HYPERGRAPH MULTI-MODAL LEARNING FOR EEG-BASED EMOTION RECOGNITION IN CONVERSATION

Zijian Kang[1,†], Yueyang Li[1,2,†], Shengyu Gong[1], Weiming Zeng[1,⋆], Hongjie Yan[3], Lingbin Bian[2,4], Zhiguo Zhang[5], Wai Ting Siok[2], and Nizhuan Wang[2,⋆]

[1]*Lab of Digital Image and Intelligent Computation, Shanghai Maritime University*, Shanghai 201306, China
[2]*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University*, Hong Kong SAR, China
[3]*Affiliated Lianyungang Hospital of Xuzhou Medical University*, Lianyungang 222002, China
[4]*The State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong*, Hong Kong SAR, China
[5]*The Institute of Computing and Intelligence, Harbin Institute of Technology Shenzhen*, Shenzhen 518000, China
[†]Co-first authors
[⋆]Correspondence: wangnizhuan1120@gmail.com; zengwm86@163.com

ABSTRACT

Emotional Recognition in Conversation (ERC) is valuable for diagnosing health conditions such as autism and depression, and for understanding the emotions of individuals who struggle to express their feelings. Current ERC methods primarily rely on semantic, audio and visual data but face significant challenges in integrating physiological signals such as Electroencephalography (EEG). This research proposes Hypergraph Multi-Modal Learning (Hyper-MML), a novel framework for identifying emotions in conversation. Hyper-MML effectively integrates EEG with audio and video information to capture complex emotional dynamics. Firstly, we introduce an Adaptive Brain Encoder with Mutual-cross Attention (ABEMA) module for processing EEG signals. This module captures emotion-relevant features across different frequency bands and adapts to subject-specific variations through hierarchical mutual-cross attention mechanisms. Secondly, we propose an Adaptive Hypergraph Fusion Module (AHFM) to actively model the higher-order relationships among multi-modal signals in ERC. Experimental results on the EAV and AFFEC datasets demonstrate that our Hyper-MML model significantly outperforms current state-of-the-art methods. The proposed Hyper-MML can serve as an effective communication tool for healthcare professionals, enabling better engagement with patients who have difficulty expressing their emotions. The official implementation codes are available at https://github.com/NZWANG/Hyper-MML.

Keywords Emotion Recognition in Conversation (ERC) · EEG-based Emotion Recognition (EER) · Hypergraph Learning · Multi-modal Fusion · Mutual-cross Attention · Electroencephalography (EEG).

1 Introduction

1.1 Emotion Recognition in Conversation (ERC)

Emotion Recognition in Conversation (ERC) holds significant potential for assessing mental conditions such as autism and depression. Recent studies suggest that individuals with these conditions frequently exhibit unique communication challenges, including speech impairments, emotional disturbances, literal interpretation of questions, and difficulty sustaining coherent dialogue [1]. Current ERC research primarily focuses on transcribed text from spoken dialogue, supplemented by visual (e.g., facial expressions) and acoustic (e.g., intonation, loudness) cues [2]. These methods typically analyze complete, uninterrupted dialogue transcripts, integrating multi-modal data to contextualize individual utterances. Context modeling in ERC generally incorporates three key elements: 1) the content of previous exchanges, 2) the timing of conversational turns, and 3) speaker-specific details such as identity and evolving emotional states [3]. However, the semantic structure of dialogue is vulnerable to disruptions such as fragmented sentences and missing

segments, which distort the logical relationships between utterances. This resulting incoherence severely degrades the performance of emotion recognition models, ultimately constraining their practical applications in clinical settings for disorders like autism and depression. Figure 1 provides a demonstration of this phenomenon, showing how such fragmentation leads to significant emotion misclassification. To address these limitations, psychotherapists require robust physiological indicators that remain reliable even with incomplete conversational data.

