Valence		Arousal		ENTER	
		BA	F1	BA	F1	BA	F1
EF-Net	EEG+fNIRS	40.52 ± 0.78	34.25 ± 8.72	36.50 ± 3.15	33.04 ± 6.58	48.83 ± 5.46	47.79 ± 6.34
EFRM	EEG+fNIRS	44.60 ± 1.31	43.49 ± 1.16	35.09 ± 3.81	32.16 ± 6.96	46.72 ± 3.66	46.46 ± 3.19
TSMMF	EEG+fNIRS	41.28 ± 3.97	41.26 ± 6.90	38.45 ± 4.28	37.06 ± 3.85	48.33 ± 4.66	47.83 ± 4.42
STA-Net	EEG+fNIRS	43.06 ± 3.61	42.03 ± 4.12	41.67 ± 4.32	41.24 ± 2.90	44.50 ± 0.75	43.56 ± 1.11
DeepSyncNet	EEG+fNIRS	45.57 ± 1.31	43.40 ± 6.08	45.04 ± 1.28	44.80 ± 3.32	43.52 ± 12.23	42.89 ± 12.50
TAFN	EEG+fNIRS	43.31 ± 3.06	43.33 ± 3.16	33.80 ± 3.48	27.69 ± 7.49	38.05 ± 9.56	33.72 ± 16.45
LaBraM	EEG	44.94 ± 4.20	42.29 ± 9.37	41.69 ± 3.26	40.87 ± 3.52	57.97 ± 7.39	57.81 ± 7.36
CBraMod	EEG	42.00 ± 0.79	42.84 ± 1.36	44.49 ± 3.72	45.21 ± 1.47	52.23 ± 6.80	51.52 ± 7.05
LUNA	EEG	39.28 ± 0.82	35.83 ± 7.10	42.08 ± 2.80	40.59 ± 5.33	48.81 ± 3.56	48.14 ± 3.10
EEG-JEPA	EEG	37.96 ± 3.59	42.49 ± 1.04	42.63 ± 2.59	43.86 ± 0.47	46.61 ± 15.52	41.34 ± 21.96
BIOT	EEG+fNIRS	44.85 ± 2.21	45.86 ± 3.44	40.36 ± 1.08	37.82 ± 2.48	48.43 ± 6.08	48.62 ± 5.92
CoLaP	EEG+fNIRS	48.36 ± 0.51	50.13 ± 3.45	46.34 ± 2.44	45.80 ± 4.11	51.85 ± 4.61	51.21 ± 5.39

Table 3: Modality ablation of CoLaP (%).

Task	CoLaP		CoLaP-EEG		CoLaP-fNIRS	
	BA	AUC-PR / F1	BA	AUC-PR / F1	BA	AUC-PR / F1
MI	56.76 ± 6.19	60.03 ± 8.72	54.26 ± 2.43	61.15 ± 7.69	53.80 ± 1.97	54.66 ± 3.70
MA	66.93 ± 4.19	73.81 ± 2.72	56.31 ± 3.26	64.10 ± 4.90	59.95 ± 2.42	67.07 ± 1.75
NB	38.27 ± 2.15	29.89 ± 5.91	40.84 ± 4.43	33.50 ± 8.94	35.08 ± 4.39	33.92 ± 4.84
WG	65.00 ± 3.51	70.62 ± 4.48	58.43 ± 3.48	62.49 ± 5.01	55.85 ± 4.45	57.92 ± 5.19
REFED-Valence	48.36 ± 0.51	50.13 ± 3.45	44.13 ± 4.28	47.04 ± 3.40	46.43 ± 2.06	45.72 ± 2.19
REFED-Arousal	46.34 ± 2.44	45.80 ± 4.11	42.45 ± 2.27	43.46 ± 2.87	38.85 ± 3.65	37.35 ± 3.61
ENTER	51.85 ± 4.61	51.21 ± 5.39	49.02 ± 6.95	48.67 ± 7.03	35.62 ± 7.46	35.80 ± 7.02

REFED and ENTER. Table 2 summarizes the results on REFED-Valence, REFED-Arousal, and ENTER. On REFED-Valence and REFED-Arousal, CoLaP is the highest in both the hybrid and the EEG-only baseline groups, with BA of 48.36% and 46.34%, respectively, and its lead over the strongest method in each group ranges from 1.3% to 3.4%. On ENTER, CoLaP achieves a BA of 51.85%, higher than all hybrid baselines and 3.0% higher than the strongest of them, EF-Net, but lower than LaBraM at 57.97% and CBraMod at 52.23%, with a gap of 6.1% relative to LaBraM.

3.3 Modality Ablation

To examine the contribution of each modality, we pretrain two variants using only EEG and only fNIRS, respectively. CoLaP-EEG and CoLaP-fNIRS use exactly the same encoder, tokenization, pretraining budget, and downstream protocol as CoLaP, with only the branch and the loss of the other modality turned off, so the three differ only in the input modality. Table 3 reports the results of CoLaP and its two variants on the 7 downstream tasks.

It can be seen from Table 3 that the full CoLaP achieves the best result on 6 of the 7 tasks, indicating that the coupling of the two modalities indeed brings a gain rather than redundant information. Between the two single-modality variants, CoLaP-EEG outperforms CoLaP-fNIRS in overall recognition performance, which is consistent with the conclusion that EEG has higher temporal resolution and stronger discriminative information in most cognitive and motor tasks.

Two clear exceptions emerge task by task. On the NB task, CoLaP-EEG at 40.84% is higher than the full model at 38.27%, which is also the only case among the three inputs in which a single modality exceeds the dual modality, indicating that introducing fNIRS does not bring complementary information on this task. On the four-class emotion recognition of ENTER, CoLaP-fNIRS achieves only 35.62%, far below CoLaP-EEG at 49.02% and the full model at 51.85%, indicating that this task relies mainly on EEG. In contrast, on the MA and REFED-Valence tasks, CoLaP-fNIRS achieves BA of 59.95% and 46.43%, higher than CoLaP-EEG at 56.31% and 44.13%, indicating that fNIRS provides complementary information beyond EEG on these two tasks.

