

Progressive Cross-Modal Attention Network for Brain-Computer Interface Decoding Using EEG and Functional Near-Infrared Spectroscopy (fNIRS)

Yunyuan Gao^{a,*}, Xiaoqi Luo^a, Zhengnan Zhang^a, Xu Shao^b, Yanhua Qin^{b,**}

^a*College of Automation, Hangzhou Dianzi University, Hangzhou, China*

^b*Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China*

Abstract

Electroencephalography (EEG) offers millisecond-level temporal resolution but limited spatial localization, whereas functional near-infrared spectroscopy (fNIRS) provides better spatial resolution but shows a hemodynamic delay of several seconds. A hybrid EEG-fNIRS brain-computer interface fuses the complementary strengths of both modalities but faces challenges from intrinsic spatiotemporal heterogeneity between electrophysiological and hemodynamic signals. To address this, this study proposes a Progressive Cross-Modal Alignment Network (PCMAN), which achieves deep EEG-fNIRS fusion and alignment through a three-layer progressive mechanism. The first layer, physical alignment, employs a Spatiotemporal correlation Dynamic Fusion framework (SDF) with sliding windows to capture hemodynamic delay and uses fNIRS-driven spatial attention to enhance EEG features. The second layer, semantic alignment, builds on physical alignment and applies a Multi-Dimensional Feature Enhancement Attention network (MDFEA) to learn deep cross-modal representations while retaining low-level cues. The third layer, task alignment, employs a Cross-modal Adaptive Decoding Network (CADN) for context-aware fusion, adaptively adjusting strategies based on previous layers. On three cognitive tasks—mental arithmetic (MA), word generation (WG), and motor imagery (MI)—PCMAN achieved accuracies of 88.18%, 81.93%, and 72.94%, outperforming state-of-the-art hybrid BCI methods. Interpretability analysis shows that PCMAN aligns temporal and spatial attention with canonical hemodynamic responses, improving accuracy, robustness, and inter-session transferability

under subject-dependent hybrid BCI settings .

Keywords:

Hybrid brain-computer interface, EEG-fNIRS fusion, Cross-modal attention, Progressive alignment, Multimodal deep learning

1. Introduction

A brain-computer interface (BCI) decodes neural activity to establish a direct communication channel between the brain and external devices, showing great application potential in neurorehabilitation, assistive control, and human-computer interaction [1, 2]. Noninvasive BCI systems, owing to their noninvasiveness, safety, and portability, have become a key direction for clinical translation. Among them, electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) are two of the most widely used neuroimaging modalities[3, 4]. EEG records scalp potential changes reflecting synchronized neuronal discharges, providing millisecond-level temporal resolution to capture rapid neural dynamics[5]. fNIRS measures changes in cortical blood oxygenation based on the tissue penetration of near-infrared light, offering subcentimeter spatial resolution and strong tolerance to motion artifacts[6]. The two modalities provide complementary information about neural activity from electrophysiological and hemodynamic dimensions, and hold promise for breaking through the performance bottlenecks of single-modality systems[7].

Despite the theoretical advantages of multimodal fusion, the practical performance of existing EEG-fNIRS systems remains below expectations. Recent reviews indicate that gains over single-modality baselines are often moderate and strongly protocol-dependent [8, 9] . The root cause is that current methods still struggle to handle intrinsic heterogeneity between the two signals. Temporally, multimodal studies report multisecond hemodynamic delays with clear task dependence [6] , and reproducibility is sensitive to preprocessing choices such as short-channel regression [10] . Spatially, fNIRS sensitivity and localization

*Corresponding author: Yunyuan Gao, gyy@hdu.edu.cn.

**Corresponding author: Yanhua Qin, 21118323@zju.edu.cn.

are affected by optode placement and probe geometry [11, 12], while EEG is constrained by volume conduction [13]. Such fundamental differences in temporal and spatial scales render simple alignment strategies insufficient to fully exploit cross-modal complementarity.

Current fusion architectures still have limitations in handling cross-modal interactions. Early decision-level fusion methods combine only at the classifier output layer and ignore feature-level cross-modal dependencies [14, 9]. Recently proposed deep-learning methods have shown improvements through various architectural designs. FGANet [15] adopts an fNIRS-guided attention mechanism to enhance cross-modal interaction. EFNet [16] employs a dual-stream architecture to process the two modalities separately before final fusion. BiMNC [17] introduces bidirectional Transformers to model long-range dependencies between modalities. EFMLNet [18] explores multi-level progressive learning to improve fusion effectiveness. STANet [19] focuses on temporal alignment between EEG and fNIRS signals. ECA-FusionNet [20] combines efficient channel attention with multimodal feature fusion for hybrid EEG–fNIRS classification. STeCANet [21] introduces a spatio-temporal cross-attention mechanism to capture inter-modal dependencies across both spatial and temporal dimensions. TSFNet [22] proposes a temporal–spatial fusion architecture that jointly models the two modalities through parallel encoding streams. Lin et al. [23] propose subject-specific neurovascular coupling modeling through task-related tensor decomposition, addressing the need for physiologically grounded EEG–fNIRS integration. Chen et al. [24] develop a transfer learning-based hybrid BCI framework with Wasserstein-metric domain selection to improve cross-subject adaptability in clinical rehabilitation. Despite these advances, several fundamental challenges persist: many methods rely on fixed time offsets that cannot accommodate inter-individual variability in hemodynamic delay; dual-stream architectures often miss opportunities for early cross-modal interaction; high computational complexity limits real-time deployment; and insufficient incorporation of physiological constraints results in poor cross-dataset generalization. Although methods such as EFMLNet [18] also adopt multi-level architectures, they mainly perform progressive feature transformation without assigning an explicit cross-modal alignment objective to each stage. In contrast, PCMAN formulates EEG–fNIRS fusion as three coordinated alignment

problems linked through shared optimization objectives, so that higher-level decoding supervision shapes earlier alignment during end-to-end training. PCMAN further introduces explicit physiological constraints through spatial correlation and temporal alignment losses, turning neurovascular consistency into a verifiable training objective rather than an implicit assumption.

In addition, the neural dynamics of cognitive tasks critically affect fusion effectiveness, yet this factor has not received adequate attention. Classical ERD/ERS findings indicate that MI-related discriminative activity is concentrated in short-latency windows [25]. Recent multimodal studies further show that the relative contribution of EEG and fNIRS varies across tasks and subjects, arguing for task-specific fusion design rather than a single universal strategy [26, 27, 28]. Transfer-learning-based hybrid decoders and newly released clinical EEG–fNIRS datasets further underscore the growing need for task-adaptive and physiologically consistent modeling across diverse paradigms and populations [29, 30]. Collectively, these observations indicate that universal, task-agnostic fusion strategies remain a key limitation of current methods.

In summary, existing EEG–fNIRS fusion methods face three major challenges: (1) intrinsic spatiotemporal heterogeneity, (2) insufficient physiological constraints and early cross-modal interaction, and (3) limited task adaptivity. To systematically address these issues, this study proposes the Progressive Cross-Modal Alignment Network (PCMAN), which decomposes the fusion process into three levels—physical alignment, semantic alignment, and task alignment—and achieves end-to-end coordinated optimization through bidirectional information flow. The main contributions of this paper are as follows:

(1) This study introduces a Spatiotemporal Dynamic Fusion (SDF) framework that models hemodynamic delays through a learnable multi-window attention mechanism and incorporates fNIRS spatial information to refine EEG representations.

(2) A Multi-Dimensional Feature Enhancement Attention (MDFEA) module is proposed to enable efficient cross-modal semantic alignment with $O(C^2)$ computational complexity.

(3) A task-adaptive dynamic fusion mechanism is developed to achieve context-aware decoding via multi-branch adaptive weighting.

2. Material and Processing

2.1. Dataset and Experimental setup

PCMAN is evaluated on two public multimodal BCI benchmark datasets. Dataset A [31] contains a word-generation (WG) task, and Dataset B [32] contains motor imagery (MI) and mental arithmetic (MA) tasks. Both datasets are publicly accessible through official TU Berlin resources: Dataset A corresponds to the WG subset of the simultaneous EEG-fNIRS dataset available at https://doc.ml.tu-berlin.de/simultaneous_EEG_fNIRS/, and Dataset B can be accessed at <https://doc.ml.tu-berlin.de/hBCI/>. Both datasets comprise synchronized EEG and fNIRS acquisitions. All methods are tested under identical data splits and evaluation protocols to ensure comparability and fairness.

Dataset A contains a word-generation task and includes 26 healthy participants (9 male, 17 female; age 26.1 ± 3.5 years). EEG signals were recorded from 30 channels at 1000 Hz, with electrode placement following the international 10–5 system. The fNIRS data were sampled at 10.4 Hz, with 16 sources and 16 detectors forming 36 measurement channels. The experiment adopted a block-design paradigm with two conditions: word generation and resting baseline. In the word-generation task, participants first viewed a letter on the screen (2s), followed by 10s of covert generation of words beginning with that letter. The baseline required central fixation with low cognitive load. Intertrial intervals were randomly set to 13–15s. The experiment consisted of three sessions, each containing 20 trials (10 word-generation and 10 baseline; randomized), for a total of 60 trials.

Dataset B contains MI and MA tasks and includes 29 healthy participants (14 male, 15 female; age 28.5 ± 3.7 years). The EEG configuration is similar to Dataset A, with minor adjustments to electrode positions. The fNIRS system operates at 12.5 Hz, with 14 sources and 16 detectors forming 36 channels. This dataset contains three conditions: left/right-hand MI, MA, and baseline. In the MI task, participants, cued by an arrow, imagined opening/closing the indicated hand at 1 Hz; in the MA task, they continuously subtracted a one-digit number from a three-digit number. Each trial contains a 2s cue, 10s task period, and 15–17s random intertrial interval. MI and MA are recorded in separate sessions, with

20 repetitions per task.

These three tasks represent cognitive processes at different temporal scales: MI involves sub-second SMR modulation [25]; MA requires sustained working-memory updating; WG comprises rapid lexical retrieval followed by slower phonological encoding. The experimental paradigms and sensor layouts are illustrated in Figure 1.

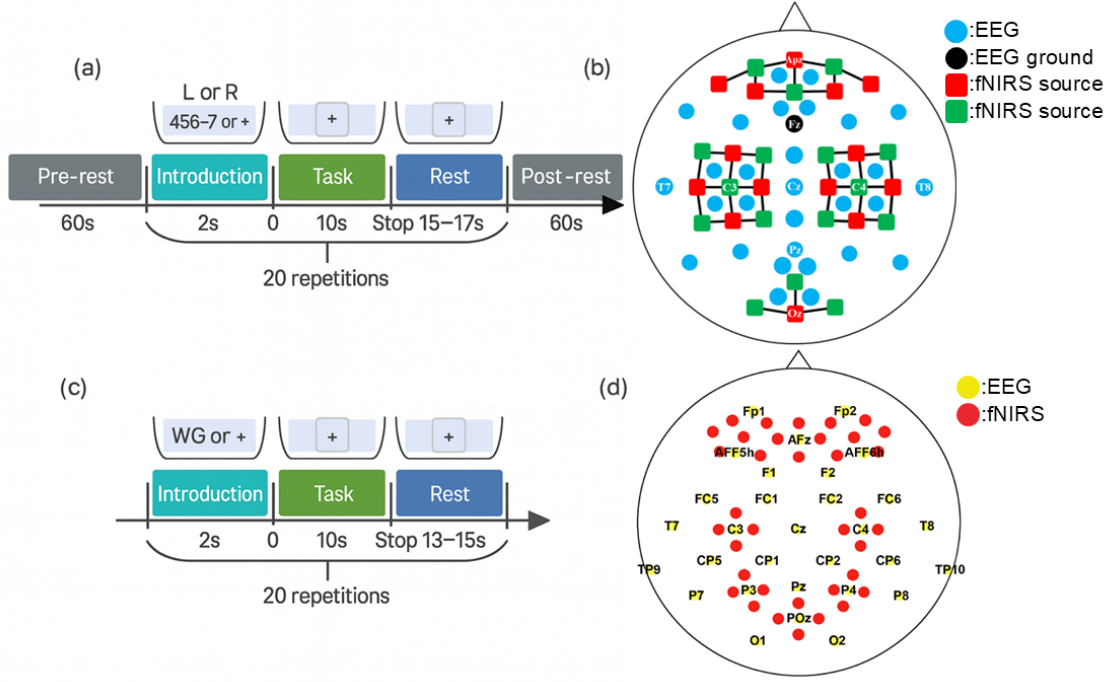


Figure 1: Paradigms for MI, MA, and WG and sensor layouts. Panels (a)/(b) show MI/MA configurations; (c)/(d) correspond to WG.