Physiological signals – particularly electroencephalography (EEG) data – provide a direct window into neural activity and emotional states, surpassing text-based methods in objectivity and immediacy [4, 5]. While textual analysis depends on extended linguistic context, EEG signals operate on shorter timescales, making them ideal for detecting transient emotional shifts (e.g., sudden frustration or momentary joy) in real time. Thus, by integrating EEG signals with text-based modalities, clinicians can address key limitations of language-driven approaches, such as distortions caused by fragmented or incomplete dialogue. EEG's intrinsic nature avoids language-related biases, enabling clearer and more objective emotion measurement. Furthermore, combining EEG with multi-modal data (e.g., audio and video) outperforms single-source EEG analysis, enhancing diagnostic accuracy [6]. This integrative framework will allow psychologists or clinicians to correlate physiological responses (e.g., brainwave patterns) with behavioural cues (e.g. voice tone, facial expressions), constructing a comprehensive emotional profile, which supports the development of tailored treatment strategies that better address individual patient needs [7].

In multi-modal ERC tasks, graph neural networks (GNNs) are often employed to model interactions by capturing contextual and multi-modal data (e.g., text, audio, visual). However, GNNs face a critical limitation: they can only model complex interactions by chaining together simple pairwise relationships (e.g., between two nodes at a time). This sequential approximation of high-order relationships – such as group dynamics or multi-modal dependencies – often leads to suboptimal accuracy. Hypergraph theory overcomes this limitation by supporting high-order connections (e.g., linking three or more nodes simultaneously), enabling direct modeling of intricate multi-modal interactions. For instance, a hyperedge could connect a speaker's utterance, their facial expression, and a listener's reaction in a single interaction step. This capability makes hypergraphs a more precise and efficient framework for multi-modal ERC tasks [8].

1.2 Our Contribution

We propose a novel Hypergraph Multi-Modal Learning (Hyper-MML) framework for ERC that achieves state-of-the-art (SOTA) performance on the EAV [9] and AFFEC [10] datasets. Our model advances ERC in following three key innovations:

1) Hypergraph Multi-Modal Learning Framework (Hyper-MML): We introduce an end-to-end architecture that integrates EEG signals with audio and visual data to model complex emotional dynamics in conversations. Unlike traditional language-centric approaches, Hyper-MML directly leverages physiological (EEG) and behavioral (audio-visual) cues, bypassing the limitations of language-based ambiguity or incomplete dialogue transcripts.

2) Adaptive Brain Encoder with Mutual-cross Attention(ABEMA): A specialized EEG encoder captures emotion-relevant patterns across frequency bands while adapting to individual brain characteristics. ABEMA's hierarchical attention mechanism models both within-band and cross-band relationships, enabling robust feature extraction from noisy EEG signals.

3) Adaptive Hypergraph Fusion Module (AHFM): A specialized fusion module enhances cross-modal interaction within hypergraph structures. This module employs adaptive weighted aggregation to dynamically prioritize the most informative modalities (e.g., emphasizing EEG during subtle emotional shifts or audio during tone-based cues). This strategy optimizes information propagation across modalities, significantly improving emotion recognition accuracy.

The remainder of this paper is organized as follows: Section 2 reviews prior work in multimodal emotion recognition and hypergraph learning. Section 3 introduces the architecture of our proposed Hyper-MML model. Section 4 describes the datasets and experimental setup. Results and discussion are presented in Section 5. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related work

2.1 Multimodal Emotion Recognition

ERC has become essential for healthcare diagnostics, particularly in autism and depression assessment [2]. This field is evolving from early text-centric approaches to sophisticated multimodal frameworks that integrate textual, acoustic, and visual information. Early approaches focused on textual analysis, incorporating conversational history, temporal dynamics, and speaker information [11, 12]. Recent advances have introduced graph-based approaches such

Figure 1: Illustration of emotion recognition in conversational text. The complete dialogue (a) shows coherent emotional transitions and logical semantic structure. Incomplete dialogue (b), such as fragmented sentences, missing dialogue segments and broken semantic structure, causes emotion label ambiguity within a conversation. This leads to emotion misclassification and contextual confusion.

