Transformer-based Multimodal Fusion for Spatiotemporal Synchronization of Rehabilitation Signals: A Precise Identification Method for Unilateral Compensatory Movements

Jiayi Shen, Luyan Shen, Sitong Ruan, Zhihui Xu, Jingyi Xu*
Artificial Intelligence College, Zhejiang Dongfang Polytechnic, Wenzhou, Zhejiang, China
**Corresponding Author*

Abstract: Unilateral compensatory movements are prevalent abnormal motor patterns in stroke patients during rehabilitation, impairing neural plasticity training and causing secondary injuries. Conventional multimodal rehabilitation systems are plagued by poor signal spatiotemporal synchronization, failure in dynamic weighting of critical information and high misdiagnosis rates of compensatory motions. This paper develops a Transformer-based multimodal fusion framework. A hardware-software co-synchronization strategy realizes millisecond-level alignment of EEG, sEMG and IMU signals. The cross-modal attention fusion model mines spatiotemporal correlations between neuromuscular and kinematic data, while a fine-grained classifier identifies six upper-limb compensatory patterns. Tests on 20 stroke patients yield a misjudgment rate of only 4.2%, far outperforming traditional algorithms. Clinically, the system delivers real-time reminders to cut compensatory movements and boost rehabilitation efficiency.

Keywords: Transformer; Multimodal Fusion; Spatiotemporal Synchronization; Unilateral Compensatory Movements; Stroke Rehabilitation; Real-Time Intervention

1. Introduction

Stroke remains the leading cause of long-term motor disability worldwide, with approximately 50% of survivors suffering from persistent upper limb motor dysfunction. During rehabilitation training, patients often adopt compensatory strategies using the trunk, shoulder girdle or contralateral healthy limbs to complete motor tasks due to weakness and poor control of the affected limb. Clinical studies have shown that long-term use of compensatory movements will strengthen abnormal neural pathways, hinder the recovery of normal motor function of the affected limb, and even cause musculoskeletal damage such as shoulder subluxation and low back pain.

Traditional compensatory movement assessment mainly relies on the subjective observation of therapists, which has problems such as poor consistency, low efficiency, and inability to quantify fine-grained movement patterns. In recent years, sensor-based multimodal rehabilitation assessment systems have received widespread attention. By integrating EEG, sEMG, IMU and other sensors, these systems can simultaneously collect neural activity, muscle activation and kinematic data, providing an objective basis for compensatory movement detection. However, existing systems still face three key challenges:

Poor spatiotemporal synchronization: Different sensors have different sampling rates and clock drifts, leading to time misalignment between multimodal signals, which seriously affects the accuracy of fusion analysis.

Ineffective fusion methods: Traditional early fusion and late fusion methods cannot dynamically adjust the weights of different modalities according to the task stage and signal quality, resulting in the loss of key information.

High misjudgment rate: Most existing methods only rely on single-modal or simple feature fusion, making it difficult to distinguish subtle differences between normal movements and compensatory movements, leading to high false positive and false negative rates.

To solve the above problems, this paper proposes a Transformer-based multimodal fusion framework for precise identification of unilateral compensatory movements. The main contributions are as follows:

Proposes a hardware-software collaborative spatiotemporal synchronization method, which

combines FPGA global clock triggering and Lab Streaming Layer (LSL) software calibration to achieve millisecond-level alignment of multimodal signals.

Designs a cross-modal attention-driven MMT-Fusion model, which dynamically weights sEMG activation timing and IMU joint angle change features, effectively capturing the spatiotemporal correlation between neuromuscular activity and kinematic data.

Constructs a fine-grained compensatory movement dataset containing 6 common upper limb compensatory movement types, and verifies that the proposed method can reduce the misjudgment rate of compensatory movements to less than 5%.

Develops a real-time intervention system that can provide visual and auditory prompts to therapists when compensatory movements are detected, helping to correct abnormal motor patterns in a timely manner.

