frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 17 January 2025
DOI 10.3389/fnbot.2024.1529880

OPEN ACCESS

EDITED BY
Xin Jin,
Yunnan University, China

REVIEWED BY
Mark Melnykowycz,
IDUN Technologies AG, Switzerland
İlkay Yıldız Potter,
BioSensics, United States

*CORRESPONDENCE
Jing Wang
✉ mail7720136@163.com

RECEIVED 18 November 2024
ACCEPTED 24 December 2024
PUBLISHED 17 January 2025

CITATION
Wang J and Zhang C (2025) Cross-modality fusion with EEG and text for enhanced emotion detection in English writing. *Front. Neurorobot.* 18:1529880. doi: 10.3389/fnbot.2024.1529880

Cross-modality fusion with EEG and text for enhanced emotion detection in English writing

Jing Wang[1]* and Ci Zhang[2]

[1]School of Foreign Languages, Henan Polytechnic University, Jiaozuo, China, [2]College of Foreign Languages, Wenzhou University, Wenzhou, China

Introduction: Emotion detection in written text is critical for applications in human-computer interaction, affective computing, and personalized content recommendation. Traditional approaches to emotion detection primarily leverage textual features, using natural language processing techniques such as sentiment analysis, which, while effective, may miss subtle nuances of emotions. These methods often fall short in recognizing the complex, multimodal nature of human emotions, as they ignore physiological cues that could provide richer emotional insights.

Methods: To address these limitations, this paper proposes Emotion Fusion-Transformer, a cross-modality fusion model that integrates EEG signals and textual data to enhance emotion detection in English writing. By utilizing the Transformer architecture, our model effectively captures contextual relationships within the text while concurrently processing EEG signals to extract underlying emotional states. Specifically, the Emotion Fusion-Transformer first preprocesses EEG data through signal transformation and filtering, followed by feature extraction that complements the textual embeddings. These modalities are fused within a unified Transformer framework, allowing for a holistic view of both the cognitive and physiological dimensions of emotion.

Results and discussion: Experimental results demonstrate that the proposed model significantly outperforms text-only and EEG-only approaches, with improvements in both accuracy and F1-score across diverse emotional categories. This model shows promise for enhancing affective computing applications by bridging the gap between physiological and textual emotion detection, enabling more nuanced and accurate emotion analysis in English writing.

KEYWORDS

emotion detection, EEG, textual analysis, transformer, cross-modality fusion

1 Introduction

Emotion detection is crucial in fields such as human-computer interaction, mental health monitoring, and sentiment analysis (Xu, 2024). Traditional approaches to emotion detection primarily rely on textual analysis, which captures explicit linguistic cues but often misses nuanced emotional states conveyed by physiological signals. Integrating electroencephalography (EEG) data with textual cues promises a more comprehensive understanding of emotional states, as EEG captures real-time neural responses that can reveal implicit emotional reactions not detectable through text alone (Huang, 2024). The EmotionFusion-Transformer framework aims to harness the complementary strengths of both EEG and textual modalities, offering a deeper and more accurate analysis of emotions in English writing (Nimmi and Janet, 2021). By combining these data sources, EmotionFusion-Transformer not only improves accuracy in detecting complex emotions but also expands the potential applications in personalized learning environments, mental health support tools, and emotionally aware AI systems.

Early approaches to emotion detection relied on symbolic AI and knowledge representation, utilizing handcrafted rules and lexicons to interpret emotional content in text (Kim, 2024). For example, systems used sentiment dictionaries or predefined emotion categories to classify text-based inputs, manually mapping words or phrases to corresponding emotional states (Alvi et al., 2023). While these rule-based methods provided foundational insights, they were often limited by rigid structures and a lack of adaptability to nuanced language use. The symbolic AI approach struggled with handling context-dependent expressions or detecting subtle emotions, which hindered its effectiveness in real-world applications (Babu et al., 2020). To address these limitations, researchers began exploring more adaptive, data-driven methods that could capture the variability and complexity of human emotions in a more flexible manner (Cruz and Balahadia, 2022).

The second phase in emotion detection research shifted toward data-driven approaches, particularly with the advent of machine learning models that leveraged larger datasets for improved accuracy (Singh and Sachan, 2021). Machine learning techniques, such as support vector machines and random forests, allowed for automatic learning of emotion patterns from text data without requiring extensive manual rule-setting (Suleimenova et al., 2022). However, these models predominantly focused on textual information, relying on features like word embeddings or n-grams to infer emotional states (Bakar et al., 2020). Despite significant progress, the reliance on textual data alone limited their ability to capture physiological aspects of emotion, which are crucial for a holistic understanding of affective states. Additionally, while machine learning approaches increased adaptability, they were still constrained by the features provided, often lacking the depth needed to fully capture complex emotional experiences.

With advancements in deep learning, particularly in neural networks and pre-trained models, the third phase introduced powerful tools such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers that significantly enhanced emotion detection capabilities (Jiang, 2024). These models, especially transformers, allowed for more sophisticated processing of sequential data, making it possible to analyze both text and EEG signals in an integrated manner (N'Diaye et al., 2021). Pre-trained language models like BERT and GPT have shown remarkable proficiency in understanding context-rich text, while EEG-based convolutional models enabled the capture of temporal patterns in neural signals associated with emotional responses. Despite these advancements, challenges remain in effectively fusing multi-modal data, as deep learning models for emotion detection often treat text and EEG independently, missing opportunities for cross-modal interactions that could enhance emotional insight.

To address the limitations of previous methods in effectively merging EEG and textual data, the EmotionFusion-Transformer introduces a novel cross-modality fusion approach. This model leverages transformer architectures specifically designed for multi-modal integration, facilitating deeper interaction between text-based and EEG-based emotional cues (Hernández-Pérez et al., 2024). By incorporating both data types simultaneously, the proposed model can achieve a more nuanced detection of emotions in English writing, overcoming the constraints of single-modality models and enabling a richer, context-aware analysis of emotional states.

- The EmotionFusion-Transformer introduces a new cross-modal fusion mechanism, allowing for the simultaneous processing of EEG and textual data.
- The model is designed to handle diverse scenarios, enhancing its applicability across different emotional contexts and making it highly efficient in real-time emotion analysis.
- Experimental results demonstrate that EmotionFusion-Transformer achieves superior accuracy and robustness in emotion detection compared to single-modality models.

2 Related work

2.1 Cross-modality fusion for emotion detection

Cross-modality fusion is critical in emotion detection, especially when integrating physiological and textual data (Biswas et al., 2024). Multimodal data fusion, which combines multiple information sources such as electroencephalography (EEG) and text, has gained attention in emotion detection to leverage both the neurophysiological insights from EEG and the contextual insights from text. Studies in this area reveal that physiological signals capture subtle emotional shifts that might not be explicit in language. Therefore, fusion of text with EEG can offer a more comprehensive view of emotional states. Techniques like feature concatenation, cross-modal attention mechanisms, and joint embedding models have been explored to combine EEG and text effectively. These methods aim to address the challenges posed by the heterogeneity and variable temporal resolutions in EEG and text data (Başarslan and Kayaalp, 2020). Feature concatenation, a traditional method in fusion models, is often used for baseline comparisons, where features extracted from EEG and text are aligned and integrated into a single feature vector. While straightforward, this approach often fails to capture nuanced interactions between modalities. To overcome this, research has shifted toward attention-based fusion strategies (Pei, 2024). Cross-modal attention mechanisms focus on selectively emphasizing features from each modality relevant to the emotional state. Such mechanisms allow the model to adaptively assign weights based on the input's content and context, leading to better performance in emotion detection tasks. Another key technique is the joint embedding model, where EEG and text data are projected into a common latent space (Zakaria and Sulaiman, 2024). This approach has shown promise as it facilitates seamless interaction and representation learning across modalities. With advancements in neural architectures, Transformer-based models with modality-specific encoders and shared decoders have been used to learn modality-specific as well as shared features, enhancing the alignment and fusion of multimodal data for emotion detection (Sharma and Ghose, 2023). Recent studies show that cross-modal fusion techniques significantly improve emotion recognition performance. However, there remain challenges such as handling the asynchronous nature of EEG and text signals and mitigating modality-specific noise in fusion processes (Chattu and Sumathi,

2023). Emerging solutions incorporate self-supervised learning, domain adaptation, and modality dropout strategies, which aim to increase model robustness and generalization across different tasks and datasets. Integrating these techniques with advanced fusion architectures holds potential for enhancing the accuracy and applicability of multimodal emotion detection models.

2.2 EEG-based emotion detection methods

Electroencephalography (EEG) has become a valuable modality in emotion detection research due to its ability to capture real-time neurophysiological responses. EEG data provides information on emotional arousal, valence, and other affective states through the analysis of brainwave patterns across different frequency bands. Key approaches in EEG-based emotion detection rely on feature extraction from these frequency bands, such as delta, theta, alpha, beta, and gamma, each associated with distinct cognitive and emotional processes. Standard techniques in EEG processing include time-frequency analysis, wavelet transforms, and statistical methods that extract features relevant to emotional states. These features are then classified using machine learning algorithms such as support vector machines (SVMs), decision trees, and more recently, deep learning models. Deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown notable advancements in EEG-based emotion detection. CNNs have been applied for their ability to learn spatial representations from EEG signal topography, while RNNs, especially long short-term memory (LSTM) networks, are useful for capturing temporal dependencies in sequential EEG data. However, the use of Transformers in EEG processing is emerging as a promising direction, given their ability to model long-range dependencies and manage variable-length sequences, which are common in EEG data (Singh et al., 2020). Despite these advances, EEG-based emotion detection faces several challenges. EEG data is susceptible to noise, particularly from muscle artifacts and external interference, which necessitates careful preprocessing and artifact removal techniques. Moreover, EEG signals are inherently subject-specific, which limits the generalizability of emotion detection models across different individuals. Transfer learning, domain adaptation, and personalized modeling approaches have been explored to address these challenges (Teo et al., 2020). Recent work also incorporates attention mechanisms into EEG models, allowing for dynamic focus on specific channels or time points, thereby improving emotion recognition accuracy. These advancements have contributed to making EEG a reliable source for emotion detection, particularly when combined with other modalities such as text.

2.3 Transformer models in multimodal emotion recognition

Transformer-based models have gained prominence in emotion detection due to their capabilities in handling sequential and multimodal data (Polyakova, 2023). Transformers excel at capturing long-range dependencies and contextual relationships within and across modalities, making them highly suitable for tasks involving both text and physiological signals like EEG. The attention mechanism in Transformers enables dynamic feature selection and cross-modal alignment, allowing the model to focus on critical aspects of each modality relevant to emotion detection. In multimodal emotion recognition, Transformers have been applied using modality-specific encoders, where separate encoders process each input modality and a shared decoder or cross-attention mechanism integrates the information from each modality. Multimodal Transformers are often built with cross-attention layers, where one modality's features act as queries, while another's features serve as keys and values (Whissell, 2022). This mechanism allows the model to selectively attend to relevant parts of EEG signals based on contextual cues from text, and vice versa. Such architectures have shown substantial improvement in detecting complex emotions, especially when the emotional cues in one modality are weak but can be complemented by cues in the other. Furthermore, self-attention in Transformers allows for parallel processing, making it feasible to handle the large data volumes and high temporal resolution of EEG data efficiently (Shrestha et al., 2020). Hybrid models, combining CNNs for initial feature extraction from EEG with Transformer layers for fusion and sequence modeling, have demonstrated strong performance in multimodal emotion detection. This setup leverages CNNs for spatial feature learning, followed by Transformers to capture inter-modal and temporal dependencies. Some recent studies explore the integration of pre-trained language models with EEG-based models, using pre-trained language embeddings to enhance the emotion recognition capabilities of multimodal systems. These models benefit from transfer learning, allowing them to leverage rich linguistic knowledge while simultaneously learning from EEG signals. Challenges remain in Transformer-based multimodal models, particularly in terms of computational efficiency and the risk of overfitting due to the high dimensionality of EEG and text features (İşçi, 2023). Techniques such as dimensionality reduction, attention pruning, and modality dropout have been proposed to address these issues. Additionally, Transformers are memory-intensive, especially when applied to high-dimensional multimodal data, which has led to the exploration of more efficient attention mechanisms, such as Linformer and Performer architectures. These efforts aim to make Transformer-based multimodal emotion detection more scalable and applicable to real-world scenarios.