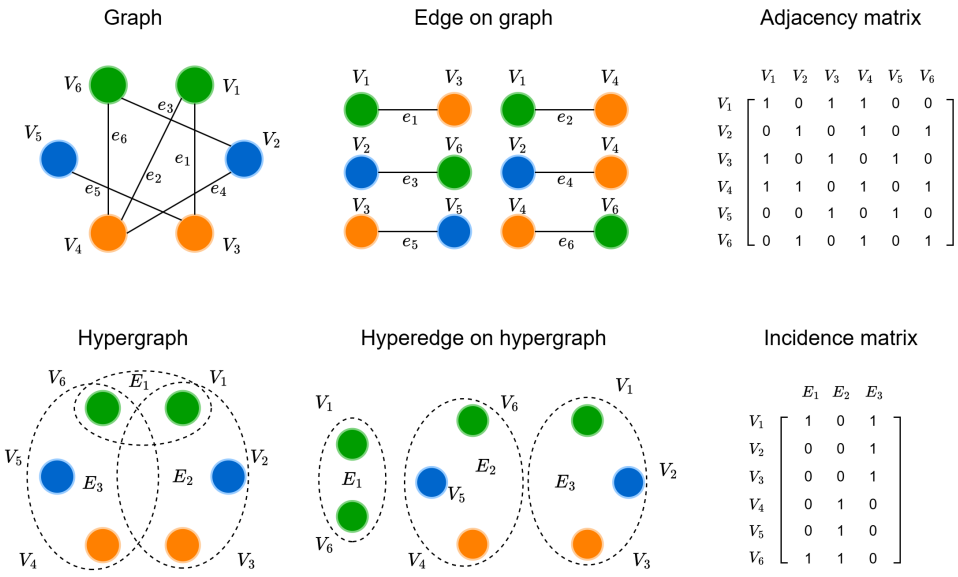

Figure 2: Representation of Graphs and Hypergraphs. The upper left section displays a graph and its edges, while the upper right section presents the adjacency matrix, which describes the connections between nodes. The lower section features a hypergraph and its hyperedges, with the right side showing the incidence matrix, which indicates the relationships between hyperedges and nodes.

as DialogueGCN [12] and MMGCN [3], which model complex conversational relationships through graph neural networks. However, these text-dependent methods struggle with fragmented dialogue data, common in clinical settings with communication-impaired patients. Cross-modal reconstruction techniques address missing modalities [13], but suffer from modality gap issues [14], where different modalities embed at distinct distances, affecting reconstruction accuracy. This limitation motivates the exploration of alternative modalities that can provide robust emotional indicators independent of textual coherence.

EEG provides direct access to neural activity and emotional states, offering advantages in objectivity, temporal resolution, and language independence compared to behavioral modalities [4]. EEG-based emotion recognition has progressed from traditional frequency analysis using Power Spectral Density (PSD) and Differential Entropy (DE) features to deep learning approaches including CNNs [15], graph methods [16, 17], and temporal modeling [18, 19]. Recent work captures spatial asymmetry and temporal dynamics [20] while modeling brain connectivity through graph networks [21]. However, existing EEG emotion recognition primarily uses passive stimuli (images, videos, music) rather than natural conversational interactions. This gap is significant because the conversational emotions exhibit different temporal evolution and contextual dependencies compared to stimulus-induced emotions.

Fortunately, recent public datasets help to address this limitation. The EAV dataset [9] provides the first public multimodal EEG-audio-video resource for conversational emotion recognition, featuring 8,400 interactions from 42 participants across five emotional states. The AFFEC dataset [10] includes EEG, eye-tracking, galvanic skin response (GSR), and facial video data from 73 participants, distinguishing felt versus perceived emotions in face-to-face interactions. Initial research demonstrates EEG integration feasibility with conventional modalities [22], though sophisticated fusion mechanisms for multimodal emotional interactions remain underdeveloped.

2.2 Hypergraph Learning

As illustrated in Figure 2, conventional graphs utilize adjacency matrices to encode binary or weighted relationships, limiting their expressive power to simple pairwise interactions. However, many real-world scenarios involve complex higher-order relationships that cannot be adequately captured by pairwise connections. Hypergraphs extend edges to hyperedges, connecting multiple nodes simultaneously to model group interactions and multi-way dependencies [23]. Hypergraph learning employs incidence matrices instead of adjacency matrices, providing enhanced flexibility for complex relational structures [8, 24]. This transition of higher-order relationship modeling in hypergraph offers enhanced capabilities for sophisticated relational reasoning.