2.2. Data preprocessing

Multimodal signal preprocessing is critical to effective training of PCMAN. For EEG, the raw signal was downsampled from 1000 Hz to 200 Hz to reduce computational cost and match the time scale of fNIRS, and then a 0.5–50 Hz band-pass filter was applied to remove low-frequency drift and high-frequency noise. Artifacts were removed via 20-component ICA with automated screening, in which ocular components were identified by Fp1/Fp2 spatial weights exceeding mean + 1.5 std and a frontal-to-posterior ratio above 2, muscular components by dominant spectral power in 20–100 Hz relative to 4–13 Hz, and spike-related

components by kurtosis above mean + 3.0 std, with at most 15% of components excluded to preserve neural activity. Finally, a common average reference (CAR) was applied to suppress common-mode noise. The fNIRS preprocessing pipeline was designed around the characteristics of hemodynamic signals. For the 29-subject MI/MA dataset, raw optical-density signals were converted to oxy- and deoxy-hemoglobin (HbO/HbR) concentration changes via the modified Beer–Lambert law, and 36 HbO channels and 36 HbR channels were extracted via even/odd indexing; for the 26-subject WG dataset, this conversion had been completed. Both datasets then used a 6th-order zero-phase Butterworth band-pass filter to effectively remove high-frequency physiological noise and low-frequency drift. Baseline correction used the pre-task baseline for signal normalization. Given the generally small amplitude of fNIRS signals, both datasets scaled the signals by 1e3 so that their value ranges matched those of EEG.

2.3. EEG-fNIRS Tensor representation design

After preprocessing, EEG and fNIRS signals have different physical properties and spatiotemporal resolutions. Designing a unified data representation that fully exploits the complementarities of the two modalities is key to multimodal BCI research. This study proposes a novel multimodal tensor representation framework. For EEG, this study adopts a 4D tensor representation:

$$X_{\text{EEG}} \in \mathbb{R}^{H \times W \times T \times C} \quad (1)$$

where H is spatial height, W is spatial width, T is temporal samples, and C is the number of channels. The core innovation is to transform the traditional 1D electrode arrangement into a 2D spatial grid. By azimuthal-equidistant projection, the 3D scalp electrode positions are mapped onto a regular 2D plane, preserving the spatial topology among electrodes. The temporal dimension contains multiple samples to capture the dynamic course of cognitive activity.

The fNIRS representation fully considers the delay characteristics of the hemodynamic response and adopts a 5D tensor structure:

$$X_{\text{fNIRS}} \in \mathbb{R}^{N \times H \times W \times T \times M} \quad (2)$$

where N is the number of temporal windows, H is spatial height, W is spatial width, T is samples within a window, and M is the hemoglobin type. The key innovation is the introduction of a multi-window temporal dimension. Each window has a fixed length and slides with a fixed stride, covering from task onset into the post-task period to capture the full hemodynamic response. The spatial dimensions match EEG for cross-modal feature fusion, and the last dimension corresponds to HbO and HbR.

To construct a complete spatial representation of brain signals, discrete EEG and fNIRS signals are converted into continuous 3D tensor representations. Using cubic-spline interpolation, discrete electrode signals are mapped to a regular grid to form a continuous spatial representation. The proposed multimodal tensor representation framework is illustrated in Figure 2, which depicts both the spatial interpolation process and the channel layouts for the two datasets. This unified tensor framework preserves the spatial topology of brain signals, allowing convolution-based feature extraction to exploit spatial locality; the multi-window design enables the model to learn complex temporal relationships between fast EEG electrophysiology and slow fNIRS hemodynamics; and the standardized tensor format simplifies the implementation of multimodal fusion, laying the foundation for end-to-end deep-learning models.

2.4. Data segmentation

To capture time-varying dynamics, both datasets use a sliding-window segmentation method: a 3s window with a 1s stride (67% overlap). EEG analysis windows span from 5s pre-task to task end (15s total). fNIRS analysis windows span from 5s pre-task to 12s post-task (27s total), fully covering preparation, execution, and recovery. Temporal correspondence between EEG and fNIRS is ensured by synchronized triggers. Within each sliding window, EEG and fNIRS features from the same time window are paired to form a complete hybrid feature vector. To improve inter-session robustness and address the high variability of brain signals, the following strategies were applied: merging multiple sessions

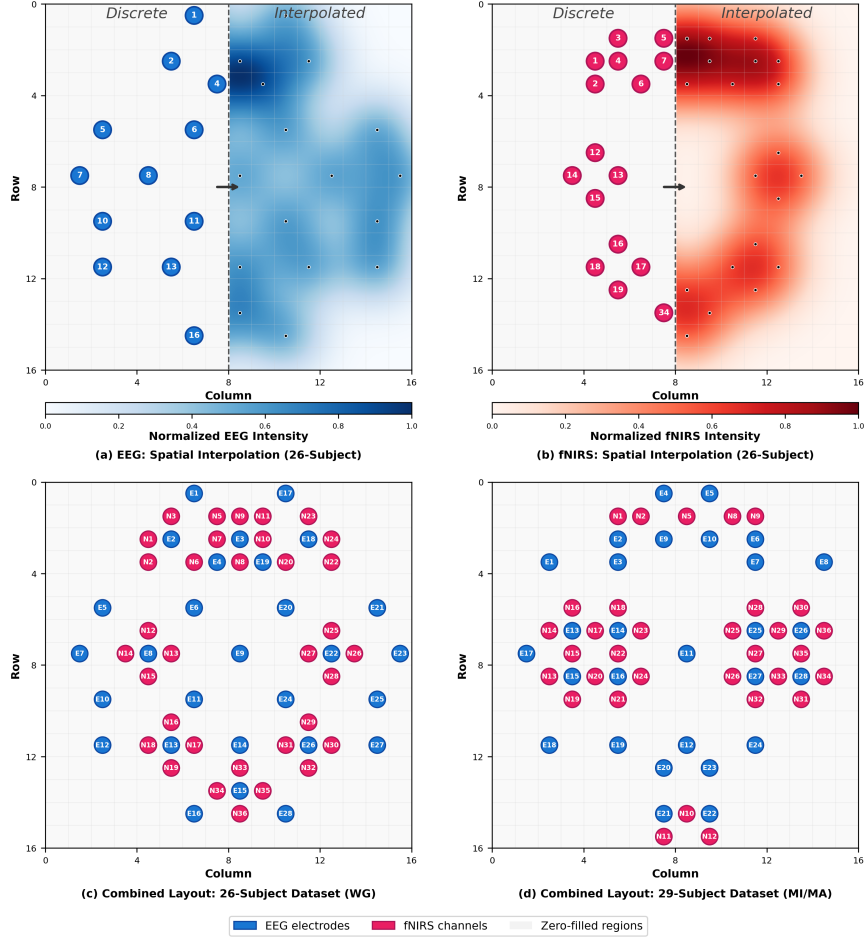


Figure 2: Spatial interpolation and channel layout. (a)–(b) Discrete electrode positions (left) and cubic-spline interpolated distributions (right) for EEG and fNIRS on a 16×16 grid. (c)–(d) Combined EEG (blue) and fNIRS (pink) layouts for the 26-subject and 29-subject datasets. Gray areas denote zero-filled regions.

per subject, generating more samples via high-overlap sliding windows [33], and extending analysis intervals. These strategies effectively increased the number and diversity of training samples.

3. Method

This study proposes the Progressive Cross-Modal Alignment Network (PCMAN) to address spatiotemporal heterogeneity in EEG-fNIRS hybrid BCI. As shown in Figure 3, PCMAN adopts a three-layer progressive architecture that integrates three core modules for deep multimodal fusion: the Spatiotemporal correlation Dynamic Fusion framework (SDF),

the Multi-Dimensional Feature Enhancement Attention network (MDFEA), and the Cross-modal Adaptive Decoding Network (CADN). SDF establishes spatiotemporal correspondence between modalities at the physical level, addressing the time-scale disparity between millisecond-scale electrophysiology and second-scale hemodynamics; MDFEA learns discriminative cross-modal representations at the semantic level; CADN dynamically adjusts the fusion strategy based on signal quality and task characteristics to produce the final classification. The progressive design lets each layer focus on a specific dimension of heterogeneity while end-to-end training ensures global optimization.

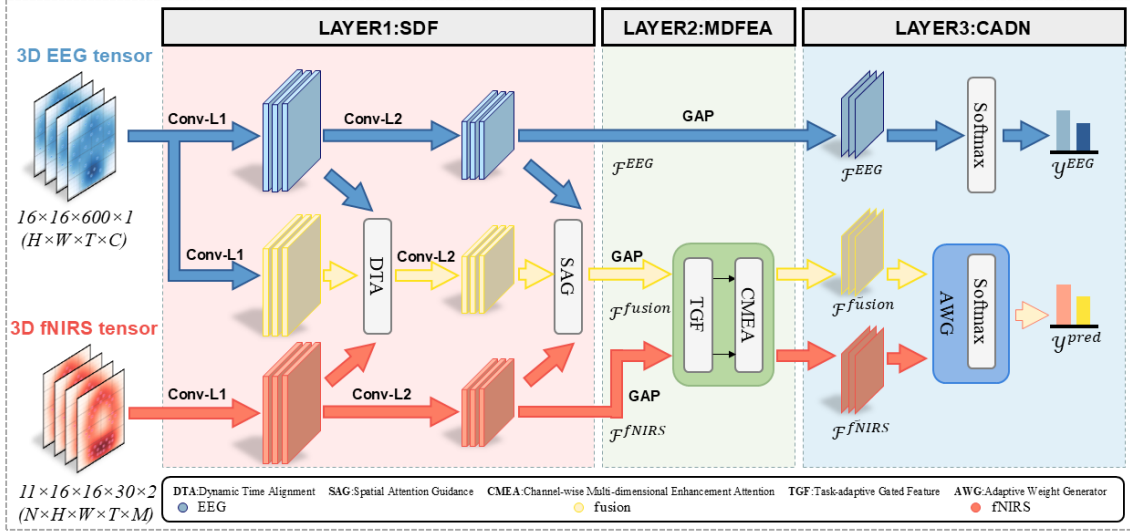


Figure 3: Overall PCMAN framework. LAYER 1 (SDF) generates three streams: a yellow fused path (aligned by DTA and SAG) and a red fNIRS path feeding LAYER 2 (MDFEA) for dual-stream enhancement, while the blue EEG path bypasses MDFEA to LAYER 3 (CADN). MDFEA outputs enhanced features for CADN’s fusion and fNIRS branches, where the Adaptive Weight Generator (AWG) performs dynamic weighting.

3.1. SDF—Spatiotemporal Dynamic Fusion

SDF is the core module enabling physical alignment in PCMAN, designed to tackle the fundamental spatiotemporal heterogeneity between EEG and fNIRS. Unlike existing methods that use fixed time offsets or simple feature concatenation, SDF achieves adaptive cross-modal physical alignment via two synergistic mechanisms: dynamic time alignment and spatial-attention-guided fusion. Its overall architecture is shown in Figure 4 .

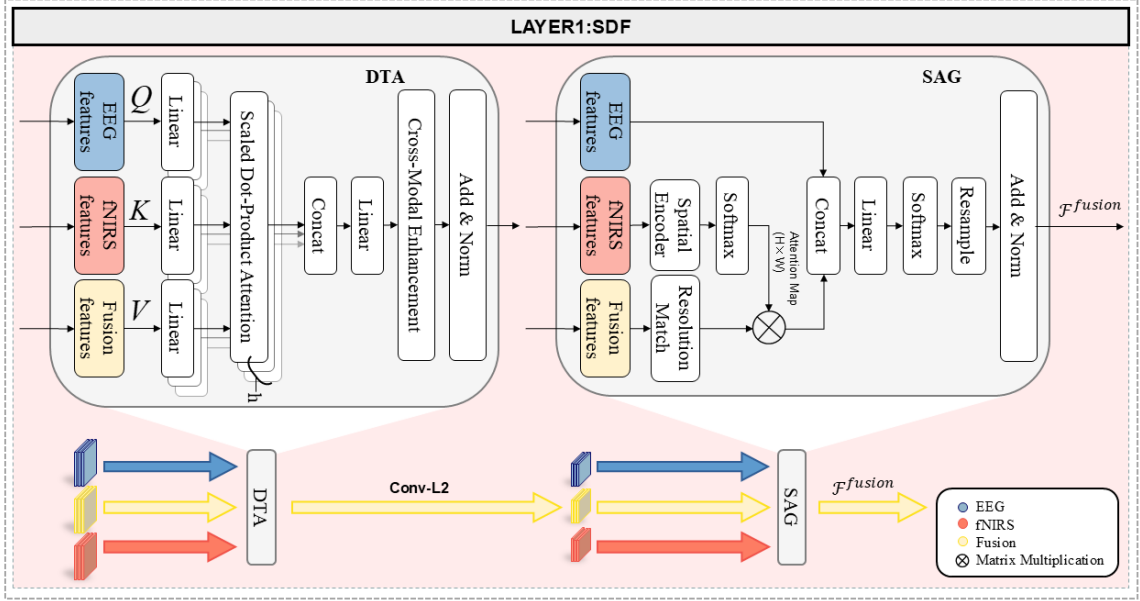


Figure 4: structure of LAYER1:SDF.