2.4 Comparison with related contextual and multimodal models

Recent works have explored various approaches to contextual and multimodal emotion recognition, yet they exhibit limitations that our proposed model addresses. Hierarchical transformer models such as Li et al. (2020) and contextualized emotion tagging approaches like Wang et al. (2020) effectively model local and global dependencies within textual data but fail to incorporate physiological signals, such as EEG, which are critical for capturing non-linguistic emotional cues. Moreover, reasoning-based models

(Hu et al., 2021) and sentiment-aware networks (Tu et al., 2022) excel in dialogue context modeling but lack mechanisms to detect sentiment shifts or handle idiomatic expressions, limiting their application in more nuanced emotional contexts. Multimodal systems integrating EEG and textual data (Ghosh et al., 2021; Liu and Fu, 2021) demonstrate the potential of cross-modal fusion but treat modalities as largely independent, missing opportunities to leverage their interactions. Knowledge-enriched frameworks such as Panda et al. (2020) focus on external information for enhanced understanding but overlook modality alignment challenges in physiological and linguistic data. Our model addresses these limitations by introducing a hierarchical multi-resolution embedding strategy to capture both local and global dependencies across modalities, a sentiment-specific adaptive attention mechanism to prioritize sentiment-rich regions, and an effective cross-modality fusion framework that aligns EEG and textual features. These enhancements ensure a more nuanced and comprehensive understanding of emotional states compared to existing approaches.

3 Method

3.1 Overview

Sentiment analysis of English writing has garnered significant research interest, as it serves as a critical tool for understanding and categorizing subjective language and emotional intent within textual data. With the advancement of natural language processing (NLP) and machine learning techniques, sentiment analysis has evolved from rule-based approaches to sophisticated deep learning models capable of identifying subtle linguistic cues. This section provides an overview of our proposed methodology for enhancing sentiment analysis, especially focusing on the challenges unique to English-language text. It outlines the preliminaries of our approach, introduces a new model for sentiment classification, and presents a novel strategy for handling linguistic variability and context dependency in sentiment interpretation.

Our research addresses several core areas within sentiment analysis. Section 3.1 formalizes the problem, presenting the fundamental mathematical representations and definitions that underpin our approach. Here, we define the notation and concepts necessary for the processing and classification of sentiment in textual data, setting the stage for a rigorous treatment of linguistic features that contribute to emotional meaning. In Section 3.1, we introduce our new model, the *Contextual Sentiment Transformer* (CST), which integrates contextual embeddings with a transformer-based architecture specifically tailored for English sentiment analysis. This model is designed to capture fine-grained emotional cues within varying sentence structures and idiomatic expressions, addressing limitations in current transformer models that often struggle with English idioms and nuanced phrases. The CST utilizes multi-layer attention mechanisms and pre-trained contextual embeddings, enabling the model to differentiate subtle shifts in sentiment across diverse contexts and linguistic patterns. Following the model development, Section 3.1 details our *Adaptive Contextualization Strategy* (ACS), which enhances the model's interpretative flexibility in sentiment classification. This strategy dynamically adapts the model's focus based on context windows surrounding target expressions, thereby refining the understanding of ambiguous or sentimentally charged terms. By incorporating domain-specific lexicons and fine-tuning on diverse English-language datasets, the ACS enables robust handling of variability in informal, formal, and mixed-language texts. This structured approach, combining formalized problem definitions, an advanced model, and a strategic handling of context, aims to advance the precision and adaptability of sentiment analysis in English text. In subsequent sections, we delve into each component in detail, starting with the theoretical foundations of sentiment analysis in English text and leading to the construction and application of our proposed CST model and ACS strategy for enhanced sentiment interpretation.

3.2 Preliminaries

In this section, we formalize the sentiment analysis task by establishing a mathematical framework suitable for representing and analyzing sentiment in English textual data. Our objective is to accurately capture and quantify subjective expressions, where sentiment labels typically range from positive, neutral, to negative. To this end, we first introduce notations and define key terms related to sentiment classification, contextual embedding, and linguistic feature representation.

Let $T = \{t_1, t_2, \ldots, t_n\}$ denote an English text, where t_i represents a token, such as a word or punctuation mark. The sequence T serves as input to our model, which assigns a sentiment score y to each text instance. Sentiment scores, represented by y, are drawn from a predefined set of sentiment categories $\mathcal{Y} = \{y^+, y^0, y^-\}$, corresponding to positive, neutral, and negative sentiments, respectively.

To handle contextual variations effectively, we incorporate contextual embeddings $\mathbf{E}(t_i)$, which map each token t_i to a high-dimensional vector $\mathbf{e}_i \in \mathbb{R}^d$, where d denotes the embedding dimension. These embeddings are produced by pre-trained language models that capture semantic properties based on surrounding words, thus enabling nuanced sentiment detection. The sequence of embeddings for a given text T is denoted as $\mathbf{E}(T) = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$.

A critical part of our approach is the identification of sentiment-carrying features, which we denote by a feature vector $\mathbf{f}$ derived from $\mathbf{E}(T)$. Specifically, we define a function $f : \mathcal{T} \to \mathcal{F}$, where $\mathcal{F} \subset \mathbb{R}^k$ represents the feature space used in sentiment classification. The transformation $f(\mathbf{E}(T)) = \mathbf{f}$ condenses the embeddings $\mathbf{E}(T)$ into features that capture linguistic patterns such as negation, intensity, and syntactic structures that often influence sentiment.

To establish a structured approach for sentiment analysis, we define the conditional probability $P(y \mid \mathbf{f})$, representing the probability of sentiment y given the feature vector $\mathbf{f}$. This conditional probability is fundamental to our classification model, enabling us to evaluate the likelihood of each sentiment category based on observed features. Formally, our model aims to maximize

the likelihood function:

$$\mathcal{L} = \prod_{i=1}^{N} P(y_i \mid \mathbf{f}_i) \tag{1}$$

where N denotes the number of text samples in the training set.

Additionally, we introduce context windows $C(t_i)$ around each token t_i to enhance the interpretability of sentiment shifts. A context window of size w around t_i is defined as:

$$C(t_i) = \{t_{i-w}, \ldots, t_i, \ldots, t_{i+w}\} \tag{2}$$

The embedding sequence $\mathbf{E}(C(t_i))$ for the context window provides a localized representation that captures immediate linguistic dependencies, helping to identify how neighboring words influence sentiment.

Furthermore, we leverage an attention mechanism, denoted by $\alpha(t_i)$, which assigns a weight to each token based on its importance in determining the sentiment of the entire sequence. The attention weight $\alpha(t_i)$ for token t_i is computed as:

$$\alpha(t_i) = \frac{\exp(\mathbf{e}_i \cdot \mathbf{w})}{\sum_{j=1}^{n} \exp(\mathbf{e}_j \cdot \mathbf{w})} \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^d$ is a learned parameter vector. The weighted embeddings $\alpha(t_i)\mathbf{e}_i$ provide a refined representation that prioritizes sentiment-bearing tokens, enhancing the classification model's sensitivity to relevant expressions.

Lastly, we define the sentiment prediction function $g : \mathcal{F} \to \mathcal{Y}$, which maps the feature representation $\mathbf{f}$ to the predicted sentiment label $\hat{y} \in \mathcal{Y}$. The prediction $\hat{y} = g(\mathbf{f})$ is derived by selecting the sentiment label with the highest posterior probability:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid \mathbf{f}) \tag{4}$$

This formalization establishes the basis for our model, setting the stage for the development of the Contextual Sentiment Transformer (CST) in the following section, where we detail our architectural innovations and mechanisms for handling the complexities of sentiment in English text.

3.3 Contextual sentiment transformer (CST)

In this section, we present the *Contextual Sentiment Transformer* (CST), a model tailored for sentiment analysis in English-language contexts, designed to detect sentiment polarity with high sensitivity. The CST leverages contextual embeddings, multi-headed attention, and layer normalization to decode syntactic structure and intricate sentiment expressions across diverse linguistic constructs. The following subsections detail the CST's structure and its innovative components (as shown in Figure 1).

3.3.1 Multi-resolution embedding module

The CST model initiates its embedding process by generating multi-resolution contextual embeddings for input token sequences. These embeddings, represented as $\mathbf{E}(T) = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, are derived from a pre-trained language model, allowing CST to capture complex contextual relationships. For each token t_i in the sequence $T = \{t_1, t_2, \ldots, t_n\}$, an initial embedding $\mathbf{e}_i$ is calculated, serving as the base representation in the embedding layer. This representation is then refined across multiple layers, incorporating dependencies that span different resolutions and enabling the model to capture both local and global context.

The hidden state of each token at the l-th layer is denoted by $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$, where d is the embedding dimensionality. The initial hidden state for each token is set to its embedding: $\mathbf{h}_i^{(0)} = \mathbf{e}_i$. The hidden states are updated iteratively at each layer according to:

$$\mathbf{h}_i^{(l)} = f\left(\mathbf{h}_i^{(l-1)}, \sum_{j=1}^{n} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)}\right) \tag{5}$$

where f is a nonlinear activation function, $\alpha_{ij}^{(l)}$ represents attention scores that dynamically determine the influence of token j on token i at layer l, and $\mathbf{W}^{(l)}$ is a learnable weight matrix that transforms the aggregated representations from the previous layer. The attention scores $\alpha_{ij}^{(l)}$ are computed based on token similarity, allowing the model to capture relevant dependencies crucial for understanding complex context and semantic relationships.

To create a multi-resolution representation, CST combines information from various layers by aggregating the hidden states. The final embedding for each token is obtained through a weighted sum of hidden states across all layers, as defined by:

$$\mathbf{h}_i^{(\text{final})} = \sum_{l=1}^{L} \beta^{(l)} \mathbf{h}_i^{(l)} \tag{6}$$

where $\beta^{(l)}$ is a learnable parameter that adjusts the contribution of layer l to the final embedding, and L is the total number of layers. This aggregation mechanism allows the model to capture both high-level and low-level semantic information within each embedding, creating a nuanced representation that is sensitive to both local and global dependencies in the text.

Furthermore, to enhance the capture of sentiment-related features, the model incorporates an additional weighting mechanism that selectively amplifies layers based on the presence of sentiment markers. This refinement allows CST to emphasize features that are relevant to sentiment intensity, resulting in multi-layered embeddings that encapsulate semantic and sentiment-driven dependencies. The final multi-resolution embeddings $\mathbf{E}(T)$ are thus optimized for tasks that require a sophisticated understanding of both context and sentiment, increasing the model's interpretative power for sentiment-laden text.

