



Multimodal Learning in Brain-Computer Interfaces: A Research Review on Applications

Rui Li

School of Electronic and Optical Engineering, Nanjing University of Science and Technology
ZiJin College, Nanjing Jiangsu 210023, China
ruili@asu.edu.pl

Abstract. Brain-Computer Interfaces (BCIs) enable human-machine interaction by decoding neural activity, demonstrating broad application potential in fields such as medical rehabilitation and neural engineering. However, traditional unimodal BCI systems face limitations due to signal noise, individual variability, and task complexity, hindering their practical utility. Multimodal learning capabilities significantly enhances decoding accuracy, robustness, and generalization capabilities through the fusion of signals from multiple data sources, including electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and eye-tracking. This paper analyzes research progress in this field, covering signal acquisition and preprocessing, fusion strategies at the data layer, feature layer, and decision layer, typical applications (such as emotion recognition, vigilance monitoring, and neural decoding), and cutting-edge directions (including zero-shot learning, federated learning, and generative augmentation). Research indicates that multimodal fusion can improve BCI system performance by 15%–25% and enhance stability in complex environments by over 30%. This review systematically summarizes the theoretical framework and technical pathways of multimodal BCI, offering key insights to facilitate its transition from laboratory research to clinical applications

Keywords: Multimodal Learning, Brain-Computer Interfaces (BCIs), Federated Learning.

1 Introduction

Brain-Computer Interface (BCI) technology facilitates human-machine interaction by decoding neural activity, demonstrating broad application prospects in medical rehabilitation, neural engineering, and intelligent control. However, traditional unimodal BCI systems are constrained by issues such as signal noise, individual variability, and task complexity, making it difficult to meet practical demands for high precision, strong robustness, and broad applicability. A single signal dimension cannot fully reflect complex cognitive states, which leads to limitations in decoding performance [1].

Multimodal learning, by integrating different types of data sources, significantly improves the decoding performance of brain-computer interface systems. In a review,

he et al. for the first time quantitatively verified that the integration of electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) could increase the accuracy of emotion recognition from a single peak of 68.2% to 83.7% [2]. The core advantages of multimodal fusion are reflected in three aspects: spatio-temporal complementarity, noise robustness, and multi-dimensional correlation [3].

In recent years, research on multimodal brain-computer interfaces has moved from theoretical verification to practical application, covering aspects such as signal acquisition and preprocessing, multi-level fusion methods, typical application scenarios, and emerging directions [4].

This paper systematically reviews the progress of multimodal learning in brain-computer interfaces, providing a theoretical basis for technical optimization and application expansion. It specifically analyzes the acquisition and preprocessing methods of multi-source signals, and focuses on the challenges faced by spatio-temporal alignment of cross-modal data. This study classified and evaluated the fusion strategies at the data, feature and decision-making levels, compared the performance differences among representative algorithms, and also summarized the benefits of fusion in practical scenarios. Through quantitative benchmarking of mainstream methods, it demonstrates significant improvements in accuracy, stability, and generalization capabilities achieved by multimodal BCIs. Addressing persistent challenges—including dynamic environment adaptation, individual variability, and computational overhead—the review proposes future directions such as neuromorphic computing and adaptive fusion strategies. This work establishes both theoretical frameworks and practical references for transitioning multimodal BCI systems from experimental validation to real-world deployment.

2 Current Methods for Multimodal BCIs

2.1 Multimodal BCIs significantly enhance system decoding performance by integrating heterogeneous data sources such as EEG, fNIRS, and eye-tracking. The core implementation workflow (as depicted in Fig. 1) comprises synchronized signal acquisition, preprocessing, multi-level fusion, and decoding output.

