Disentangled Multimodal Spatiotemporal Learning for Hybrid EEG-fNIRS Brain-Computer Interface

Yun Xu, Chi-Man Vong, *Senior Member*, *IEEE*, Zihao Xu, Jianlin Fu, Junhua Li, *Senior Member*, *IEEE*, and Chuangquan Chen, *Member*, *IEEE*

***Abstract*—The hybrid EEG-fNIRS Brain-computer interface (BCI) combines the high temporal resolution of electroencephalography (EEG) with the high spatial resolution of functional near-infrared spectroscopy (fNIRS) to enable comprehensive brain activity detection. However, integrating these modalities to obtain highly discriminative features remains challenging. Most existing methods fail to effectively capture the spatiotemporal coupling features and correlations between EEG and fNIRS signals. Furthermore, these methods adopt a holistic learning paradigm for the representation of each modality, leading to unrefined and redundant multimodal representations. To address these challenges, we propose a disentangled multimodal spatiotemporal learning (DMSL) method for hybrid EEG-fNIRS BCI systems, which simultaneously performs multimodal spatiotemporal coupling and disentangled representation learning within a unified framework. Specifically, DMSL utilizes a compact convolutional module with one-dimensional temporal and spatial convolution layers to extract complex spatiotemporal patterns from each modality and introduces a multimodal attention interaction module to comprehensively capture the inter-modality correlations, enhancing the representations for each modality. Subsequently, DMSL designs an adaptive multi-branch graph convolutional module based on reconstructed channels to effectively capture the spatiotemporal coupling features, incorporating modality consistency and disparity constraints to disentangle common and modality-specific representations for each modality. These disentangled representations are finally adaptively fused to perform different task predictions. The proposed DMSL demonstrates state-of-the-art performance on publicly available datasets for mental arithmetic, motor imagery, and emotion recognition tasks, exceeding the best baselines by 2.34%, 0.59%, and 1.47%, respectively. These results demonstrate the effectiveness of DMSL in improving EEG-fNIRS decoding and its strong generalization ability in BCI applications.**

***Index Terms*—BCI, EEG, fNIRS, multimodal representation learning.**

I. Introduction

BRAIN-computer interfaces (BCIs)[1] facilitate human-computer interaction by establishing direct communication between external devices and the brain, bypassing peripheral muscles. BCIs have found applications in a wide range of fields, including robot control, workload detection, and emotion recognition, playing a crucial role in helping patients with limb disability or other neuromuscular degeneration [2], [3], [4].

BCIs utilize a variety of modalities, including stereoelectroencephalography (SEEG), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and functional near-infrared spectroscopy (fNIRS). Among these modalities, EEG and fNIRS have become prominent in both research and practical applications [5], [6], [7], [8], [9] due to their non-intrusive nature and relative ease of operation. Combining these two modalities enables a more comprehensive understanding of brain activity. EEG records brain activity via scalp electrodes [10], has good temporal resolution but limited spatial precision, and is susceptible to artifacts and noise [11]. In contrast, fNIRS measures cerebral blood flow and metabolism [12], [13], has better spatial resolution and less noise than EEG, but poor temporal resolution. Since near-infrared light does not interfere with electrical signals, the synchronous measurement of EEG and fNIRS has become increasingly popular [14]. Studies have shown that hybrid EEG-fNIRS BCI systems achieve improved classification accuracy compared to single-modality BCI systems by employing appropriate fusion methods [15].

Despite progress in EEG-fNIRS BCI research achieved through traditional machine learning [16], [17], [18], [19], [20], [21], [22] and deep learning methods [23], [24], two core challenges persist. First, current models still struggle to effectively capture the spatiotemporal coupling features and cross-modal correlations between EEG and fNIRS signals.

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011978, and in part by the Hong Kong and Macau Joint Research and Development Fund of Wuyi University under Grant 2021WGALH19. (Corresponding author: *Chuangquan Chen*).

Yun Xu, Zihao Xu, Jianlin Fu, and Chuangquan Chen are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China. (e-mail: chenchuangquan87@163.com).

Chi-Man Vong is with the Department of Computer and Information Science, University of Macau, Taipa 999078, Macao. (e-mail: cmvong@um.edu.mo).

Junhua Li is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: junhua.li@essex.ac.uk).

Second, they treat the representation of each modality in a holistic learning manner without decomposing the features into distinct components (e.g., common and modality-specific representations). This oversight often results in noise or redundancy due to the inherent heterogeneity across modalities, which ultimately leads to suboptimal and unrefined multimodal representations.

To address the abovementioned challenges, we propose a novel disentangled multimodal spatiotemporal learning (DMSL) method for hybrid EEG-fNIRS BCI systems. DMSL is designed to simultaneously capture multimodal spatiotemporal coupling features and disentangle modality-specific and common representations, thus providing more refined and effective multimodal representations. Notably, DMSL adopts a hybrid fusion method that synergistically integrates the inherent capability of early fusion in modeling inter-modality relationships with the distinctive strength of late fusion in preserving intra-modality information. In contrast to previous methods [25], [26], which apply disentangled representation learning to raw modalities, we are the first to apply this technique to enhanced modalities—those enriched with cross-modality information. This innovative approach enables richer and more precise feature extraction, as demonstrated by our experimental results.

The main contributions of our work can be summarized as follows:

1) We propose a novel multimodal learning framework that simultaneously performs multimodal spatiotemporal coupling and disentangled representation learning within a unified structure.
2) We design a multimodal attention module that uses one modality to iteratively extract complementary features from the hybrid modality, integrating them via residuals to improve robustness.
3) We develop an adaptive graph convolution module with multi-branch mapping and attention fusion to capture complex spatiotemporal patterns.
4) Experiments demonstrate that DMSL achieves state-of-the-art performance on mental arithmetic, motor imagery, and emotion recognition tasks.

The remainder of this article is organized as follows. Section II presents a review of the related work. Section III briefly introduces the adopted datasets for evaluation. Section VI presents the proposed DMSL in detail. The experimental results and analysis are discussed in Section V. Finally, the conclusions are summarized in Section VI.

II. Related work

Given the complexity of EEG and fNIRS signals—which originate from multiple brain regions and exhibit temporal fluctuations—researchers have developed various approaches to harness their spatiotemporal information [16], [17]. Traditional machine learning methods, such as Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), are widely adopted in hybrid EEG-fNIRS BCI systems. These methods primarily rely on feature-level or decision-level fusion, where manually extracted features from EEG and fNIRS, or their respective decision scores, are combined to boost classification performance [18], [19], [20], [21], [22]. However, their heavy reliance on manual feature engineering limits their ability to effectively capture the intricate spatiotemporal dynamics inherent in EEG and fNIRS signals.

The rise of deep learning has driven significant breakthroughs in hybrid EEG-fNIRS systems. Many current deep learning methods employ late fusion, where features extracted from EEG and fNIRS are merged in the later stages [23], [24]. Yet, late fusion often fails to capture the inherent temporal and spatial correlations between the two modalities, resulting in suboptimal classification performance. As a result, recent efforts have shifted toward early fusion techniques, which integrate modalities at an earlier stage (e.g., raw signals or low-level features) to more effectively model inter-modality relationships [27], [28], [29]. Notably, methods like [28] and [29] address the alignment challenge posed by EEG and fNIRS (e.g., differing temporal resolutions and recording locations) using interpolation-based approaches. Unfortunately, this introduces additional noise, increases model complexity, and may ultimately degrade overall performance.

Recent progress in spatiotemporal and multimodal methodologies has further advanced this field. For example, STA-Net [30] focused on spatial-temporal alignment to improve hybrid EEG-fNIRS decoding; Bunterngchit et al. [31] introduced selective channel representation combined with spectrogram imaging for simultaneous EEG-fNIRS classification; and ASTDF-net [32] demonstrated the value of attention-based spatial-temporal dual-stream fusion for EEG-only emotion recognition—providing valuable insights for multimodal learning design. These advances have promoted the development of EEG-fNIRS BCIs, but several critical challenges remain. STA-Net [30] fails to capture deep spatiotemporal coupling features, Bunterngchit et al. [31] adopt a holistic learning approach without feature decomposition that introduces redundancy, and ASTDF-net [32], which only targets EEG, cannot model inter-modal correlations.