3.4 Adaptation Protocol Comparison

We compare four adaptation protocols, namely training from scratch, linear probing, nonlinear probing (MLP probing), and fine-tuning. Training from scratch and fine-tuning both update all parame-

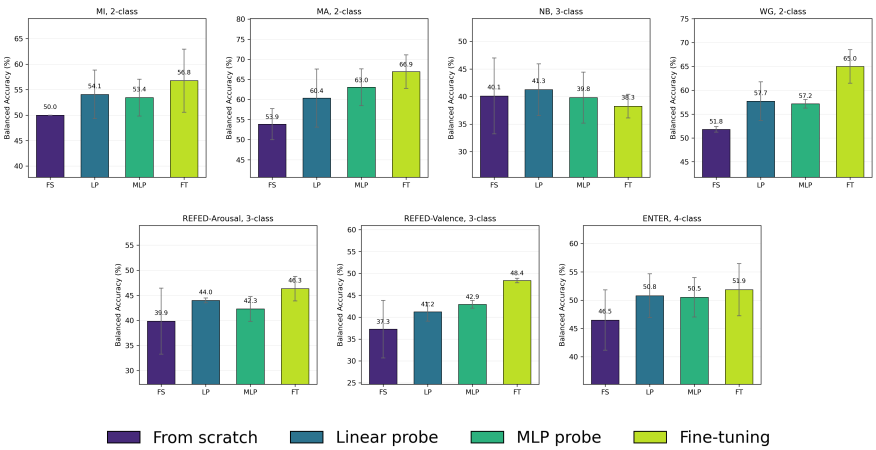


Figure 3: Results of CoLaP under the four adaptation protocols on the 7 tasks (%).

ters, and differ only in the initial weights, where the former uses random initialization and the latter loads the pretrained weights, so their difference is the net gain brought by pretraining. The latter two protocols freeze the encoder and the fusion module and train only the classification head, where linear probing uses a linear head and nonlinear probing uses a two-layer MLP head. The difference between them indicates whether the discriminative information in the representation is already linearly separable.

Figure 3 shows the BA of CoLaP on the 7 downstream tasks under different adaptation protocols. Fine-tuning achieves the highest result on 6 of the 7 tasks, and the improvements are most pronounced on WG, MA, and REFED-Valence. The two probing protocols with a frozen backbone also consistently outperform training from scratch, and linear probing is higher than training from scratch on all 7 tasks, indicating that the pretrained representation already contains a large amount of linearly separable information. After replacing the linear head with a two-layer MLP head, the recognition performance of CoLaP is roughly on par with linear probing, indicating that this information has been largely linearized and that deepening the classification head brings no additional benefit. In addition, fine-tuning on the NB task achieves only 38.27%, lower than training from scratch at 40.12% and linear probing at 41.26%, making NB the only task on which fine-tuning is not advantageous. We attribute this to the weak discriminative information of the task itself. On NB, the BA of all methods falls between 33% and 42%, close to the chance level of three-class classification, so the task signal is limited and the gradients are more easily dominated by noise under full-parameter fine-tuning, whereas the two probing protocols with a frozen backbone fit only a simple classification head on top of the representation and are therefore less affected by noise.

4 Conclusion

This paper proposes CoLaP, a foundation model for synchronized EEG-fNIRS. CoLaP processes EEG and fNIRS with two modality-specific encoders, and the encoders decompose attention along the channel axis and the time axis without depending on the number or the order of channels, so the same set of weights can be applied directly to datasets with different electrode and optode layouts. The pretraining uses no task labels. The intra-modality masked latent prediction learns the structure of each modality, and the cross-modality temporal latent prediction learns the coupling between the two modalities on the fused representation produced by joint self-attention. The pretraining corpus aggregates 16 public synchronized EEG-fNIRS datasets with 96,305 ten-second unlabeled windows. Among the 7 downstream tasks, CoLaP ranks among the top three on 6 tasks and further outperforms all compared baselines on the MA, REFED-Valence, and REFED-Arousal tasks. The modality ablation provides additional evidence of the information complementarity between EEG and fNIRS. We hope that this cross-modal coupling pretraining approach can offer some inspiration for building brain foundation models and thereby promote their application to more tasks and populations.

AI USE STATEMENT

Generative AI tools were used to assist with polishing the Chinese text and debugging part of the experimental code. The main model code, research questions, method design, experimental protocols, and all conclusions were produced and confirmed by the authors. The authors have reviewed all AI-assisted output and take full responsibility for the final content of this paper.

ETHICS STATEMENT

This study uses only publicly released synchronized EEG-fNIRS datasets and involves no new data collection or experiments on human subjects. These datasets were released together with statements that ethical approval had been obtained from the corresponding institutions and informed consent from the participants, and that the data had been de-identified. Their use in this paper follows the original license terms of each dataset. The goal of this work is general representation learning, and the tasks involve routine brain-computer interface scenarios such as motor imagery, cognitive load, and emotion recognition, and are not used for individual identity recognition or clinical diagnosis. The sources of the pretraining corpus are given in Appendix C, and the downstream datasets in Appendix D.

REPRODUCIBILITY STATEMENT

The hyperparameters for pretraining and downstream adaptation are listed in Appendix B, the per-dataset statistics of the pretraining data in Appendix C, the subject counts, channel counts, window counts, and task definitions of the downstream tasks in Appendix D, and the metric definitions and the implementation of each baseline in Appendix E. All datasets used are public. The windowing scheme, the 8:1:1 subject split rule, and the 3 random seeds (7, 42, 2026) are given in Appendix D, and all results are the mean and standard deviation over these 3 seeds. The code, the subject split script, and the pretrained weights are submitted with this paper as an anonymous archive and will be released in a public repository after acceptance.

REFERENCES

Z. Ajra, G. Vergotte, S. Perrey, L. Evra, S. Pla, G. Dray, J. Montmain, and B. Xu. Multimodal dataset from the CMx7-MM experiment. OpenNeuro ds007554, 2026.

A. Arif, Y. Wang, R. Yin, X. Zhang, and A. Helmy. EF-Net: Mental state recognition by analyzing multimodal EEG-fNIRS via CNN. *Sensors*, 24(6):1889, 2024.

A. Aygun, G. Blaney, Z. Haga, T. McWilliams, J. Mertens, J. P. de Ruiter, N. Ward, and M. Scheutz. A multimodal virtual reality data acquisition platform and dataset to assess systemic human cognitive states. *Scientific Data*, 13:76, 2026.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

A. P. Buccino, H. O. Keles, and A. Omurtag. Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks. *PLoS ONE*, 11(1):e0146610, 2016.

C. Bunterngchit, J. Wang, J. Su, et al. Temporal attention fusion network with custom loss function for EEG-fNIRS classification. *Journal of Neural Engineering*, 2024.

G. Chen, Y. Liu, and X. Zhang. EEG-fNIRS-based emotion recognition using graph convolution and capsule attention network. *Brain Sciences*, 14(8):820, 2024.

B. Döner, T. M. Ingolfsson, L. Benini, and Y. Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Y. El Ouahidi, J. Lys, P. Thölke, et al. 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 (NeurIPS)*, 2025.