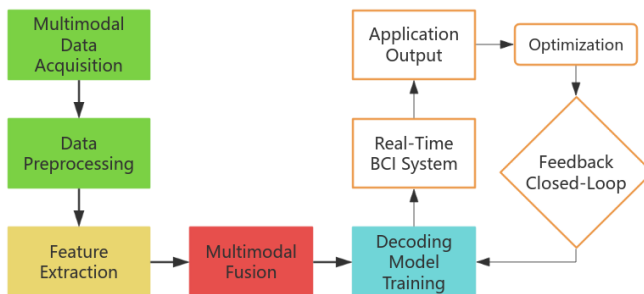


Fig. 1. Multimodal learning implementation workflow for BCIs.

During the signal acquisition and preprocessing stage, as shown in Table 1, distinct characteristics exist across multi-source signals: EEG detects neural electrical activity with high temporal resolution but limited spatial resolution; fNIRS monitors hemodynamic changes with moderate spatial resolution yet significant time delays (2–5 seconds); eye-tracking directly reflects attentional states but does not constitute direct neural physiological signals; while EMG is commonly employed to validate actual execution of motor intentions [1]. To enable effective fusion, priority must be given to resolving spatiotemporal alignment challenges across modalities. Common methodologies include hardware synchronization (e.g., Lab Streaming Layer protocol with <2 ms error), event marker synchronization, and signal resampling techniques (e.g., cubic spline interpolation) [5].

Table 1. Characteristics of Multi-Source Signals.

Modality	Detected Physical Quantity	Temporal Resolution	Spatial Resolution	Advantages
EEG	Neuronal Electrical Activity	Millisecond-Level	Low (>10 cm ²)	High Temporal Resolution
fNIRS	Hemodynamic Changes	Second-Level	Moderate (3-5 cm ²)	Electromagnetic Interference Resistance
Eye-Tracking	Pupillary Motion Trajectory	Millisecond-Level	-	Direct Attention Reflection
EMG	Muscle Electrical Activity	Millisecond-Level	-	Validation of Motor Intent Execution

At the technical architecture level, multi-level fusion strategies constitute the core of multimodal BCI systems, as summarized in Table 2. Data-level fusion (e.g., spatiotemporal registration) preserves maximal raw information but faces limited adoption due to high computational complexity, achieving only a 19.3% mutual information increase in experimental settings. Feature-level fusion, as the mainstream approach, typically selects highly correlated features through mutual information frameworks (e.g., EEG wavelet packet entropy and fNIRS HbO/HbR slopes), elevating F1-scores by 23% to 0.91 in motor imagery tasks [3]. Furthermore, deep learning architectures (e.g., multi-stream CNNs) attain 79.3% Top-5 accuracy in visual neural decoding via cross-modal attention mechanisms, while adaptive Transformers enhance correlation coefficients to 0.81 in affective recognition. Decision-level fusion often employs adaptive weighting strategies—for instance, assigning EEG and eye-tracking weights of 0.6 and 0.3 respectively in vigilance monitoring, effectively reducing prediction error to RMSE 0.18 [4].

Table 2. Performance Comparison of Multimodal BCI Fusion Methods.

Fusion Level	Representative Methods	Advantages	Performance Metrics	Application Scenarios
Data-Level	Spatiotemporal Registration	Preserves raw information integrity	+19.3% Mutual Information	Laboratory Environments
Feature-Level	Mutual Information Framework	Strong interpretability	F1-score 0.91	Motor Imagery
	Multi-stream CNNs	Automatic feature learning	Top-5 Acc 79.3%	Neural Decoding
	Adaptive Transformers	Dynamic handling of missing modalities	Corr 0.81	Affective Recognition
Decision-Level	Weighted Voting	Simple implementation	RMSE 0.18	Vigilance Monitoring
Generative Models	DDPM Data Augmentation	Mitigates data scarcity	F1+7.2%	Few-shot Learning

At the application level, multimodal integration significantly enhances system performance, as detailed in Table 3. For affective recognition tasks, EEG-fNIRS fusion elevates accuracy to 83.6% (an 11.5% improvement over unimodal EEG at 72.1%); further incorporating eye-tracking and facial expression data achieves 89.3% [6]. In vigilance monitoring, combining EEG θ/α power ratios with eye-tracking PERCLOS metrics maintains high stability (RMSE 0.18) during 2-hour continuous experiments. For neural decoding, spatial alignment of brain signals and visual features via CLIP models increases grasp success rates to 76.4% in spinal cord injury patients.