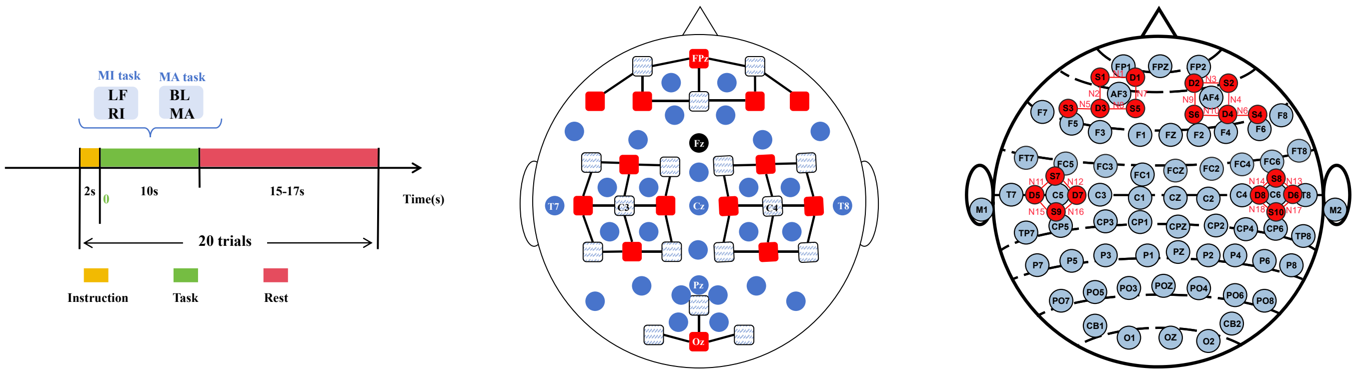
III. Datasets

This study adopted the 2017 Berlin Open Dataset HBCI [20] and the ENTER dataset [33] to evaluate our model.

(a) HBCI Dataset

The HBCI dataset includes 29 subjects (28.5±3.7 years). The experimental paradigm is presented in Figure 1 (a). Every subject completed 6 sessions, including 3 motor imagery (MI) sessions and 3 mental arithmetic (MA) sessions, and each session was composed of 20 trials. Each trial began with a 2-second visual introduction of the task, then there was a 10 s task period and concluded with a randomly allocated rest period of 15 to 17 seconds.

EEG and fNIRS data were collected simultaneously, with sampling frequencies of 1000 Hz for EEG and 12.5 Hz for fNIRS. The electrode positions are illustrated in Figure 1 (b). EEG signals were recorded from 30 channels, while fNIRS signals



(a) (b) (c)

Fig. 1. (a) The paradigm of the experiment, in which MI task includes left-hand motor imagery (LF) and right-hand motor imagery (RI), and MA task includes baseline (BL) and mental arithmetic (MA). (b) The positions of the EEG electrodes (blue and black dots) and fNIRS channels (black lines) in the MA and MI datasets. (c) The positions of EEG electrodes (blue dots) and fNIRS channels (red lines) in the ENTER dataset.

were collected from 36 channels, consisting of 14 transmitters and 16 receivers.

We split the data into two datasets based on the completed tasks: the MA dataset (baseline and mental arithmetic) and the MI dataset (left-hand motor imagery and right-hand motor imagery). We downsampled the EEG to 200Hz and the fNIRS to 10Hz, then divided both signals into 3-second segments starting from the task onset, with a 1-second time step. Thus, the sizes of the EEG signals and fNIRS signals were 30 channels×600 times and 72 channels×30 times, respectively, where the fNIRS data were stacked by HbO and HbR in the channel dimension, and there were a total of 600 samples (10 segments×30 trials×2 tasks) for each subject.

(b) ENTER Dataset

The ENTER dataset, collected and organized by researchers from Taiyuan University of Technology (TYUT), is an EEG-fNIRS dataset for emotion recognition. The channel distributions are shown in Figure 1 (c). It utilizes 64 EEG channels and 18 fNIRS channels, and was collected from 50 college students (25 males, aged 22.92±1.71; 25 females, aged 24.12±1.67) while they watched emotional video clips. The EEG data were recorded at a sampling rate of 1000 Hz with 64 channels, and the fNIRS data were recorded at a sampling rate of 11 Hz with 18 channels. The task is a four-class emotion recognition: sad, happy, calm and fear.

Similarly, we downsampled the EEG to 200Hz and the fNIRS to 10Hz, then divided both signals into 3-second segments starting from the task onset, with a 1-second time step. Thus, the sizes of the EEG signals and fNIRS signals were 62 channels×600 times and 36 channels × 30 times, respectively, where the fNIRS data were stacked by HbO and HbR in the channel dimension, and we extract a portion of the data, with a total of 600 samples (10 segments×60 trials) for each subject.

IV. Methods

A. Overview

Figure 2 depicts the overall framework of the proposed DMSL. The DMSL consists of four modules: 1) a channel reconstruction module that generates rich spatiotemporal patterns for each modality; 2) a multimodal attention module that comprehensively captures inter-modality correlations and enhances the representations for each modality; 3) a multi-branch graph convolutional module with modality consistency and disparity constraints for disentangled representation learning and effective spatiotemporal coupling feature capture; and 4) a classification module that adaptively fuses the disentangled representations and performs task predictions.

B. Channel Reconstruction Module

The variation in cognitive processes in the brain is reflected in the activation levels across different timestamps and brain regions. Taking motor imagery as an example, there are different stable patterns when imagining the left or right hand at specific time nodes or brain regions [34]. To effectively explore intricate cognitive patterns, we design the feature extraction module by decomposing the 2D convolution operation into two 1D layers.

Assume $\boldsymbol{E}^{(0)}$ is the input from the EEG, and $\boldsymbol{F}^{(0)}$ is the input from the fNIRS, where $\boldsymbol{E}^{(0)} \in \mathbb{R}^{C_e \times S_e}$, $\boldsymbol{F}^{(0)} \in \mathbb{R}^{2C_f \times S_f}$. Here, $C_{(\cdot)}$ is the number of channels, and $S_{(\cdot)}$ is the number of sampling points. We reconstruct channels by adding a dimension of depth, with each reconstructed channel getting different information from the original inputs.

For the EEG sub-model, as outlined in Table I, the specific formulas of the two convolution layers are as follows:

$$\boldsymbol{E}_1^k(i,j) = \sum_{t=0}^{T-1} \boldsymbol{W}_1^k(1,j)\boldsymbol{E}^{(0)}(i,j+t) + \boldsymbol{b}_1^k, \qquad (1)$$

$$\boldsymbol{E}_2^l(1,j) = \sum_{k=1}^{K}\sum_{i=1}^{C_e} \boldsymbol{W}_2^{l,k}(i,1)\boldsymbol{E}_1^k(i,j) + \boldsymbol{b}_2^l, \qquad (2)$$

$$i = 1,2,3,\dots,C_e; j = 1,2,3,\dots,S_e - T + 1;$$
$$k,l = 1,2,3,\dots,K.$$

Here, $\boldsymbol{W}_1^k \in \mathbb{R}^{1\times T}$ denotes the one-dimensional filter used for temporal convolution (with $T = 25$ as in [34]). $\boldsymbol{W}_2^{l,k} \in \mathbb{R}^{C_e \times 1}$ represents the one-dimensional filter for spatial convolution across the C_e channels. $\boldsymbol{b}_1^k$ and $\boldsymbol{b}_2^l$ are the bias terms for each respective layer. The output of the first layer is denoted as $\boldsymbol{E}_1=\{\boldsymbol{E}_1^k\}_{k=1}^K \in \mathbb{R}^{K\times C_e \times (S_e - T + 1)}$, and the final output is given by $\boldsymbol{E}_2 = \{\boldsymbol{E}_2^l\}_{l=1}^K \in \mathbb{R}^{K\times 1 \times (S_e - T + 1)}$.

The first two layers focus on time dimension and electrode channel interactions, respectively, followed by batch normalization and the use of ELUs for nonlinearity. The third

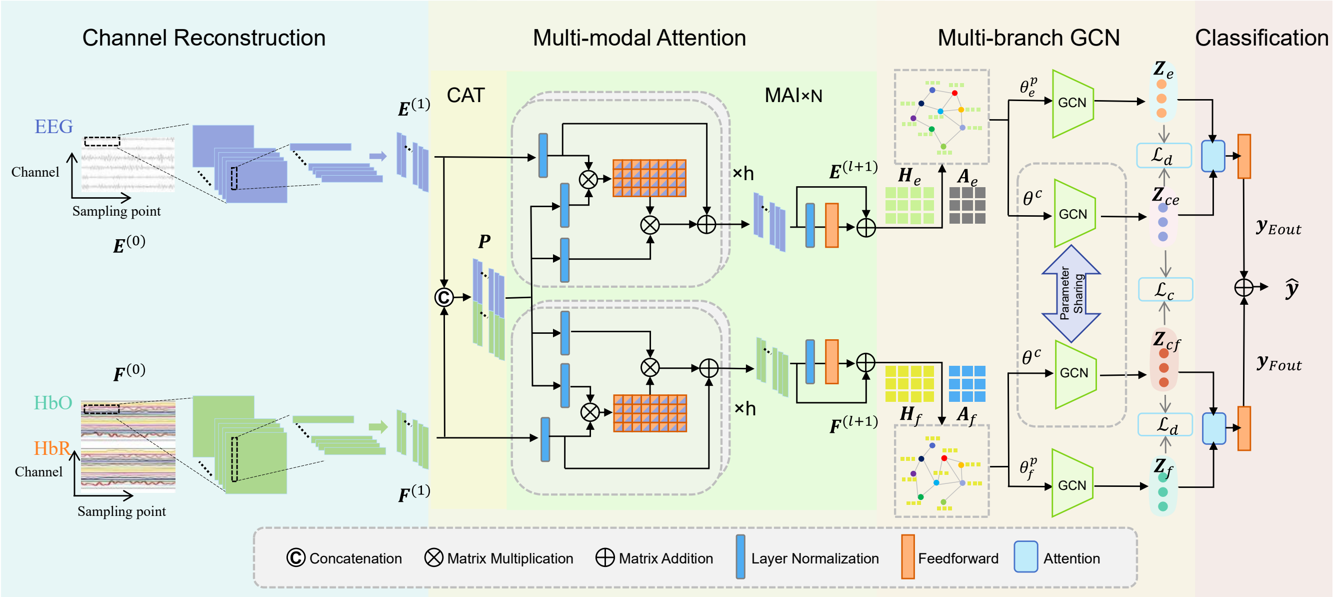


Fig. 2. Framework of the proposed DMSL model. The entire framework consists of four parts: (1) The Channel Reconstruction Module generates features with rich spatiotemporal patterns for each modality. (2) The Multi-modal Attention Module comprehensively captures inter-modal correlations and enhances the representation of each modality. (3) The Multi-branch GCN Module with modality consistency and disparity constraints is used for disentangled representation learning and capturing effective spatiotemporal coupled features. (4) And the Classification Module adaptively fuses the disentangled representations and performs task prediction.