3.1.1. Dynamic time alignment (DTA)

The primary challenge in EEG–fNIRS fusion is the time-scale disparity between the signals. To address this, SDF replaces fixed compensation with dynamic alignment via a multi-head attention mechanism. The core idea is to let the network automatically learn the optimal correspondence between EEG features at the current time and fNIRS features at different delayed times. Specifically, given EEG features and fNIRS sequence features, the spatiotemporal alignment first performs feature projection. EEG features serve as queries, representing current neural activity; fNIRS multi-window features serve as key–value pairs, representing hemodynamic responses at different delays:

$$Q = \text{vec}(F_{\text{EEG}})W_Q, \quad W_Q \in \mathbb{R}^{(HWC) \times d} \quad (3)$$

$$K = [F_{\text{fNIRS}}^1; \dots; F_{\text{fNIRS}}^{11}] W_K \quad (4)$$

$$V = [F_{\text{fNIRS}}^1; \dots; F_{\text{fNIRS}}^{11}] W_V \quad (5)$$

The temporal-window configuration was designed to reflect the canonical time course of event-related hemodynamic responses. Prior studies have shown that the response typically begins about 2 s after neural activity and reaches its main peak around 4.5–6 s [34, 35], while the exact latency varies across subjects and modalities. Accordingly, a 3-second window with a 1-second stride was adopted to capture local hemodynamic evolution, and 11 windows were used to cover approximately 0–13 s after task onset, spanning the main rise, peak, and later response interval including the post-stimulus undershoot [34, 35]. This physiological rationale motivates dynamic rather than fixed temporal alignment, and its practical validity is further supported by the sensitivity analysis in Figure 13, where fewer windows under-cover the effective response interval whereas excessive windows introduce redundant noise. Learnable positional encodings, unlike fixed encodings in standard Transformers, allow the model to adapt the time representation to the data, capturing task- and subject-specific delay patterns.

The design of multi-head attention [36] has physiological significance: different heads learn different temporal preferences, enabling parallel learning of multiscale correspondences. Individual heads can focus on fast, typical, or slow hemodynamic responses, ensuring adaptation to different physiological characteristics.

$$\text{Attention}^{(i)} = \text{softmax} \left(\frac{Q^{(i)} (K^{(i)})^T}{\sqrt{d_h}} \right) \quad (6)$$

$$F_{\text{aligned}} = \text{Attention} \cdot V \quad (7)$$

A cross-modal enhancement mechanism further optimizes spatiotemporal alignment:

$$F_{\text{cross}} = \text{CrossAttention} \left(\text{Reshape} \left(Q, \left[\sqrt{d}, \sqrt{d} \right] \right), \text{Reshape} \left(\text{MultiHead}, \left[\sqrt{d}, \sqrt{d} \right] \right) \right) \quad (8)$$

A 2D reshape exploits spatial structure, and local convolutions enable fine-grained feature interaction.

To support dynamic time alignment, SDF’s base feature extractor adopts a parallel-branch architecture. Each extractor block integrates information progressively through three branches: an EEG branch that captures fast electrophysiological dynamics, a fNIRS branch that extracts slow hemodynamic features, and a fusion branch that realizes initial cross-modal interaction under guided attention. The feature transform of the ℓ -th layer is:

$$H_i^l = \text{Conv3D}(\text{BN}(H_i^{l-1})) + \text{Skip}(H_i^{l-1}), \quad i \in \text{EEG, fNIRS, fusion} \quad (9)$$

where σ denotes the activation function, BN denotes batch normalization, and residual connections ensure stable gradient propagation.

Kernels are tailored to signal characteristics. The EEG branch uses longer temporal kernels to capture transient events such as ERPs; the fNIRS branch uses shorter temporal kernels to accommodate slow hemodynamics; the fusion branch lies in between to bridge the two (Table 1). After two stages, each branch’s features are downsampled to compact representations, reducing computation while retaining key spatiotemporal patterns for subsequent spatial-attention guidance.

Table 1: Feature extractor parameter configuration

Layer	Branch	Kernel	Stride	Out-Channels
L1	EEG	(2,2,13)	(2,2,6)	16
L1	fNIRS	(2,2,5)	(2,2,2)	16
L1	Fusion	(2,2,13)	(2,2,6)	16
L2	EEG	(2,2,5)	(2,2,2)	32
L2	fNIRS	(2,2,3)	(2,2,2)	32
L2	Fusion	(2,2,5)	(2,2,2)	32

3.1.2. Spatial-attention-guided fusion (SAG)

Although EEG has millisecond-level temporal advantages, its spatial localization is limited. The Guided-Attention-Fusion module exploits the spatial resolution of fNIRS to guide multimodal fusion via four cascaded steps:

(1) Attention-map generation. A multi-level aggregation strategy generates a spatial attention map from fNIRS’s multi-channel features. A single-channel output (filters=1)

achieves channel pooling; the temporal kernel is adjusted by level. Cascaded aggregation over temporal and channel dimensions extracts global spatial patterns, ensuring that the attention map preserves spatial specificity while remaining temporally stable.

(2) Resolution matching. Bilinear interpolation resolves spatial-resolution differences between fNIRS and EEG, ensuring smooth transitions in the attention map and avoiding blocky artifacts.

(3) Adaptive fusion. Fusion combines attention weighting with adaptive residual learning to ensure robustness:

$$H_{\text{output}} = A' \odot H_{\text{fusion}} + H_{\text{residual}} + \text{Sigmoid}(\gamma) \times H_{\text{EEG}} + (1 - \text{Sigmoid}(\gamma)) \times H_{\text{fusion}} \quad (10)$$

where γ is a learnable residual weight. This design lets the network dynamically balance the relative contributions of the direct EEG path and the fusion path.

(4) Correlation regularization. A correlation loss imposes neurophysiological constraints, ensuring that learned spatial attention is significantly correlated with EEG activation patterns, guiding learning while retaining flexibility.

The cross-modal interaction module realizes bidirectional information exchange between EEG and fNIRS. Given complementary spatial patterns across modalities, cross-attention mutually enhances features. ReLU (rather than ELU) is used because subsequent element-wise products require nonnegativity to preserve the physical meaning of attention. 3×3 convolutions extract local context for attention computation. A similarity-driven mechanism leverages strong co-activation at the same spatial locations across modalities as an indicator of task relevance. Residual connections ensure robust information flow, allowing raw information to pass through even if attention fails.

SDF outputs three feature sets: (1) Fused features that have undergone both DTA and SAG, encoding fast EEG dynamics, precise fNIRS localization, and their spatiotemporal correspondences, for deep semantic learning in MDFEA; (2) fNIRS-aligned features that retain spatial resolution advantages for direct input to CADN’s spatially enhanced branch; and (3) EEG baseline features as a single-modality performance benchmark for direct input

to CADN’s auxiliary branch.

3.2. MDFEA—Multi-Dimensional Feature Enhancement Attention

MDFEA receives SDF’s fused features $\mathcal{F}_{fusion}$ and spatially informative fNIRS features $\mathcal{F}_{fNIRS}$ and performs deep semantic learning on top of physical alignment, constituting the second layer of progressive alignment. The MDFEA adopts a dual-stream parallel processing architecture, which respectively enhances the coupled expression between fused features and fNIRS feature channels, thereby improving the semantic representation capability. Its overall architecture is shown in Figure 5, using TGF and CMEA to enhance each input and output two enhanced feature sets for CADN’s fusion $\mathcal{F}_{fusionenhanced}$ and fNIRS branches $\mathcal{F}_{fNIRSenhanced}$, respectively. By contrast, EEG baseline features bypass MDFEA and go directly to CADN to preserve their role as a single-modality benchmark. The design is inspired by Restormer-style architectures in image restoration and customized for brain-signal processing[37]. Each channel of brain signals corresponds to a functional region; inter-channel interaction patterns reflect functional connectivity. MDFEA achieves deep enhancement via channel-dimension multidimensional enhanced attention (CMEA) and a task-gated feature (TGF) network.

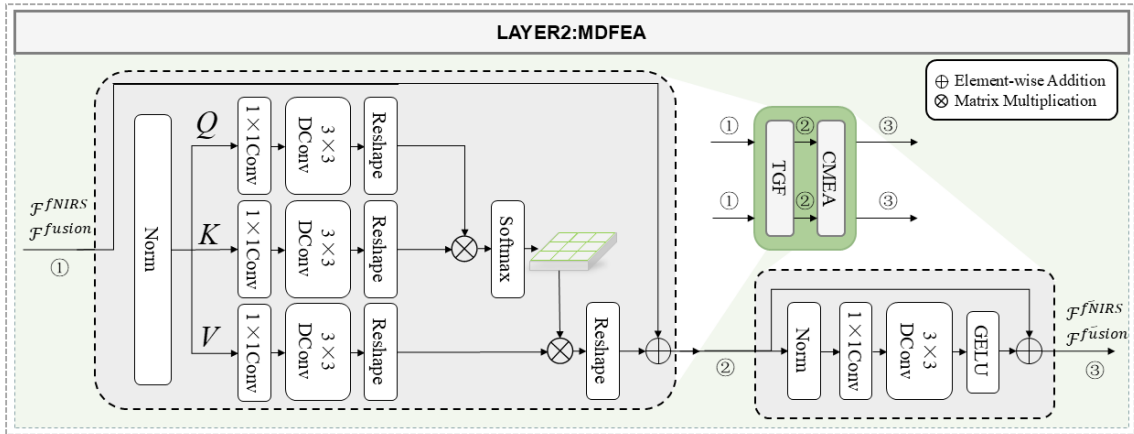


Figure 5: structure of LAYER2:MDFEA.

3.2.1. Channel-Dimension Multidimensional Enhanced Attention (CMEA)

The computational complexity of traditional spatial self-attention grows quadratically with resolution, creating a bottleneck for high-resolution brain-map processing. CMEA transposes attention to the channel dimension, preserving representational power while improving efficiency. Given brain-signal characteristics, inter-channel functional connectivity often carries more task-relevant information than local spatial pixel correlations.

CMEA enhances the alignment features obtained from the SDF via a three-stage processing procedure. After layer normalization, 1×1 convolutions perform channel aggregation to generate initial Q, K, V projections. Depthwise separable convolutions then enhance local context: depthwise convolutions apply spatial filtering per channel, capturing neighborhood information while preserving channel independence. Because brain-signal spatial patterns exhibit locally continuous activation, depthwise convolutions effectively encode such continuity, providing rich context for global inter-channel interaction.

CMEA’s core is transposed attention. After reshaping Q, K, V, inter-channel attention is computed via a transposed dot product:

$$A_{\text{channel}} = \text{Softmax} \left(\frac{K \cdot Q}{\sqrt{C}} \right) \quad (11)$$

The attention map encodes global inter-channel correlations and captures cross-modal, cross-region functional coordination in EEG–fNIRS fused features. Enhanced features are obtained via attention weighting with residual connections to ensure gradient flow:

$$X_{\text{enhanced}} = \text{Reshape}(V \cdot A_{\text{channel}}) + X \quad (12)$$

CMEA’s complexity is $O(C^2)$, compared with $O(HW \cdot HW)$ for spatial attention; when spatial resolution is much larger than channel count, this yields significant efficiency gains, enabling real-time operation.

Fused and fNIRS features both enter CMEA; the two streams share the same architecture but have independent parameters to learn targeted inter-channel dependencies. The fusion stream emphasizes cross-modal coordination; the fNIRS stream focuses on preserving and

enhancing spatial localization. After CMEA, the two streams output:

$$F_{\text{fusion}}^{\text{CMEA}} = \text{CMEA}_{\text{fusion}}(F_{\text{fusion}}) \quad (13)$$

$$F_{\text{fNIRS}}^{\text{CMEA}} = \text{CMEA}_{\text{fNIRS}}(F_{\text{fNIRS}}) \quad (14)$$

3.2.2. Task-Gated Feature (TGF)

After CMEA establishes global inter-channel dependencies, TGF performs nonlinear transformation and selective information routing. Inspired by neural gating, TGF dynamically modulates information flow to enhance task-relevant components and suppress irrelevant ones. For EEG–fNIRS fused features, this mechanism adaptively adjusts the contributions of different modality components according to task type.