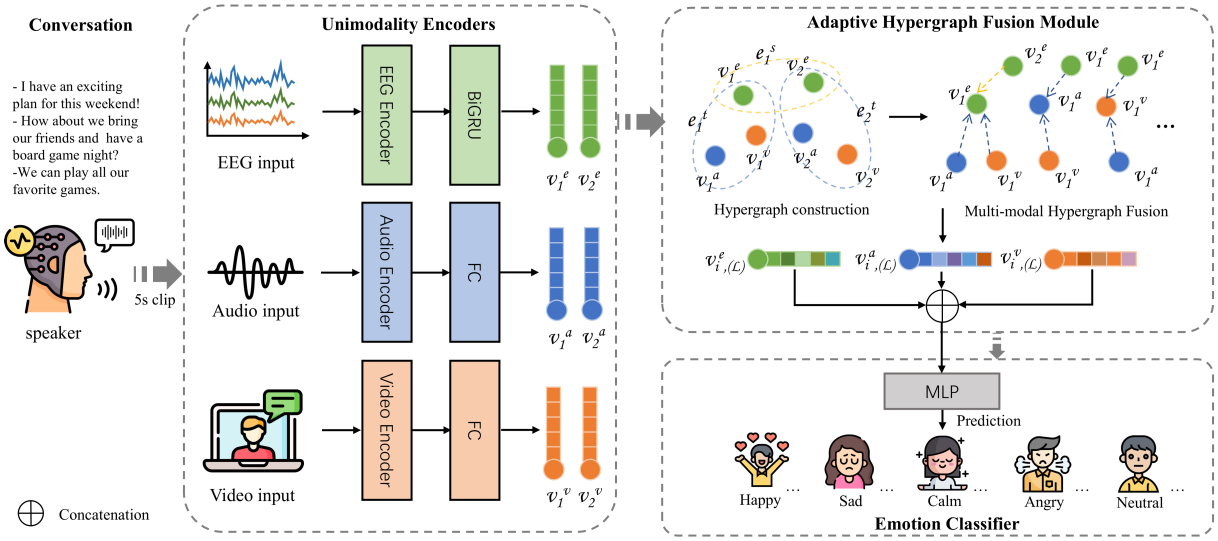

Figure 3: Overall framework of the Hyper-MML.

In the context of EEG-based multimodal ERC, hypergraphs offer several compelling advantages over traditional graph-based approaches. The simultaneous interaction among EEG signals, audio features, and visual cues during emotional expression represents a natural higher-order relationship that cannot be effectively decomposed into pairwise connections without losing critical information. Recent applications of hypergraph learning in multimodal tasks have demonstrated superior performance in capturing complex correlations [25], with successful implementations in diverse domains including recommender systems [26], sleep stage classification [27], and drug-target interaction prediction [28]. In terms of ERC, hypergraphs enable the direct modeling of multi-way dependencies between different modalities within the same temporal context, as well as cross-temporal relationships that span multiple conversation turns. The recent emergence of hypergraph-based approaches in ERC [29] has shown promising results, though this method primarily focuses on traditional modalities and has not fully explored the integration of physiological signals like EEG. Furthermore, existing approaches often employ fixed hypergraph structures, limiting their adaptability to the dynamic nature of emotional states and the varying importance of different modalities across different emotional contexts, highlighting the need for adaptive hypergraph learning mechanisms that can dynamically adjust to the complex interplay of multimodal emotional cues in conversational settings.

3 Methodology

3.1 Problem Formulation of Multimodal ERC

Multimodal ERC aims to identify participants' emotional states from multimodal interactions. Generally, we segment each utterance u_i in a conversation sequence $u_i (i = 1, \ldots, N)$ into fragments based on fixed time intervals Δt: $u_i = \{s_{i1}, s_{i2}, ..., s_{ik_i}\}$. Each fragment s_{ij} contains three modalities: EEG signals $E_{ij} \in \mathbb{R}^{C \times L}$, where C is the number of channels and L is the length of time window), audio features $A_{ij} \in \mathbb{R}^{d_a}$, and video features $V_{ij} \in \mathbb{R}^{d_v}$. Our objective is to learn a mapping function $f : (E_{ij}, A_{ij}, V_{ij}) \rightarrow y_i$, where $y_i \in Y$ represents the emotion category.