layer does average pooling to reduce overfitting and complexity (with $K = 40$ as in [34]). Finally, the feature map from the convolution module is reordered by squeezing and transposing dimensions, and all feature channels at each time point are input as tokens into the subsequent module.

For the fNIRS sub-model, we perform a similar operation, as shown in Table I. We design the kernel sizes to match the feature dimensions of fNIRS with those of EEG, simplifying the calculation. Since fNIRS has a low time resolution and does not contain much information in the frequency domain, only two convolution modules and rearrangement are used for fNIRS branches in the embedding layer.

After passing through the feature extractor of the two sub-models, we obtain $\boldsymbol{E}^{(1)}$ and $\boldsymbol{F}^{(1)}$ for the two modalities. Here, $\boldsymbol{E}^{(1)} \in \mathbb{R}^{M\times K}$ and $\boldsymbol{F}^{(1)} \in \mathbb{R}^{M\times K}$ serve as the preliminary extracted features input to the subsequent multimodal attention module, with M denoting the number of time points.

C. Multi-modal Attention Module

Considering that the primary feature of these two modalities contains redundant and complementary information, a simple concatenation may introduce irrelevant information and result in suboptimal performance. Inspired by MulT [35], we propose a Multi-modal Attention module to facilitate data exchange following the feature extraction of the channel reconstruction module. Notably, MulT is mainly applied to unaligned multi-modal language sequences, while our module is for hybrid EEG-fNIRS BCI systems and aims to enhance the single-modal feature representation by leveraging the hybrid modality in time series.

Specifically, the Multi-modal Attention Module consists of two stages: one is the feature concatenation (CAT) layer, and the other is the Modality Attention Interaction (MAI) modules connected in series, as shown in Figure 2. In the MAI module, the multimodal features initially fused by concatenation will be transformed into a set of different key/value pairs, so as to conduct dynamic attention interactions with the corresponding single modalities (queries). This design allows the queries to focus on the intrinsic properties of the single modalities, while guiding the model to complement fNIRS with temporal details and EEG with spatial localization from the hybrid modalities. At each layer of the MAI module, the single-modal branches will update their sequences through the multi-head attention and residual structures.

TABLE I
THE STRUCTURE OF THE FEATURE EXTRACTION LAYER

Layer	Layer	In	Out	kernel	stride
EEG embedding	Temporal Conv	1	K	(1,25)	(1,1)
	Spatial Conv	K	K	$(C_e,1)$	(1,1)
	Avg Pooling	K	K	(1,75)	(1,25)
	Rearrange	$(K,1,M) \rightarrow (M,K)$			
fNIRS embedding	Temporal Conv	1	K	(1,10)	(1,1)
	Spatial Conv	K	K	$(2C_f,1)$	(1,1)
	Rearrange	$(K,1,M) \rightarrow (M,K)$			

After passing through the channel reconstruction module, we obtain a pair of preliminary features of the two modalities, represented as $\boldsymbol{U}^{(1)} \in \{\boldsymbol{E}^{(1)}, \boldsymbol{F}^{(1)}\} \in \mathbb{R}^{M\times K}$. By concatenating them, we obtain the hybrid features of the two modalities, $\boldsymbol{P} = [\boldsymbol{E}^{(1)}, \boldsymbol{F}^{(1)}] \in \mathbb{R}^{2M\times K}$. We then embed $\boldsymbol{P}$ into two spaces, denoted as $\boldsymbol{K} = LN(\boldsymbol{P})$ and $\boldsymbol{V} = LN(\boldsymbol{P})$, while defining the query for each block l as $\boldsymbol{Q}^{(l)} = LN(\boldsymbol{U}^{(l)})$, where LN represents layer normalization, $l = 1, \dots, N$ and N is the number of blocks. The attention interaction for the l-th block is defined as follows:

$$\boldsymbol{U}_1^{(l)} = Concat(head_1^{(l)}, head_2^{(l)}, \dots, head_h^{(l)})\boldsymbol{W}_o^{(l)}, \qquad (3)$$

$$head_i^{(l)} = softmax\left(\frac{\boldsymbol{Q}^{(l)}\boldsymbol{W}_{q,i}^{(l)}(\boldsymbol{K}\,\boldsymbol{W}_{k,i}^{(l)})^T}{\sqrt{d_k}}\right)\left(\boldsymbol{V}\boldsymbol{W}_{v,i}^{(l)}\right), \quad (4)$$

where h represents the number of heads, $\boldsymbol{W}_o^{(l)} \in \mathbb{R}^{K\times K}$ is the output projection matrix of the l-th block, $\boldsymbol{W}_{q,i}^{(l)}$, $\boldsymbol{W}_{k,i}^{(l)}$ and $\boldsymbol{W}_{v,i}^{(l)} \in \mathbb{R}^{K\times d_k}$ are the projection matrices for queries, keys, and values, respectively, and $d_k = K/h$ is the dimension of the key vector. Specifically, the attention score matrix $softmax(\cdot) \in \mathbb{R}^{M\times 2M}$ calculated in formula (3) is utilized to gauge the extent of attention that the i-th time step within the single modality directs towards the j-th time step of the hybrid modality. Accordingly, the i-th time step of $\boldsymbol{U}_1^{(l)}$ in formula (3) represents the weighted aggregate of the elements in $\boldsymbol{V}$, where the weight is ascertained by the i-th row of $softmax(\cdot)$.

Next, the two residual processes in the l-th block are represented by the forward propagation:

$$\boldsymbol{U}_2^{(l)} = LN\left(\boldsymbol{U}^{(l)}\right) + \boldsymbol{U}_1^{(l)}, \quad (5)$$

$$\boldsymbol{U}^{(l+1)} = FC\left(LN\left(\boldsymbol{U}_2^{(l)}\right)\right) + \boldsymbol{U}_2^{(l)}, \quad (6)$$

where FC is the feedforward layer, and $\boldsymbol{U}^{(l+1)} \in \{\boldsymbol{E}^{(l+1)}, \boldsymbol{F}^{(l+1)}\} \in \mathbb{R}^{M\times K}$.

Finally, the output $\boldsymbol{H} \in \{\boldsymbol{H}_e, \boldsymbol{H}_f\} \in \mathbb{R}^{K\times M}$ of the two modalities is represented as:

$$\{\boldsymbol{H}_e, \boldsymbol{H}_f\} = \left\{\left(\boldsymbol{E}^{(N+1)}\right)^T, \left(\boldsymbol{F}^{(N+1)}\right)^T\right\}. \quad (7)$$

The advanced features learned through the MAI module have a more comprehensive fusion feature representation, which can improve the expression ability and generalization ability of features.

D. Multi-branch Graph Convolution Module

At the end of the MAI module, the results of the two modalities have potential relevance and differences in spatiotemporal patterns. We further use a multi-branch graph convolution module to explicitly extract the complex features contained in $\boldsymbol{H} \in \{\boldsymbol{H}_e, \boldsymbol{H}_f\}$. This enables us to balance these features, assigning greater weights to more conducive features, and less weights to those that have confusing properties [36], [37].

Specifically, at this stage, the output representations of the MAI module are mapped to two distinct GCN branches: the common GCN branch and the private GCN branch, as depicted in Figure 2. The output dimensions of GCN satisfy $M_{out} = round(0.85 \times M)$, as this value yields the best performance during hyperparameter tuning.

The common GCN branch employs a parameter-sharing encoding function $C_{(E,F)}(\cdot)$ to learn the common representations $\boldsymbol{Z}_{ce} \in \mathbb{R}^{K\times M_{out}}$ and $\boldsymbol{Z}_{cf} \in \mathbb{R}^{K\times M_{out}}$. The formulas are as follows:

$$\boldsymbol{Z}_{ce} = C_{(E,F)}(\boldsymbol{H}_e; \theta^c), \quad (8)$$

$$\boldsymbol{Z}_{cf} = C_{(E,F)}\left(\boldsymbol{H}_f; \theta^c\right), \quad (9)$$

where $C_{(E,F)}(\cdot)$ is based on a graph convolution network, and θ^c represents its shared parameters.