TGF likewise adopts a dual-stream parallel architecture to process the two CMEA outputs. Ultimately, MDFEA outputs two independent enhanced feature vectors:

$$F_{\text{fusion}}^{\text{enhanced}} = \text{GAP}(\text{TGF}_{\text{fusion}}(F_{\text{fusion}}^{\text{CMEA}})) \in \mathbb{R}^{256} \quad (15)$$

$$F_{\text{fNIRS}}^{\text{enhanced}} = \text{GAP}(\text{TGF}_{\text{fNIRS}}(F_{\text{fNIRS}}^{\text{CMEA}})) \in \mathbb{R}^{256} \quad (16)$$

These are input to CAD’s fusion and fNIRS branches to achieve fine-grained, semantic-level fusion. Channel expansion provides larger representational capacity; depthwise convolutions continue to encode local context. Gating signals produced via GELU yield smooth nonlinear transformation with gradients near zero, better approximating gradual neuronal activation. Element-wise multiplication implements soft selection, assigning adaptive weights per feature for task-driven filtering.

TGF provides three key functions: dynamic feature selection to adjust weights according to input content; task-specific adaptation to activate differentiated gating patterns across tasks; and residual paths to optimize gradient propagation and ease deep-network training.

3.3. CADN—Cross-Modal Adaptive Decoding

The Cross-modal Adaptive Decoding Network (CADN) serves as the top-level module of PCMAN, designed to integrate three groups of features from SDF and MDFEA: progressively aligned fusion features, semantically enhanced fNIRS features, and baseline EEG features without deep processing. Because the reliability and importance of each modality vary dynamically across different cognitive tasks, time phases, and signal quality conditions, CADN employs a dual-path weighted fusion architecture combined with an auxiliary EEG supervision mechanism to achieve flexible and adaptive multimodal integration. The detailed network configuration and branch structure of CADN are summarized in Table 2.

Table 2: CADN architecture

Branch	Input features	Network	Output
Fusion	$F_{fusion}^{enhanced}$	FC(256→256)→ELU→ FC(256→2)→Softmax	$\hat{y}_{fusion}$
fNIRS spatially enhanced	$F_{fNIRS}^{enhanced}$	FC(256→256)→ELU→ FC(256→2)→Softmax	$\hat{y}_{fNIRS}$
Adaptive Weight Generator (AWG)	$[F_{fusion}^{enhanced}, F_{fNIRS}^{enhanced}]$	FC(512→256)→ELU→ FC(256→2)→Softmax	$[W_{fusion}, W_{fNIRS}]$
EEG baseline	$F_{EEG}(H \times W)$	Flatten→FC(→256)→ELU→ FC(256→2)→Softmax	$\hat{y}^{EEG}$

CADN consists of three parallel branches. The fusion branch receives the fully aligned features from MDFEA and generates the main prediction vector through nonlinear transformations. The fNIRS branch independently processes the temporally aligned fNIRS features, utilizing their high spatial resolution to maintain system performance when EEG quality decreases. The EEG branch retains the unimodal predictive ability of raw EEG features, serving both as a reference for evaluating fusion gain and as a fail-safe mechanism when multimodal fusion fails.

At the core of the dynamic weighting mechanism lies the Adaptive Weight Generator (AWG). By concatenating the enhanced fusion and fNIRS features, AWG automatically learns task-dependent scalar weights to determine the contribution of each branch to the final prediction:

$$y^{\text{pred}} = w_{\text{fusion}} \cdot y_{\text{fusion}} + w_{\text{fNIRS}} \cdot y_{\text{fNIRS}} \quad (17)$$

Unlike fixed or empirically assigned weights, AWG adaptively optimizes the fusion strategy according to the feature distribution, offering three advantages: context awareness, lightweight design, and interpretability. During training, the evolution of the learned weights exhibits clear task-specific patterns: in motor imagery tasks, fusion weights gradually increase as the hemodynamic response develops; in mental arithmetic tasks, weights remain relatively stable; and in word generation tasks, a bimodal distribution appears, corresponding to lexical retrieval and phonological encoding stages.

The EEG branch is used only during training, where it provides auxiliary loss supervision that quantifies the unimodal baseline and prevents the optimizer from over-relying on the slower hemodynamic pathway. During inference, only the AWG-weighted fusion output is used. The EEG branch is not included in the final decision because its information has already been integrated into the fusion features through SDF and MDFA; adding it again at the decision level would introduce redundancy rather than new complementary evidence.

More than a simple decision layer, CADN functions as the coordination center of the entire progressive alignment framework. Through backpropagation, its task loss guides the optimization direction of preceding layers, ensuring that both physical and semantic alignments evolve toward improving final decoding performance. Experimental results show that the dynamic weighting mechanism significantly outperforms fixed-weight fusion, and under motion artifacts, the system automatically reduces the weight of the affected modality, maintaining stable decoding performance.

3.4. Loss Function Design

PCMAN’s optimization objective adopts a multi-task learning framework with hierarchical supervision to coordinate progressive alignment. The total loss contains four complementary terms corresponding to different levels of the fusion pipeline, and their respective weights and roles are summarized in Table 3.

$$L_{\text{total}} = L_{\text{class}} + \lambda_{\text{eeg}}L_{\text{eeg}} + \lambda_{\text{corr}}L_{\text{corr}} + \lambda_{\text{align}}L_{\text{align}} \quad (18)$$

The main classification loss L_{class} uses cross-entropy to supervise the AWG-weighted final

Table 3: Loss weight configuration

Term	Weight	Role	Objective
L_{class}	1.0	Task supervision	Maximize accuracy
L_{eeg}	0.5	Auxiliary supervision & regularization	Prevent overreliance on fNIRS
L_{corr}	0.1	Physiological constraint	Ensure biological plausibility
L_{align}	0.1	Temporal-alignment constraint	Promote cross-modal consistency

output y^{pred} , ensuring end-to-end optimization:

$$L_{\text{class}} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^C y_{ij} \log(y_{ij}^{\text{pred}}) \quad (19)$$

The auxiliary EEG loss L_{eeg} independently supervises the EEG baseline branch y^{EEG} :

$$L_{\text{eeg}} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^C y_{ij} \log(y_{ij}^{\text{EEG}}) \quad (20)$$

where B is batch size, C is the number of classes, and y_{ij} denotes the ground-truth label. L_{eeg} has two functions: providing an independent single-modality baseline to quantify fusion gain, and acting as a regularizer to prevent overreliance on slow hemodynamics in the main path. L_{eeg} is used only during training; at inference the EEG branch does not contribute to the final prediction.

A correlation loss L_{corr} introduces physiological constraints to ensure that SDF’s learned spatial-attention maps have neurobiological plausibility:

$$L_{\text{corr}} = \frac{1}{L} \sum_{l=1}^L (1 - \rho_l) \quad (21)$$

where ρ_l is the Pearson correlation between the ℓ -th layer’s spatial attention and EEG activation. This soft constraint prevents attention from degenerating into arbitrary distributions while preserving task-specific flexibility. A temporal alignment loss L_{align} maximizes cross-modal correlation after alignment to promote semantic consistency:

$$L_{\text{align}} = 1 - \frac{1}{B} \sum_{i=1}^B \rho(Q_i, A_i V_i) \quad (22)$$

where Q_i is the query vector, A_i is the attention weights, and V_i is the value matrix. The constraint losses L_{corr} and L_{align} are weighted at 0.1, compared with 1.0 for the main classification loss, so they act as soft regularizers rather than hard constraints. They are not intended to force the learned attention to follow a fixed physiological template exactly. Instead, they discourage solutions that depart too far from known neurovascular structure. This design still allows moderate deviations in timing or spatial correspondence when these are needed to capture subject-specific, task-specific, or region-specific variation. In practice, the learned attention on held-out data remains physiologically meaningful, with temporal attention aligned with the canonical HRF envelope and spatial attention concentrated in task-relevant cortical regions. In this sense, the constraint terms serve as mild priors that improve biological plausibility without restricting discriminative optimization too strongly.

4. Experimental Results and Analysis

4.1. Experimental Setup

This study follows the benchmark datasets and segmentation strategy described in Section 2. Considering the balanced distribution of samples and the binary classification nature of the tasks, accuracy (ACC) and Cohen’s kappa coefficient are used as the primary performance evaluation metrics. The Kappa coefficient is calculated as follows:

$$\text{Kappa} = \frac{\text{acc} - p_0}{1 - p_0} \quad (23)$$

where acc denotes the observed accuracy, and p_0 is the probability of chance agreement.

This study adopts a subject-dependent evaluation protocol. Each subject’s model is trained and tested exclusively on that subject’s own data via Leave-One-Session-Out (LOSO) cross-validation, where one session serves as the test set in each iteration. Supplementary cross-subject experiments are reported in Section 4.8. A two-stage training process is used in each iteration: the training set is split into training and validation sets in a 4:1 ratio for model fitting and hyperparameter tuning. An early stopping mechanism is employed to mitigate overfitting.

All models are trained using the Adam optimizer with an initial learning rate of 1×10^{-3} , adjusted dynamically using cosine annealing. The batch size is set to 32, the maximum number of epochs to 200, and the early-stopping patience to 20. He initialization is used for model weights. Experiments are conducted using PyTorch 1.12 on a workstation equipped with an NVIDIA RTX 4060 GPU. To ensure a fair comparison, all baseline models received the same preprocessed data from the pipeline in Section 2.2, including the identical 4D EEG and 5D fNIRS tensors produced by cubic-spline interpolation, the same data augmentation strategy, the same subject-dependent LOSO splits, and individually tuned hyperparameters on the validation set.

4.2. Performance Comparison Analysis

Table 4 presents the overall performance (mean $\pm$ standard deviation) of PCMAN and 12 comparative methods across three tasks: Motor Imagery (MI), Mental Arithmetic (MA), and Word Generation (WG). The methods include unimodal models (e.g., EEG-only, fNIRS-only) and various hybrid fusion structures. All models are tested using identical data splits and evaluation protocols to ensure fairness and comparability.

Table 4: Overall performance (mean $\pm$ std)

Methods	Signal	MI		MA		WG	
		Accuracy (%)	Kappa	Accuracy (%)	Kappa	Accuracy (%)	Kappa
Conformer [38]	EEG	65.89 $\pm$ 12.96	0.32 $\pm$ 0.26	82.98 $\pm$ 9.05	0.66 $\pm$ 0.18	77.88 $\pm$ 9.48	0.56 $\pm$ 0.19
LMDA [39]	EEG	65.17 $\pm$ 12.41	0.30 $\pm$ 0.25	81.68 $\pm$ 8.79	0.63 $\pm$ 0.18	75.75 $\pm$ 10.27	0.52 $\pm$ 0.21
fNIRSNet [40]	fNIRS	62.63 $\pm$ 6.26	0.25 $\pm$ 0.13	73.79 $\pm$ 9.61	0.48 $\pm$ 0.19	67.78 $\pm$ 7.95	0.36 $\pm$ 0.16
fNIRS-Transformer [41]	fNIRS	59.83 $\pm$ 5.52	0.20 $\pm$ 0.11	69.50 $\pm$ 8.32	0.39 $\pm$ 0.17	64.56 $\pm$ 6.93	0.29 $\pm$ 0.14
FGANet [15]	EEG+fNIRS	64.95 $\pm$ 8.65	0.30 $\pm$ 0.17	76.58 $\pm$ 9.00	0.53 $\pm$ 0.18	71.33 $\pm$ 9.01	0.43 $\pm$ 0.18
EF-Net [16]	EEG+fNIRS	61.80 $\pm$ 8.04	0.24 $\pm$ 0.16	75.59 $\pm$ 9.55	0.51 $\pm$ 0.19	68.51 $\pm$ 9.30	0.37 $\pm$ 0.19
BiMNC [17]	EEG+fNIRS	67.25 $\pm$ 7.37	0.34 $\pm$ 0.15	82.36 $\pm$ 9.24	0.65 $\pm$ 0.18	76.86 $\pm$ 10.22	0.54 $\pm$ 0.20
M2NN [33]	EEG+fNIRS	66.21 $\pm$ 9.85	0.32 $\pm$ 0.20	82.72 $\pm$ 7.30	0.65 $\pm$ 0.15	76.07 $\pm$ 9.87	0.52 $\pm$ 0.20
EFMLNet [18]	EEG+fNIRS	66.92 $\pm$ 8.99	0.34 $\pm$ 0.18	83.14 $\pm$ 7.64	0.66 $\pm$ 0.15	77.49 $\pm$ 9.05	0.55 $\pm$ 0.18
Dual-EEGNet [42]	EEG+fNIRS	66.22 $\pm$ 7.86	0.32 $\pm$ 0.16	78.27 $\pm$ 8.22	0.57 $\pm$ 0.16	73.53 $\pm$ 10.14	0.47 $\pm$ 0.20
E-FNet [43]	EEG+fNIRS	64.91 $\pm$ 7.51	0.30 $\pm$ 0.15	83.66 $\pm$ 7.87	0.67 $\pm$ 0.16	76.74 $\pm$ 9.89	0.53 $\pm$ 0.20
STA-Net [19]	EEG+fNIRS	69.65 $\pm$ 9.52	0.39 $\pm$ 0.19	85.14 $\pm$ 7.17	0.70 $\pm$ 0.14	79.03 $\pm$ 8.41	0.58 $\pm$ 0.17
PCMAN	EEG+fNIRS	72.94 $\pm$ 11.71	0.46 $\pm$ 0.23	88.18 $\pm$ 7.28	0.76 $\pm$ 0.15	81.93 $\pm$ 8.29	0.64 $\pm$ 0.17

From both accuracy and Kappa coefficient perspectives, PCMAN achieves the best performance across all tasks, consistently outperforming state-of-the-art methods such as STA-Net and BiMNC. This demonstrates the effectiveness, robustness, and inter-session transfer capability of the proposed progressive cross-modal alignment strategy.