Table 3. Accuracy Comparison of Multimodal Systems vs. Unimodal Systems.

Modality Combination	Accuracy	Improve ment
EEG	72.1%	-
EEG + fNIRS	83.6%	+11.5%
EEG + Eye-Tracking + Facial Expressions	89.3%	+17.2%

Emerging technologies are continuously overcoming traditional limitations: Zero-shot learning utilizes semantic embeddings and prototype networks to recognize untrained affective states, achieving an unweighted average recall (UAR) of 68.3% [6]. Federated learning combined with local differential privacy ensures multi-institutional data security while limiting model AUC degradation to only 1.8% [7]. Generative data augmentation methods (e.g., sample generation via diffusion models like DDPM) enhance classification task F1-scores by 7.2% under few-sample conditions [8]. Furthermore, neuromorphic computing architectures (e.g., NeuCube-based spiking neural networks, SNNs) compress power consumption to one-twentieth of traditional

CNNs while maintaining modeling capabilities, enabling viable solutions for low-power edge deployment [5].

3 Limitations of Current Methods

Although multimodal brain-computer interface technology has made progress in decoding performance, it still faces challenges when transitioning to practical applications.

Poor adaptability to dynamic environments is a major limiting factor. The performance optimized in the laboratory cannot be reflected in real-world scenarios. Some interference factors caused by movement, can seriously affect highly sensitive neural signals. In this way, the signal will deteriorate significantly and the accuracy will be greatly reduced. The current artifact suppression methods have relatively limited suppression effects on complex time-varying interferences under intense motion conditions. Under dynamic conditions, sensors are prone to failure. Although the emerging Mobile Brain/Body Imaging (MoBI) paradigm provides an empirical foundation for dynamic neural signal research, integrating its processing techniques into real-time wearable multimodal BCI systems presents substantial engineering challenges [9].

Weak individual generalization capability represents a core bottleneck: Significant inherent physiological variations in neural signals prevent models trained on one cohort from effectively transferring to new individuals, resulting in universally low cross-subject transfer accuracy. While collaborative frameworks like federated learning mitigate data silo issues, privacy-preserving noise injection degrades model performance and fundamentally fails to overcome inherent inter-subject physiological differences. This generalization deficit becomes particularly acute when addressing special pathological conditions, such as baseline drift in Parkinson's disease patients with tremor-induced artifacts. To address this challenge, Li et al. proposed semi-supervised meta-learning, improving cross-subject transfer accuracy by 10.2% in motor imagery tasks. Although partially alleviating generalization deficiencies, its applicability to complex cognitive states requires further validation, and the model's dependency on source-domain data quality/diversity demands urgent optimization [10].

Computational efficiency bottlenecks similarly hinder real-time edge deployment: Mainstream parallel deep learning architectures require massive parameterization for effective multimodal fusion, resulting in heavy computational loads that fail to meet low-latency response demands in resource-constrained embedded devices or edge computing platforms. Although neuromorphic computing architectures can compress power consumption to one-twentieth of traditional CNNs while preserving modeling capabilities, existing technologies still face difficulties in simultaneously achieving millisecond-level response times and high-accuracy decoding requirements.

Additionally, low robustness against modality loss poses a hidden risk: during real-world operation, partial signal sources may experience temporary interruptions or quality degradation. Traditional fusion strategies cannot dynamically adapt to such

modality absence or signal quality fluctuations, causing unstable performance or even system failure. While generative data augmentation techniques have been employed to simulate missing signals or expand datasets, the fundamental stability issues in generated sample quality remain unresolved, potentially introducing additional uncertainties.