The private branch for the EEG modality uses the private encoding function $P_E(\cdot)$ with the parameter θ_e^p to learn the private representation $\boldsymbol{Z}_e \in \mathbb{R}^{K\times M_{out}}$. Similarly, the private encoding function $P_F(\cdot)$ with the parameter θ_f^p for the fNIRS modality learns the private representation $\boldsymbol{Z}_f \in \mathbb{R}^{K\times M_{out}}$. The formulas are as follows:

$$\boldsymbol{Z}_e = P_E\left(\boldsymbol{H}_e; \theta_e^p\right), \quad (10)$$

$$\boldsymbol{Z}_f = P_F\left(\boldsymbol{H}_f; \theta_f^p\right), \quad (11)$$

where $P_E(\cdot)$ and $P_F(\cdot)$ are also realized by graph convolution networks.

The specific implementation of the graph convolution network is as follows. For the feature representation $\boldsymbol{H} = [\boldsymbol{h}^1, \boldsymbol{h}^2, \ldots, \boldsymbol{h}^K]^T \in \mathbb{R}^{K\times M}$, we dynamically construct a graph structure $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{H})$ for each sample. This graph structure enables the model to learn the associations between different convolutional kernel channels. Following [38], the adjacency matrix is defined as:

$$\boldsymbol{A} = \Phi_{ReLU}(\boldsymbol{A}_{base} \circ \boldsymbol{MASK}), \quad (12)$$

where the ReLU activation function is deployed to guarantee the non-negative property of the adjacency matrix, ∘ denotes element-wise multiplication. We presuppose that the node linkage is undirected, and the elemental adjacency matrix of the graph structure $\boldsymbol{A}_{base} \in \mathbb{R}^{K\times K}$ is symmetric, represented by:

$$\boldsymbol{A}_{base} = \begin{bmatrix} \boldsymbol{h}^1 \cdot \boldsymbol{h}^1 & \cdots & \boldsymbol{h}^1 \cdot \boldsymbol{h}^K \\ \vdots & \ddots & \vdots \\ \boldsymbol{h}^K \cdot \boldsymbol{h}^1 & \cdots & \boldsymbol{h}^K \cdot \boldsymbol{h}^K \end{bmatrix}, \quad (13)$$

and · is the dot product. From a neuroscience perspective, the dot-product of feature vectors from reconstructed channels (integrating spatiotemporal info) quantifies similarity between spatiotemporal synergetic patterns, aligning with dynamic neural coupling in cognitive tasks [28], [38]. $\boldsymbol{MASK} \in \mathbb{R}^{K\times K}$ is a trainable symmetric matrix, which is initialized as $\boldsymbol{MASK} = \frac{1}{2}(\boldsymbol{\xi} + \boldsymbol{\xi}^T)$, where $\boldsymbol{\xi} \sim \boldsymbol{\mathcal{N}}(0, \sigma^2)$.

We adopt the normalization of the adjacency matrix, which is denoted as:

$$\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A} + \boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}, \quad (14)$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_K)$ is the degree matrix of $\boldsymbol{A}$, with $d_m = \sum_n \boldsymbol{A}(m,n)$, and $\boldsymbol{I}$ is the identity matrix.

With the feature matrix $\boldsymbol{H}$ and the normalized adjacency matrix $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{K\times K}$ constructed, the GCN layer is formulated as:

$$\boldsymbol{Z} = \Phi_{ReLU}(\widetilde{\boldsymbol{A}}(\boldsymbol{HW} - \boldsymbol{b})), \quad (15)$$

where $\boldsymbol{W} \in \mathbb{R}^{M\times M_{out}}$ is the weight matrix, $\boldsymbol{b}$ is the bias vector.

We interpret the two types of representations from a neuroscientific perspective as follows.

Common representations primarily capture the cross-modal neural synergy that is simultaneously expressed in EEG and fNIRS signals. These representations may reflect neural processes occurring in brain regions where the two modalities exhibit task-related co-activation [39], [40], such as frontal lobe activation during mental arithmetic [30], sensorimotor activation during motor imagery [41], and the frontoparietal activation during emotion recognition [42].

In contrast, private representations are extracted from modality-specific GCN branches and are expected to complement the shared subspace by encoding unique physiological signatures dominated by each respective modality. Given the high temporal resolution of electrophysiology, EEG-based private representations are likely to emphasize rapid neural oscillation dynamics across specific frequency bands (e.g., beta, or gamma) [10]. Conversely, fNIRS-based representations are likely to capture spatially distributed hemodynamic changes, specifically the distinct concentration shifts of oxygenated (HbO) and deoxygenated (HbR) hemoglobin associated with sustained cortical activation [12].

E. Classification Module

Now we have two specific embeddings $\boldsymbol{Z}_e \in \mathbb{R}^{K\times M_{out}}$ and $\boldsymbol{Z}_f \in \mathbb{R}^{K\times M_{out}}$, as well as two general embeddings $\boldsymbol{Z}_{ce} \in \mathbb{R}^{K\times M_{out}}$ and $\boldsymbol{Z}_{cf} \in \mathbb{R}^{K\times M_{out}}$. We further combine their channel dimension and feature dimension to convert the data into vectors with a length of $d = K \times M_{out}$. Considering that sample labels can be associated with one or even a combination of them, we use the attention mechanism to learn their corresponding importance, as follows:

$$(\alpha_e, \alpha_{ce}) = att(\boldsymbol{v}_e, \boldsymbol{v}_{ce}), \tag{16}$$

$$(\alpha_f, \alpha_{cf}) = att(\boldsymbol{v}_f, \boldsymbol{v}_{cf}), \tag{17}$$

where $\boldsymbol{v}_e, \boldsymbol{v}_{ce} \in \mathbb{R}^{1\times d}$ represent the normalized embeddings obtained by applying L_2-normalization to the flattened $\boldsymbol{Z}_e$ and $\boldsymbol{Z}_{ce}$, and α_e and α_{ce} represent the attention values between them. Similarly, α_f and α_{cf} follow the same principle.

We use a shared weight vector $\boldsymbol{q} \in \mathbb{R}^{d\times 1}$ to obtain the attention value ω_e as follows:

$$\omega_e = \boldsymbol{v}_e \boldsymbol{q}. \tag{18}$$

Similarly, we can get the attention value ω_{ce} of the embedded vector $\boldsymbol{v}_{ce}$. Then, we use the softmax function to normalize the attention values ω_e and ω_{ce} to obtain the final weight:

$$\alpha_e = softmax(\omega_e) = \frac{exp(\omega_e)}{exp(\omega_e)+exp(\omega_{ce})}. \tag{19}$$

A larger α_e indicates that the corresponding embedding is more important. Similarly, we can get $\alpha_{ce} = softmax(\omega_{ce})$. Then, we combine these two embeddings to get the final embedding $\boldsymbol{v}_{Eout}$:

$$\boldsymbol{v}_{Eout} = \alpha_e \cdot \boldsymbol{v}_e + \alpha_{ce} \cdot \boldsymbol{v}_{ce}. \tag{20}$$

After obtaining $\boldsymbol{v}_{Eout} \in \mathbb{R}^{1\times d}$, we further perform a linear transformation to obtain the class predictions:

$$\boldsymbol{y}_{Eout} = softmax(\boldsymbol{v}_{Eout}\boldsymbol{W}_e + \boldsymbol{b}_e), \tag{21}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{d\times C}$, $\boldsymbol{y}_{Eout} \in \mathbb{R}^{1\times C}$, and C is the number of classes. Similarly, given $\boldsymbol{v}_f$ and $\boldsymbol{v}_{cf}$, we obtain $\boldsymbol{y}_{Fout} \in \mathbb{R}^{1\times C}$ by following the same steps in Eqs. (18)-(21).

Subsequently, we add the two predictions together to obtain the final output:

$$\hat{\boldsymbol{y}} = (\boldsymbol{y}_{Eout} \oplus \boldsymbol{y}_{Fout})/2, \tag{22}$$

where $\oplus$ represents the element-wise addition. Here, $\hat{\boldsymbol{y}} = [\hat{y}_c] \in \mathbb{R}^{1\times C}$, with $\hat{y}_c$ representing the probability of belonging to category c.

F. Objective Optimization

1) Task loss

In the case of fNIRS and EEG data, it is easy to over-match the results of classification, and label smoothing [43] can prevent the model from becoming over-confident and improve generalization.

During training, we calculate the cross-entropy loss for each batch of samples with size B (represented as $\hat{\boldsymbol{Y}} = [\hat{y}_{b,c}] \in \mathbb{R}^{B\times C}$) and take the average within the batch:

$$\mathcal{L}_{task} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\sum_{c=1}^{C} -y_{b,c}^{LS}\, log(\hat{y}_{b,c}), \tag{23}$$

where $\hat{y}_{b,c}$ is the predicted probability of the b-th sample in the batch for category c, $y_{b,c}^{LS}$ is the smoothed true probability for the same sample and category.