Further analysis attributes the performance improvement to three key mechanisms:

(1) Dynamic Temporal Alignment (SDF) effectively addresses variability in hemodynamic response delay under the subject-dependent LOSO setting .

(2) Spatial Attention-Guided Fusion leverages fNIRS’s superior spatial resolution to compensate for EEG’s limitations in source localization.

(3) The three-tiered progressive alignment structure (physical–semantic–task) enables hierarchical enhancement of cross-modal information.

Compared with STA-Net’s dual-branch parallel alignment architecture, PCMAN introduces inter-layer information flow and interaction, significantly improving fusion depth and representational capacity. Feature visualization further validates this: as the network deepens, inter-class distances increase and intra-class variance decreases, highlighting the robustness of the multi-layer fusion structure. Figure 6 provides a comprehensive performance comparison across the three cognitive tasks, clearly demonstrating PCMAN’s superiority over all baseline methods in terms of both accuracy and Cohen’s κ .

4.3. Ablation Study

To systematically evaluate the individual and combined contributions of key PCMAN modules, multiple ablation experiments were conducted. Figure 7 shows performance variations under different module combinations, and Table 5 summarizes the detailed accuracy and Kappa changes for each configuration.

The results show that removing the Spatial Attention-Guided (SAG) module leads to a significant performance decline across all tasks, particularly for MI (-6.6%, $p < 0.001$), indicating its critical role in enhancing cross-modal spatial interaction for fast sensorimotor decoding. Attention visualization results demonstrate more focused spatial attention after SAG integration, validating its feature enhancement effect.

Removing the DTA module results in a substantial performance decline (average -4.5%), particularly in WG tasks, confirming its critical role in aligning delayed hemodynamic signals between modalities. Kappa drops consistently with accuracy across all three tasks, indicating that performance degradation is not metric-specific. In contrast, removing the MDFEA module causes a smaller accuracy drop (-2.8%) but significantly degrades feature

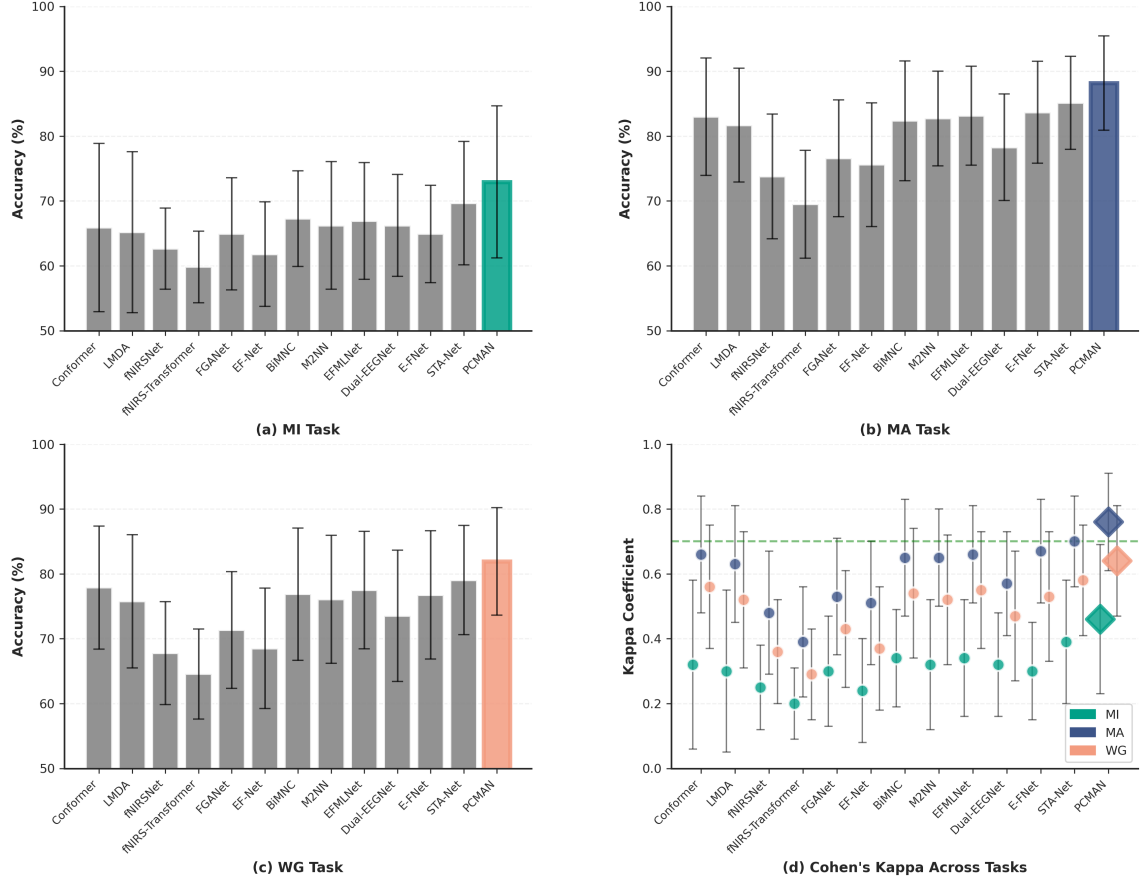


Figure 6: Performance comparison across three cognitive tasks (MI, MA, and WG). PCMAN consistently outperforms all baseline models in both accuracy and Cohen’s κ . Error bars denote ± 1 standard deviation (STD) across subjects and sessions.

Table 5: Module-level ablation results of PCMAN across MI, MA, and WG under the subject-dependent LOSO protocol (Accuracy and Cohen’s κ).

Ablation Type	MI Acc (%)	MI Kappa	MA Acc (%)	MA Kappa	WG Acc (%)	WG Kappa	Avg. Acc Decrease
Full Model	72.94	0.46	88.18	0.76	81.93	0.64	-
w/o SAG	66.34(-6.6)***	0.33(-0.13)	85.48(-2.7)**	0.71(-0.05)	77.33(-4.6)**	0.55(-0.09)	-4.6%
w/o DTA	65.34(-7.6)***	0.31(-0.15)	86.28(-1.9)**	0.73(-0.03)	77.83(-4.1)**	0.56(-0.08)	-4.5%
w/o MDFEA	70.14(-2.8)*	0.40(-0.06)	85.38(-2.8)**	0.71(-0.05)	79.13(-2.8)*	0.58(-0.06)	-2.8%
w/o CADN	68.74(-4.2)**	0.37(-0.09)	83.98(-4.2)**	0.68(-0.08)	77.73(-4.2)**	0.55(-0.09)	-4.2%

Note: “w/o” denotes removing the corresponding module from the full model. Values in parentheses indicate the change relative to the full model. Significance: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. For balanced binary tasks, $\kappa = 2 \times \text{ACC} - 1$.

quality (e.g., lower SNR and Fisher ratio), highlighting its importance in enhancing representational stability. Eliminating the CADN module—thereby switching to a single-branch decoding structure—reduces performance by 4.2% on average, demonstrating the comple-

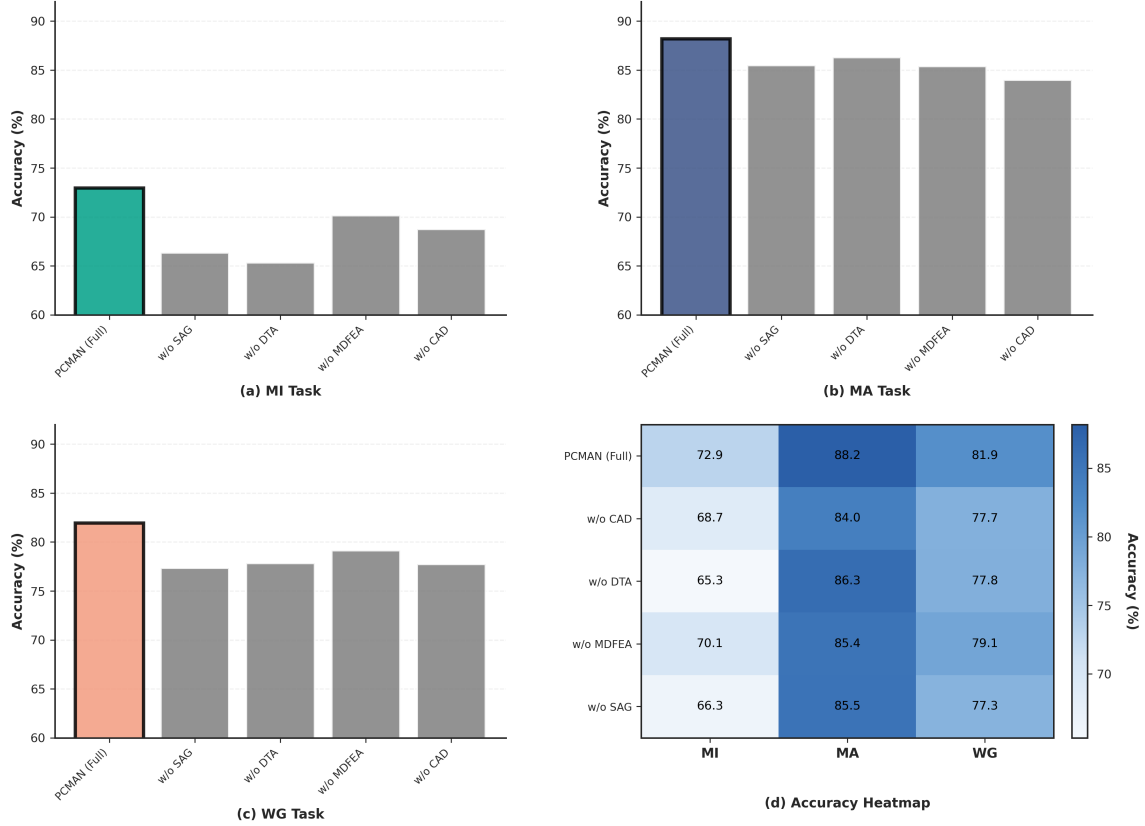


Figure 7: Task-wise accuracy comparison for module-level ablations of PCMAN. The bar plots and heatmap summarize the accuracy changes across MI, MA, and WG under different module-removal settings. The corresponding Cohen’s κ results are reported in Table 5.

mentary advantage of PCMAN’s multi-branch adaptive decoding design.

Figure 8 visualizes the progressive feature evolution of PCMAN. Raw features are highly entangled across classes, while progressive alignment through SDF, MDFEA, and CADN progressively expands inter-class distance and reduces intra-class variance, with the final-stage features achieving clear class boundaries.

The magnitude of DTA’s contribution varies across tasks, consistent with the expectation that temporal alignment benefits are larger for sustained cognitive paradigms (MA, WG) than for transient motor activation (MI).