L. Feng, B. Xu, H. Zhang, et al. A multimodal fNIRS–EEG dataset for unilateral limb motor imagery. *Scientific Data*, 13, 2026.

A. Hamann, R. J. Kneffel, D. Cyrol, et al. Examining physiological parameters of mental workload in the cockpit: A multimethod approach. *Frontiers in Physiology*, 2026.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). arXiv:1606.08415, 2016.

T. Hrnčírová, M. Augustynek, J. Kubicek, and J. A. Knocikova. Data for: Synchronized multimodal dataset for central–autonomic coupling in meditation and rest. Zenodo, 2025.

Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. CCNet: Criss-cross attention for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

E. Jung and J. An. EFRM: A multimodal EEG–fNIRS representation-learning model for few-shot brain-signal classification. *Computers in Biology and Medicine*, 199:111292, 2025.

S. Lee, H. R. Jung, I.-N. Wang, M.-K. Jung, H. Kim, and D.-J. Kim. Multimodal EEG and fNIRS biosignal acquisition during motor imagery tasks in patients with orthopedic impairment. OpenNeuro ds004022 / NEMAR on004022, 2022.

H. Li. DeepSyncNet: Deep synchronized fusion network for EEG-fNIRS multimodal brain-computer interfaces, 2025.

H. Li, X. Zhang, G. Chen, Y. Zhou, Z. Wu, L. Huang, and Y. Sun. Spatiotemporal alignment for EEG-fNIRS emotion recognition. *IEEE Transactions on Affective Computing*, 2026a.

J. Li, Z. Ma, X. Han, et al. EEG-JEPA: Structured latent prediction for EEG foundation models. arXiv, 2026b.

M. Liu et al. STA-Net: Spatial–temporal alignment network for hybrid EEG-fNIRS decoding. *Information Fusion*, 119:103023, 2025.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

A. Medvedeva, N. Syrov, L. Yakovlev, et al. Multisession fNIRS-EEG data of post-stroke motor recovery: Recordings during intact and paretic hand movements. *Scientific Data*, 2026.

MER 2026 challenge. MER-PS: Second-by-second valence and arousal annotation. CodaBench competition 15992, 2026. URL `https://www.codabench.org/competitions/15992/`.

H. Nazeer, N. Naseer, and M. J. Khan. Noninvasive brain-computer interfaces using fNIRS, EEG, and hybrid EEG-fNIRS. In *Brain-Computer Interfaces*. 2025.

X. Ning, J. Wang, Z. Feng, et al. REFED: A subject real-time dynamic labeled EEG-fNIRS synchronized recorded emotion dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

T. Phukhachee, T. Angsuwatanakul, K. Iramina, and B. Kaewkamnerdpong. A simultaneous EEG-fNIRS dataset of the visual cognitive motivation study in healthy adults. *Data in Brief*, 53:110260, 2024.

M. Rybář, R. Poli, and I. Daly. Simultaneous EEG and fNIRS recordings for semantic decoding of imagined animals and tools. *Scientific Data*, 12, 2025.

J. Shin, A. von Lühmann, B. Blankertz, D.-W. Kim, J. Jeong, H.-J. Hwang, and K.-R. Müller. Open access dataset for EEG+NIRS single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2017.

J. Shin, K.-R. Müller, and H.-J. Hwang. Eyes-closed hybrid brain-computer interface employing frontal brain activation. *PLoS ONE*, 13(5):e0196359, 2018a.

J. Shin, A. von Lühmann, B. Blankertz, et al. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, 5:180003, 2018b.

X. Si, S. Zhang, Z. Yang, et al. A bidirectional cross-modal transformer representation learning model for EEG-fNIRS multimodal fusion. *Expert Systems with Applications*, 2025.

K. A. Usal and M. P. Çakır. Modulation of interbrain and intrabrain connectivity due to social presence and task difficulty: A dual EEG/fNIRS study. *European Journal of Neuroscience*, 61 (7):e70091, 2025.

A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

J. Wang, S. Zhao, H. Jiang, et al. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025.

S. Watanabe, T. Suzuki, and Y. Ono. Predicting the intensity of the flow state using EEG and fNIRS biomarkers. *Advanced Biomedical Engineering*, 13:335–342, 2024.

Q. Xiao, Z. Cui, C. Zhang, et al. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

C. Yang, M. B. Westover, and J. Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

W. Yi, J. Chen, D. Wang, X. Hu, et al. A multi-modal dataset of electroencephalography and functional near-infrared spectroscopy recordings for motor imagery of multi-types of joints from unilateral upper limb. *Scientific Data*, 12:953, 2025.

B. Yu, L. Cao, J. Jia, C. Fan, Y. Dong, and C. Zhu. E-FNet: An EEG-fNIRS dual-stream model for brain-computer interfaces. *Biomedical Signal Processing and Control*, 100:106943, 2025.

H. Zhang, Y. Hu, Y. Li, S. Zhang, X. Li, and C. Zhao. Simultaneous dataset of brain, eye and hand during visuomotor tasks. *Scientific Data*, 12:189, 2025.

A Related Work

A.1 EEG Foundation Models

EEG modeling has evolved from handcrafted features and task-specific networks to foundation models. The former relies on domain priors to extract features such as band power and differential entropy, and designs the network architecture for a single task, whereas the latter first pretrains a general encoder on large-scale unlabeled EEG signals in a self-supervised manner and then transfers it to various downstream tasks through linear probing or fine-tuning. According to the pretraining objective, existing methods can be roughly divided into two categories. Masked reconstruction is currently the mainstream (He et al., 2022). For example, LaBraM (Jiang et al., 2024) discretizes channel patches into neural codes and predicts the masked tokens, CBraMod (Wang et al., 2025) models spatial and temporal dependencies separately with cross-attention and an asymmetric conditional positional encoding, REVE (El Ouahidi et al., 2025) introduces four-dimensional positional encoding to accommodate arbitrary lengths and electrode layouts, LUNA (Döner et al., 2025) compresses arbitrary channel sets into a fixed latent space through learnable queries so that the computational cost grows linearly with the channel count, BrainOmni (Xiao et al., 2025) pretrains with temporal masked reconstruction on EEG and magnetoencephalography (MEG) and maps input channels to the three-dimensional coordinates of standard electrodes through a sensor encoder to adapt to different channel layouts, and BIOT (Yang et al., 2023) uses a unified tokenized representation to handle channel mismatch, variable length, and missing values. In contrast, fewer works perform prediction in the latent space. EEG-JEPA (Li et al., 2026b) uses a momentum encoder to provide the prediction target, avoiding direct fitting of the background noise and artifacts in the raw waveform. The two families share the same boundary on the input modality, as all the aforementioned models take single-modality EEG as input and none of them models fNIRS or characterizes the relationship between EEG and fNIRS under synchronized acquisition.