2) Consistency loss

For the two output embeddings $\boldsymbol{Z}_{ce}$ and $\boldsymbol{Z}_{cf}$ of $C_{(E,F)}$, we use a consistency constraint to further enhance their commonness.

Firstly, the embedded matrices are normalized to $\boldsymbol{v}_{ce}$ and $\boldsymbol{v}_{cf}$. Let $\boldsymbol{V}_{ce} \in \mathbb{R}^{B\times d}$ and $\boldsymbol{V}_{cf} \in \mathbb{R}^{B\times d}$ be the matrices whose rows denote $\boldsymbol{v}_{ce}$ and $\boldsymbol{v}_{cf}$, where B denotes the batch size. Then, two matrices, $\boldsymbol{S}_e \in \mathbb{R}^{B\times B}$ and $\boldsymbol{S}_f \in \mathbb{R}^{B\times B}$, are utilized to capture the similarity between samples:

$$\boldsymbol{S}_e = \boldsymbol{V}_{ce}\boldsymbol{V}_{ce}^T, \tag{24}$$

$$\boldsymbol{S}_f = \boldsymbol{V}_{cf}\boldsymbol{V}_{cf}^T. \tag{25}$$

Consistency means that the distance between the two similarity matrices is narrowed to obtain the following constraints:

$$\mathcal{L}_c = \frac{\|\boldsymbol{S}_e - \boldsymbol{S}_f\|_F^2}{B^2}, \tag{26}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm.

3) Difference loss

We apply orthogonal constraints to ensure that the graph convolution of private channels and common channels captures different aspects of the input. Let $\boldsymbol{V}_e \in \mathbb{R}^{B\times d}$ and $\boldsymbol{V}_f \in \mathbb{R}^{B\times d}$ be the matrices whose rows correspond to $\boldsymbol{v}_e$ and $\boldsymbol{v}_f$, where B is the batch size. The topological relationships between the representations of the private and common channel outputs are aligned by minimizing the square F-norm of the difference matrix, normalized by the matrix size:

$$\mathcal{L}_d = \frac{\|\boldsymbol{V}_e\boldsymbol{V}_{ce}^T\|_F^2}{B^2} + \frac{\left\|\boldsymbol{V}_f\boldsymbol{V}_{cf}^T\right\|_F^2}{B^2}. \tag{27}$$

4) Overall objective function

Combining the task loss $\mathcal{L}_{task}$, the consistency loss $\mathcal{L}_c$, and the difference loss $\mathcal{L}_d$, the final calculation formula is:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha\mathcal{L}_c + \beta\mathcal{L}_d, \tag{28}$$

where α, β are trade-off parameters.

V. Experiments

A. Experimental Setup

To assess our proposed DMSL, benchmark datasets involving MI, MA and emotion recognition tasks are employed, as described in Section III. We adopt standard protocols [9], [29], [33] to evaluate the model performance. Specifically, we utilize leave-one-subject-out cross-validation (LOSO-CV) to verify individual differences and model generalization [44]. In each fold of the cross-validation, the data of one subject served as the test set, while the remaining subjects formed the training set. This process was repeated until all subjects had been tested.

TABLE II
RESULTS OF SUBJECT INDEPENDENT COMPARISON ALGORITHMS ON THE MA, MI, AND ENTER DATASET

Model	Signal Type	MA		MI		ENTER	
		ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
LSTM [47]	EEG	61.46** ± 6.35	60.63**	52.70** ± 2.45	52.35**	33.30** ± 8.49	31.28**
	fNIRS	65.35** ± 6.39	64.86**	59.32** ± 5.06	59.07**	38.30** ±13.32	36.74**
EEGNet [48]	EEG	67.66** ± 9.17	65.30**	65.32 ±11.60	64.71	41.25* ± 8.31	39.24*
	fNIRS	69.29** ± 6.92	68.92**	60.47** ± 5.81	59.97**	37.20** ± 8.13	33.62**
TSCeption [49]	EEG	67.90** ± 8.34	66.88**	63.12** ± 8.66	62.01**	36.50** ± 7.26	31.93**
	fNIRS	66.11** ± 6.12	65.74**	58.81** ± 4.47	58.24**	40.20** ±11.66	37.52**
fNIRSNet [9]	fNIRS	69.76** ± 7.41	69.48*	60.14** ± 6.49	59.74**	38.25** ± 9.45	35.29**
EF-Net [50]	EEG+fNIRS	66.78** ± 7.04	66.16**	57.47** ± 3.81	56.94**	43.73 ±11.26	41.02*
pth-PF [24]	EEG+fNIRS	71.29** ± 6.07	71.02*	62.08** ± 5.83	61.50**	43.93 ± 8.22	41.71
M2NN [51]	EEG+fNIRS	67.16**± 7.27	66.75**	56.76** ± 5.58	52.02**	41.80* ± 4.86	38.03**
EFMLNet [27]	EEG+fNIRS	71.03**± 3.51	70.86**	62.10** ± 4.69	61.41**	39.23** ± 5.70	35.31**
Dual-EEGNet [52]	EEG+fNIRS	73.09* ± 7.25	72.61*	67.65 ± 9.64	67.10	42.68* ± 7.40	39.81*
DMSL (ours)	EEG+fNIRS	75.43 ± 7.28	75.04	68.24 ± 7.86	67.75	45.40 ± 7.74	43.60

1)"*" represents the significant differences ($p < 0.05$, Wilcoxon signed-rank test with Holm-Bonferroni correction) compared to the DMSL.
2)"**" represents the significant differences ($p < 0.001$, Wilcoxon signed-rank test with Holm-Bonferroni correction) compared to the DMSL.

The reported results are the average of all subjects. We utilize accuracy (ACC) and macro F1-score (F1) [9] to measure the model's performance.

B. Parameter Settings

To evaluate our model more comprehensively, it is necessary to control certain variable factors. We set the batch size of the training set to 40 and the number of epochs to 60. Our model employs the AdamW optimizer [45], initialized with a learning rate of 0.001. To expedite training and enhance performance, we mitigate flooding levels during specific periods to maintain a continuous regularization effect [46]. We choose the consistency coefficient α and the difference constraint coefficient β from the candidate sets $\{0.01, 0.1, 1, 3, 10, 30\}$. Through hyperparameter tuning, the optimal values are determined as α=10 and β=3 for both MA and MI datasets, and α=10 and β=0.1 for the ENTER dataset. This configuration effectively balances modality consistency $\mathcal{L}_c$ and disparity constraints $\mathcal{L}_d$.

C. Baselines

To rigorously evaluate the proposed DMSL, we compare it against established unimodal deep learning networks and state-of-the-art (SOTA) hybrid EEG-fNIRS frameworks.

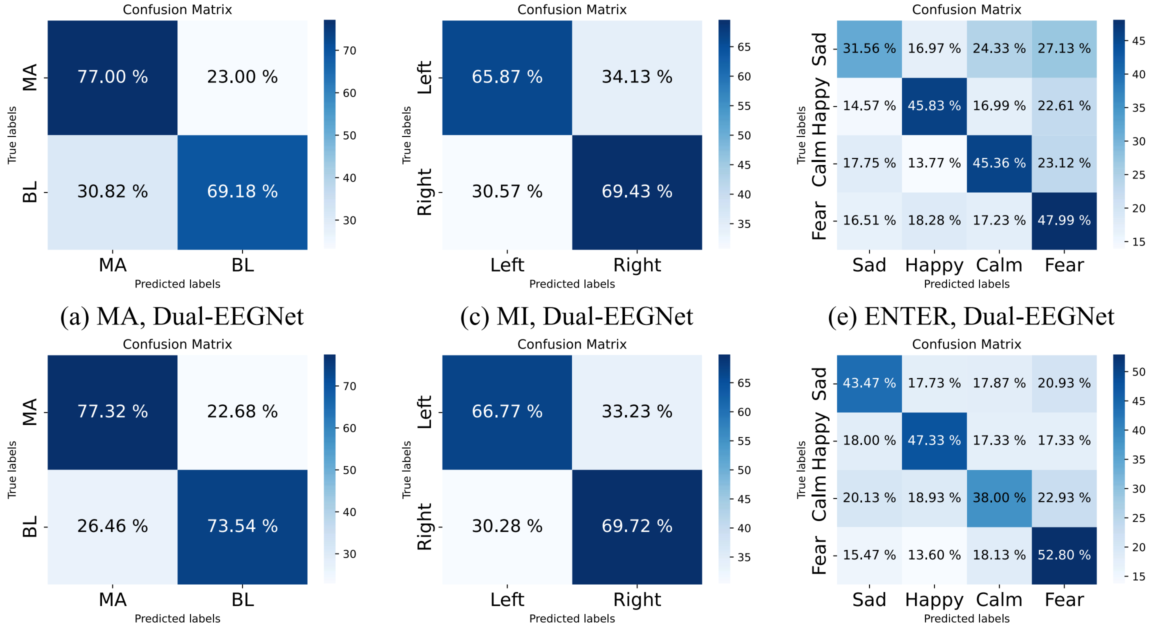
For unimodal comparisons, we selected **LSTM** [47], which utilizes gating mechanisms for sequence dependency modeling; **EEGNet** [48], a compact architecture employing depthwise separable convolutions; and **TSCeption** [49], which fuses multi-scale convolutions with attention mechanisms. For fNIRS-specific analysis, we include **fNIRSNet** [9], designed with DHR and DWS convolutions to capture hemodynamic patterns.