Further analysis of loss function contributions shows that the auxiliary EEG supervision term effectively stabilizes training and prevents overreliance on fNIRS features, while the spatial correlation loss and temporal alignment loss enhance biological plausibility and ensure

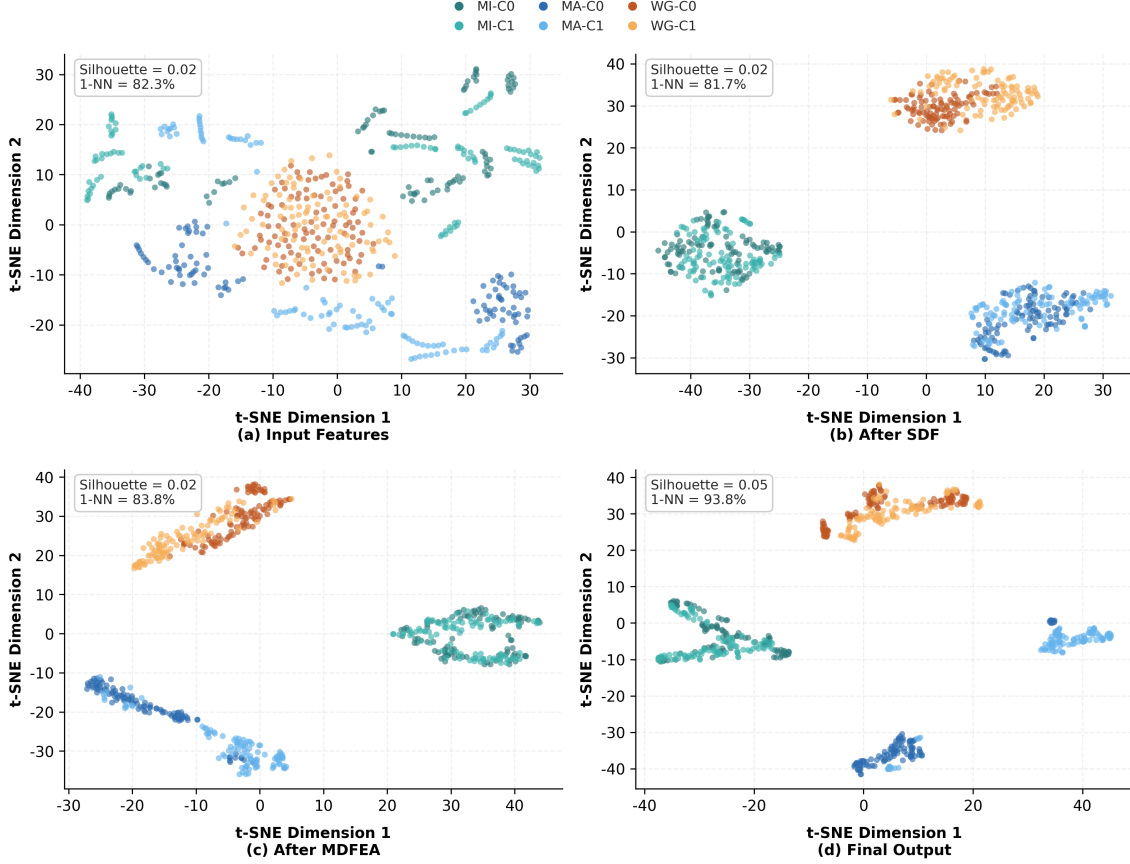


Figure 8: Progressive feature evolution visualized by t-SNE on pooled held-out samples. The four panels show the input representation, the SDF representation, the semantic-alignment proxy, and the final decision-stage representation. Feature separation increases progressively across stages.

consistent cross-modal alignment. Specifically, removing the spatial correlation constraint leads to more scattered and physiologically implausible attention maps, whereas removing the temporal alignment term reduces cross-modal consistency, especially under asynchronous neural–hemodynamic conditions.

4.4. Analysis of Alternative Design Choices

To examine whether the learnable attention mechanism performs better than standard statistical alignment, the multi-head attention in the DTA module was replaced with three alternative cross-modal alignment methods: Pearson correlation, canonical correlation analysis (CCA), and a mutual-information-inspired scoring network. The rest of the architecture remained unchanged. Table 6 reports the results across the three tasks.

Table 6: Comparison of cross-modal alignment strategies (mean $\pm$ std %)

Alignment method	MI (%)	MA (%)	WG (%)
Pearson correlation	70.85 $\pm$ 11.48	85.94 $\pm$ 7.96	79.96 $\pm$ 8.55
CCA-inspired projection	71.36 $\pm$ 11.31	86.47 $\pm$ 7.81	80.58 $\pm$ 8.43
MI-inspired scoring	71.22 $\pm$ 11.27	86.88 $\pm$ 7.69	80.41 $\pm$ 8.39
Multi-head attention (ours)	72.94 $\pm$ 11.71	88.18 $\pm$ 7.28	81.93 $\pm$ 8.29

The attention-based alignment achieves the highest average accuracy across the three tasks at 81.02%. The corresponding averages are 78.92% for Pearson correlation, 79.47% for CCA, and 79.50% for the MI-inspired scoring method. The improvement is most evident on MA and WG, where hemodynamic delays vary more across trials and benefit from a flexible nonlinear alignment.

Cubic-spline interpolation was also compared with zero-padding for spatial grid construction. Cubic-spline interpolation reduces accuracy loss caused by zero-padding, yielding improvements of 0.76 percentage points on MI, 0.83 on MA, and 1.13 on WG. Zero-padding introduces artificial discontinuities near the grid boundaries, which can weaken spatial patterns in edge brain regions. Cubic-spline interpolation is applied once during preprocessing and adds negligible computational cost.

The learnable AWG was further replaced with fixed fusion-to-fNIRS weight ratios of 0.3:0.7, 0.5:0.5, and 0.7:0.3. No single fixed ratio performs best across MI, MA, and WG. AWG avoids task-specific weight tuning while maintaining competitive accuracy and lower inter-subject variability.

4.5. Feature Representation and Interpretability Analysis

Figure 9 illustrates the temporal and spatial attention weight distributions learned by the model, further validating its physiological interpretability. The temporal attention exhibits a distinct unimodal pattern, with a pooled peak near 5.5s after task onset, showing moderate correspondence to the canonical hemodynamic response function (HRF). The task-wise peak-latency distributions reveal clear variability across MI, MA, and WG, indicating that the learned weighting is physiologically grounded but not reducible to a single fixed delay.

The spatial attention maps display clear task-specific activation patterns: for the Motor

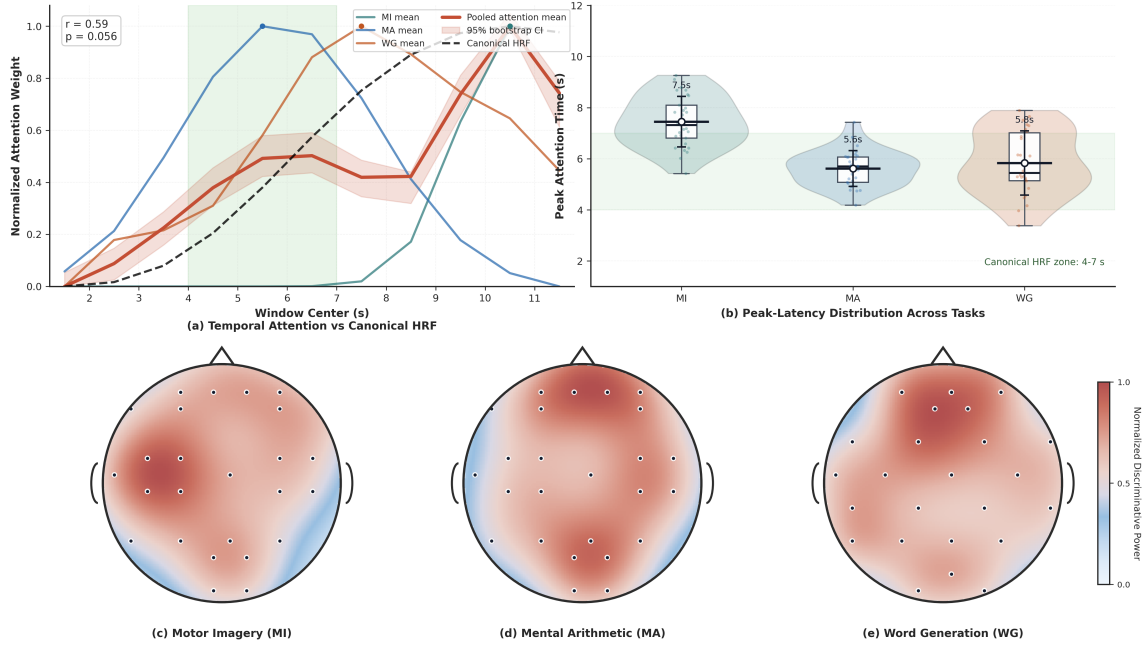


Figure 9: Temporal and spatial attention patterns learned by PCMAN from held-out data. (a) Pooled temporal attention profiles across representative MI, MA, and WG folds, shown together with the canonical HRF. (b) Task-wise distributions of temporal attention peak latency. (c)–(e) Spatial attention maps for MI, MA, and WG, respectively.

Imagery (MI) task, activations are focused around the central sulcus in the primary motor cortex (M1) and supplementary motor area (SMA); for the Mental Arithmetic (MA) task, activations are concentrated in the middle frontal gyrus and prefrontal cortex, reflecting the engagement of working memory and cognitive control; for the Word Generation (WG) task, high-intensity activation emerges in the left inferior frontal gyrus (Broca’s area), consistent with known language production regions. These activation distributions align well with cortical localizations reported in prior fMRI and fNIRS literature, further supporting the neurophysiological plausibility of the attention learned by PCMAN.

Overall, these analyses demonstrate that PCMAN not only surpasses existing methods in decoding accuracy but also exhibits high interpretability of its internal representations and attention mechanisms, suggesting its potential evolution from a purely data-driven model toward a neurophysiologically coupled framework.

Figure 10 depicts the evolution of the CADN dynamic weighting mechanism throughout task execution, revealing the system’s adaptive fusion strategy. At the early task stage, the

fNIRS branch dominates, reflecting the temporal advantage of the EEG signal; as the hemodynamic response gradually develops, the fusion branch weight increases linearly; during the stable phase, an optimal combination of fast EEG dynamics and steady fNIRS responses emerges.

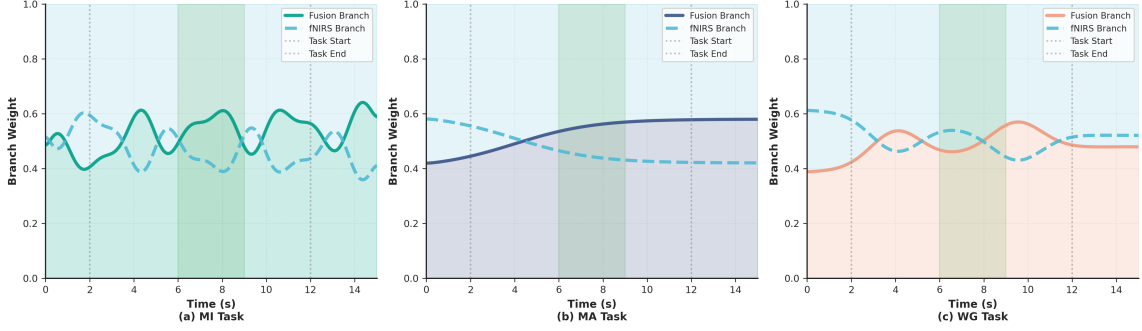


Figure 10: Temporal evolution of dynamic fusion weights during task execution. Fusion-branch weights adaptively vary across MI, MA, and WG tasks.

4.6. Robustness Evaluation

To evaluate PCMAN’s robustness in practical subject-dependent settings, perturbation experiments were conducted from two perspectives: resistance to signal noise and adaptability to motion artifacts.

Figure 11 presents the classification performance trends under varying signal-to-noise ratio (SNR) conditions. Gaussian white noise was added after preprocessing directly to the held-out model-input tensors, with channel-wise SNR defined identically for both EEG and fNIRS to ensure comparable perturbation intensity across modalities. The results show that while all methods experience performance degradation as noise increases, PCMAN maintains the highest accuracy and the slowest rate of decline.

For the motion robustness analysis, artifacts were synthetically generated using a spike-and-shift template and applied to HbO channels only, with 20% of channels randomly perturbed at $5\times$, $10\times$, and $20\times$ the baseline standard deviation. Results show that PCMAN experiences the smallest accuracy drop under high-intensity motion disturbance, confirming that the model effectively leverages fNIRS’s inherent robustness to motion artifacts. Mean-

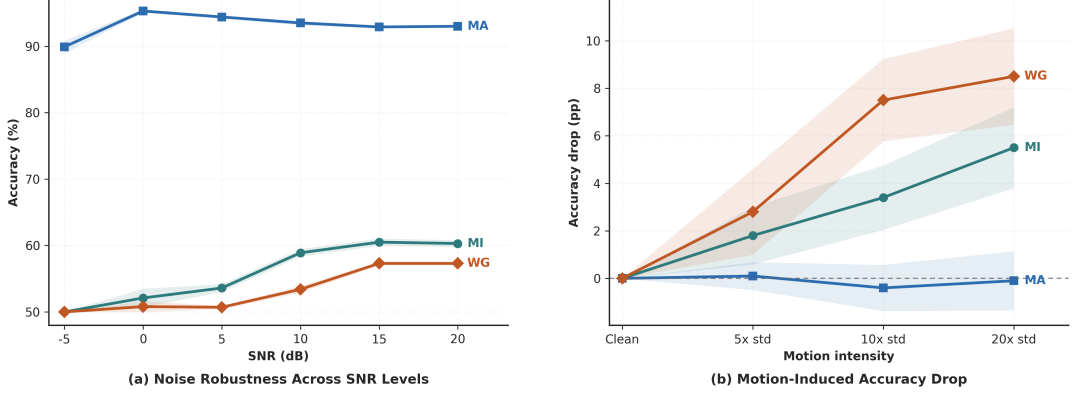


Figure 11: Robustness evaluation of PCMAN on representative held-out folds. (a) Accuracy under Gaussian-noise perturbation at multiple SNR levels. (b) Accuracy drop under motion-artifact perturbation at increasing intensity levels.

while, CADN’s dynamic fusion mechanism can detect degraded EEG quality and automatically increase the weighting of the fNIRS branch, achieving stable multimodal integration.