2 Overall System Architecture

The proposed multimodal compensatory movement identification system adopts a three-layer architecture, including the multimodal signal acquisition layer, spatiotemporal synchronization and fusion layer, and compensatory movement identification and feedback layer [1].

2.1 Multimodal Signal Acquisition Layer

The acquisition layer integrates three types of sensors to comprehensively capture the neuromuscular and kinematic information of patients during rehabilitation training:

EEG acquisition module: Uses a 16-channel dry electrode EEG headset with a sampling rate of 250 Hz to collect brain neural activity signals related to motor intention [2].

SEMG acquisition module: Uses 8-channel wireless sEMG sensors with a sampling rate of 1000 Hz, placed on the biceps brachii, triceps brachii, deltoid, and trapezius muscles of both upper limbs to collect muscle activation signals.

IMU acquisition module: Uses 6-axis IMU sensors with a sampling rate of 200 Hz, placed on the wrist, elbow, shoulder and trunk to collect joint angle, acceleration and angular velocity data.

All sensors are connected to the edge computing gateway via Bluetooth 5.0, ensuring low-latency data transmission [3].

2.2 Spatiotemporal Synchronization and Fusion Layer

This layer is the core of the system, responsible for aligning multimodal signals to a unified time axis and extracting and fusing deep features. First [4], the hardware-software collaborative synchronization method is used to correct the time offset and clock drift between different sensors. Then, lightweight feature extraction networks are used to extract temporal features from EEG, sEMG and IMU signals respectively. Finally, the MMT-Fusion model based on cross-modal attention is used to dynamically fuse multimodal features to obtain a comprehensive feature representation containing neuromuscular and kinematic information [5].

2.3 Compensatory Movement Identification and Feedback Layer

This layer uses the fused features to identify compensatory movements through a fully connected classification network. The system can identify 6 common upper limb compensatory movements: trunk forward lean, trunk rotation, shoulder elevation, shoulder abduction, elbow hyperextension, and contralateral limb assistance [6]. When a compensatory movement is detected, the system immediately sends a prompt signal to the VR headset worn by the patient and the tablet computer held by the therapist, providing real-time intervention guidance. At the same time, the system records the type, frequency and duration of compensatory movements, generating a quantitative rehabilitation assessment report [7].

3 Key Technology Implementation

3.1 Hardware-Software Collaborative Millisecond-Level Spatiotemporal Synchronization

To solve the problem of time misalignment between multimodal signals caused by different sampling rates and clock drifts, this paper proposes a hardware-software collaborative spatiotemporal synchronization method.

3.1.1 Hardware global clock triggering

An FPGA-based synchronous trigger device is designed to generate a unified global clock signal. The device sends synchronous trigger pulses to all sensors at a frequency of 100 Hz, and each sensor records the local timestamp

when receiving the trigger pulse. This method eliminates the initial time offset between sensors and provides a hardware reference for subsequent software calibration [8].

3.1.2 LSL-based software time calibration

The Lab Streaming Layer (LSL) protocol is used for software-level time synchronization. Each sensor acts as an LSL outlet to send data streams with local timestamps, and the edge gateway acts as an LSL inlet to receive all data streams. The LSL protocol continuously measures the clock offset and drift between the gateway and each sensor through an NTP-like packet exchange mechanism, and dynamically corrects the timestamps of the data streams [9].

3.1.3 Resampling and alignment

After time calibration, the signals of different sampling rates are resampled to a unified sampling rate of 200 Hz using linear interpolation. For each time point, the corresponding data of EEG, sEMG and IMU signals are aligned to form a multimodal data sample [10]. Experimental results show that the synchronization accuracy of this method reaches ±1.8 ms, which fully meets the requirements of fine-grained motor pattern analysis.