A.2 Hybrid EEG-fNIRS Research

Joint EEG-fNIRS modeling was initially dominated by shallow fusion, in which the two modalities extract frequency-domain or time-domain features separately and then are combined through concatenation or weighted fusion. Deep learning moves the fusion into the network, and the structures that emerged afterwards differ mostly in where and how the fusion is performed. EF-Net (Arif et al., 2024) uses convolutional networks to learn the representations of EEG and fNIRS separately and evaluates on a word-generation task under a subject-independent setting. E-FNet (Yu et al., 2025) adopts a two-stream structure in which each modality passes through its own extraction pathway before fusion. TAFN (Bunterngchit et al., 2024) extracts temporal features from each modality using temporal convolution, temporal attention, and long short-term memory (LSTM), and then introduces an asymmetric loss and temporal smoothing regularization on a binary classification task. DeepSyncNet (Li, 2025) maps the two modalities onto the same spatial grid, performs gated fusion and cross-modality attention at multiple stages, and finally adds the fused output and the EEG output with learnable weights. These models merge the two modalities at the feature level. Other works design the interaction between modalities separately. For example, TSMMF (Si et al., 2025) takes raw EEG and HbO as input and performs spatiotemporal fusion with a bidirectional cross-modality Transformer, in which cross-attention allows one modality to access the information of the other during encoding. STA-Net (Liu et al., 2025) maps the two modalities onto a two-dimensional grid,

extracts spatiotemporal features with three-dimensional convolution, and selects channels and time points through feature-guided spatial attention and EEG-guided temporal attention. Some works model the coordination among channels with graph structures rather than encoding each channel independently (Chen et al., 2024). EFRM (Jung & An, 2025) pushes fusion to the representation level by first pretraining the individual representations of the two modalities and a shared representation, and then transferring to downstream tasks with a small number of labeled samples. Beyond how to fuse, the temporal correspondence between the two modalities also needs to be handled. The hemodynamic response lags the neural electrical activity by several seconds, and this delay varies across subjects and tasks, so most works perform an alignment before fusion, either approximating it with a fixed temporal offset or estimating the frame-by-frame correspondence with attention mechanisms and learnable pointers. For instance, ENSTAN (Li et al., 2026a) uses lagged cross-attention to let EEG actively search for the corresponding response frames in fNIRS. Such alignment is designed for specific acquisition configurations and tasks, and requires readjustment once the channel layout or the magnitude of the delay changes.

B Hyperparameter Settings

This appendix provides all hyperparameter settings used in pretraining and downstream adaptation. The pretraining hyperparameters are listed in Table 4, and the settings shared by the four downstream adaptation protocols are listed in Table 5. Values not listed in the tables follow the default implementation described in Section 2. At each pretraining step, one batch of paired windows was drawn from a single dataset, and the momentum coefficient of the EMA encoder was increased linearly from 0.999 to 0.9999 over the first 500 steps and then kept constant, so that the target representation follows the encoder more closely in the early stage of training and becomes stable later.

C Pretraining Dataset Description

The pretraining corpus comes from 16 public synchronized EEG-fNIRS datasets, covering tasks including motor imagery and control, cognitive load, emotion and semantic processing, and meditation. All signals were resampled to 200 Hz for EEG and 10 Hz for fNIRS and divided into non-overlapping ten-second windows, yielding 96,305 windows in total, each composed of synchronously acquired EEG and fNIRS signals. The task labels provided by each dataset were not used during pretraining. Each window was further divided into patches of one second, and the two modalities were divided in the same way, so that patches with the same index are aligned in time. The four downstream datasets used for evaluation do not overlap with these 16 datasets, and none of the subjects involved in evaluation appears in the pretraining data. The corpus amounts to 8,676.8 EEG channel-hours and 17,990 fNIRS channel-hours. Across all datasets, the number of subjects ranges from 7 to 48, the number of EEG channels from 6 to 66, and the number of fNIRS channels from 16 to 140.

- FJMI (Yi et al., 2025). Multi-joint motor imagery of the unilateral upper limb, with 8 movement classes involving the hand, wrist, elbow, and shoulder. 18 healthy subjects, 66 EEG channels and 48 fNIRS channels, 2,070.6 EEG channel-hours and 1,505.9 fNIRS channel-hours.
- Hybrid SMR BCI (Buccino et al., 2016). Sensorimotor-rhythm motor imagery of 4 classes, right arm, left arm, right hand, and left hand, with 15 subjects, 21 EEG channels and 68 fNIRS channels, 122.0 EEG channel-hours and 395.2 fNIRS channel-hours.
- MIND (Feng et al., 2026). Motor imagery of the unilateral upper limb in four directions, 25 subjects, 42 EEG channels and 102 fNIRS channels, 252.2 EEG channel-hours and 612.6 fNIRS channel-hours.
- Orthopedic MI (Lee et al., 2022). 7 patients with upper-limb orthopedic impairment performing four motor imagery classes, reach, grasp, lift, and twist, 17 EEG channels and 24 fNIRS channels, 74.8 EEG channel-hours and 105.6 fNIRS channel-hours.
- Stroke recovery (Medvedeva et al., 2026). 16 stroke patients performing movements with the healthy hand and the affected hand in response to visual cues during rehabilitation, 11 EEG channels and 140 fNIRS channels, 206.8 EEG channel-hours and 2,631.6 fNIRS channel-hours.

Table 4: Hyperparameter settings for CoLaP pretraining.

Hyperparameter	Value
Data and input	
Number of pretraining datasets	16 (96,305 ten-second windows)
Sampling rate	EEG 200 Hz; fNIRS 10 Hz
Patch partitioning	$P = 10$, one second per patch (200 EEG points, 10 fNIRS frames)
Masking ratio	EEG 0.7 (block masking), fNIRS 0.5 (per-patch random masking)
Data augmentation	Gaussian noise strength 0.1, channel dropout probability 0.1
Network architecture	
Latent dimension d	384
Encoder	12 criss-cross layers, 8 attention heads, feed-forward expansion ratio 4.0
Channel pooling	Concatenation of mean, maximum, and standard deviation followed by a linear projection
Fusion stream	One temporal self-attention block after joint self-attention
Attention heads in joint self-attention and temporal predictor	8
Pretraining objective	
Masked prediction loss weights	$\lambda_E = 1.0$, $\lambda_F = 0.3$
Temporal prediction loss weights	$\lambda_t = 0.5$, $\gamma = 1.0$, $\lambda_p = 0.5$
VCReg weights	$\lambda_c = 0.5$ (shared by all four terms), $\lambda_v = 0.2$, variance lower bound $\gamma_v = 0.5$
Covariance sampling limit	4,096 tokens
Cross-modal loss proportion	target 0.35, upper bound 0.40, updated every 50 steps; scaling coefficient clipped to 0.05–3.0
Cross-modal gradient scaling and output zeroing	EEG 0.5, fNIRS 1.0, cross-modal output zeroing probability 0.15
Optimization	
Optimizer and learning rate	AdamW, 1×10^{-4}
Weight decay	0.05
Batch size	64
Training budget	80 epochs $\times$ 810 steps
EMA momentum	linearly increased from 0.999 to 0.9999 over the first 500 steps
Gradient clipping threshold	5.0
Mixed precision and gradient checkpointing	mixed precision enabled, gradient checkpointing enabled when the channel count is at least 48