As illustrated in Figure 3, we present the Hyper-MML framework to achieve this goal, which effectively models higher-order relationships among multimodal signals through hypergraph structures, with particular emphasis on fusing EEG signals with traditional modalities to achieve accurate emotion recognition in conversation fragments.

3.2 ABEMA for EEG Embedding

EEG directly reflects neural activity and emotional states, but effectively encoding them for emotion recognition remains challenging due to high dimensionality, low signal-to-noise ratio, and subject-specific variations. To address these challenges, we propose the Adaptive Brain Encoder with Mutual-cross Attention (ABEMA) of EEG (Figure 4), which captures emotion-relevant features across different frequency bands while adapting to individual differences. ABEMA transforms raw EEG signals into compact, information-rich embeddings optimized for multi-modal feature fusion.

3.2.1 Subject-specific Layer

EEG signals exhibit significant variations across different subjects due to differences in brain structures, electrode placement, and neurophysiological characteristics. This inter-subject variability reduces model generalization capability. To address this issue, we adopt the subject-specific layer from our previous work [30] as the first component of ABEMA:

$$E_s = \text{SubjectLayer}(E_i, s) = \mathbf{M}_s E_i, \tag{1}$$

where $E_i \in \mathbb{R}^{C \times L}$ represents the raw EEG signal for utterance fragment i, with C channels and time window length L; s is the subject identifier; and $\mathbf{M}_s \in \mathbb{R}^{C \times C}$ is a learnable weight matrix that captures subject-specific channel interaction patterns. We initialize $\mathbf{M}_s$ as identity matrices, preserving critical signal features during early training while gradually adapting to individual characteristics.

3.2.2 Temporal-Spectral Transformer Block

Emotional states exhibit dynamic changes over time, requiring effective modeling of temporal dependencies in EEG signals. After subject-specific processing, we employ an improved Transformer architecture [31] for temporal feature extraction:

$$E_{\text{t}} = \text{iTransformer}(E_s) \in \mathbb{R}^{B \times C \times L}, \tag{2}$$

where B is the batch size. Unlike traditional Transformers that apply attention across time steps, iTransformer applies self-attention to the channel dimension, capturing relationships between brain regions while preserving temporal information.

3.2.3 Hierarchical Mutual-Cross Attention Block

The core innovation of ABEMA lies in its hierarchical mutual-cross attention mechanism, which integrates complementary features from different frequency bands. Emotional states are closely associated with specific EEG frequency band activities; for instance, alpha bands correlate with relaxation states, while beta bands relate to alertness and concentration. Our approach captures complex relationships both within and between frequency bands through a two-level attention structure.

Frequency Band Decomposition and Feature Extraction: We first transform the temporal EEG signals into the frequency domain using Fast Fourier Transform (FFT):

$$\mathit{\Gamma} = \mathcal{F}[E_{\text{t}}] \in \mathbb{R}^{B \times C \times F}, \tag{3}$$

where B is the batch size, C is the number of channels, and F is the number of frequency points. We then extract signals from five standard frequency bands using band selection masks:

$$\mathit{\Gamma}_b = \mathit{\Gamma} \odot M_b(f), \quad b \in \{\delta, \theta, \alpha, \beta, \gamma\} \tag{4}$$

where $M_b(f)$ is the band selection mask, and δ(0.5-4Hz), θ(4-8Hz), α(8-13Hz), β(13-30Hz), and γ(30-50Hz) represent the five standard EEG frequency bands.

For each band, we extract DE and PSD:

$$D_b = \frac{1}{2}\log(2\pi e \sigma_b^2) \in \mathbb{R}^{B \times C} \tag{5}$$

$$P_b = \frac{1}{F}\sum_{f \in b} |\mathit{\Gamma}_f|^2 \in \mathbb{R}^{B \times C} \tag{6}$$

where σ_b^2 is the variance of the signal in band b. DE measures signal complexity and uncertainty, while PSD quantifies energy distribution across frequencies, jointly providing comprehensive emotional state representation.