3.2 Transformer-Based Multimodal Fusion Model (MMT-Fusion)

To dynamically fuse multimodal features and capture the spatiotemporal correlation between different modalities, this paper proposes the MMT-Fusion model based on cross-modal attention mechanism.

The model consists of three unimodal feature extraction branches and a cross-modal fusion module:

EEG feature branch: Uses a 1D convolutional neural network (CNN) to extract local temporal features from EEG signals, followed by a bidirectional LSTM network to capture long-term temporal dependencies.

SEMG feature branch: Uses a depthwise separable CNN to extract muscle activation pattern features, focusing on the timing and intensity changes of muscle contraction.

IMU feature branch: Uses a temporal convolutional network (TCN) to extract kinematic features such as joint angle, velocity and acceleration changes.

The outputs of the three branches are input to the cross-modal fusion module, which is the core of the MMT-Fusion model. The module adopts a multi-head cross-attention mechanism, which can calculate the attention weight between each modality and other modalities, and dynamically adjust the contribution of different modalities according to the task stage and signal quality. Specifically, during the initial stage of movement, the model will assign higher weights to EEG signals reflecting motor intention; during the execution stage of movement, the model will focus more on sEMG signals reflecting muscle activation and IMU signals reflecting kinematic changes.

The output of the cross-modal fusion module is passed through a global average pooling layer and a fully connected layer to obtain the final classification result of compensatory movements. To adapt to the computing power constraints of edge devices, the model is quantized using INT8, which reduces the model size by 75% and improves the inference speed by 3 times while the accuracy loss is less than 1%.

3.3 Fine-Grained Compensatory Movement Identification Algorithm

Based on the MMT-Fusion model, this paper proposes a fine-grained compensatory movement identification algorithm, which can accurately identify 6 common upper limb compensatory movements.

First, a sliding window technique with a window size of 500 ms and a step size of 100 ms is used to segment the continuous multimodal signals. For each window, the MMT-Fusion model is used to extract fused features and predict the probability of each compensatory movement type.

Then, a post-processing strategy based on temporal smoothing is adopted to reduce the influence of noise and transient misjudgment. Specifically, the prediction results of consecutive 3 windows are averaged, and only when the average probability of a certain compensatory movement type exceeds 0.7, it is determined that the compensatory movement occurs.

Finally, the system records the type, start time, end time and duration of the compensatory movement, and calculates the frequency and proportion of each compensatory movement type in the training process. This quantitative information can help therapists understand the patient's motor pattern defects and formulate personalized rehabilitation plans.

4 Experiments and Result Analysis

4.1 Experimental Setup

To comprehensively evaluate the performance of the proposed system, three groups of experiments were designed: spatiotemporal synchronization accuracy test, compensatory movement identification performance test, and clinical intervention effect verification.

Spatiotemporal synchronization accuracy test: A high-precision signal generator was used to generate synchronous square wave signals, which were simultaneously input to the EEG, sEMG and IMU sensors. The time difference between the signals received by different sensors was measured, and the average synchronization error and standard deviation were calculated. A total of 1000 tests were conducted.

Compensatory movement identification performance test: 20 stroke patients (12 males and 8 females, aged 42-68 years, disease duration 2-18 months) and 10 healthy subjects (5 males and 5 females, aged 23-35 years) were recruited to participate in the experiment. All patients had upper limb motor dysfunction and used compensatory movements during rehabilitation training. The experimental tasks included 5 common upper limb rehabilitation movements: reaching forward, reaching sideways, lifting the arm, drinking water, and grasping objects. Each subject performed each task 20 times, including both normal movements and compensatory movements. Two experienced therapists labeled all data, and the consensus labeling result was used as the ground truth.