- Visuomotor (Zhang et al., 2025). A joystick target-tracking task alternating with rest, 48 subjects, 33 EEG channels and 88 fNIRS channels, 2,577.5 EEG channel-hours and 6,873.3 fNIRS channel-hours.

- Eyes-closed hybrid BCI (Shin et al., 2018a). Mental arithmetic and rest in the eyes-closed state, 12 subjects, 23 EEG channels and 18 fNIRS channels, 128.8 EEG channel-hours and 100.8 fNIRS channel-hours.

- Pilot workload (Hamann et al., 2026). An N-back cognitive load task during flight simulation, 13 pilots, 22 EEG channels and 50 fNIRS channels, 123.6 EEG channel-hours and 281.0 fNIRS channel-hours.

- CMx7-MM (Ajra et al., 2026). Each subject sequentially performed 7 task types including motor execution, motor imagery, working memory, mental arithmetic, and their combinations, 30 subjects, 32 EEG channels and 32 fNIRS channels, 955.3 EEG channel-hours and 955.3 fNIRS channel-hours.

- Hyperscanning (Usal & Çakır, 2025). Dyadic cooperative N-back with difficulty covering 0-back to 3-back, 29 subject pairs recorded simultaneously, 20 EEG channels and 32 fNIRS channels, 237.4 EEG channel-hours and 379.9 fNIRS channel-hours.

Table 5: Hyperparameter settings for the downstream adaptation protocols.

Hyperparameter	Value
Data and split	
Windowing	same as pretraining (ten-second windows, EEG 200 Hz, fNIRS 10 Hz)
Subject split	8:1:1 by subject, repeated over 3 random seeds (7, 42, 2026)
Adaptation protocol	
Training from scratch	all parameters randomly initialized
Linear probing	backbone frozen, only the linear classification head trained
Nonlinear probing	backbone frozen, only a two-layer MLP classification head trained
Fine-tuning	all parameters initialized from pretrained weights
Optimization	
Optimizer and learning rate	AdamW, 1×10^{-4}, cosine annealing
Weight decay	0.05
Batch size and epochs	64 / 30
Per-layer learning rate decay	not used
Label smoothing	0.1
Data augmentation	Gaussian noise strength 0.1, channel dropout probability 0.1
Gradient clipping threshold	5.0
Model selection	AUC-PR (binary classification) or balanced accuracy (multi-class classification) on the validation set

Table 6: Overview of the downstream datasets and tasks.

Dataset	Subjects	Channels (EEG/fNIRS)	Windows	Task	Classes
EEG-NIRS BCI	29	30 / 72	3,483	MA	2-class
			3,480	MI	2-class
EEG-NIRS cognitive tasks	26	30 / 72	4,212	NB	3-class
			3,616	WG	2-class
REFED	32	64 / 102	4,768	REFED-Valence	3-class
			4,768	REFED-Arousal	3-class
ENTER	50	62 / 36	23,579	ENTER	4-class

- Semantic decoding (Rybář et al., 2025). Silent naming of animal and tool images, supplemented with visual, auditory, and tactile imagery, 12 subjects, 55 EEG channels and 28 fNIRS channels, 550.2 EEG channel-hours and 280.1 fNIRS channel-hours.

- MER-PS (MER 2026 challenge, 2026). Second-by-second valence and arousal annotation under emotional video elicitation, 24 subjects, 64 EEG channels and 102 fNIRS channels, 652.6 EEG channel-hours and 1,040.1 fNIRS channel-hours.

- Meditation and rest (Hrnčírová et al., 2025). Guided breathing meditation and rest, continuously recorded without trial structure, 8 subjects, 16 EEG channels and 16 fNIRS channels, 58.8 EEG channel-hours and 58.8 fNIRS channel-hours.

- Tetris flow (Watanabe et al., 2024). 13 subjects completing a Tetris game at hard, optimal, and easy difficulty levels, 6 EEG channels and 66 fNIRS channels, 27.9 EEG channel-hours and 307.4 fNIRS channel-hours.

- VR driving (Aygun et al., 2026). Virtual-reality driving with a superimposed secondary task, 37 subjects, 8 EEG channels and 42 fNIRS channels, 226.0 EEG channel-hours and 1,186.3 fNIRS channel-hours.

- Visual cognitive motivation (Phukhachee et al., 2024). Deciding whether to remember a scene after viewing it, followed by a recognition test, 16 subjects, 31 EEG channels and 96 fNIRS channels, 412.2 EEG channel-hours and 1,276.5 fNIRS channel-hours.

Table 7: Subject splits of the four downstream datasets under the 3 random seeds. Subject indices follow the order in which each dataset was released, and the training set consists of the remaining subjects.

Dataset	Random seed	Validation subjects	Test subjects
EEG-NIRS BCI	7	4, 20, 23	5, 16, 26
	42	8, 11, 21	7, 15, 20
	2026	14, 25, 26	2, 7, 27
EEG-NIRS cognitive tasks	7	4, 20, 23	5, 16
	42	8, 11, 15	7, 20
	2026	14, 21, 24	2, 7
REFED	7	4, 20, 23	5, 16, 26
	42	8, 11, 15	7, 20, 29
	2026	14, 25, 28	2, 7, 27
ENTER	7	15, 24, 29, 40, 47	4, 5, 20, 26, 48
	42	11, 19, 21, 23	8, 15, 29, 39, 43
	2026	6, 13, 29, 30, 46	2, 7, 14, 27, 49

D More Details for Downstream Tasks

D.1 Tasks and Datasets

The four datasets provide 7 tasks in total. The subject count, channel count, window count, and task definition of each dataset are given below.

EEG-NIRS BCI (Shin et al., 2017). 29 subjects, 30 EEG channels and 72 fNIRS channels. Two binary classification tasks, left-right hand motor imagery and mental arithmetic, with 3,480 windows and 3,483 windows, respectively.