This robustness primarily stems from the synergistic effects of three mechanisms:

- (1) The dynamic weighting mechanism in CADN adaptively adjusts each modality’s contribution, automatically reducing the weight of low-quality signals;
- (2) Redundant information between EEG and fNIRS provides complementary protection, allowing one modality to compensate when the other degrades;
- (3) The cross-modal attention mechanism functions as a noise filter, dynamically focusing on high-SNR channels and suppressing the propagation of interference features.

4.7. Computational Efficiency and Parameter Sensitivity

Figure 12 reports the accuracy-efficiency trade-off at the module level. The full model has 2.10M parameters, 27.2ms inference latency, and 12.6 MB GPU memory. As the figure shows, SAG and DTA are the most cost-effective modules: both cause large accuracy drops when removed yet account for only a small fraction of the total parameters and latency. CADN carries the heaviest computational load, so future lightweight variants could start by simplifying the multi-branch decoder. Training completes in approximately 4.5 hours, with the model reaching 80% of its final accuracy within 20 epochs. Overall, PCMAN maintains a practical computational profile suitable for desktop-GPU or edge deployment.

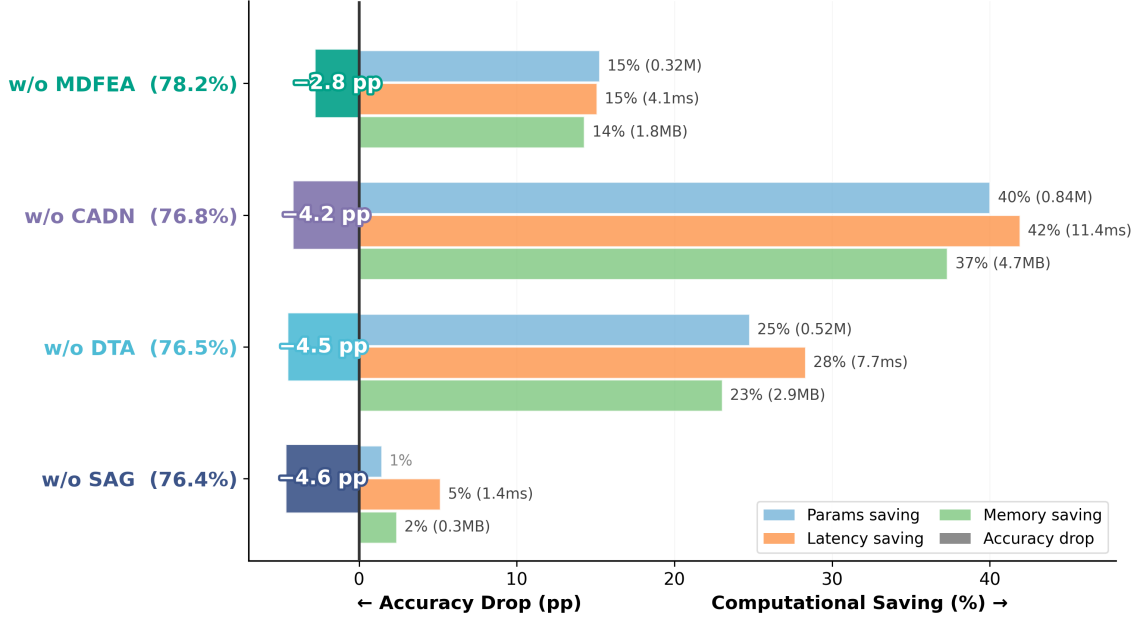


Figure 12: Module-level efficiency analysis of PCMAN. Left bars show the accuracy drop when each module is removed, and right bars show the corresponding savings in parameters, inference latency, and GPU memory relative to the full model.

Figure 13 examines sensitivity to four key hyperparameters. Too few attention heads limit the capacity for multi-scale temporal correspondence, while too many introduce redundancy and slow convergence; a moderate count yields the best trade-off. Similarly, 11 temporal windows cover the main hemodynamic response interval from 0 to 13s without introducing excessive noise, consistent with the physiological rationale in Section 3.1. Dropout in the range 0.2 to 0.5 effectively controls overfitting. The loss weights are set to 1.0 for the main classification term, 0.5 for the EEG auxiliary term, and 0.1 each for the spatial correlation and temporal alignment constraints, balancing decoding accuracy with physiologically guided regularization.

Training reduction experiments further show that PCMAN retains over 90% of its full-data accuracy when the training set is halved, indicating efficient data utilization. This property is valuable for practical BCI deployment, where per-subject calibration cost is a key concern.

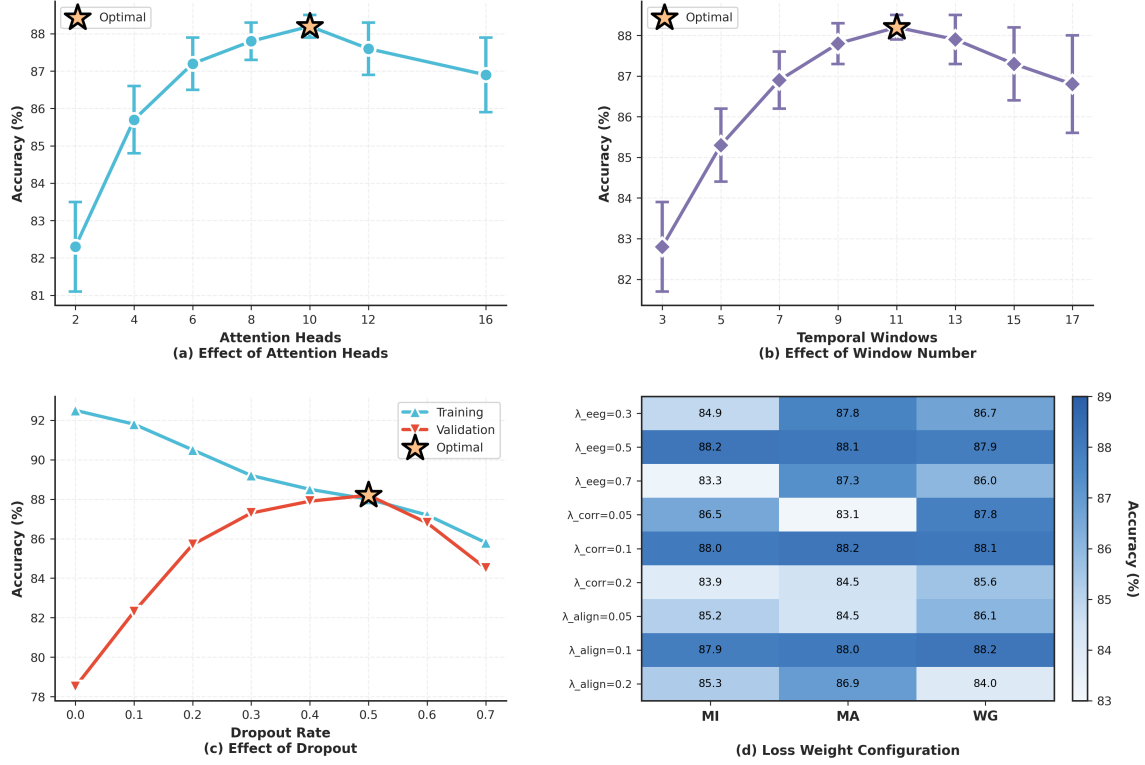


Figure 13: Influence of Key Hyperparameters on Performance

4.8. Subject-Independent Transfer

To further evaluate model generalization beyond the subject-dependent setting used in the main benchmark, additional experiments were conducted under a subject-independent leave-one-subject-out protocol. In this setting, models were trained on all but one subject and evaluated on the held-out subject. The preprocessing pipeline, tensor representation, and training procedure remained identical to those used in the subject-dependent experiments to ensure a fair comparison.

The results are summarized in Table 7. Compared with the subject-dependent benchmark in Table 4, all methods exhibit a performance decline under the cross-subject protocol, although the magnitude varies across models. This reduction reflects the substantial variability in anatomical structure, sensor placement, and neurovascular coupling patterns across subjects, which makes cross-subject decoding considerably more challenging than inter-session transfer within the same individual.

Table 7: Cross-subject decoding accuracy under the leave-one-subject-out protocol (mean $\pm$ std %). Drop: average decrease relative to Table 4.

Method	MI (%)	MA (%)	WG (%)	Avg. (%)	Drop (pp)
FGANet [15]	57.92 $\pm$ 11.83	74.02 $\pm$ 10.34	69.11 $\pm$ 9.76	67.02	3.94
EF-Net [16]	58.67 $\pm$ 11.41	75.13 $\pm$ 9.98	69.98 $\pm$ 9.45	67.93	0.71
BiMNC [17]	60.05 $\pm$ 9.88	76.10 $\pm$ 9.12	70.41 $\pm$ 9.68	68.85	6.64
M2NN [33]	59.48 $\pm$ 10.72	75.44 $\pm$ 9.57	70.03 $\pm$ 9.51	68.32	6.68
EFMLNet [18]	60.82 $\pm$ 10.11	76.97 $\pm$ 8.96	71.18 $\pm$ 9.03	69.66	6.19
Dual-EEGNet [42]	58.94 $\pm$ 10.37	74.58 $\pm$ 9.74	69.47 $\pm$ 9.61	67.66	5.01
E-FNet [43]	59.76 $\pm$ 10.06	76.12 $\pm$ 9.22	70.88 $\pm$ 9.34	68.92	6.18
STA-Net [19]	61.92 $\pm$ 10.87	77.86 $\pm$ 8.46	72.31 $\pm$ 9.02	70.70	7.24
PCMAN (ours)	65.71 $\pm$ 10.26	82.04 $\pm$ 8.31	75.88 $\pm$ 8.72	74.54	6.47

Despite this increased difficulty, PCMAN consistently achieves the best performance across all three tasks, reaching 65.71%, 82.04%, and 75.88% accuracy on MI, MA, and WG respectively, with the highest overall average of 74.54%. The corresponding average drop relative to the subject-dependent benchmark is 6.47 percentage points.

These results suggest that the progressive alignment mechanism captures partially transferable cross-subject structure rather than relying solely on subject-specific fitting. Nevertheless, a measurable performance gap between subject-dependent and subject-independent settings remains, indicating that low-level spatiotemporal differences across subjects still pose challenges. Further improvements may benefit from lightweight subject-specific calibration or domain-adaptation strategies.

5. Discussion

PCMAN appears to gain its advantage not simply from greater model complexity, but from progressively addressing the main sources of EEG–fNIRS heterogeneity at three levels. At the physical level, it establishes cross-modal correspondence in both time and space. At the semantic level, it refines the shared representation across modalities. At the task level, it adapts decoding to the signal structure of each paradigm. Taken together, these results suggest that effective hybrid decoding requires jointly handling delayed hemodynamics and spatial mismatch, rather than relying on a late fusion of loosely connected modality-specific features.

The interpretability analyses further suggest that the learned alignment is physiologically meaningful, but not invariant across tasks. Temporal attention remains concentrated around the expected hemodynamic response, with a pooled peak near 5.5s and a moderate correspondence to the canonical HRF, while the peak latency differs across MI, MA, and WG. Spatial attention also changes with the task. MI places greater emphasis on central motor regions and appears to depend more on early sensorimotor EEG dynamics, whereas MA and WG rely more heavily on delayed hemodynamic information and show stronger prefrontal or language-related patterns. These findings indicate that temporal and spatial alignment play complementary, but not identical, roles across paradigms. They also suggest that future hybrid BCI models may benefit from task-adaptive fusion strategies rather than a single fixed fusion prior.

These results should also be interpreted in light of the evaluation protocol. The main experiments measure subject-dependent inter-session generalization under a LOSO design, rather than subject-independent decoding. Because all overlapping windows from a given session are assigned to the same fold, direct temporal leakage is avoided. At the same time, adjacent samples are still not fully independent, so the reported performance is better understood as transfer under dense segmentation rather than under strictly non-overlapping sampling. The more challenging cross-subject experiments further show that substantial subject shift remains. Under the leave-one-subject-out protocol, PCMAN achieves the highest average accuracy of 74.54%, but this is still 6.47 percentage points lower than the subject-dependent benchmark. This gap indicates that low-level spatiotemporal mismatch across subjects remains an important source of error.