Clinical intervention effect verification: The 20 stroke patients were randomly divided into an experimental group and a control group, with 10 patients in each group. The experimental group used the proposed system for 4 weeks of rehabilitation training, 5 days a week, 30 minutes per day. The system provided real-time prompts when compensatory movements were detected, and the therapist corrected the patient's movements in a timely manner. The control group received conventional rehabilitation training without real-time compensatory movement detection and intervention. Before and after training, the Fugl-Meyer Assessment (FMA) upper limb score and the Modified Ashworth Scale (MAS) were used to evaluate the patients' motor function and muscle tension.

4.2 Spatiotemporal Synchronization Accuracy Test Results

Table 1 shows the synchronization accuracy comparison between the proposed method and the traditional software-only synchronization method.

Table 1. The Synchronization Accuracy Comparison

Synchronization Method	EEG-sEMG Error	EEG-IMU Error	sEMG-IMU Error	Average Error
Traditional Software Method	12.6 ± 3.2 ms	15.8 ± 4.1 ms	10.3 ± 2.7 ms	12.9 ms
Proposed Method	1.7 ± 0.5 ms	1.9 ± 0.6 ms	1.8 ± 0.4 ms	1.8 ms

As can be seen from Table 1, the average synchronization error of the proposed method is only 1.8 ms, which is an order of magnitude lower than that of the traditional software method. This high-precision spatiotemporal synchronization lays a solid foundation for subsequent multimodal fusion and compensatory movement identification.

4.3 Compensatory Movement Identification Performance Test Results

Table 2 shows the performance comparison between the proposed method and three traditional fusion methods (early fusion, late fusion, CNN-LSTM fusion) in compensatory movement identification.

Table 2. The Performance Comparison

Method	Accuracy	Precision	Recall	F1-Score	Misjudgment Rate
Early Fusion	76.3%	72.5%	68.9%	70.7%	23.7%
Late Fusion	81.2%	78.6%	75.3%	76.9%	18.8%
CNN-LSTM Fusion	85.4%	83.1%	80.7%	81.9%	14.6%
Proposed Method	**95.8%**	**94.2%**	**93.5%**	**93.8%**	**4.2%**

As can be seen from Table 2, the results show that the proposed method achieves the best performance in all evaluation indicators, with an accuracy of 95.8% and a misjudgment rate of only 4.2%, which is significantly lower than the traditional methods. This is mainly because the MMT-Fusion model can dynamically weight key signals and effectively capture the spatiotemporal correlation between multimodal

signals, thereby improving the accuracy of compensatory movement identification.

4.4 Clinical Intervention Effect Verification Results

Table 3 shows the changes in FMA upper limb scores and MAS scores of patients in the experimental group and the control group before and after training.

Table 3. The Changes

Group	Time	FMA Upper Limb Score	MAS Score
Experimental Group	Before Training	25.3 ± 5.1	2.1 ± 0.6
Experimental Group	After Training	39.7 ± 4.8	1.2 ± 0.5
Control Group	Before Training	24.8 ± 4.9	2.2 ± 0.5
Control Group	After Training	31.5 ± 5.3	1.7 ± 0.4

As can be seen from Table 3, after 4 weeks of training, the FMA upper limb score of the experimental group increased by 14.4 points on average, and the MAS score decreased by 0.9 points on average, which were significantly better than those of the control group (increased by 6.7 points and decreased by 0.5 points respectively). This indicates that real-time detection and intervention of compensatory movements can effectively improve the efficiency of rehabilitation training and promote the recovery of motor function in stroke patients.

5 Discussion

This paper proposes a Transformer-based multimodal fusion framework for precise identification of unilateral compensatory movements, which successfully solves the key problems of poor spatiotemporal synchronization of multimodal signals, ineffective fusion methods and high misjudgment rate of compensatory movements in existing rehabilitation systems.

Compared with existing studies, the proposed system has the following advantages:

High-precision spatiotemporal synchronization: The hardware-software collaborative method achieves millisecond-level alignment of multimodal signals, ensuring the consistency of time information between different modalities.

Adaptive multimodal fusion: The cross-modal attention mechanism can dynamically adjust the weights of different modalities according to the task stage and signal quality, making full use of the complementary information between neuromuscular and kinematic data.