For multimodal comparisons, we benchmark against five SOTA methods representing diverse fusion strategies: **EF-Net** [50] and **pth-PF** [24] both employ CNN-based extraction with late fusion, where the latter utilizes kernel tensor multiplication for EEG, HbO, and HbR streams. **M2NN** [51] introduces an end-to-end framework integrating spatio-temporal learning with multi-task capabilities. **EFMLNet** [27] leverages Transformer-based multi-head attention to model cross-modal mutual learning between temporal EEG and spatial fNIRS features. Finally, **Dual-EEGNet** [52] implements an early fusion strategy via a Y-shaped architecture rooted in the EEGNet framework.

D. Classification Results

In the LOSO-CV experiment, the relevant results of the average accuracy (avg ± std%) and F1 score of the test set are shown in Table II. In addition, the Wilcoxon signed-rank test with Holm-Bonferroni correction was used to analyze the differences between DMSL and the baseline methods. We define "*" to indicate a p-value less than 0.05, and "**" to indicate a p-value less than 0.001. For single-modal EEG and fNIRS models, the EEGNet and fNIRSNet models demonstrated strong learning capabilities, and their classification results were significantly better than those of the traditional LSTM and CNN classifiers. For the multimodal EEG+fNIRS model, the average accuracies of the DMSL model on the MA, MI, and ENTER datasets were 75.43%, 68.24%, and 45.40% respectively. Its classification performance was significantly better than that of the vast majority of baseline methods, with significant differences (Wilcoxon signed-rank test with Holm-Bonferroni correction, $p < 0.05$). While statistical significance was not observed for a few baselines, DMSL consistently maintained the highest mean accuracy, indicating superior robustness. Compared with Dual-EEGNet, which ranked second in accuracy, DMSL improved the accuracy by 2.34% and 0.59% on the MA and MI datasets, respectively; on the ENTER dataset, compared with pth-PF, which ranked second in performance, DMSL improved the accuracy by 1.47%. This indicates that multimodal fusion and our proposed strategy have enhanced the classification accuracy and generalization ability of the model.

The confusion matrices in Figure 3 demonstrate that, compared with Dual-EEGNet, the DMSL method significantly enhances the classification accuracy of hybrid EEG-fNIRS data. In the MA task, DMSL achieves higher accuracy in predicting



(a) MA, Dual-EEGNet (c) MI, Dual-EEGNet (e) ENTER, Dual-EEGNet

(b) MA, DMSL (ours) (d) MI, DMSL (ours) (f) ENTER, DMSL (ours)

Fig. 3. The confusion matrices of Dual-EEGNet and DMSL, where each column represents the true labels and each row represents the predicted labels of the model. The results of the MA, MI, and ENTER datasets are shown, where MA and BL respectively denote mental arithmetic and baseline, and Left and Right respectively denote the left hand and right hand in MI.

MA and BL states by integrating EEG and fNIRS modalities, which compensates for the limitations of single-modal approaches. In the MI task, DMSL also improves the classification accuracy for left- and right-hand motor imagery, with increased true positive (TP) and true negative (TN) rates, showcasing its superior classification performance. For emotion classification within the ENTER dataset context, DMSL effectively discriminates between emotional states like Sad, Happy, Calm, and Fear. It reduces misclassifications across these affective categories, and thus proves its capability in handling multi-dimensional emotional data analysis based on hybrid modalities.

E. Ablation Studies

1) Importance of Modality.

Firstly, we delete modalities separately to explore the performance of DMSL, as shown in Table III. To ensure comparability, we retained the feature extractors and graph convolutions of private channels in DMSL, only changed the interactive attention to self-attention, and removed the graph convolution of public channels, the attention mechanism, as well as the consistency and difference loss functions. When the EEG modality is deleted, the performance of the model significantly declines, indicating that the EEG modality

TABLE III
RESULTS OF ABLATION RESEARCH ON LOSO VALIDATION METHOD

Model	Dataset MA	Dataset MI	Dataset ENTER
	ACC ± std	ACC ± std	ACC ± std
DMSL	75.43 ± 7.28	68.24 ± 7.86	45.40 ± 7.74
Importance of Modality			
w/o EEG	67.71 ± 6.20	59.98 ± 3.98	38.75 ±10.20
w/o fNIRS	73.71 ± 8.72	66.28 ±12.96	44.15 ± 9.49
Importance of Constraint			
w/o $\mathcal{L}_d$	74.56 ± 7.49	67.61 ± 7.39	45.23 ± 8.08
w/o $\mathcal{L}_c$	74.37 ± 7.00	67.86 ± 8.19	44.97 ± 8.33
w/o $\mathcal{L}_c$+$\mathcal{L}_d$	73.05 ± 7.38	65.93 ± 7.46	44.53 ± 8.13
Importance of Different Components			
w/o Phase 1	73.48 ± 6.99	60.26 ± 3.99	41.20 ± 9.10
w/o Phase 2	74.43 ± 7.18	66.76 ± 6.95	44.60 ± 9.92
w/o CAT (Variant 1)	71.98 ± 6.62	61.85 ± 4.32	39.65 ± 9.46
w/o CAT (Variant 2)	74.57 ± 7.29	68.29 ± 8.42	42.08 ± 9.52

"w/o" represents removal for the mentioned factors.

dominates the multi-modal task. However, under the direction of the hybrid features, the discriminative features contained in it can be extracted effectively. In addition, compared with multi-modal DMSL, the performance of single-modal is always poorer. This indicates that our model can effectively extract the complementary features among different modalities, which is beyond the capabilities of single-modal models.

2) The importance of constraint.

We individually remove losses to verify the impact of different constraints. When the difference loss ($\mathcal{L}_d$) is absent, the model relies on the consistency loss ($\mathcal{L}_c$) to learn various multimodal representations, and the model performance will deteriorate. In addition, we observe that the consistency constraint enhances model performance. When both constraints are absent, the worst performance highlights the crucial role of constraints in multi-branch representation learning.

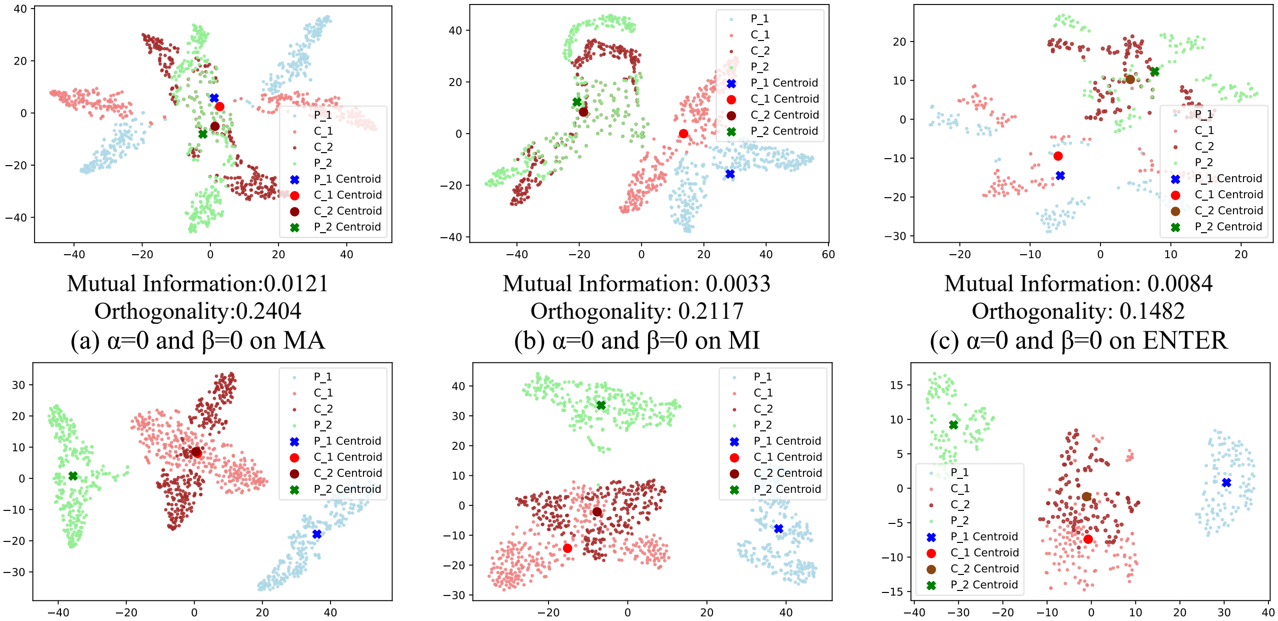
3) The importance of different components.