Intra-band Mutual-Cross Attention: For each frequency band, we adopt the mutual-cross attention mechanism [6] to fuse DE and PSD features. We extend this into a hierarchical structure, first applying it within frequency bands and subsequently establishing connections between bands. We generate query (Q), key (K), and value (V) matrices according to the following equations:

$$\begin{aligned} Q_b^D &= D_b \mathbf{W}_b^{QD}, \quad K_b^P = P_b \mathbf{W}_b^{KP}, \quad V_b^P = P_b \mathbf{W}_b^{VP}, \\ Q_b^P &= P_b \mathbf{W}_b^{QP}, \quad K_b^D = D_b \mathbf{W}_b^{KD}, \quad V_b^D = D_b \mathbf{W}_b^{VD}, \end{aligned} \tag{7}$$

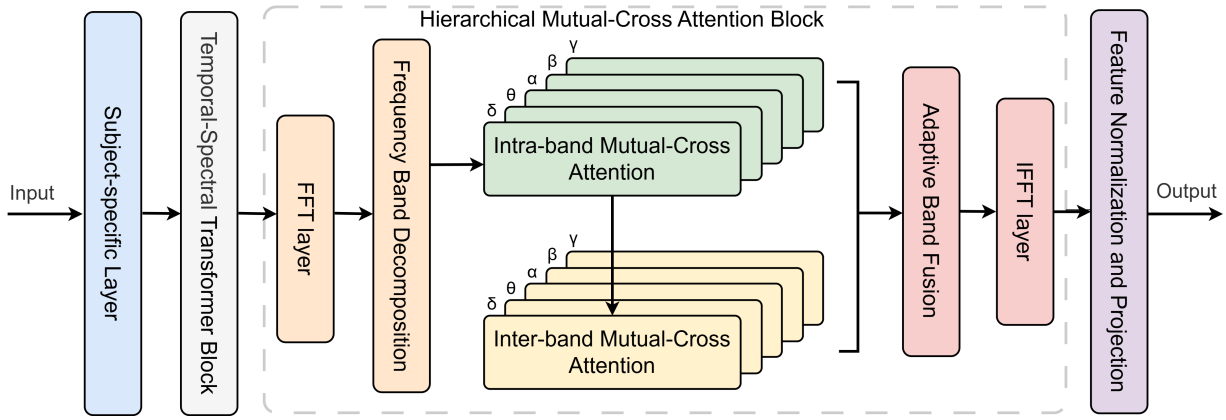

Figure 4: ABEMA for EEG embedding

where $\mathbf{W}_b^{QD}, \mathbf{W}_b^{KP}, \mathbf{W}_b^{VP}, \mathbf{W}_b^{QP}, \mathbf{W}_b^{KD}, \mathbf{W}_b^{VD} \in \mathbb{R}^{C\times d_k}$ are learnable parameter matrices, and d_k is the hidden dimension. Then, we calculate bidirectional mutual-cross attention:

$$A_b^{DP} = \text{softmax}\left(\frac{Q_b^D {K_b^P}^T}{\sqrt{d_k}}\right) V_b^P, A_b^{PD} = \text{softmax}\left(\frac{Q_b^P {K_b^D}^T}{\sqrt{d_k}}\right) V_b^D, \tag{8}$$

Finally, we combine the outputs from both directions:

$$F_b = A_b^{DP} + A_b^{PD} \in \mathbb{R}^{B\times d_k}, \tag{9}$$

Inter-band Mutual-Cross Attention: After intra-band feature fusion, we design a second layer of mutual-cross attention to model relationships between frequency bands. Emotional states typically manifest as coordinated changes across multiple frequency bands; for instance, anxiety may simultaneously present as decreased alpha and increased beta band activity. Firstly, we concatenate the fused features from all frequency bands as follow:

$$F_{all} = [F_\delta; F_\theta; F_\alpha; F_\beta; F_\gamma] \in \mathbb{R}^{B\times 5d_k} \tag{10}$$