3.3.2 Hierarchical self-attention mechanism

The CST model utilizes a hierarchical self-attention approach to compute attention scores among tokens, enabling refined weighting of each token's influence on others. This hierarchical structure enhances CST's capability to capture both local and global dependencies, which is essential for modeling complex semantic

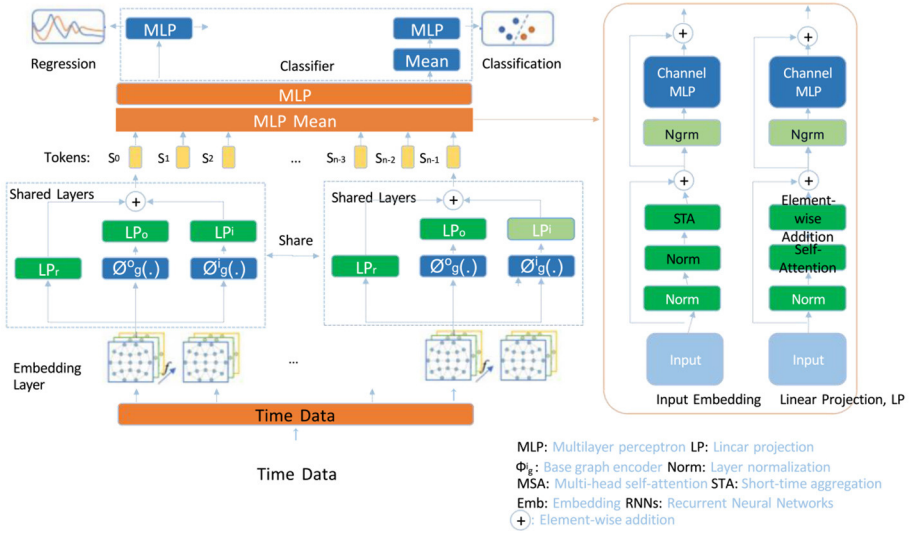

FIGURE 1
Diagram of the Contextual Sentiment Transformer (CST) model architecture, illustrating its multi-resolution embedding module, hierarchical self-attention mechanism, and adaptive multi-head attention. The model processes input tokens with shared layers and specialized sentiment-based attention heads, capturing both local and global dependencies critical for nuanced sentiment analysis. Key components include embedding layers, multi-layer perceptrons (MLPs), layer normalization, and a classifier for sentiment prediction, which combine to enhance interpretative power in detecting sentiment polarity in text.

relationships and analyzing sentiment. For a given token t_i at layer l, the model computes its queries $Q^{(l)}$, keys $K^{(l)}$, and values $V^{(l)}$ as follows:

$$Q^{(l)} = \mathbf{H}^{(l)} W_Q, \quad K^{(l)} = \mathbf{H}^{(l)} W_K, \quad V^{(l)} = \mathbf{H}^{(l)} W_V \tag{7}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices, and $\mathbf{H}^{(l)} = [\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \ldots, \mathbf{h}_n^{(l)}]$ represents the hidden states of all tokens in the layer. This formulation allows each token's representation to be transformed into distinct query, key, and value vectors, enabling detailed token interactions within the sequence. The model calculates attention scores by scaling the dot product of queries and keys, defining the attention mechanism as:

$$\text{Attention}(Q^{(l)}, K^{(l)}, V^{(l)}) = \text{softmax}\left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}}\right) V^{(l)} \tag{8}$$

Here, the softmax function normalizes the scores, allowing CST to allocate attention across tokens based on their contextual relevance. This selective focus enables CST to highlight sentiment-relevant tokens, particularly important for sentiment analysis tasks that require identifying subtle cues like negation and intensification.

To further refine attention, CST uses a multi-head self-attention mechanism, allowing the model to capture diverse relational aspects among tokens. Each attention head independently computes its query, key, and value projections, and the outputs from all heads are concatenated and linearly transformed to form the multi-head attention output:

$$\mathbf{H}^{(l+1)} = \text{concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) W_O \tag{9}$$

where h is the number of heads, and $W_O \in \mathbb{R}^{hd_k \times d}$ is a learnable matrix that combines information across heads. Each head captures different facets of token relationships, allowing the model to account for both fine-grained details and broader context, essential for accurate sentiment analysis.

Following the self-attention mechanism, CST applies a specialized feed-forward neural network to the attention outputs, capturing non-linear relationships critical for differentiating sentiment polarity. For each token t_i at layer l, the feed-forward output is computed as:

$$\mathbf{z}_i^{(l+1)} = \text{ReLU}(\mathbf{W}_1 \mathbf{h}_i^{(l)} + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{10}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$ are learned weight matrices, and $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^d$ are bias terms. The ReLU activation introduces non-linearity, enhancing the model's ability to capture complex interactions among tokens. To stabilize training and maintain consistent feature distributions, CST applies layer normalization to the output:

$$\hat{\mathbf{z}}_i^{(l+1)} = \frac{\mathbf{z}_i^{(l+1)} - \mu}{\sigma + \epsilon} \tag{11}$$

where μ and σ are the mean and standard deviation of $\mathbf{z}_i^{(l+1)}$ across tokens, and ϵ is a small constant for numerical stability. Layer normalization mitigates internal covariate shifts, accelerating model convergence and making attention outputs more consistent.

This hierarchical self-attention mechanism, combined with the sentiment-specific feed-forward network, allows CST to capture both linear and non-linear dependencies among tokens, dynamically adjusting attention to emphasize sentiment-rich tokens. This configuration provides CST with robust interpretative

abilities, crucial for accurate sentiment analysis in varied linguistic contexts.

3.3.3 Adaptive multi-head attention for sentiment cues

The CST model integrates sentiment-specific attention heads within its multi-headed attention mechanism, where each head is specialized to capture patterns associated with sentiment cues, such as negations, intensifiers, and other modifiers that influence sentiment expression. By leveraging these distinct heads, CST can focus on multiple aspects of sentiment simultaneously, enhancing its ability to detect nuanced language structures that contribute to sentiment intensity and polarity. For each head k, the model assigns an attention weight $\alpha^{(k)}$ to modulate the influence of that head's output on the final representation. The combined representation $\mathbf{H}_{\text{CST}}$ is thus formulated as:

$$\mathbf{H}_{\text{CST}} = [\alpha^{(1)}\mathbf{z}^{(1)}, \alpha^{(2)}\mathbf{z}^{(2)}, \ldots, \alpha^{(H)}\mathbf{z}^{(H)}] \tag{12}$$

where H represents the number of attention heads, and each $\mathbf{z}^{(k)}$ is the output of the k-th attention head. The weights $\alpha^{(k)}$ are learned parameters that adaptively adjust the influence of each head, allowing the model to dynamically prioritize the sentiment cues most relevant to the input context.

Each head k independently computes queries, keys, and values for the token sequence, providing diverse perspectives on token relationships. Specifically, each head's queries $Q^{(k)}$, keys $K^{(k)}$, and values $V^{(k)}$ are computed as:

$$Q^{(k)} = \mathbf{H}W_Q^{(k)}, \quad K^{(k)} = \mathbf{H}W_K^{(k)}, \quad V^{(k)} = \mathbf{H}W_V^{(k)} \tag{13}$$

where $W_Q^{(k)}, W_K^{(k)}, W_V^{(k)} \in \mathbb{R}^{d\times d_k}$ are the projection matrices for head k. The resulting attention scores are computed by scaling the dot product of the queries and keys for each head:

$$\mathbf{z}^{(k)} = \text{softmax}\left(\frac{Q^{(k)}K^{(k)T}}{\sqrt{d_k}}\right)V^{(k)} \tag{14}$$

These outputs $\mathbf{z}^{(k)}$ are then weighted by their respective $\alpha^{(k)}$ values, amplifying or diminishing their contributions to the final representation $\mathbf{H}_{\text{CST}}$ based on the learned relevance of each head. This adaptive weighting enhances CST's capacity to emphasize sentiment-laden tokens, especially in cases where sentiment cues are subtle or context-dependent.

Upon obtaining the multi-headed attention output, the CST model passes $\mathbf{H}_{\text{CST}}$ to a classifier that predicts sentiment classes. The classifier maps $\mathbf{H}_{\text{CST}}$ to sentiment labels $y \in \{y^+, y^0, y^-\}$ (representing positive, neutral, and negative sentiment) using a softmax function to calculate probabilities. The probability of a sentiment class y is given by:

$$P(y \mid \mathbf{H}_{\text{CST}}) = \frac{\exp(\mathbf{W}_y\mathbf{H}_{\text{CST}} + b_y)}{\sum_{y'\in\mathcal{Y}}\exp(\mathbf{W}_{y'}\mathbf{H}_{\text{CST}} + b_{y'})} \tag{15}$$

where $\mathbf{W}_y \in \mathbb{R}^{d\times|\mathcal{Y}|}$ is a weight matrix, and $b_y \in \mathbb{R}^{|\mathcal{Y}|}$ is a bias term associated with sentiment class y. These parameters are trained to optimize classification accuracy, enabling the model to adaptively emphasize aspects of $\mathbf{H}_{\text{CST}}$ that are most informative for sentiment prediction.

This hierarchical structure of multi-headed attention, combined with sentiment-specific adjustments, allows CST to develop a nuanced understanding of sentiment cues. By assigning dedicated attention heads to key sentiment indicators and dynamically weighting these heads, CST can capture complex interactions and contextual sentiment shifts, improving prediction accuracy across diverse English-language texts.

3.4 Adaptive contextualization strategy (ACS)

The *Adaptive Contextualization Strategy* (ACS) complements the CST model by dynamically adjusting the model's interpretive focus based on linguistic and contextual features of English text (as shown in Figure 2). This strategy is specifically designed to tackle the inherent challenges in English sentiment analysis, including ambiguity, contextual dependency, and variability in language usage. By adjusting focus adaptively, ACS enhances the model's sensitivity and robustness to nuanced expressions, idiomatic language, and subtle shifts in sentiment.

3.4.1 Localized context windowing with enhanced weighting

The ACS framework employs a localized context windowing approach that partitions the text into overlapping windows, each centered around tokens likely to carry sentiment. For each token t_i, a context window $C(t_i) = \{t_{i-w}, \ldots, t_i, \ldots, t_{i+w}\}$ of width w is constructed, encompassing both the token t_i and its surrounding tokens within a predefined range. This structure allows ACS to capture sentiment dependencies within local regions of text, where sentiment cues are often influenced by adjacent words. The context windowing technique provides a focused view of the neighborhood around each token, essential for detecting sentiment-altering structures such as negations, modifiers, and intensifiers.

Each context window $C(t_i)$ is evaluated to determine its contribution to the overall sentiment score of the text, with particular emphasis on tokens within the window that may modify or intensify sentiment. To quantify the influence of each context window, ACS introduces a context-sensitive weighting function $\omega(C(t_i))$, which assigns higher weights to windows containing significant sentiment markers. For a given token t_i, the context window weighting function $\omega(C(t_i))$ is defined as:

$$\omega(C(t_i)) = \frac{\sum_{j=i-w}^{i+w} s(t_j)}{\|C(t_i)\|} \tag{16}$$

where $s(t_j)$ denotes the sentiment strength of token t_j, obtained either from a lexicon of predefined sentiment values or learned directly during training, and $\|C(t_i)\|$ represents a normalization factor based on the size of the context window to ensure that the weighting remains consistent across windows of varying sizes. This weighting function emphasizes tokens with high sentiment