EEG-NIRS cognitive tasks (Shin et al., 2018b). 26 subjects, 30 EEG channels and 72 fNIRS channels. Two tasks, working memory, which distinguishes the three load levels 0-back, 2-back, and 3-back and contains 4,212 windows, and language generation, which contrasts generation with rest and contains 3,616 windows.

REFED (Ning et al., 2025). 32 subjects, 64 EEG channels and 102 fNIRS channels. Valence and arousal each form a three-class classification task, with 4,768 windows per task.

ENTER (Chen et al., 2024). 50 subjects, 62 EEG channels and 36 fNIRS channels. Four-class emotion recognition with 23,579 windows.

D.2 Preprocessing and Split

All tasks share the same window definition as pretraining. EEG was resampled to 200 Hz and fNIRS to 10 Hz, and the signals were divided into non-overlapping ten-second windows within a single recording session of the same subject. The split was performed by subject, so that the windows of a given subject appear in only one of the training, validation, or test subsets, with an approximate ratio of 8:1:1. Evaluation was repeated over 3 random seeds (7, 42, 2026), the subjects were re-partitioned under each seed, and the results were averaged over the 3 seeds. The subject splits of the four datasets are listed in Table 7, and the remaining subjects form the training set.

E Baselines and Metrics

This appendix describes the implementation of the baselines and the evaluation metrics.

E.1 Baselines

The baselines are described in two groups. The first six are hybrid EEG-fNIRS models, followed by the EEG foundation models, and finally the general biosignal model BIOT. Unless otherwise stated, every method uses the same ten-second windows and subject splits as CoLaP.

- EF-Net (Arif et al., 2024) learns representations of synchronized EEG and fNIRS with convolutional networks.
- EFRM (Jung & An, 2025) first learns the individual representations of the two modalities and a shared representation through pretraining, and then transfers to downstream tasks in the few-shot setting.
- TSMMF (Si et al., 2025) takes raw EEG and HbO as input, performs spatiotemporal fusion with a bidirectional cross-modal Transformer, and exchanges information between the two modalities through cross-attention.
- STA-Net (Liu et al., 2025) maps the two modalities onto a two-dimensional grid, extracts spatiotemporal features with three-dimensional convolution, and combines feature-guided spatial attention with EEG-guided temporal attention.
- TAFN (Bunterngchit et al., 2024) extracts temporal features from each modality with temporal convolution, temporal attention, and LSTM, concatenates them for fusion, and applies an asymmetric loss and temporal smoothing regularization on binary classification tasks.
- DeepSyncNet (Li, 2025) maps the two modalities onto the same spatial grid, obtains three outputs, EEG, fNIRS, and fused, through RFB, gated early fusion, FAM, and STA, and then combines the fused output and the EEG output with learnable weights. The original version has about 153M parameters. The evaluation in this paper uses a lightweight configuration of this implementation, retaining the gated early fusion and the three-output structure, adjusting the pooling kernel according to the input duration, and shrinking the classification head to about 17M parameters.
- LaBraM (Jiang et al., 2024) uses vector-quantized neural spectrum prediction as the pretraining task and models masked patches with a vision Transformer (ViT) encoder.
- CBraMod (Wang et al., 2025) models channel and temporal dependencies separately with criss-cross attention and supplements positional information with an asymmetric conditional position encoding.
- LUNA (Döner et al., 2025) compresses arbitrary channel sets into a fixed latent space through learnable queries, so the computational cost grows linearly with the channel count.
- EEG-JEPA (Li et al., 2026b) uses a momentum encoder to provide the prediction target and performs prediction in the latent space rather than on the raw waveform.
- BIOT (Yang et al., 2023) is a general biosignal model evaluated on EEG, electrocardiography, and human activity sensing data, and is not EEG-specific. It divides multi-channel signals into sentence-like token sequences and handles channel mismatch, variable length, and missing values. Its pretrained version accepts at most 18 EEG channels, and channels beyond this limit are fed in groups by channel.

E.2 METRICS

The evaluation metrics are defined as follows. Let TP_k, FP_k, and FN_k denote the true positives, false positives, and false negatives of the k-th class, K the number of classes, n_k the number of samples of the k-th class, and N the total number of samples.

$$\mathrm{BA} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{16}$$

$$\text{AUC-PR} = \sum_{n=1}^{N_p} (R_n - R_{n-1})\, P_n \tag{17}$$

$$\mathrm{F1} = \sum_{k=1}^{K} \frac{n_k}{N} \cdot \frac{2 P_k R_k}{P_k + R_k} \tag{18}$$

where BA is the arithmetic mean of the recall of each class and still reflects the recognition performance of every class under class imbalance. AUC-PR is the area under the precision-recall curve,

Table 8: The unfreezing curve from linear probing to full-parameter fine-tuning (%).

Task	Head only	Last 4 layers	All parameters
MI	54.07 ± 4.78	54.17 ± 4.03	56.76 ± 6.19
MA	60.35 ± 7.22	62.99 ± 2.67	66.93 ± 4.19
NB	41.26 ± 4.68	39.71 ± 1.08	38.27 ± 2.15
WG	57.72 ± 4.03	63.03 ± 1.85	65.00 ± 3.51
REFED-Valence	41.21 ± 2.11	48.99 ± 2.24	48.36 ± 0.51
REFED-Arousal	44.00 ± 0.49	45.08 ± 1.00	46.34 ± 2.44
ENTER	50.79 ± 3.87	49.23 ± 3.15	51.85 ± 4.61
Trainable parameters	0.20 M	29.74 M	53.41 M

where N_p is the number of threshold points and P_n and R_n are the precision and recall at the n-th point after sorting by threshold; it is used only for binary classification tasks. F1 is weighted by the number of samples of each class, where $P_k = TP_k/(TP_k + FP_k)$ and $R_k = TP_k/(TP_k + FN_k)$, and it is used only for multi-class classification tasks.

F Partial Fine-tuning

This appendix presents the unfreezing curve from linear probing to full-parameter fine-tuning, in order to explore how large an unfreezing scope is sufficient to fully exploit the pretrained representation. Table 8 reports the BA of the 7 tasks under three fine-tuning strategies, which are, in order, training only the classification head, unfreezing the last 4 layers, and unfreezing all parameters. The three settings share the same downstream hyperparameters and the same readout, and differ only in the number of trainable layers in the encoder. The last row gives the number of trainable parameters of each setting.