4 Future Improvement Directions

To address the aforementioned core challenges, future research should prioritize the following key directions to facilitate the transition of multimodal BCIs from laboratory validation to practical deployment:

Enhancing dynamic environment adaptability requires developing intelligent adaptive anti-interference technologies. This involves leveraging inertial sensor data to predictively model motion artifacts, dynamically adjusting filtering parameters or fusion strategies. This paper combines neuromorphic computing, which have event-driven processing capabilities, ultra-low power consumption, and achieve efficient dynamic noise suppression. This significantly enhances the system's robustness in complex motion scenarios.

To address the challenge of individual generalization, it is necessary to promote the coordinated progress of all aspects. On the one hand, a shared representation space capable of capturing cross-disciplinary universal physiological patterns should be established. On the other hand, generative data improvement techniques should be advanced to produce high-quality and diverse synthetic samples. These samples are customized for different physiological states. By adopting this dual approach, the adaptive ability of the model can be significantly enhanced.

Overcoming computational efficiency bottlenecks remains critical for achieving real-time edge deployment. The optimization strategy can incorporate model architecture compression techniques to reduce the size of parameters and memory usage. It is also possible to explore some new hardware acceleration solutions to accelerate the inference of spike neural networks. Moreover, an event-driven asynchronous processing architecture is adopted to replace the traditional synchronous paradigm. This trident approach works together to eliminate computational redundancy, reduce power consumption and latency, thus providing a feasible and low-latency deployment path for edge devices with relatively limited resources.

Enhancing robustness against modality loss necessitates fusion strategies incorporating dynamic adaptability. This requires developing intelligent mechanisms capable of real-time assessment of per-modality signal quality to autonomously adjust fusion weights or strategies. Concurrent optimization of generative models should focus on improving stability and synthetic quality for reliable prediction-based compensation of missing modalities, though vigilant mitigation of their inherent uncertainty risks remains essential.

Based on the information presented in Table 4, this paper outlines the existing limitations of multimodal BCI and provides corresponding suggestions for improvement.

Table 4. Mapping of Multimodal BCI Limitations to Improvement Recommendations.

Source	Existing Limitations/Shortcomings	Specific Manifestations	Corresponding Recommendations
Modality	EEG: Susceptible to EMG Interference	Motion artifacts contaminate signals; recognition accuracy degrades in dynamic environments	Develop dynamic adaptive anti-interference techniques
	fNIRS: Significant Time Delays	Hemodynamic response lag hinders real-time interaction	Optimize edge computing efficiency
	Eye-Tracking: Non-neurophysiological Signals	Cannot directly reflect neural activity; prone to interruption during intense motion	Design dynamic fusion mechanisms
Modality Fusion	Data-Level Fusion: High Computational Complexity	Spatiotemporal registration requires high computing power; only feasible in labs	Optimize edge computing efficiency
	Feature-Level Fusion: Feature Engineering Dependency	Poor generalization of handcrafted features; cross-subject transfer failure	Construct cross-subject universal physiological representations
	Feature-Level Fusion: Large Data Requirements	Massive parameters consume excessive training resources	Generative data augmentation
	Decision-Level Fusion: Ignores Inter-modal Correlations	Fixed-weight voting fails to adapt to modality loss	Design dynamic fusion mechanisms
	Generative Models: Unstable Output Quality	Low reliability of synthetic samples may introduce uncertainty	Optimize generative models

Overall, in future research, three core directions should be advanced simultaneously. These three core directions are improving dynamic robustness, conducting cross-domain physiological modeling, and optimizing edge computing. If breakthroughs can be made in these key technologies, the main obstacles currently faced by multi-modal brain-computer interface systems can be well overcome. It enables multi-mode brain-computer interface systems to transition more quickly from controlled laboratory environments to complex scenarios in the real world.