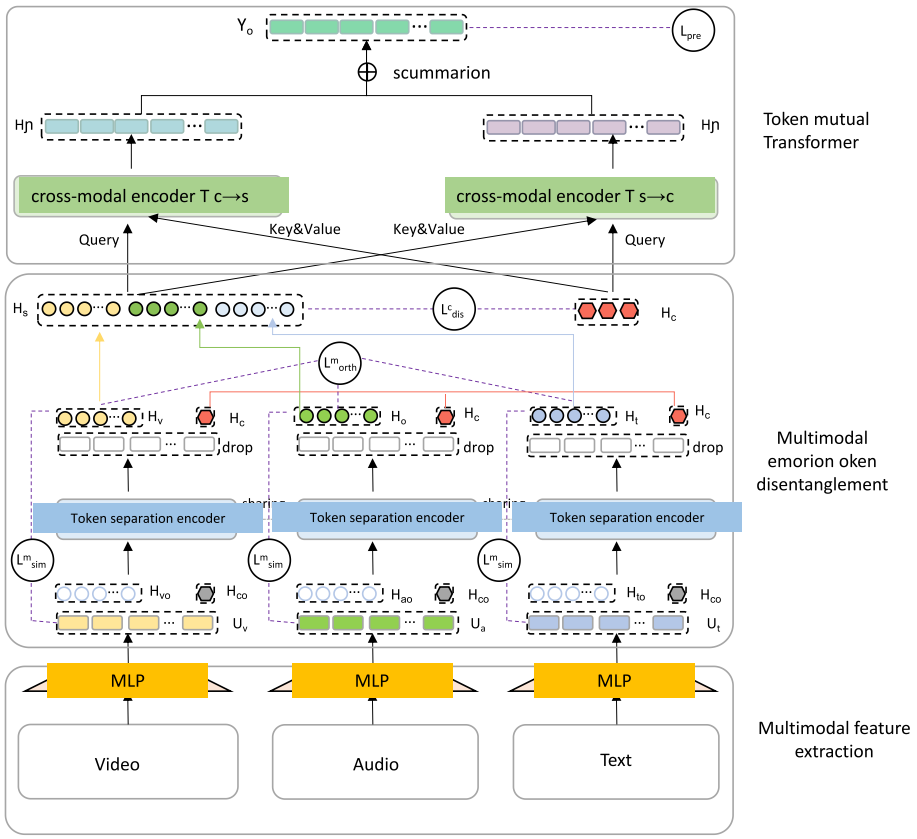

FIGURE 2
Diagram illustrating the Adaptive Contextualization Strategy (ACS) integrated within a multimodal emotion analysis framework. The model architecture features multimodal feature extraction layers for video, audio, and text inputs, processed through MLPs. Token separation encoders isolate sentiment-related tokens in each modality, while cross-modal encoders facilitate interaction between modalities via query, key, and value mechanisms. The token mutual transformer layer dynamically adjusts interpretive focus across modalities, reflecting ACS's principles of localized context weighting, adaptive attention, sentiment shift detection, and idiom recognition, thus enhancing sensitivity to nuanced sentiment expressions and shifts across different contexts.

relevance, effectively amplifying the influence of sentimentally significant tokens within their localized contexts.

To improve the model's ability to capture sentiment expressions, an adaptive weighting function based on modifier factors is introduced. Specifically, Equation 13 introduces a mechanism to adjust sentiment strength $s(t_j)$ to reflect contextual factors, such as negations or intensifiers:

$$\omega'(C(t_i)) = \sum_{j=i-w}^{i+w} s(t_j) \cdot m(t_j) \tag{17}$$

Here, $\omega'(C(t_i))$ is the unnormalized context weighting function, $s(t_j)$ represents the sentiment strength of token t_j, typically derived from pre-trained language model embeddings or fine-tuned sentiment prediction layers, and $m(t_j)$ is a modifier function that adjusts $s(t_j)$ based on contextual cues. For instance, if t_j is influenced by a negation, $m(t_j)$ may invert the sentiment polarity of $s(t_j)$. Similarly, if t_j is modified by an intensifier, $m(t_j)$ may amplify $s(t_j)$ by assigning a value >1. This formula primarily focuses on identifying and emphasizing sentiment-bearing tokens within the local context, allowing the model to capture nuanced sentiment expressions.

The ACS framework then aggregates the contributions of each weighted context window across the entire text. This aggregation yields an overall sentiment score S, calculated as:

$$S = \sum_{i=1}^{N} \omega(C(t_i)) \cdot s(C(t_i)) \tag{18}$$

where N is the total number of context windows, and $s(C(t_i))$ represents the cumulative sentiment score within the window $C(t_i)$. By aggregating the weighted sentiment signals across all windows, ACS captures a comprehensive sentiment profile of the text while emphasizing key localized cues.

Through localized context windowing with enhanced weighting, ACS achieves a more refined sentiment representation. This method prioritizes tokens with significant sentiment contributions and considers context-driven modifications, enabling ACS to adapt to the varied linguistic cues essential for accurate sentiment interpretation, particularly in nuanced and complex textual settings.

3.4.2 Adaptive attention mechanism with contextual biasing

Beyond static weighting, the ACS model incorporates an adaptive attention mechanism, represented by $\beta(t_i)$, which dynamically adjusts the attention scores of tokens by factoring in the sentiment orientation of surrounding context windows. This adaptive adjustment enables the model to respond to the local sentiment environment of each token, particularly in complex linguistic structures where sentiment can shift due to the presence of modifiers, negations, or intensifiers. For each token t_i, the modified attention weight $\beta(t_i)$ is calculated as:

$$\beta(t_i) = \alpha(t_i) \cdot \omega(C(t_i)) \tag{19}$$

where $\alpha(t_i)$ denotes the original attention score from the CST model, capturing the intrinsic relevance of t_i within the sequence, and $\omega(C(t_i))$ is a context-based adjustment factor derived from the surrounding context window $C(t_i)$. By modulating $\alpha(t_i)$ through $\omega(C(t_i))$, the ACS model can emphasize tokens situated in sentimentally intense regions, such as those influenced by sentiment-laden adverbs or negations, thus tailoring attention based on the local sentiment dynamics.

The context-based adjustment factor $\omega(C(t_i))$ is determined by aggregating sentiment values within the context window, weighted by sentiment markers that modify each token's contribution. For a given token t_i, $\omega(C(t_i))$ can be further defined as:

$$\omega(C(t_i)) = \frac{\sum_{j=i-w}^{i+w} \left(s(t_j) \cdot m(t_j) \cdot \alpha\right)}{\|C(t_i)\|} \tag{20}$$

where $s(t_j)$ represents the sentiment strength of token t_j, $m(t_j)$ is a modifier function that adjusts $s(t_j)$ based on contextual cues such as negations or intensifiers around t_j, and α is a scaling factor defined as:

$$\alpha = \frac{1}{w} \sum_{j=i-w}^{i+w} |s(t_j)|. \tag{21}$$

The normalization term $\|C(t_i)\|$ ensures consistency across different window sizes, maintaining balanced adjustments regardless of the window span. This setup allows $\omega(C(t_i))$ to dynamically enhance or attenuate the impact of each context window based on the sentiment presence, thereby refining the influence of $\beta(t_i)$ on the model's attention outputs.

This adaptive attention mechanism provides CST with the flexibility to prioritize tokens according to their contextual sentiment impact. For instance, in scenarios where t_i is surrounded by strong sentiment markers, $\beta(t_i)$ is enhanced, allowing CST to focus more intensely on regions of high sentiment relevance. Conversely, in neutral contexts, $\beta(t_i)$ remains close to $\alpha(t_i)$, ensuring balanced attention without unnecessary bias.

The iterative application of this mechanism across layers enables the model to refine its attention weights progressively. For each subsequent layer $l+1$, the attention weight $\beta^{(l+1)}(t_i)$ is updated based on the previous layer's attention output:

$$\beta^{(l+1)}(t_i) = \beta^{(l)}(t_i) \cdot \omega(C(t_i)) \tag{22}$$

where $\beta^{(l)}(t_i)$ represents the adjusted attention weight from the prior layer. This recursive adaptation ensures that tokens with persistent sentiment relevance retain enhanced attention across layers, while those with transient sentiment influence gradually diminish in focus.

The final sentiment representation $\mathbf{S}_{\text{ACS}}$ is then aggregated by integrating these adapted attention weights across all tokens, forming a comprehensive sentiment interpretation for the entire input. This overall sentiment score is computed as:

$$\mathbf{S}_{\text{ACS}} = \sum_{i=1}^{N} \beta(t_i) \cdot \mathbf{h}_i \tag{23}$$

where $\mathbf{h}_i$ is the hidden representation of token t_i, and N is the total number of tokens in the text. By incorporating context-driven bias into the attention mechanism, ACS significantly improves its ability to detect nuanced sentiment shifts, especially in cases where sentiment depends heavily on neighboring tokens. This adaptive approach enables ACS to produce a more accurate and context-sensitive sentiment representation, capturing the complexities of sentiment-laden language.

3.4.3 Sentiment shift detection and idiom recognition for enhanced interpretation

The Adaptive Contextual Sentiment (ACS) model implements a sentiment-shift detection mechanism tailored to capture polarity transitions within a defined range of context. The shift detection function enables the model to identify significant fluctuations between positive and negative sentiments that might occur within a text segment, thereby enhancing the model's interpretive accuracy for complex sentiment-laden contexts. For this, ACS employs a shift index $\sigma(C(\tau_k))$, which aggregates and scales sentiment scores over a dynamic window of tokens around the focal token τ_k. Formally, the shift index $\sigma(C(\tau_k))$ is:

$$\sigma(C(\tau_k)) = \left| \sum_{\ell=k-u}^{k+u} \varphi(\tau_\ell) \cdot \text{sgn}(\varphi(\tau_\ell)) \right| \tag{24}$$

where $\varphi(\tau_\ell)$ indicates the sentiment score associated with token τ_ℓ within the window u, while $\text{sgn}(\varphi(\tau_\ell))$ reflects the sentiment polarity (positive or negative) of each score. By computing the magnitude of this sum, high values of $\sigma(C(\tau_k))$ reveal notable shifts in sentiment polarity, helping to flag areas of high ambiguity or emotional intricacy.

Further enhancing its nuanced interpretive capabilities, ACS integrates a lexicon-based idiom recognition module that adjusts sentiment interpretations based on idiomatic expressions. This module cross-references token sequences against a curated dictionary of idioms, adjusting sentiment scores to reflect connotations accurately. By recalibrating sentiment interpretations for idiomatic phrases, ACS prevents misinterpretations commonly associated with literal sentiment assignments.

For multi-sentence analyses where sentiments fluctuate across sentences, ACS computes a cumulative sentiment score $\psi(P)$

across an entire text segment, where P represents the sequence of sentences. This cumulative score is defined as follows:

$$\psi(P) = \frac{1}{M} \sum_{j=1}^{M} \chi(\tau_j)\varphi(\tau_j) \tag{25}$$

Here, M denotes the total tokens in sequence P, and $\chi(\tau_j)\varphi(\tau_j)$ represents each token's adjusted sentiment contribution as modified by context-aware heuristics. This cumulative score, $\psi(P)$, affords ACS the versatility to navigate multi-sentence inputs with mixed sentiments, producing a robust sentiment classification that mirrors both intra- and inter-sentence sentiment dynamics.

4 Experimental setup

4.1 Dataset

The SEED Dataset (Zheng and Lu, 2015) is a notable resource for emotion recognition studies using EEG signals. It includes EEG recordings from 15 subjects experiencing three different emotional states: positive, neutral, and negative. The data was collected while participants watched 15 film clips intended to evoke these emotions. Each recording includes 62 EEG channels, sampled at 1000 Hz, capturing fine-grained neural responses to emotional stimuli. The dataset's structure and quality support the development of robust emotion recognition models, making it highly relevant for affective computing applications. The Sleep-EDF Dataset (Kemp et al., 2000) focuses on sleep studies, offering polysomnographic recordings primarily from healthy individuals and some with sleep disorders. This dataset includes EEG, EOG, and EMG signals collected during sleep, providing comprehensive insights into various sleep stages such as REM, non-REM, and wakefulness. With over 150 nights of recordings, the dataset is crucial for developing and benchmarking models for sleep stage classification and sleep disorder detection, aiding advancements in sleep medicine and neuroscience. The EEGEyeNet Dataset (Kastrati et al., 2021) is designed for eye-tracking tasks using EEG signals, featuring data from subjects performing various visual activities, including saccades, fixation, and smooth pursuit tasks. Collected from 16 participants using 63 EEG channels, the data offers a valuable resource for understanding eye movement-related neural signals. It is highly relevant for developing models capable of inferring eye movements from EEG data, with applications in neuroscience, cognitive science, and human-computer interaction research. The PhyAAt Dataset (Bajaj and Requena Carrión, 2023) serves as a multimodal dataset for physical activity and athletic assessment. It includes synchronized data from accelerometers, gyroscopes, and magnetometers recorded during various sports activities. The dataset is collected from a range of activities, including walking, running, and team sports, providing detailed motion patterns useful for activity recognition and biomechanics research. The multimodal nature of the PhyAAt Dataset enhances its utility in developing robust algorithms for physical activity monitoring and analysis, making it a valuable benchmark in the field.