We discuss the importance of different components. The key improvement of the DMSL method is the addition of an improved Multi-modal Attention module to learn fusion features and a Multi-branch GCN module for disentangled representation learning to capture complex spatiotemporal relationships inherent in signals. Therefore, we conducted ablation studies on the dataset, as shown in Table III, in which the Multi-modal Attention module (Phase 1) and Multi-branch GCN module (Phase 2) were deleted, respectively. It can be seen that when Phase 1 is removed, the classification performance of the model drops most significantly, with decreases of 1.95%, 7.98%, and 4.20% on the MA, MI, and ENTER datasets, respectively. The results indicate that the multi-head attention mechanism is beneficial. Applying disentanglement to the enhanced modalities, which are refined through the multi-head attention mechanism, rather than directly to the raw modalities, enables richer and more precise feature extraction. When Phase 2 is deleted, the experimental results will decline. This indicates that applying disentangled representation learning can enhance the performance of feature extraction.

In addition, we replace hybrid modal guidance with its single modal version (denoted as 'w/o CAT (Variant 1)' in Table III). The observed performance declines of 3.45%, 6.39%, and 5.75% on the MA, MI, and ENTER datasets, respectively, further confirm that without hybrid guidance, the extracted features lack cross-modal complementarity. We also replace the hybrid modal guidance with another single modal guidance (denoted as 'w/o CAT (Variant 2)' in Table III). The observed performance declines of 0.86% and 3.32% on the MA and ENTER datasets, respectively, further confirms that our improved strategy of enhancing single-modal branches with hybrid modalities plays a more guiding role than the single-modal enhancement strategy, thereby demonstrating the superiority of our proposed MAI over MulT [35].

F. Visualization

To evaluate different methods' capability in extracting highly distinct features from EEG and fNIRS signals, we employ the t-SNE technique [53] to visualize the features generated in a 2D embedding space. Utilizing t-SNE, the high-dimensional output from the final fully connected layer of all training models is transformed into a two-dimensional feature space for visual analysis. Figure 4 shows the t-SNE



Mutual Information:0.0121
Orthogonality:0.2404
(a) α=0 and β=0 on MA

Mutual Information: 0.0033
Orthogonality: 0.2117
(b) α=0 and β=0 on MI

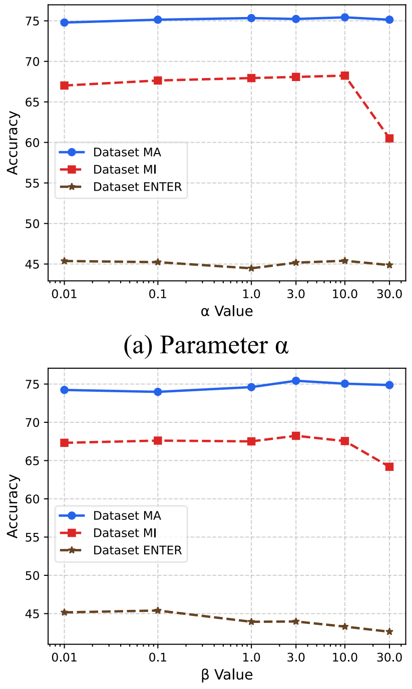
Mutual Information: 0.0084
Orthogonality: 0.1482
(c) α=0 and β=0 on ENTER

Mutual Information: 0.0998
Orthogonality: 0.0017
(d) α≠0 and β≠0 on MA

Mutual Information: 0.1450
Orthogonality: 0.0013
(e) α≠0 and β≠0 on MI

Mutual Information: 0.0269
Orthogonality: 0.0092
(f) α≠0 and β≠0 on ENTER

Fig. 4. Scatter Visualize both common and private representations in the test set on three tasks. α=0, β=0 indicates no consistency and difference constraints, and vice versa. C_1, C_2, P_1 and P_2 correspond to Z_{ce}, Z_{cf}, Z_e and Z_f, respectively. The colored dots (red for C_1 centroid, dark red for C_2 centroid, blue for P_1 centroid, green for P_2 centroid) mark the centroids of each cluster, aiding in assessing the central tendency of common and private representation distributions.



(a) Parameter α

(b) Parameter β

Fig. 5. Analysis of parameters α and β.

visualization (with class centroids clearly marked) of the centralized common and private representations on the test set for the three tasks. α = 0, β = 0 means there is no consistency constraint and difference constraint, and vice versa. Red and pink dots correspond to common representations, while green and blue dots correspond to private representations. We visualized common representations $C\_ = \left\{Z_{ce}, Z_{cf}\right\}$ and private representations $P\_ = \left\{Z_e, Z_f\right\}$ of learning without or with consistency loss and difference loss on three tasks.

When $\alpha = 0, \beta = 0$, as seen in Figure 4 (a)-(c), the distributions of $C\_$ and $P\_$ sometimes overlap, and no clear common representation cluster is formed. Conversely, when $\alpha \neq 0, \beta \neq 0$, as shown in Figure 4 (d)-(f), where class centroids of C_ are positioned to reflect integrated common information, and the indistinct boundary reflects tight cross-modal neural synergy. At the same time, for private representations, as evidenced by the separate clusters of green (Z_e) and blue (Z_f) dots and their distinct centroids, each subspace specific to a mode is separable, where the difference constraint eliminates the potential representation of redundancy.

To quantitatively demonstrate the disentanglement performance, we introduce mutual information and orthogonality metrics as shown in Figure 4. Mutual Information measures the statistical dependence between common representations, where a higher mutual information indicates a greater degree of aggregation. Orthogonality quantifies the angular separation between common and private representations, with values closer to 0 reflecting high orthogonality. As shown in Figure 4, incorporating consistency and difference constraints results in higher mutual information and lower orthogonality across all tasks, demonstrating improved feature disentanglement.

G. Parameter Sensitivity Analysis

We conduct a sensitivity analysis of parameters α and β to gain a deeper understanding of their effects on model performance, thus providing a strong reference for parameter selection in practical applications. The results are presented in Figure 5.

Parameter α: To evaluate the effect of α, we fixed β at 3 for the MA and MI datasets, and at 0.1 for the ENTER dataset—these values correspond to the optimal β identified during hyperparameter tuning. As α varies from 0.01 to 10.0, the accuracies of MA and MI remain relatively stable, exhibiting small fluctuations and an upward trend. However, when α reaches 30.0, the accuracy of MI drops sharply, suggesting that overly strong modality consistency constraints may interfere with effective optimization of the task loss (i.e., classification loss). In contrast, the ENTER dataset shows consistently low and stable accuracy across all α values, indicating limited sensitivity to this parameter.

Parameter β: To evaluate the effect of β, we fix α at its optimal value of 10 for all datasets. For the MA dataset, as β increases from 0.01 to 10.0, the accuracy first decreases slightly and then increases and stabilizes. For the MI dataset, the accuracy rises steadily when β ranges from 0.01 to 3.0, and shows a downward trend when β exceeds 3.0. Regarding the ENTER dataset, the model achieves optimal accuracy when β =0.1. These results reflect that different datasets have varying degrees of sensitivity to changes in β, and an appropriate range of β values needs to be selected in practical applications to ensure optimal model performance.

H. Computational Complexity Analysis

We rigorously analyze the model's computational complexity (based on one batch), comparing it with existing BCI models. As can be seen from Table IV, the DMSL model ranks second in terms of the number of parameters and third in terms of FLOPs among the compared models. Specifically, while maintaining a relatively reasonable number of parameters and FLOPs, DMSL (Ours) outperforms models such as Dual-EEGNet and pth-PF in classification performance, achieving a better balance between computational efficiency and classification effectiveness. Moreover, compared with models with a large number of parameters and FLOPs, DMSL (Ours) has a significant advantage in terms of computational resource consumption and is more suitable for deployment in resource-constrained devices.

TABLE IV
THE NUMBER OF MODEL PARAMETERS

Model	Parameters	FLOPs
EF-Net	4.00M	27.18G
pth-PF	0.63M	0.50G
M2NN	5.36M	4.35G
EFMLNet	3.76M	14.80G
Dual-EEGNet	0.05M	0.77G
DMSL（Ours）	0.21M	1.96G

VI. Conclusion

This paper introduces DMSL, a novel multimodal learning framework that enables multimodal spatiotemporal coupling and disentangled representation learning within a unified structure. We propose a multimodal attention module to comprehensively capture inter-modality correlations and enhance the representations for each modality. Additionally, we present a multi-branch graph convolutional module based on reconstructed channels, incorporating modality consistency and disparity constraints to facilitate disentangled representation learning and effective spatiotemporal coupling feature capture. Experimental results show that DMSL outperforms the state-of-the-art EEG-fNIRS fusion method, exceeding the best baseline by 2.34%, 0.59% and 1.47% on the MA, MI, and ENTER datasets, respectively. Furthermore, ablation studies demonstrate the effectiveness of our fusion strategy and the importance of consistency constraints. The t-SNE visualization further indicates that our model has an excellent ability to feature learning.

Overall, DMSL offers a flexible framework that can be extended to other hybrid EEG-fNIRS BCI tasks, providing a promising foundation for future research in multimodal and spatiotemporal learning.