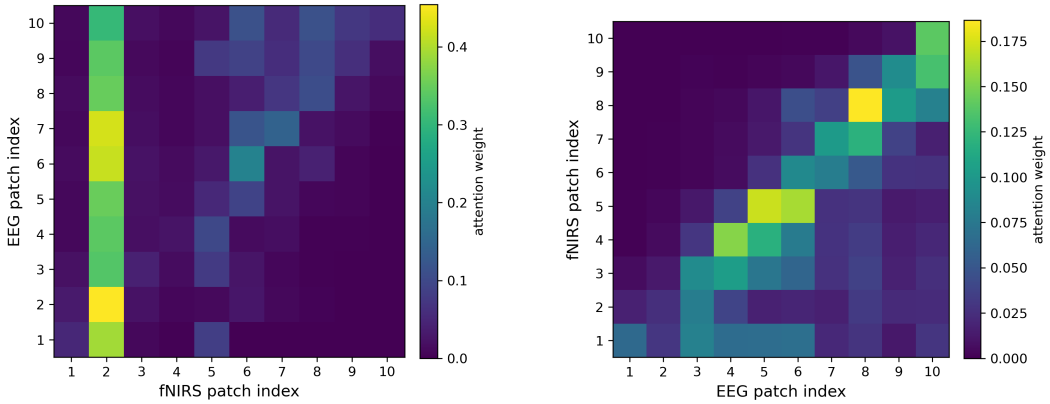
Table 8 reports the comparison of the three fine-tuning strategies. Linear probing trains only the classification head, partial fine-tuning unfreezes the last 4 layers of the encoder, and full-parameter fine-tuning releases all parameters. The number of trainable parameters of the three settings is 0.20 M, 29.74 M, and 53.41 M, respectively. Most tasks improve as the unfreezing scope expands, but the gain diminishes. REFED-Valence and WG are already close to the full-parameter setting when only the last 4 layers are unfrozen, reaching 48.99% and 63.03%, which differ from the full-parameter results of 48.36% and 65.00% by 0.63% and 1.97%, respectively, indicating that the discriminative information required by these two tasks is concentrated in the later layers of the encoder. MA, ENTER, and MI behave differently, as their improvement over training only the classification head becomes evident only when all parameters are unfrozen, among which MA rises from 62.99% with the last 4 layers unfrozen to 66.93% with all parameters unfrozen. NB is the only exception. As the unfreezing scope expands, the results of the three settings are 41.26%, 39.71%, and 38.27%, declining step by step, which is consistent with the observation in Section 3.4 that the BA of full-parameter fine-tuning is lower than that of training from scratch and linear probing.

G Scaling the Pretraining Budget

This appendix reports the influence of the pretraining budget on downstream performance. All 7 tasks were evaluated over 3 random seeds, and Table 9 shows the BA of each task at 16, 40, and 80 epochs. The two tasks of EEG-NIRS BCI behave differently. MI is nearly unchanged under the three budgets, at 56.48%, 56.76%, and 56.76%, while MA rises continuously with the budget, from 59.62% through 62.39% to 66.93%. The EEG-NIRS cognitive tasks are likewise inconsistent internally. WG shows the largest improvement, rising from 53.10% to 65.00%, whereas NB moves in the opposite direction, falling from 41.15% to 38.27%. The two tasks of REFED do not benefit from longer pretraining. REFED-Valence reaches its highest value of 51.67% at 16 epochs and then falls to 48.36%, and REFED-Arousal is highest at 40 epochs, at 46.84%. The three results of ENTER are nearly flat, differing by about 1%. Among the four datasets, only EEG-NIRS BCI and EEG-NIRS cognitive tasks each have one task that keeps improving with the budget, whereas REFED and ENTER are already close to their best level at 16 or 40 epochs.

Table 9: Results under different pretraining budgets (%).

Task	16 epochs	40 epochs	80 epochs
MI	56.48 ± 5.99	56.76 ± 5.67	56.76 ± 6.19
MA	59.62 ± 4.49	62.39 ± 5.49	66.93 ± 4.19
NB	41.15 ± 4.78	40.74 ± 3.25	38.27 ± 2.15
WG	53.10 ± 1.46	56.64 ± 1.72	65.00 ± 3.51
REFED-Valence	51.67 ± 1.70	48.62 ± 1.45	48.36 ± 0.51
REFED-Arousal	44.55 ± 1.88	46.84 ± 2.05	46.34 ± 2.44
ENTER	52.90 ± 6.19	51.98 ± 5.03	51.85 ± 4.61



(a) EEG→fNIRS (b) fNIRS→EEG

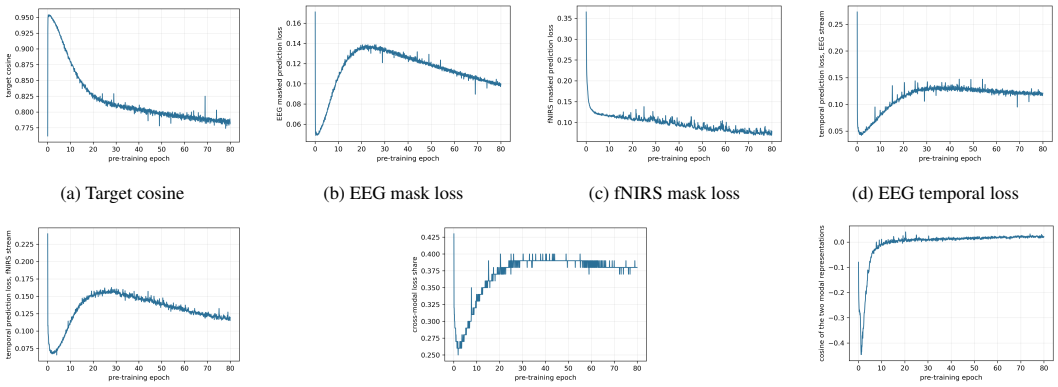
Figure 4: The cross-modal coupling matrices.

H Cross-modal Coupling Analysis

Figure 4 shows two cross-modal attention matrices of CoLaP on the test set of the language generation task, where Figure 4a is the EEG-to-fNIRS direction and Figure 4b the fNIRS-to-EEG direction. In the EEG-to-fNIRS direction, the attention is highly concentrated, and each EEG patch aggregates only a few fNIRS patches. All 10 EEG patches attain significant coupling with the second fNIRS patch. The fNIRS-to-EEG direction does not show the same prominent concentrated structure, as the attention of each fNIRS patch is distributed fairly evenly over the 10 EEG patches, with the highest weight of only 0.19.