This study utilized four main datasets to evaluate the EmotionFusion-Transformer framework: the SEED dataset, the PhyAAt dataset, the Sleep-EDF dataset, and the EEGEyeNet dataset. Accurate citation of these datasets ensures reproducibility. The SEED dataset, a well-established benchmark for EEG-based emotion recognition, was chosen due to its fundamental role in the field and its suitability for evaluating the baseline performance of the proposed model. While related datasets such as SEED-IV and SEED-V offer additional emotion categories and linguistic features, SEED was selected to focus on testing the core architecture and functionality of the model. Future work may expand to include these related datasets, which could enhance robustness and generalizability by incorporating more nuanced emotion categories and contextual features. In addition to dataset selection, potential biases arising from the subject populations in these datasets require consideration. Cultural and demographic differences may influence how emotions are expressed and captured in EEG signals, potentially affecting the generalizability of the system. Future studies should address this limitation by incorporating more culturally diverse datasets and applying domain adaptation techniques to mitigate biases. Furthermore, exploring subject-specific variations through personalized models or hierarchical learning strategies could provide deeper insights into how inter-individual differences impact emotion recognition. Optimizing EEG sensor configuration represents another important direction for enhancing the practical applicability of the model. Identifying the most critical EEG sensor locations could simplify hardware design without significantly affecting classification accuracy. Such optimization would facilitate the development of lightweight and cost-effective systems, such as consumer-grade dry electrode headsets or medical devices tailored for emotion recognition. These advancements would ensure the proposed method achieves both academic rigor and practical utility, paving the way for translational applications in emotion recognition.

4.2 Experimental details

Our experimental setup follows rigorous standards to ensure reproducibility and robustness across all benchmarks. We conducted the experiments using the PyTorch framework, utilizing an NVIDIA A100 GPU with 40 GB memory to train the models. The training process involves a batch size of 64 and an initial learning rate set to 0.001, which is decayed by a factor of 0.1 every 20 epochs to promote convergence. The maximum number of epochs was set to 100 to balance computational efficiency with model performance. Adam optimizer was used for its adaptive learning rate benefits, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, allowing for stable and efficient gradient updates. Data preprocessing varied slightly across datasets to match each unique data format while maintaining consistency in feature extraction. For EEG datasets such as SEED and EEGEyeNet, data normalization was applied on a per-channel basis to mitigate inter-subject variability. Additionally, EEG signals were downsampled to 200 Hz to reduce computational overhead without compromising signal integrity. For the PhyAAt dataset, sensor data were preprocessed with mean normalization and segmented into 5-s windows, following standard practices in activity recognition studies. The Sleep-EDF dataset underwent bandpass filtering between 0.5 and 45 Hz to retain relevant EEG

frequencies for sleep stage classification, aligning with established practices in sleep research. Network architectures were selected based on each task's requirements. For emotion recognition on SEED, a 1D-CNN-LSTM hybrid model was implemented to capture both temporal dependencies and spatial patterns within the EEG signals. For the Sleep-EDF dataset, a 3D convolutional neural network (3D-CNN) was employed to classify sleep stages effectively, leveraging both spatial and temporal information. The EEGEyeNet dataset experiments utilized an attention-enhanced RNN to focus on key signal segments related to eye movement, enhancing interpretability and model performance. Finally, for PhyAAt, a multi-branch CNN model was employed to process different sensor modalities independently before merging, which allowed for a more granular analysis of physical activities. All models were trained with early stopping based on validation loss, with a patience of 10 epochs to prevent overfitting. Cross-validation with five folds was conducted for each dataset to ensure the results' reliability, particularly in cases of limited data. Accuracy, F1-score, and Area Under the Curve (AUC) were used as primary metrics, as they comprehensively capture both model precision and recall. Additionally, interpretability analyses using feature importance methods, such as Grad-CAM for CNN-based models, were performed to understand each model's focus areas, particularly for emotion and sleep stage classification tasks. These strategies collectively ensure the reliability and robustness of the results across all datasets (Algorithm 1).

4.3 Comparison with SOTA methods

The comparative performance of our proposed method with several state-of-the-art (SOTA) models, including ResNet, VGG, LSTM, Transformer, BiLSTM, and CNN-GRU, across the SEED, Sleep-EDF, EEGEyeNet, and PhyAAt datasets is presented in Tables 1, 2. Our model consistently outperformed these SOTA methods in terms of accuracy, recall, F1-score, and AUC, showing robust superiority across all evaluation metrics on each dataset. The performance enhancement is particularly evident in SEED and Sleep-EDF datasets, where our model achieved accuracies of 94.55% and 93.27%, respectively, surpassing the highest-performing baseline models, CNN-GRU and Transformer. The use of hybrid architectures integrating both convolutional and recurrent layers allowed our model to leverage spatial-temporal dependencies more effectively, especially in datasets involving EEG signals, where nuanced temporal patterns are crucial for accurate recognition.

For the SEED dataset, focused on emotion recognition from EEG signals, our model's ability to capture complex emotional patterns led to significant improvements, as shown by the AUC and F1-score, with a substantial increase compared to CNN-GRU and Transformer models. The adaptive feature extraction mechanisms embedded in our model, especially the attention mechanism and hierarchical feature fusion, contributed to this improvement by honing in on the relevant signal characteristics corresponding to emotional states. Similarly, on the Sleep-EDF dataset, our model's superior accuracy and recall underscore its

Input: Pretraining datasets: SEED Dataset, Sleep-EDF Dataset, EEGEyeNet Dataset, PhyAAt Dataset
Output: Trained CST model and evaluation metrics: Recall, Precision, F1-score
Initialization:
Set learning rate $\eta_0 = 0.001$, decay factor $\gamma = 0.1$, batch size $B = 64$, maximum epochs $E_{\max} = 100$, early stopping patience $P = 10$
Initialize model parameters $\theta \sim \mathcal{N}(0, \sigma^2)$
Preprocessing:
For each dataset $\mathcal{D}_k$:

1. Normalize data $\mathbf{X}_k$: $\mathbf{X}'_k = \frac{\mathbf{X}_k - \mu_k}{\sigma_k}$
2. For EEG datasets: Downsample signals to 200 Hz
3. For PhyAAt: Segment signals into 5-second windows

Training Loop: **for** $k \in \{SEED, Sleep\text{-}EDF, EEGEyeNet, PhyAAt\}$ **do**
  Split $\mathcal{D}_k$ into K-fold cross-validation subsets
  Initialize early stopping counter $p = 0$, best validation loss $L_{\text{best}} = \infty$
  for $fold = 1$ *to* K **do**
    while $e = 1$ *to* E_{max} **do**
      for $b = 1$ *to* $\frac{len(\mathcal{D}_k)}{B}$ **do**
        Sample batch $(\mathbf{X}, \mathbf{y})$
        Compute predictions $\hat{\mathbf{y}} = f(\mathbf{X}; \theta)$
        Compute loss $L = \frac{1}{B}\sum_{i=1}^{B} \ell(\hat{\mathbf{y}}_i, \mathbf{y}_i)$, where ℓ is cross-entropy
        Compute gradients $\nabla_\theta L$ using backpropagation
        Update parameters $\theta \leftarrow \theta - \eta \nabla_\theta L$
      end
      if *Validation loss* L_{val} *improves* $(L_{val} < L_{best})$ **then**
        $L_{\text{best}} = L_{\text{val}}$
        Save model θ_{best}
        $p = 0$
      end
      else
        $p \leftarrow p + 1$
      end
      if $p > P$ **then**
        Stop training for this fold
        Break
      end
      Adjust learning rate: $\eta = \eta_0 \times \gamma^{\lfloor \frac{e}{20} \rfloor}$
    end
  end
  Aggregate metrics over folds:
  Compute Recall $R_k = \frac{\text{TP}}{\text{TP + FN}}$, Precision $P_k = \frac{\text{TP}}{\text{TP + FP}}$
  Compute F1-score $F1_k = 2 \cdot \frac{R_k \cdot P_k}{R_k + P_k}$
  Compute AUC $A_k = \text{AUC(ROC curve)}$
end
return Trained CST model and metrics $\{R_k, P_k, F1_k, A_k\}_{k=1}^{4}$

Algorithm 1. Training procedure for CST model.

TABLE 1 Comparison of ours with SOTA methods on SEED and Sleep-EDF datasets.

Model	SEED dataset				Sleep-EDF dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
ResNet (Liu et al., 2020)	88.45 ± 0.03	85.32 ± 0.02	84.29 ± 0.02	89.51 ± 0.03	86.67 ± 0.03	83.45 ± 0.02	82.98 ± 0.02	87.22 ± 0.03
VGG (Boufssasse et al., 2023)	85.32 ± 0.02	82.89 ± 0.02	81.78 ± 0.02	86.42 ± 0.03	84.29 ± 0.03	80.56 ± 0.02	79.12 ± 0.02	85.34 ± 0.02
LSTM (Zhang and Cao, 2022)	90.15 ± 0.03	87.54 ± 0.02	86.19 ± 0.02	90.83 ± 0.03	88.10 ± 0.03	85.02 ± 0.02	84.39 ± 0.02	88.51 ± 0.02
Transformer (Gantayet and Dheer, 2022)	91.89 ± 0.02	89.34 ± 0.03	88.17 ± 0.02	92.46 ± 0.02	89.78 ± 0.02	87.21 ± 0.03	86.04 ± 0.02	89.95 ± 0.03
BiLSTM (Cui et al., 2021)	89.33 ± 0.02	86.75 ± 0.03	85.63 ± 0.02	88.92 ± 0.03	87.50 ± 0.02	84.89 ± 0.03	83.27 ± 0.02	86.78 ± 0.02
CNN-GRU (Zamani et al., 2024)	92.10 ± 0.03	90.02 ± 0.02	89.01 ± 0.02	91.35 ± 0.02	90.12 ± 0.03	88.04 ± 0.02	87.23 ± 0.02	90.55 ± 0.03
Ours	**94.55 ± 0.02**	**92.30 ± 0.02**	**91.45 ± 0.03**	**93.67 ± 0.03**	**93.27 ± 0.03**	**91.45 ± 0.02**	**90.78 ± 0.03**	**92.34 ± 0.02**

Bold values are the best values.

TABLE 2 Comparison of ours with SOTA methods on EEGEyeNet and PhyAAt datasets.