VII. References

[1] B. Graimann, B. Allison, and G. Pfurtscheller, “Brain–Computer Interfaces: A Gentle Introduction,” *Brain-Computer Interfaces*, *The Frontiers Collection*, 2009, pp. 1–27.

[2] K. Värbu, N. Muhammad, and Y. Muhammad, “Past, Present, and Future of EEG-Based BCI Applications,” *Sensors*, vol. 22, no. 9, p. 3331, Apr. 2022.

[3] M. Tariq, P. M. Trivailo, and M. Simic, “EEG-Based BCI Control Schemes for Lower-Limb Assistive-Robots,” *Frontiers in Human Neuroscience*, vol. 12, p. 312, Aug. 2018.

[4] L. Zhang et al, “Enhancing Visual-Guided Motor Imagery Performance via Sensory Threshold Somatosensory Electrical Stimulation Training,” *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 2, pp. 756-765, Feb. 2023.

[5] J. Huang et al, “HSA-Former: Hierarchical Spatial Aggregation Transformer for EEG-Based Emotion Recognition,” *IEEE Transactions on Computational Social Systems*, 2025.

[6] M. Pang et al, “Multi-Scale Masked Autoencoders for Cross-Session Emotion Recognition,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 1637-1646, 2024.

[7] C. Chen et al, “Self-Attentive Channel-Connectivity Capsule Network for EEG-Based Driving Fatigue Detection,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3152–3162, Jan. 2023.

[8] T. Xu et al, “Motor Imagery Decoding Enhancement Based on Hybrid EEG-fNIRS Signals,” *IEEE Access*, pp. 65277–65288, Jan. 2023.

[9] Z. Wang, J. Fang, and J. Zhang, “Rethinking Delayed Hemodynamic Responses for fNIRS Classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4528–4538.

[10] G. Buzsáki, C. A. Anastassiou, and C. Koch, “The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes,” *Nature Reviews Neuroscience*, pp. 407–420, Jun. 2012.

[11] E. Eldele et al, “An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 809–818, Jan. 2021.

[12] V. Quaresima and M. Ferrari, “Functional Near-Infrared Spectroscopy (fNIRS) for Assessing Cerebral Cortex Function During Human Behavior in Natural/Social Situations: A Concise Review,” *Organizational Research Methods*, pp. 46–68, Jan. 2019.

[13] Y. Li et al, “Improved dilation CapsuleNet for motor imagery and mental arithmetic classification based on fNIRS.” *Brain-Apparatus Communication: A Journal of Bacomics*, vol. 3, no. 1, p. 2335886, 2024.

[14] Md. A. Rahman et al, “A Narrative Review on Clinical Applications of fNIRS,” *Journal of Digital Imaging*, pp. 1167–1184, Oct. 2020.

[15] R. J. Deligani et al, “Multimodal fusion of EEG-fNIRS: a mutual information-based hybrid classification framework.,” *Biomedical Optics Express*, p. 1635, Mar. 2021.

[16] C. H. Chuang et al, “Brain Electrodynamic and Hemodynamic Signatures Against Fatigue During Driving,” *Frontiers in Neuroscience*, vol. 12, Mar. 2018.

[17] A. Omurtag, H. Aghajani, and H. O. Keles, “Decoding human mental states by whole-head EEG+fNIRS during category fluency task performance,” *Journal of Neural Engineering*, p. 066003, Dec. 2017.

[18] J. Cao, E. M. Garro, and Y. Zhao, “EEG/fNIRS Based Workload Classification Using Functional Brain Connectivity and Machine Learning,” *Sensors*, p. 7623, Oct. 2022.

[19] Y. Gao et al, "Hybrid EEG-fNIRS Brain Computer Interface Based on Common Spatial Pattern by Using EEG-Informed General Linear Model," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-10, 2023

[20] J. Shin et al, “Open Access Dataset for EEG+NIRS Single-Trial Classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1735–1745, Oct. 2017.

[21] E. Ergün and O. Aydemir, “A Hybrid BCI Using Singular Value Decomposition Values of the Fast Walsh–Hadamard Transform Coefficients,” *IEEE Transactions on Cognitive and Developmental Systems*, pp. 454–463, Jun. 2023.

[22] X. Jiang et al, “Independent Decision Path Fusion for Bimodal Asynchronous Brain-Computer Interface to Discriminate Multiclass Mental States,” *IEEE Access*, pp. 165303–165317, Jan. 2019.

[23] A. M. Chiarelli et al, “Deep learning for hybrid EEG-fNIRS brain–computer interface: application to motor

imagery classification," *Journal of Neural Engineering*, p. 036028, Jun. 2018.

[24] Z. Sun et al, "A novel multimodal approach for hybrid brain-computer interface," *Cornell University*, Apr. 2020.

[25] D. Yang et al. "Disentangled representation learning for multimodal emotion recognition." *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[26] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," *Cornell University - arXiv, Cornell University*, May 2020.

[27] L. Qiu et al, "EFMLNet: Fusion Model Based on End-to-End Mutual Information Learning for Hybrid EEG-fNIRS Brain-Computer Interface Applications." *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.

[28] C. Peng et al, "mBGT: Encoding Brain Signals With Multimodal Brain Graph Transformer." *IEEE Transactions on Consumer Electronics*, vol. 71, no.2, pp. 5812-5823, May 2025.

[29] Y. Kwak, W. J. Song, and S. E. Kim, "FGANet: fNIRS-guided Attention Network for Hybrid EEG-fNIRS Brain-Computer Interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 329–339, Jan. 2022.

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

[31] C. Bunterngchit et al, "EEG-fNIRS data classification through selective channel representation and spectrogram imaging," *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.

[32] P. Gong et al, "ASTDF-net: attention-based spatial-temporal dual-stream fusion network for EEG-based emotion recognition," *Proceedings of the 31st ACM international conference on multimedia*, pp. 883-892, 2023.

[33] G. Chen et al, "EEG–fNIRS-Based Emotion Recognition Using Graph Convolution and Capsule Attention Network," B*rain Sciences*, vol. 14, no. 8, p. 820, 2024.

[34] Y. Song et al, "EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 710-719, 2023

[35] Y. H. H. Tsai et al, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[36] X. Wang et al, "AM-GCN: Adaptive Multi-channel Graph Convolutional Networks," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, pp. 1243-1253, 2020.

[37] D. Liu et al, "Brain-Machine Coupled Learning Method for Facial Emotion Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10703-10717, 2023.

[38] Y. Ding et al, "LGGNet: Learning from Local-Global-Graph Representations for Brain-Computer Interface," *Cornell University - arXiv,Cornell University - arXiv*, May 2021.

[39] H. Xu et al, "EFDFNet: A multimodal deep fusion network based on feature disentanglement for attention state classification," *Biomedical Signal Processing and Control*, vol. 109, p. 108042, 2025.

[40] W. C. Su et al, "Simultaneous multimodal fNIRS-EEG recordings reveal new insights in neural activity during motor execution, observation, and imagery," *Scientific Reports*, vol. 13, no.1, p.5151, 2023.

[41] R. Blanco, C. Koba, and A. Crimi, "Investigating the interaction between EEG and fNIRS: A multimodal network analysis of brain connectivity," J Comput Sci, vol. 82, p. 102416, 2024, doi: https://doi.org/10.1016/j.jocs.2024.102416.

[42] X. Si, Y. Han, S. Li, S. Zhang, and D. Ming, "The cortical spatial responses and decoding of emotion imagery towards a novel fNIRS-based affective BCI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

[43] R. Müller, S. Kornblith, and GeoffreyE. Hinton, "When does label smoothing help," *Neural Information Processing Systems*, Jun. 2019.

[44] X. Zhao et al, "A Multi-Branch 3D Convolutional Neural Network for EEG-Based Motor Imagery Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 2164–2177, Oct. 2019.

[45] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *Learning,Learning*, Nov. 2017.

[46] Z. Wang et al, "Transformer Model for Functional Near-Infrared Spectroscopy Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2559–2569, Jun. 2022.

[47] U. Asgher et al, "Enhanced Accuracy for Multiclass Mental Workload Detection Using Long Short-Term Memory for Brain-Computer Interface," *Frontiers in Neuroscience*, Jun. 2020.

[48] V. J. Lawhern et al, "EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces," *Journal of Neural Engineering*, p. 056013, Oct. 2018.

[49] Y. Ding et al, "TSception: Capturing Temporal Dynamics and Spatial Asymmetry from EEG for Emotion Recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, Jan. 2022.

[50] A. Arif et al, "EF-Net: Mental State Recognition by Analyzing Multimodal EEG-fNIRS via CNN." *Sensors*, vol. 24, no. 6, 2024.

[51] Q. He et al, "Multimodal Multitask Neural Network for Motor Imagery Classification With EEG and fNIRS Signals," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20695–20706, Nov. 2022.

[52] Y. Li, X. Zhang, and D. Ming, "Early-stage fusion of EEG and fNIRS improves classification of motor imagery," *Frontiers in Neuroscience*, Jan. 2023.

[53] L.Maaten and GeoffreyE. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, Jan. 2008.