Low misjudgment rate: The fine-grained identification algorithm reduces the misjudgment rate of compensatory movements to less than 5%, which can meet the needs of clinical real-time intervention.

Real-time intervention capability: The system can provide real-time prompts to therapists and patients, helping to correct abnormal motor patterns in a timely manner and improve rehabilitation effects.

This study also has some limitations:

Currently, the system only focuses on upper limb compensatory movements, and future work can extend to lower limb and trunk compensatory movements.

The sample size of clinical experiments is relatively small, and larger-scale multi-center clinical trials are needed to further verify the effectiveness of the system.

The system currently only integrates EEG, sEMG and IMU signals, and future work can fuse visual signals and pressure distribution signals to further improve the comprehensiveness and accuracy of assessment.

6 Conclusion and Future Work

This paper proposes a Transformer-based multimodal fusion framework for precise identification of unilateral compensatory movements. By designing a hardware-software collaborative spatiotemporal synchronization method, millisecond-level alignment of multimodal signals is achieved. The MMT-Fusion model based on cross-modal attention can dynamically weight key signals, effectively capturing the spatiotemporal correlation between neuromuscular activity and kinematic data. Experimental results show that the misjudgment rate of the proposed method for compensatory movements is reduced to 4.2%, and clinical intervention experiments demonstrate that the system can effectively improve the efficiency of rehabilitation training.

Future work will focus on the following aspects:

Expanding the types of compensatory movements identified by the system to cover lower limb and trunk compensatory movements.

Fusing visual signals and pressure distribution signals to further improve the accuracy and comprehensiveness of assessment.

Developing a home-based rehabilitation

training system based on the proposed method, enabling patients to receive professional rehabilitation guidance at home.

Conducting large-scale multi-center clinical trials to verify the long-term effectiveness of the system in different patient populations.

References

[1] Zhang X, Wang Y, Jin J, et al. A lightweight CNN-based EEG classification model for edge computing. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 1234-1243.

[2] Li Y, Wang H, Zhang L, et al. EdgeBCI: An edge computing framework for low-latency brain-computer interfaces. IEEE Internet of Things Journal, 2024, 11(8): 13456-13467.

[3] Gao S, Chen J, Chen X, et al. Temporal dynamics and physical prior multimodal network for rehabilitation physical training evaluation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2024, 32: 1897-1908.

[4] Chen X, Wang L, Liu C. Application of multithreaded asynchronous scheduling in real-time brain-computer interface systems. Computer Engineering and Applications, 2023, 59(12): 189-195.

[5] Zhao J, Rao Y. Adaptive Multimodal Temporal Transformer for Bio Signal Driven Stroke Rehabilitation Prognosis. bioRxiv, 2025: 2025.10.18.683234.

[6] Leeb R, Friedman D, Müller-Putz G R, et al. Towards brain-computer interface driven virtual reality systems: Design and implementation. Presence: Teleoperators and Virtual Environments, 2007, 16(3): 290-307.

[7] Wolpaw J R, Birbaumer N, McFarland D J, et al. Brain-computer interfaces for communication and control. Clinical Neurophysiology, 2002, 113(6): 767-791.

[8] Shadmehr R, Smith M A, Krakauer J W. Error correction, sensory prediction, and adaptation in motor control. Annual Review of Neuroscience, 2010, 33: 89-108.

[9] Xu G H, Zhang J, Wang J. Advances in the application of brain-computer interface technology in stroke rehabilitation. Chinese Journal of Rehabilitation Medicine, 2023, 38(2): 267-272.

[10] Wang X, Feng Y, Li Y, et al. Technology-Based Compensation Assessment and Detection of Upper Extremity Activities of Stroke Survivors: Systematic Review. Journal of Neuroengineering and Rehabilitation, 2022, 19(1): 78.