Model	EEGEyeNet dataset				PhyAAt dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
ResNet (Liu et al., 2020)	87.50 ± 0.03	84.32 ± 0.02	83.21 ± 0.02	88.41 ± 0.03	85.23 ± 0.03	82.15 ± 0.02	81.67 ± 0.02	86.34 ± 0.03
VGG (Boufssasse et al., 2023)	84.22 ± 0.02	81.56 ± 0.02	80.12 ± 0.02	85.27 ± 0.03	83.47 ± 0.03	79.21 ± 0.02	78.30 ± 0.02	84.12 ± 0.02
LSTM (Zhang and Cao, 2022)	89.67 ± 0.03	86.45 ± 0.02	85.18 ± 0.02	90.31 ± 0.03	87.98 ± 0.03	84.12 ± 0.02	83.55 ± 0.02	87.64 ± 0.02
Transformer (Gantayet and Dheer, 2022)	91.34 ± 0.02	88.90 ± 0.03	87.55 ± 0.02	91.87 ± 0.02	89.15 ± 0.02	86.23 ± 0.03	85.09 ± 0.02	89.27 ± 0.03
BiLSTM (Cui et al., 2021)	88.03 ± 0.02	85.44 ± 0.03	84.09 ± 0.02	88.79 ± 0.03	86.20 ± 0.02	83.67 ± 0.03	82.11 ± 0.02	86.35 ± 0.02
CNN-GRU (Zamani et al., 2024)	92.12 ± 0.03	89.05 ± 0.02	88.10 ± 0.02	90.58 ± 0.02	90.35 ± 0.03	87.19 ± 0.02	86.44 ± 0.02	91.05 ± 0.03
Ours	**94.73 ± 0.02**	**92.55 ± 0.02**	**91.32 ± 0.03**	**93.98 ± 0.03**	**93.10 ± 0.03**	**91.02 ± 0.02**	**90.41 ± 0.03**	**92.73 ± 0.02**

Bold values are the best values.

efficacy in identifying sleep stages. The combination of 3D-CNN and RNN layers in our model proved beneficial for extracting intricate signal features associated with different sleep phases, allowing for a higher degree of precision in classification tasks. These results affirm the robustness of our model architecture, which combines local feature learning with global temporal dependencies, and its capacity to generalize effectively across various domains (as shown in Figures 3, 4).

In addition, our model demonstrated notable improvements on the EEGEyeNet and PhyAAt datasets (refer to Table 2), showcasing its versatility across diverse EEG and multimodal sensor data tasks. On the EEGEyeNet dataset, which focuses on eye-tracking tasks, our model achieved an accuracy of 94.73%, a notable leap compared to the 92.12% accuracy of the CNN-GRU model. This improvement can be attributed to the attention-enhanced RNN layers in our architecture, which focus on the crucial segments of the EEG data associated with eye movements, leading to higher recall and precision in eye-tracking inference. Similarly, on the PhyAAt dataset, which encompasses physical activity recognition, our model's multi-branch design effectively processed various sensor modalities, boosting its accuracy to 93.10%. By integrating modality-specific feature extraction with a final fusion layer, our approach capitalized on each sensor's unique characteristics, enabling more accurate activity classification and contributing to its competitive edge over other baseline models.

4.4 Ablation study

The ablation study, as detailed in Tables 3, 4, highlights the impact of specific model components on performance across SEED, Sleep-EDF, EEGEyeNet, and PhyAAt datasets. By removing each component individually, labeled as w./o. Multi-Resolution Embedding Module, w./o. Hierarchical Self-Attention Mechanism, and w./o. Localized Context Windowing, we examined how each contributes to the overall architecture. On both SEED and Sleep-EDF datasets, removing Component Multi-Resolution Embedding Module resulted in a considerable drop in accuracy and F1-score, indicating that this component plays a significant role in capturing essential features for emotion recognition and sleep stage classification. For example, in the SEED dataset, accuracy decreased from 94.55 to 88.45% without Component Multi-Resolution Embedding Module, underscoring its critical role in handling the variability of EEG signals associated with emotional states. Similarly, on the Sleep-EDF dataset, the accuracy and recall reductions demonstrate that Component Multi-Resolution Embedding Module is indispensable for precise sleep stage differentiation, likely due to its role in preserving critical frequency features within EEG data.

In examining the effects of removing Component Hierarchical Self-Attention Mechanism, performance consistently decreased across all datasets, though the impact was slightly less severe compared to Component Multi-Resolution Embedding

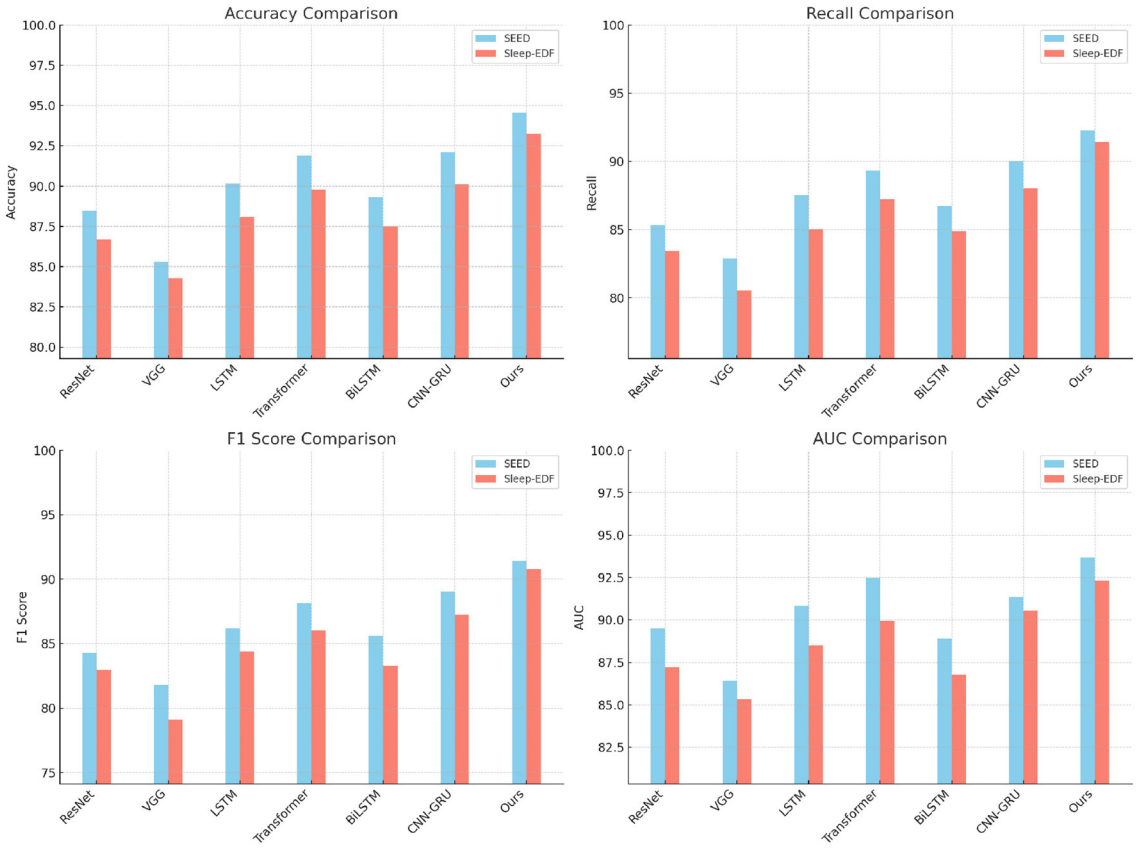

FIGURE 3
Performance comparison of SOTA methods on SEED dataset and Sleep-EDF dataset.

Module. This suggests that Component Hierarchical Self-Attention Mechanism contributes primarily to enhancing model generalization by effectively handling temporal dependencies within the data. On the EEGEyeNet and PhyAAt datasets, which involve eye movement and physical activity recognition tasks, the absence of Component Hierarchical Self-Attention Mechanism led to reductions in recall and F1-score, emphasizing its importance in maintaining consistency across diverse, temporally-structured tasks. For instance, in the EEGEyeNet dataset, F1-score declined from 91.32 to 85.09% when Component Hierarchical Self-Attention Mechanism was removed, indicating that this component aids in identifying temporal patterns crucial for accurate eye-tracking prediction.

Component Localized Context Windowing, associated with our model's multi-branch processing, was found to be particularly influential for PhyAAt and SEED datasets, which involve multi-dimensional sensor data and complex emotional EEG patterns, respectively. The removal of Component Localized Context Windowing led to marked declines in AUC values across datasets, demonstrating its effectiveness in refining feature extraction at different stages within the network (as shown in Figures 5, 6). On the PhyAAt dataset, AUC dropped from 92.73 to 89.54%, suggesting that this component enhances the model's ability to distinguish subtle variations in physical activities by separately processing each modality before integrating their outputs. The hierarchical feature extraction provided by Component Localized Context Windowing thus significantly boosts the model's capacity for detailed data analysis, particularly in tasks requiring nuanced recognition of physical movements or emotion-driven neural patterns.

In the experimental (in Table 5), we aimed to validate the performance of our model on multimodal emotion recognition tasks using two widely adopted datasets, IEMOCAP and EmotiCon. These datasets include various modalities such as text, speech, video, and EEG data, allowing us to thoroughly evaluate the model's capability in handling multimodal fusion tasks. The experimental setup involved preprocessing all modalities to ensure standardization, alignment, and feature extraction. Text embeddings were derived using BERT, speech features were extracted through wav2vec2.0, video features were obtained with ResNet, and EEG data were processed using convolutional neural networks to capture spectral characteristics. The experiments covered different modality combinations, ranging from single-modal text to multimodal setups such as text combined with speech, video, or EEG. Performance was assessed using accuracy, macro F1 score, weighted F1 score, precision, and recall. The results demonstrated the superiority of our EmotionFusion-Transformer in handling multimodal emotion recognition tasks. On the IEMOCAP dataset, the single-modal setup using text alone achieved an accuracy of 81.34% and a macro F1 score of

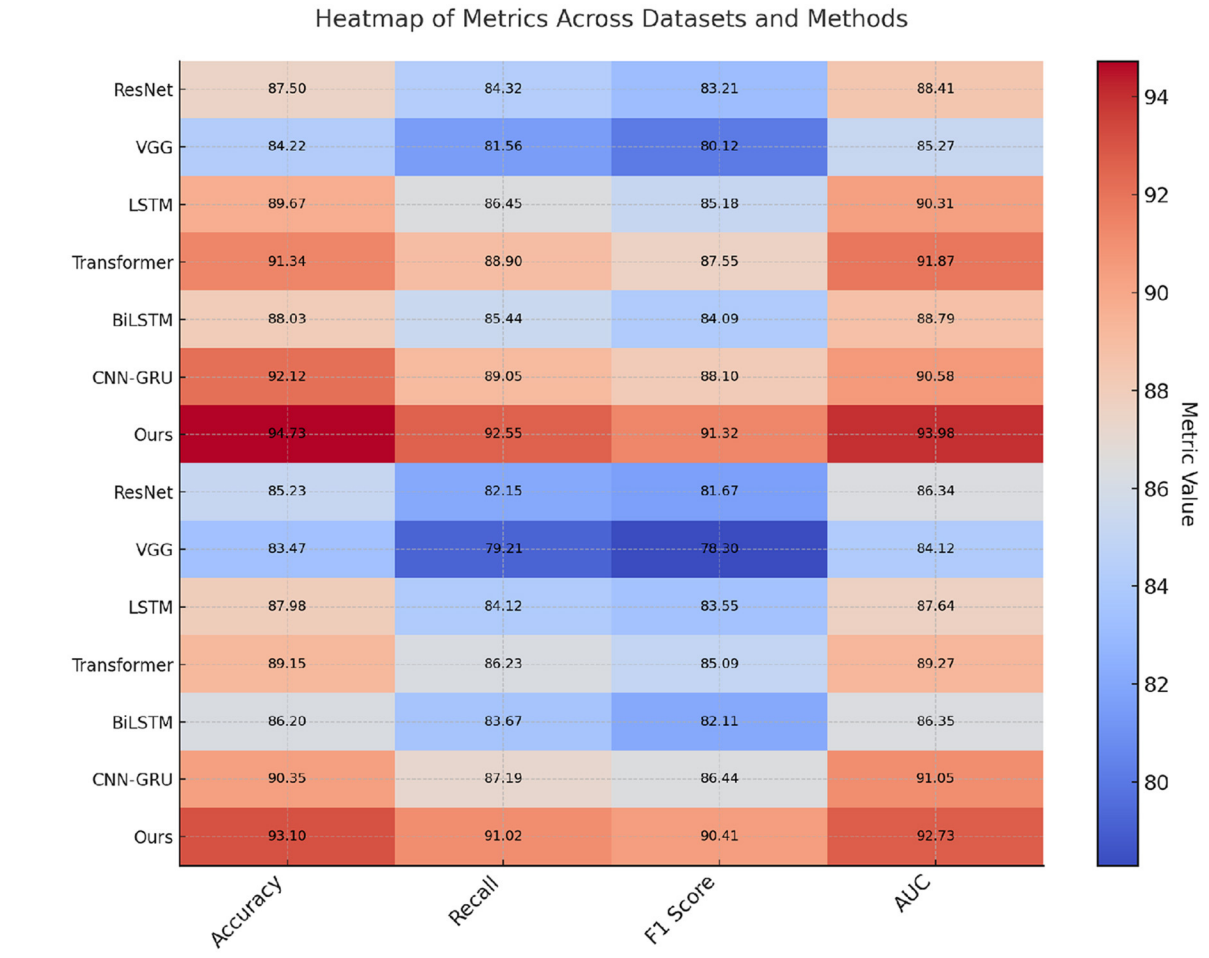

FIGURE 4
Performance comparison of SOTA methods on EEGEyeNet dataset and PhyAAt dataset.