5 Conclusion

Multimodal learning integrates diverse signals such as EEG, fNIRS, and eye-tracking to significantly improve the decoding accuracy of brain-computer interface systems, enhance their stability across varied environments, and extend their applicability to different tasks. Quantitative studies indicate that multimodal fusion yields an average performance improvement of 15% to 25% in applications such as emotion recognition,

vigilance monitoring, and neural decoding compared to single-modal approaches. It also strengthens system stability by over 30% under complex conditions. These advances help alleviate typical limitations of traditional brain-computer interfaces, including spatiotemporal resolution constraints, sensitivity to noise, and limited capacity for cognitive information representation.

However, several systematic challenges remain when implementing multimodal brain-computer interfaces in real-world settings. Motion artifacts in dynamic environments can degrade signal quality and impair overall system performance. Variability in individual physiological characteristics may also restrict generalizability across user groups. Furthermore, multi-stream deep learning models require substantial computational resources, limiting their deployment on edge devices. Adopting a fixed fusion strategy introduces additional risk—if one modality is lost, system performance can decline by more than 35%.

Addressing these bottleneck issues and achieving future breakthroughs requires focused efforts along three collaborative directions. First, developing a dynamic adaptive framework that integrates inertial sensing and neuromorphic spiking neural networks could enhance anti-interference capabilities through real-time motion artifact modeling and event-driven processing. Second, constructing a cross-domain physiological general representation using semantic embedding prototype networks and generative refinement may help overcome generalization limitations caused by individual differences. Third, establishing collaborative edge optimization pathways—combining model pruning, memristor-based acceleration, and asynchronous architecture—can enable millisecond-level responsiveness while reducing power consumption. Additionally, designing dynamic fusion mechanisms with signal quality awareness is essential to minimize performance degradation during modality loss. Together, these innovations, through deep integration of neural engineering, edge computing, and generative AI, are expected to advance multimodal BCIs from “lab-oriented precision” toward “scenario-adaptive capability balance,” thereby offering innovative and practical theoretical frameworks for high-dynamic applications such as neural rehabilitation.

References

1. Liu, Z., Shore, J., Wang, M., et al.: A systematic review on hybrid EEG/fNIRS in brain-computer interface. *Biomedical Signal Processing and Control*, 68(102595), 1-8 (2021).
2. He, Z., Li, Z., Yang F., et al.: Advances in multimodal emotion recognition based on brain-computer interfaces. *Brain Sciences*, 10(10), 687, 1-29 (2020).
3. Deligani, R., J., Borgheai, S., B., McLinden, J., Shahriari, Y.: Multimodal fusion of EEG-fNIRS: a mutual information-based hybrid classification framework. *Biomed Optics Express* 12(3), 1635-1650 (2021).
4. Wang, K., Qiu, S., Wei, W., et al.: A multimodal approach to estimating vigilance in SSVEP-based BCI. *Expert Systems with Applications* 225(120177), 1-16 (2023).

5. Garcia-Palencia, O., Fernandez, J., Shim, V., et al.: Spiking neural networks for multimodal neuroimaging: a comprehensive review of current trends and the NeuCube brain-inspired architecture. *Bioengineering* 12(6), 628 (2025).
6. Xu, X., Deng, J., Cummins, N. et al.: Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes. *IEEE Transactions on Multimedia* 24, 2752-2765 (2022).
7. Truex, S., Baracaldo, N., Anwar, A., et al.: A hybrid approach to privacy-preserving federated learning. In: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec'19)*, 1–11. Association for Computing Machinery, New York, NY, USA (2019).
8. Chen, L., Yin, Z., Gu, X., et al.: Neurophysiological data augmentation for EEG-fNIRS multimodal features based on a denoising diffusion probabilistic model. *Computer Methods and Programs in Biomedicine*. 261(108594), 1-18 (2025).
9. Tait, P.J., Timm, E.C., O'Keefe, J. et al.: *Locomotion and Posture in Older Adults*. 2nd edn. Springer, Cham, Swiss Confederation (2024).
10. Li, J., Wang, F., Huang, H., et al.: A novel semi-supervised meta learning method for subject-transfer brain-computer interface. *Neural Networks*, 163, 195-204 (2023).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