I Pretraining Process Diagnostics

Figure 5 shows the diagnostic curves of the pretraining process. The horizontal axis of each subfigure is the pretraining epoch, and the vertical axis is the corresponding diagnostic quantity. Figure 5(a) shows the target cosine, the cosine similarity between the prediction output at the masked positions and the EMA target. This value rises to 0.885 around the 10th epoch, then declines slowly to 0.785 at the end and remains stable, indicating that the masked latent prediction learns the ability to infer the latent representation from the visible context early in training, while the slight decline in the later stage stems from the continuous update of the EMA target. Figure 5(b) shows the masked prediction loss of EEG, and Figure 5(c) the masked prediction loss of fNIRS. Both drop rapidly within ten epochs, EEG from 0.171 to 0.100 and fNIRS from 0.365 to 0.076, indicating that the encoders enter a stable latent space quickly. Figure 5(d) shows the temporal prediction loss on the EEG coupling stream, and Figure 5(e) that on the fNIRS coupling stream. Both curves fall from about 0.25 to 0.08 and 0.11 and then rebound and stabilize near 0.12. This fall-then-rise pattern arises from the target moving continuously during training, indicating that the predictors track the target rather than diverge. Figure 5(f) shows the proportion of the cross-modal loss in the total loss. This proportion falls back after reaching 0.43 and then remains within the target range, indicating that the relative weights of the two objectives are fairly stable. Figure 5(g) shows the cosine similarity between the pooled representations of the two modalities. This value stays near zero throughout, indicating that the two encoders do not collapse into the same representation. Taken together, both



(a) Target cosine (b) EEG mask loss (c) fNIRS mask loss (d) EEG temporal loss

(e) fNIRS temporal loss (f) cross-modal loss proportion (g) inter-modality similarity

Figure 5: Training diagnostic curves of the pretraining process.

objectives converge to stable plateaus, the two modalities maintain independent representations, and the cross-modal interaction is established early in training and remains stable afterwards, indicating a healthy pretraining process.

J Discussion

Synchronously acquired EEG and fNIRS have previously been used mainly for supervised fusion on a single dataset. Once the acquisition protocol or the channel layout changes, both the network structure and the weights need to be readjusted, and the same model can hardly be reused across data from different sources. After aggregating the synchronized signals of 16 datasets into a pretraining corpus, CoLaP learns representations transferable across datasets without any task labels, applies the same set of weights directly to downstream data with different electrode and optode layouts, ranks among the top three on 6 of the 7 tasks, and outperforms all compared methods on MA, REFED-Valence, and REFED-Arousal. The modality ablation further shows that the full model outperforms either single-modality variant on 6 of the 7 tasks, and that on MA and REFED-Valence, fNIRS alone is even higher than EEG alone, so the information carried by the two modalities does not overlap. The adaptation protocols and the unfreezing curve in Appendix F show that the pretrained representation has a low adaptation cost on downstream tasks. Both linear probing and nonlinear probing outperform training from scratch, full-parameter fine-tuning improves further on this basis, most discriminative information is already linearly separable, and WG and REFED-Valence are close to full-parameter fine-tuning when only the last 4 layers are unfrozen. The cross-modal attention in Appendix H and the pretraining diagnostics in Appendix I provide process-level evidence, as the attention in the EEG-to-fNIRS direction is highly concentrated on the WG task, and the two modalities maintain independent representations during pretraining.

The current limitation of this research direction comes mainly from the scale of the corpus. Public data capable of synchronously recording EEG and fNIRS amount to about 267.5 hours in total, which is one to two orders of magnitude smaller than the corpora used by single-modality EEG foundation models, so it is not yet possible to judge whether the representational capacity continues to improve with the amount of data. Future work will center on the data corpus. The synchronized corpus is still the main factor determining the upper bound of the representation, so continuing to collect synchronized EEG-fNIRS data across tasks and acquisition devices, and incorporating physiological and behavioral signals into pretraining, can test whether this representation yields sustained gains as the data grow. The readout problem caused by the channel layout needs to be solved separately. Injecting the position information of electrodes and optodes explicitly into the downstream head, or adopting parameter-efficient adaptation, is more likely to narrow the gap with single-modality models than further expanding the pretraining budget. In addition, the evaluation tasks should be extended to clinical and non-healthy populations and to online decoding scenarios, to examine the robustness of the pretrained representation on patient data and under real-time conditions.

K TOKENIZATION AND ENCODER DETAILS

K.1 TOKENIZATION

Let $b^E(c,p) \in \mathbb{R}^B$ denote the band energy of the p-th patch of the c-th EEG channel, and let x_l^F, x_v^F, and $x_a^F \in \mathbb{R}^{(L_F)}$ denote the raw segment of the same fNIRS segment and its first-order and second-order differences. The patch representations of the two modalities are

$$z^E(c,p) = W_f^E \left[W_r^E x^E(c,p) \,;\, W_s^E b^E(c,p) \right] \tag{19}$$

$$z^F(c,p) = W_f^F \left[W_l^F x_l^F(c,p) \,;\, W_v^F x_v^F(c,p) \,;\, W_a^F x_a^F(c,p) \right] \tag{20}$$

where W denotes a linear projection matrix, $[\cdot]$ denotes concatenation along the feature dimension, and the output dimension is d.

K.2 ATTENTION DETAILS

Let the input to the l-th layer be $H^{(l)} \in \mathbb{R}^{(C\times P\times d)}$ and its normalized form be $\hat{H} = \mathrm{LN}(H^{(l)})$, which is split along the feature dimension into $\hat{H}_s \in \mathbb{R}^{(C\times P\times d/2)}$ and $\hat{H}_t \in \mathbb{R}^{(C\times P\times d/2)}$. Channel-axis attention is computed across the C channels at the same patch position, and for the p-th patch position

$$A_s(p) = \mathrm{softmax}\left(\frac{Q_s(p) K_s(p)^\top}{\sqrt{d_k}} \right), \qquad H_s(p) = A_s(p) V_s(p) \tag{21}$$

where $Q_s(p)$, $K_s(p)$, and $V_s(p) \in \mathbb{R}^{(C\times d_k)}$ are obtained by linear projection from the C channels of $\hat{H}_s$ at that position, and $A_s(p) \in \mathbb{R}^{(C\times C)}$ is the attention weight matrix between channels. Time-axis attention is computed across the P patches on the same channel, and for the c-th channel

$$A_t(c) = \mathrm{softmax}\left(\frac{Q_t(c) K_t(c)^\top}{\sqrt{d_k}} \right), \qquad H_t(c) = A_t(c) V_t(c) \tag{22}$$

where $Q_t(c)$, $K_t(c)$, and $V_t(c) \in \mathbb{R}^{(P\times d_k)}$ are obtained by linear projection from the P patches of that channel in $\hat{H}_t$, and $A_t(c) \in \mathbb{R}^{(P\times P)}$ is the attention weight matrix between patches.