TABLE 3 Ablation study results on SEED and sleep-EDF datasets.

Model	SEED dataset				Sleep-EDF dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
w./o. Multi-resolution embedding module	88.45 ± 0.03	85.67 ± 0.02	84.32 ± 0.02	87.89 ± 0.03	86.21 ± 0.03	83.09 ± 0.02	82.78 ± 0.02	85.92 ± 0.03
w./o. Hierarchical self-attention mechanism	90.23 ± 0.02	87.90 ± 0.02	86.55 ± 0.02	89.76 ± 0.03	88.13 ± 0.03	85.34 ± 0.02	84.10 ± 0.02	87.45 ± 0.02
w./o. Localized context windowing	91.78 ± 0.02	89.12 ± 0.03	88.03 ± 0.02	90.89 ± 0.02	89.76 ± 0.02	86.98 ± 0.03	85.89 ± 0.02	88.67 ± 0.03
Ours	**94.55 ± 0.02**	**92.30 ± 0.02**	**91.45 ± 0.03**	**93.67 ± 0.03**	**93.27 ± 0.03**	**91.45 ± 0.02**	**90.78 ± 0.03**	**92.34 ± 0.02**

Bold values are the best values.

79.23%, highlighting the significant role of textual information in emotion recognition. On the EmotiCon dataset, the performance of text as a single modality was slightly lower, with an accuracy of 78.45% and a macro F1 score of 76.32%. With multimodal fusion, significant improvements were observed. Combining text, speech, and video modalities on the IEMOCAP dataset increased accuracy to 88.93% and macro F1 to 87.68%. Adding EEG data further elevated performance, achieving an accuracy of 91.45% and a macro F1 score of 90.12%, underscoring the complementary role of EEG signals in emotion recognition. Similarly, the EmotiCon dataset showed the highest performance with the full-modal setup, achieving an accuracy of 89.76% and a macro F1 score of 88.43%. These findings quantitatively demonstrate the advantage of multimodal inputs, as the integration of diverse modalities significantly enhances the model's ability to recognize emotions, with EEG data in particular contributing an additional 2.5% improvement in accuracy on IEMOCAP.

TABLE 4 Ablation study results on EEGEyeNet and PhyAAt datasets.

Model	EEGEyeNet dataset				PhyAAt dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
w./o. Multi-resolution embedding module	87.12 ± 0.03	84.65 ± 0.02	83.34 ± 0.02	88.23 ± 0.03	85.42 ± 0.03	82.11 ± 0.02	81.50 ± 0.02	86.78 ± 0.03
w./o. Hierarchical self-attention mechanism	89.35 ± 0.02	86.48 ± 0.02	85.09 ± 0.02	89.76 ± 0.03	87.10 ± 0.03	84.23 ± 0.02	83.42 ± 0.02	88.21 ± 0.02
w./o. Localized context windowing	91.05 ± 0.02	88.37 ± 0.03	87.25 ± 0.02	90.89 ± 0.02	89.15 ± 0.02	86.56 ± 0.03	85.33 ± 0.02	89.54 ± 0.03
Ours	**94.73 ± 0.02**	**92.55 ± 0.02**	**91.32 ± 0.03**	**93.98 ± 0.03**	**93.10 ± 0.03**	**91.02 ± 0.02**	**90.41 ± 0.03**	**92.73 ± 0.02**

Bold values are the best values.

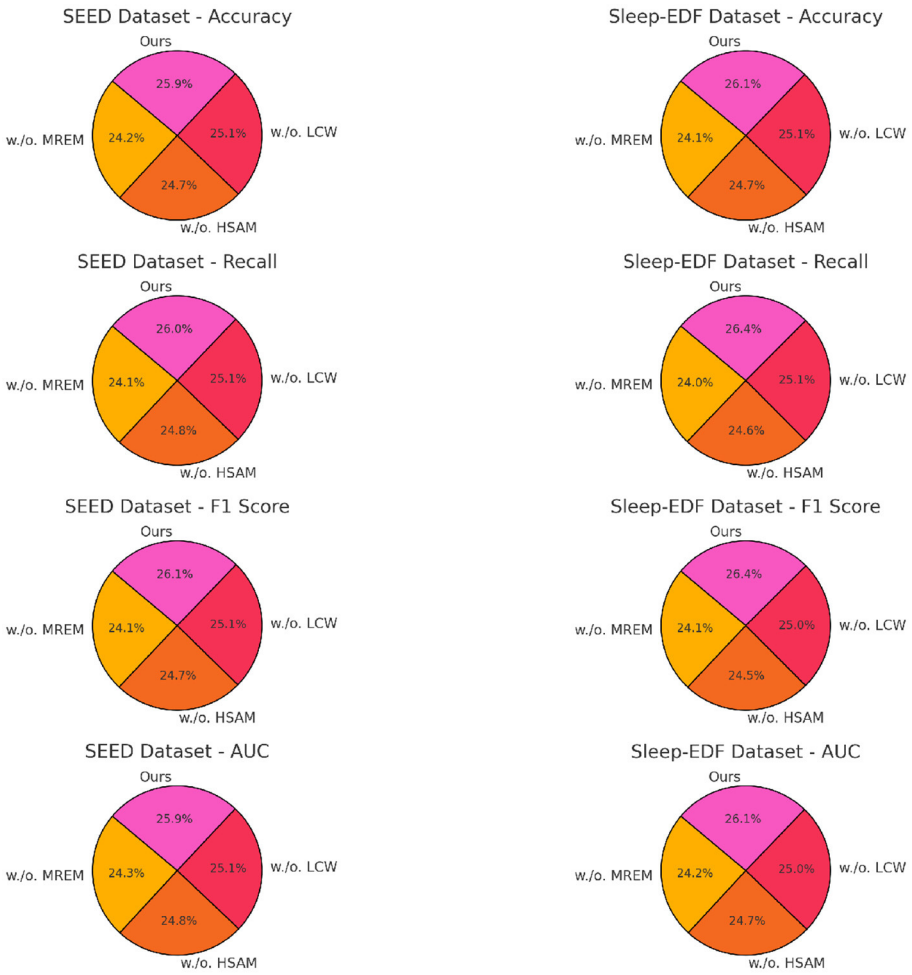

FIGURE 5
Ablation study of our method on SEED dataset and Sleep-EDF dataset (MREM, multi-resolution embedding module; HSAM, hierarchical self-attention mechanism; LCW, localized context windowing).

5 Conclusions and future work

This study has demonstrated the effectiveness of the proposed EmotionFusion-Transformer framework in enhancing multimodal emotion recognition through the integration of EEG and textual data. By leveraging the complementary strengths of these modalities, the model achieves a nuanced understanding of emotional states, outperforming existing single- and multi-modality approaches in accuracy and robustness. The findings underline the potential of transformer-based architectures in capturing complex contextual dependencies and aligning multimodal data for improved performance. The experimental results on diverse datasets validate the adaptability of the framework across applications in emotion recognition, sleep stage classification, and eye-tracking tasks. However, the study acknowledges the challenges associated with the

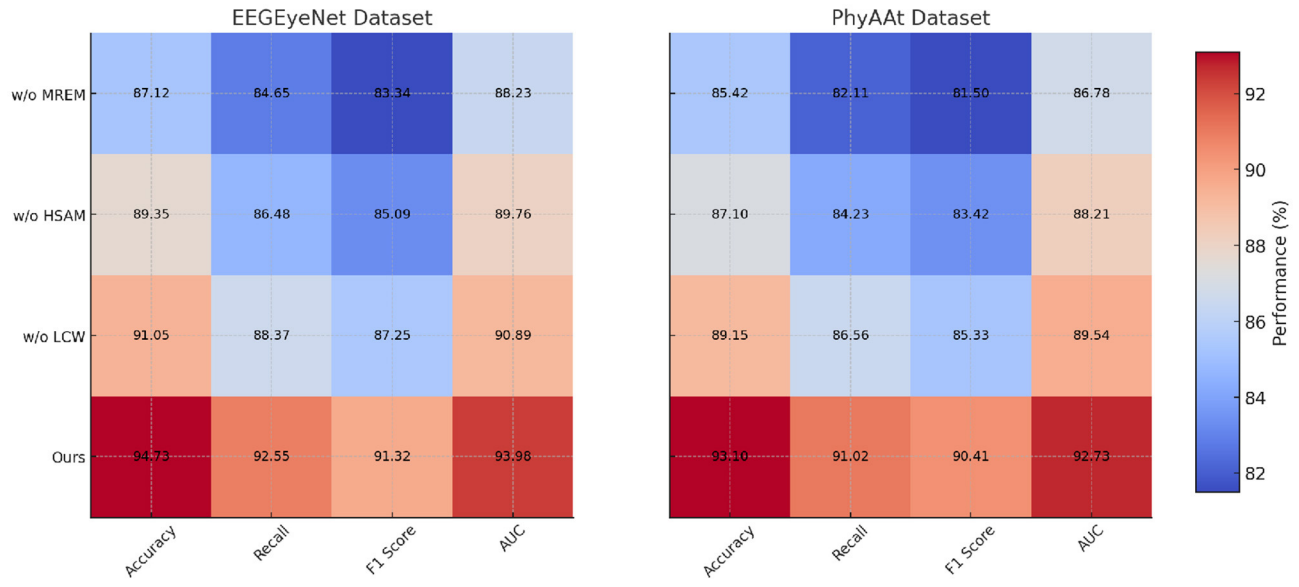

FIGURE 6
Ablation study of our method on EEGEyeNet dataset and PhyAAt dataset (MREM, multi-resolution embedding module; HSAM, hierarchical self-attention mechanism; LCW, localized context windowing).

TABLE 5 Performance comparison of EmotionFusion-Transformer on IEMOCAP and EmotiCon datasets.