Several practical limitations also remain. Although the model itself incurs little computational latency, online responsiveness is constrained mainly by the temporal context required for fNIRS modeling. Under the current 11-window setting, the effective horizon extends to about 13s after task onset, which is acceptable for neurofeedback or rehabilitation monitoring, but less suitable for rapid control interfaces. In addition, the attention modules account for a large fraction of the parameters, which may limit deployment on embedded devices, and strong performance still benefits from user-specific calibration. More generally,

the idea of progressive alignment may extend beyond EEG–fNIRS, but its components are not equally transferable. Higher-level semantic and task alignment are likely to generalize more broadly, whereas the physical alignment stage remains closely tied to the properties of a given modality pair.

6. Conclusion

The proposed PCMAN framework fuses EEG’s millisecond-scale dynamics with fNIRS’s centimeter-scale localization through multidimensional feature enhancement, dynamic spatiotemporal alignment, and adaptive decoding. Progressive processing with bidirectional flow yields cohesive optimization. On two public datasets and three tasks, PCMAN significantly outperforms prior art; learned representations align with known neurovascular coupling, supporting biological plausibility. PCMAN offers an effective and interpretable solution for multimodal neural decoding, advancing hybrid BCIs toward practice. Future directions include extending progressive cross-modal alignment to other modality pairs and developing online adaptive alignment.

Acknowledgement

This work was supported in part by the Natural Science Foundation of Zhejiang Province (LZ25F010006), the National Nature Science Foundation of China (62371171, 62371178).

References

- [1] JR Wolpaw, N Birbaumer, DJ McFarland, et al. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. [https://doi.org/10.1016/S1388-2457\(02\)00057-3](https://doi.org/10.1016/S1388-2457(02)00057-3).
- [2] JDR Millán, R Rupp, GR Müller-Putz, et al. Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in Neuroscience*, 4:161, 2010. <https://doi.org/10.3389/fnins.2010.00161>.
- [3] LF Nicolas-Alonso and J Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012. <https://doi.org/10.3390/s120201211>.
- [4] N Naseer and KS Hong. fnirs-based brain-computer interfaces: a review. *Frontiers in Human Neuroscience*, 9:3, 2015. <https://doi.org/10.3389/fnhum.2015.00003>.
- [5] F Lotte, L Bougrain, A Cichocki, et al. A review of classification algorithms for eeg-based brain-computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. <https://doi.org/10.1088/1741-2552/aab2f2>.
- [6] F Scholkmann, S Kleiser, AJ Metz, et al. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, 85:6–27, 2014. <https://doi.org/10.1016/j.neuroimage.2013.05.004>.

- [7] KS Hong, U Ghafoor, and MJ Khan. Brain-machine interfaces using functional near-infrared spectroscopy: a review. *Artificial Life and Robotics*, 25:204–218, 2020. <https://doi.org/10.1007/s10015-020-00592-9>.
- [8] Z Liu, J Shore, M Wang, F Yuan, A Buss, and X Zhao. A systematic review on hybrid eeg/fnirs in brain-computer interface. *Biomedical Signal Processing and Control*, 68:102595, 2021. <https://doi.org/10.1016/j.bspc.2021.102595>.
- [9] R Li, D Yang, F Fang, et al. Concurrent fnirs and eeg for brain function investigation: A systematic methodological review. *Sensors*, 22(15):5865, 2022. <https://doi.org/10.3390/s22155865>.
- [10] DG Wyser, O Lamercy, F Scholkmann, et al. Characterizing reproducibility of cerebral hemodynamic responses when applying short-channel regression in functional near-infrared spectroscopy. *Neurophotonics*, 9(1):015004, 2022. <https://doi.org/10.1117/1.NPh.9.1.015004>.
- [11] A Machado, Z Cai, G Pellegrino, et al. Optimal positioning of optodes on the scalp for personalized functional near-infrared spectroscopy investigations. *Journal of Neuroscience Methods*, 309:91–108, 2018. <https://doi.org/10.1016/j.jneumeth.2018.08.006>.
- [12] T Yamada, S Umeyama, and K Matsuda. Multidistance probe arrangement to eliminate artifacts in functional near-infrared spectroscopy. *Journal of Biomedical Optics*, 14(6):064034, 2009. <https://doi.org/10.1117/1.3275469>.
- [13] S Fazli, J Mehnert, J Steinbrink, et al. Enhanced performance by a hybrid nirs-eeg brain computer interface. *NeuroImage*, 59(1):519–529, 2012. <https://doi.org/10.1016/j.neuroimage.2011.07.084>.
- [14] MJ Khan and KS Hong. Hybrid eeg-fnirs-based eight-command decoding for bci: application to quadcopter control. *Frontiers in Neurorobotics*, 11:6, 2017. <https://doi.org/10.3389/fnbot.2017.00006>.
- [15] Y Kwak, WJ Song, and SE Kim. Fganet: fnirs-guided attention network for hybrid eeg-fnirs brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:329–339, 2022. <https://doi.org/10.1109/TNSRE.2022.3149899>.
- [16] B Yu, L Cao, J Jia, and C Fan. E-fnet: A eeg-fnirs dual-stream model for brain-computer interfaces. *Biomedical Signal Processing and Control*, 98:106943, 2025. <https://doi.org/10.1016/j.bspc.2024.106943>.
- [17] L Chen, Y Zhou, Z Wu, and H Li. A bidirectional cross-modal transformer representation learning model for eeg-fnirs multimodal affective bci. *Expert Systems with Applications*, 266:126081, 2025. <https://doi.org/10.1016/j.eswa.2024.126081>.
- [18] L Qiu, Y Zhong, Z He, and J Pan. Improved classification performance of eeg-fnirs multimodal brain-computer interface based on multi-domain features and multi-level progressive learning. *Frontiers in Human Neuroscience*, 16:973959, 2022. <https://doi.org/10.3389/fnhum.2022.973959>.
- [19] C Bunternghit, J Wang, J Su, et al. Temporal attention fusion network with custom loss function for eeg-fnirs classification. *Journal of Neural Engineering*, 21(6):066016, 2024. <https://doi.org/10.1088/1741-2552/ad8e86>.
- [20] Y Qin, B Li, W Wang, X Shi, C Peng, X Wang, and H Wang. Eca-fusionnet: a hybrid eeg-fnirs signals network for mi classification. *Journal of Neural Engineering*, 22(1):016030, 2025. <https://doi.org/10.1088/1741-2552/adaf58>.
- [21] M Faisal, S Sahoo, and J Hazarika. Stecanet: spatio-temporal cross attention network for brain computer interface systems using eeg-fnirs signals. *Journal of Neural Engineering*, 22(6):066035, 2025. <https://doi.org/10.1088/1741-2552/ae2954>.
- [22] Y Zhang, B Yin, and X Yuan. Tsfnet: Temporal-spatial fusion network for hybrid brain-computer interface. *Sensors*, 25(19):6111, 2025. <https://doi.org/10.3390/s25196111>.
- [23] Jianeng Lin, Jiewei Lu, Zhilin Shu, Jianda Han, and Ningbo Yu. Subject-specific modeling of EEG-fNIRS neurovascular coupling by task-related tensor decomposition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:452–461, 2024. <https://doi.org/10.1109/TNSRE.2024.3355121>.
- [24] Danyang Chen, Jian Shi, Bo Tao, Xingwei Zhao, Zhixian Zhao, Shengjie Li, Yeguang Xu, Tao Ding, Peng Zhang, Qing Ye, Kuiyou Chen, Zhuojin Wu, Yingxin Tang, Wei Jiang, Kai Shu, Li Huang, Zheng You, Ping Zhang, and Zhouping Tang. A novel transfer learning-based hybrid EEG-fNIRS brain-computer interface for intracerebral hemorrhage rehabilitation. *Advanced Science*, 12(43):e2505426, 2025. <https://doi.org/10.1002/adv.202505426>.
- [25] G Pfurtscheller and FH Lopes da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. [https://doi.org/10.1016/S1388-2457\(99\)00141-8](https://doi.org/10.1016/S1388-2457(99)00141-8).
- [26] SB Borgheai, AH Zisk, J McLinden, et al. Multimodal pre-screening can predict bci performance variability: A novel subject-specific experimental scheme. *Computers in Biology and Medicine*, 168:107658, 2024. <https://doi.org/10.1016/j.combiomed.2023.107658>.
- [27] F Herold, P Wiegel, F Scholkmann, and NG Müller. Applications of functional near-infrared spectroscopy (fnirs) neuroimaging in exercise-cognition science: A systematic methodology-focused review.

- Journal of Clinical Medicine*, 7(12):466, 2018. <https://doi.org/10.3390/jcm7120466>.
- [28] J Shin, KR Müller, CH Schmitz, et al. Evaluation of a compact hybrid brain-computer interface system. *BioMed Research International*, 2017:6820482, 2017. <https://doi.org/10.1155/2017/6820482>.
 - [29] D Chen, J Shi, B Tao, X Zhao, Z Zhao, S Li, Y Xu, T Ding, P Zhang, Q Ye, K Chen, Z Wu, Y Tang, W Jiang, K Shu, L Huang, Z You, P Zhang, and Z Tang. A novel transfer learning-based hybrid eeg-fnirs brain-computer interface for intracerebral hemorrhage rehabilitation. *Advanced Science*, 12(43), 2025. <https://doi.org/10.1002/adv.202505426>.
 - [30] J Shi, D Chen, X Zhao, Z Zhao, S Li, Y Xu, T Ding, Z Zhu, P Zhang, Q Ye, Y Tang, P Zhang, B Tao, and Z Tang. Hefmi-ich: a hybrid eeg-fnirs motor imagery dataset for brain-computer interface in intracerebral hemorrhage. *Scientific Data*, 12(1), 2025. <https://doi.org/10.1038/s41597-025-06100-7>.
 - [31] AP Buccino, HO Keles, and A Omurtag. Hybrid eeg-fnirs asynchronous brain-computer interface for multiple motor tasks. *PLoS ONE*, 11(1):e0146610, 2016. <https://doi.org/10.1371/journal.pone.0146610>.
 - [32] J Shin, A von Lüthmann, B Blankertz, DW Kim, J Jeong, HJ Hwang, and KR Müller. Open access dataset for eeg+nirs single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2017. <https://doi.org/10.1109/TNSRE.2016.2628057>.
 - [33] AM Chiarelli, P Croce, A Merla, and F Zappasodi. Deep learning for hybrid eeg-fnirs brain-computer interface: application to motor imagery classification. *Journal of Neural Engineering*, 15(3):036028, 2018. <https://doi.org/10.1088/1741-2552/aaaf82>.
 - [34] F M Miezin, L Maccotta, J M Ollinger, S E Petersen, and R L Buckner. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11(6):735–759, 2000. <https://doi.org/10.1006/nimg.2000.0568>.
 - [35] T J Huppert, R D Hoge, S G Diamond, M A Franceschini, and D A Boas. A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, 29(2):368–382, 2006. <https://doi.org/10.1016/j.neuroimage.2005.08.065>.
 - [36] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. <https://arxiv.org/abs/1706.03762>.
 - [37] SW Zamir, A Arora, S Khan, M Hayat, FS Khan, and MH Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5729. IEEE, 2022. <https://doi.org/10.1109/CVPR52688.2022.00564>.
 - [38] A Gulati, J Qin, CC Chiu, N Parmar, Y Zhang, J Yu, W Han, S Wang, Z Zhang, and Y Wu. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040. ISCA, 2020. <https://doi.org/10.21437/Interspeech.2020-3015>.
 - [39] Z Miao, M Zhao, X Zhang, and D Ming. Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability. *NeuroImage*, 276:120209, 2023. <https://doi.org/10.1016/j.neuroimage.2023.120209>.
 - [40] Z Wang, J Fang, and J Zhang. Rethinking delayed hemodynamic responses for fnirs classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:4528–4538, 2023. <https://doi.org/10.1109/TNSRE.2023.3330911>.
 - [41] Z Wang, J Zhang, X Zhang, P Chen, and B Wang. Transformer model for functional near-infrared spectroscopy classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2559–2569, 2022. <https://doi.org/10.1109/JBHI.2022.3140531>.
 - [42] Z Wang, Y Lin, X Chen, and Y Wang. Dual-eegnet: Dual-path convolutional neural network for eeg-based emotion recognition. *Sensors*, 21(10):3367, 2021. <https://doi.org/10.3390/s21103367>.
 - [43] 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. <https://doi.org/10.3390/s24061889>.