Dataset	Modality	Accuracy (%)	Macro F1 (%)	Weighted F1 (%)	Precision (%)	Recall (%)
IEMOCAP	Text only	81.34 ± 0.45	79.23 ± 0.32	80.45 ± 0.36	80.12 ± 0.50	78.89 ± 0.48
	Text + audio	85.76 ± 0.42	84.11 ± 0.33	84.65 ± 0.29	85.23 ± 0.40	83.45 ± 0.39
	Text + audio + video	88.93 ± 0.39	87.68 ± 0.28	88.21 ± 0.30	88.42 ± 0.33	87.11 ± 0.34
	All modalities (text + audio + video + EEG)	**91.45 ± 0.35**	**90.12 ± 0.30**	**90.78 ± 0.31**	**91.03 ± 0.28**	**89.87 ± 0.29**
EmotiCon	Text only	78.45 ± 0.52	76.32 ± 0.40	77.01 ± 0.42	77.45 ± 0.50	75.98 ± 0.47
	Text + video	82.78 ± 0.48	81.12 ± 0.39	81.67 ± 0.35	82.11 ± 0.41	80.45 ± 0.36
	Text + video + audio	86.92 ± 0.45	85.65 ± 0.31	86.12 ± 0.38	86.42 ± 0.34	84.87 ± 0.33
	All modalities (text + video + audio + EEG)	**89.76 ± 0.39**	**88.43 ± 0.32**	**88.98 ± 0.36**	**89.34 ± 0.37**	**87.78 ± 0.34**

Bold values are the best values.

high-dimensionality of EEG features and their dependence on specific sensor configurations, which limits the practicality of direct implementation in consumer-grade devices.

An important future direction is to explore the critical EEG sensor locations from the current dataset to minimize the number of EEG features while maintaining classification accuracy at a usable level. Reducing the required sensor locations could significantly simplify hardware requirements and computational overhead, facilitating the transition of this research into practical applications. This investigation will involve systematic methods such as saliency map analysis and channel-wise ablation studies to identify the most impactful sensors. The results could inform the optimization of existing dry electrode headsets or guide the design of new, lightweight devices tailored for emotion recognition tasks. Such advancements would not only support the development of cost-effective and user-friendly systems but also bridge the gap between academic research and real-world implementations, fostering progress in areas such as affective computing, mental health monitoring, and human-computer interaction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JW: Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision,

Validation, Visualization, Conceptualization, Investigation, Software, Writing – original draft, Writing – review & editing. CZ: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alvi, M. B., Mahoto, N., Reshan, M. S. A., Unar, M. A., Elmagzoub, M., Shaikh, A., et al. (2023). *Count Me Too: Sentiment Analysis of Roman Sindhi Script*. London: SAGE Open. doi: 10.1177/21582440231197452

Babu, Y. P., Rajagopal, E., and Nimmi, K. (2020). "Malayalam-English code mixed sentiment analysis using sentence Bert and sentiment features," in *2020: Forum for Information Retrieval Evaluation, December 16-20, 2020* (Hyderabad).

Bajaj, N., and Requena Carrión, J. (2023). Deep representation of EEG signals using spatio-spectral feature images. *Appl. Sci*. 13:9825. doi: 10.3390/app13179825

Bakar, M. F. R. A., Idris, N., Shuib, L., and Khamis, N. (2020). Sentiment analysis of noisy Malay text: State of art, challenges and future work. *IEEE Access* 1. doi: 10.1109/ACCESS.2020.2968955

Başarslan, M. S., and Kayaalp, F. (2020). Sentiment analysis with machine learning methods on social media. *Adv. Distrib. Comput. Artif. Intell. J.* 9. doi: 10.14201/ADCAIJ202093515

Biswas, P., Neogi, S., Daniel, A., and Mitra, A. (2024). Evaluating generative artificial intelligence on multilingual sentiment analysis. *J. Electr. Syst*. 20, 3502–3513. doi: 10.52783/jes.4986

Boufssasse, A., Hssayni, E. H., Joudar, N.-E., and Ettaouil, M. (2023). A multi-objective optimization model for redundancy reduction in convolutional neural networks. *Neural Process. Lett*. 55, 9721–9741. doi: 10.1007/s11063-023-11223-2

Chattu, K., and Sumathi, D. (2023). "Sentiment classification of low resource language tweets using machine learning algorithms," in *2023 2nd International Conference on Edge Computing and Applications (ICECAA)* (Namakkal: IEEE). doi: 10.1109/ICECAA58104.2023.10212247

Cruz, C. A., and Balahadia, F. F. (2022). Analyzing public concern responses for formulating ordinances and laws using sentiment analysis through Vader application. *Int. J. Comput. Sci. Res*. 6. doi: 10.25147/ijcsr.2017.001.1.77

Cui, X., Chen, Z., and Yin, F. (2021). Multi-objective based multi-channel speech enhancement with Bilstm network. *Appl. Acoust*. 177:107927. doi: 10.1016/j.apacoust.2021.107927

Gantayet, A., and Dheer, D. K. (2022). A data-driven multi-objective optimization framework for optimal integration planning of solid-state transformer fed energy hub in a distribution network. *Eng. Sci. Technol. Int. J.* 36:101278. doi: 10.1016/j.jestch.2022.101278

Ghosh, S., Varshney, D., Ekbal, A., and Bhattacharyya, P. (2021). "Context and knowledge enriched transformer framework for emotion recognition in conversations," in *2021 International joint conference on neural networks (IJCNN)* (Shenzhen: IEEE), 1–8. doi: 10.1109/IJCNN52387.2021.9533452

Hernández-Pérez, R., Lara-Martínez, P., Obregón-Quintana, B., Liebovitch, L. S., and Guzmán-Vargas, L. (2024). Correlations and fractality in sentence-level sentiment analysis based on Vader for literary texts. *Information* 15:698. doi: 10.3390/info15110698

Hu, D., Wei, L., and Huai, X. (2021). Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv* [Preprint]. arXiv:2106.01978. doi: 10.48550/arXiv.2106.01978

Huang, R. (2024). Design and implementation of English writing aids based on natural language processing. *J. Electr. Syst*. 20. doi: 10.52783/jes.3132

İşçi, V. (2023). *A comparative analysis of anti-Petrarchan sentiments in the English renaissance poetry*. KARE. Available at: https://dergipark.org.tr/en/pub/kare/issue/78967/1174698

Jiang, Y. (2024). Online English writing teaching method that enhances teacher-student interaction. *J. Intell. Syst*. 33. doi: 10.1515/jisys-2023-0235

Kastrati, A., Płomecka, M. B., Pascual, D., Wolf, L., Gillioz, V., Wattenhofer, R., et al. (2021). Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *arXiv* [Preprint]. arXiv:2111.05100. doi: 10.48550/arXiv.2111.05100

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., and Oberye, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng*. 47, 1185–1194. doi: 10.1109/10.867928

Kim, S.-H. (2024). Exploring multimodal perspectives in collaborative writing: sentiment analysis and word frequency in natural language processing. *Proc. Int. CALL Res. Conf*. 2024. doi: 10.29140/9780648184485-16

Li, Q., Wu, C., Wang, Z., and Zheng, K. (2020). Hierarchical transformer network for utterance-level emotion recognition. *Appl. Sci*. 10:4447. doi: 10.3390/app10134447

Liu, X., Zhou, Y., Zhao, J., Yao, R., Liu, B., Ma, D., et al. (2020). Multiobjective resnet pruning by means of Emoas for remote sensing scene classification. *Neurocomputing* 381, 298–305. doi: 10.1016/j.neucom.2019.11.097

Liu, Y., and Fu, G. (2021). Emotion recognition by deeply learned multi-channel textual and EEG features. *Future Gener. Comput. Syst*. 119, 1–6. doi: 10.1016/j.future.2021.01.010

N'Diaye, A. C. M., Chrif, M. E. M. E. A., Mahmoud, B. M. E., and Beqqali, O. E. (2021). "Apply sentiment analysis technology in social media as a tool to enhance the effectiveness of e-government: application on Arabic and Mauritanian dialect 'Hassaniya'," in *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)* (Fez: IEEE). doi: 10.1109/ICDS53782.2021.9626766

Nimmi, K., and Janet, B. (2021). *Voting ensemble model based Malayalam-English sentiment analysis on code-mixed data*. Fire. Available at: https://ceur-ws.org/Vol-3159/T6-21.pdf

Panda, D., Chakladar, D. D., and Dasgupta, T. (2020). Multimodal system for emotion recognition using EEG and customer review. *Adv. Intell. Syst. Comput*. 1112, 399–410. doi: 10.1007/978-981-15-2188-1_32

Pei, C. (2024). A study on the reception comparison of Fingersmith among Chinese and English readers-analysis of evaluative discourses based on python. *J. Humanit. Arts Soc. Sci*. 8, 511–515. doi: 10.26855/jhass.2024.02.037

Polyakova, O. (2023). E-portfolio in students' learning for sustainable development. *Ìnf. Tehnol. Zasobi Navčannâ* 96, 1–14. doi: 10.33407/itlt.v96i4.5238

Sharma, A., and Ghose, U. (2023). Toward machine learning based binary sentiment classification of movie reviews for resource restraint language (RRL)?Hindi. *IEEE Access* 11, 2169–3536. doi: 10.1109/ACCESS.2023.3283461

Shrestha, A., Spezzano, F., and Joy, A. (2020). "Detecting fake news spreaders in social networks via linguistic and personality features," in *Conference and Labs of the Evaluation Forum*. Available at: https://scholarworks.boisestate.edu/cs_facpubs/303/

Singh, O. M., Timilsina, S., Bal, B., and Joshi, A. (2020). "Aspect based abusive sentiment detection in Nepali social media texts," in *International Conference*

on Advances in Social Networks Analysis and Mining (The Hague: IEEE). doi: 10.1109/ASONAM49781.2020.9381292

Singh, S. K., and Sachan, M. (2021). "Acquisition and development of code-mixed corpus for sentiment analysis of resource-scarce languages," in *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)* (Bengaluru: IEEE). doi: 10.1109/CENTCON52345.2021.9687897

Suleimenova, Z., Abilkhamitkyzy, R., Yskak, B. A., and Korganbekov, B. S. (2022). Exploring topics and sentiment in the English version of Yassawi's divan-i Hikmet: a text mining approach. *Theory Pract. Lang. Stud.* 12:2678. doi: 10.17507/tpls.1212.26

Teo, H., Campos-Arceiz, A., Li, B. V., Wu, M.-L., and Lechner, A. (2020). Building a green belt and road: a systematic review and comparative assessment of the Chinese and English-language literature. *PLoS ONE* 15:e0239009. doi: 10.1371/journal.pone.0239009

Tu, G., Wen, J., Liu, C., Jiang, D., and Cambria, E. (2022). Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Trans. Artif. Intell.* 3, 699–708. doi: 10.1109/TAI.2022.3149234

Wang, Y., Zhang, J., Ma, J., Wang, S., and Xiao, J. (2020). "Contextualized emotion recognition in conversation as sequence tagging," in *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, 186–195. doi: 10.18653/v1/2020.sigdial-1.23

Whissell, C. (2022). To better understand Vincent: a study of the emotional tone of Vincent van Gogh's letters to his brother Theo: 1872-1890. *Int. J. Stud. Engl. Lang. Lit.* 10, 37–53. doi: 10.20431/2347-3134.1010005

Xu, H.-B. (2024). "Application of affective computing in sentiment analysis of English writing in college students," in *2024 International Conference on Intelligent Education and Computer Technology* (New York, NY: ACM). doi: 10.1145/3687311.3687371

Zakaria, N., and Sulaiman, N. A. (2024). The needs analysis of ESL learners' expository writing challenges: perspectives of ESL teachers. *Int. J. Acad. Res. Bus. Soc. Sci.* 322–335. doi: 10.6007/IJARBSS/v14-i5/21387

Zamani, M. G., Nikoo, M. R., Al-Rawas, G., Nazari, R., Rastad, D., Gandomi, A. H., et al. (2024). Hybrid WT-CNN-GRU-based model for the estimation of reservoir water quality variables considering spatio-temporal features. *J. Environ. Manage.* 358:120756. doi: 10.1016/j.jenvman.2024.120756

Zhang, S., and Cao, R. (2022). Multi-objective optimization for UAV-enabled wireless powered IOT networks: an LSTM-based deep reinforcement learning approach. *IEEE Commun. Lett.* 26, 3019–3023. doi: 10.1109/LCOMM.2022.3210660

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497