Then, we generate query vectors for each frequency band and shared key and value matrices:

$$Q_b^f = F_b \mathbf{W}_b^{Qf} \in \mathbb{R}^{B\times d_k} \tag{11}$$

$$K^f = F_{all}\mathbf{W}^{Kf}, \quad V^f = F_{all}\mathbf{W}^{Vf} \in \mathbb{R}^{B\times d_k} \tag{12}$$

where $\mathbf{W}_b^{Qf} \in \mathbb{R}^{d_k\times d_k}$ and $\mathbf{W}^{Kf}, \mathbf{W}^{Vf} \in \mathbb{R}^{5d_k\times d_k}$ are learnable parameter matrices. We calculate the attention of each band to all bands:

$$A_b^f = \text{softmax}\left(\frac{Q_b^f (K^f)^T}{\sqrt{d_k}}\right) V^f \in \mathbb{R}^{B\times d_k} \tag{13}$$

Finally, we integrate the original features and attention outputs through residual connections:

$$\hat{F}_b = F_b + A_b^f \in \mathbb{R}^{B\times d_k} \tag{14}$$

Adaptive Band Fusion and Feature Reconstruction: We calculate the importance of each frequency band through a learnable weight vector:

$$\alpha_b = \frac{\exp(\mathbf{w}^T \hat{F}_b)}{\sum_{i\in\{\delta,\theta,\alpha,\beta,\gamma\}} \exp(\mathbf{w}^T \hat{F}_i)}, \tag{15}$$

where $w \in \mathbb{R}^{d_k}$ is a learnable weight vector. We introduce a balance parameter $\alpha \in [0, 1]$ to control the proportion between processed features and the original signal:

$$E_n = \alpha \cdot \text{LN}\left(\mathcal{F}^{-1}\left(\sum_b \alpha_b \hat{F}_b\right)\right) + (1-\alpha)\cdot E_t, \tag{16}$$

where $\mathcal{F}^{-1}$ represents the inverse FFT and LN denotes layer normalization. The resulting $E_n \in \mathbb{R}^{B\times C\times L}$ serves as an optimized EEG embedding for subsequent multi-modal fusion.

3.2.4 Feature Normalization and Projection

After adaptive band fusion, we apply layer normalization to the EEG features and linear projection to map the normalized features to a shared embedding space, generating EEG embeddings that align dimensionally with audio and video features for subsequent hypergraph multi-modal fusion.

3.3 Unimodal Encoders

To enhance the representation capability of individual modality features and facilitate subsequent cross-modal information fusion, we encode each modality into a unified d-dimensional semantic space using specialized strategies tailored for different modality characteristics.

Following common practice in multimodal learning, we employ fully connected networks for audio and video features, while for EEG modality, we adopt Bidirectional Gated Recurrent Units (BiGRU) to capture bidirectional temporal dependencies due to the complexity of temporal characteristics and dynamic nature of emotional state changes. The specific encoding process is formulated as:

$$\begin{aligned} v_i^a &= \mathbf{W}_a A_{ij} + b_a \in \mathbb{R}^d, \\ v_i^v &= \mathbf{W}_v V_{ij} + b_v \in \mathbb{R}^d, \\ v_i^e &= \text{BiGRU}(E_n, v_{i(+,-)}^e) \in \mathbb{R}^d \end{aligned} \tag{17}$$

where $W_a \in \mathbb{R}^{d \times d_a}$ and $W_v \in \mathbb{R}^{d \times d_v}$ are the learnable weight matrices for audio and video modalities respectively, with corresponding bias vectors b_a and b_v. The raw audio features A_{ij} are extracted using the openSMILE toolkit with the IS10 configuration [32] from the audios, while the raw facial expressions features V_{ij} are extracted using a pre-trained MA-NET [33] from the videos. For EEG encoding, E_n represents the EEG features processed by ABEMA, and $v_{i(+,-)}^e$ represents the bidirectional processing result that fuses forward and backward temporal information.

This differentiated encoding strategy ensures that features from each modality are mapped into a unified semantic space while preserving their inherent characteristics, providing a solid foundation for subsequent multi-modal fusion.

3.4 Adaptive Hypergraph Fusion Module (AHFM)

Current approaches to ERC often simplify cross-modal interactions by modeling them as pairwise relationships (e.g., audio-text or video-text). In our study, our AHFM uses hypergraphs to directly capture complex higher-order relationships (e.g., simultaneous EEG-audio-video dependencies), which better reflect the group dynamics of multi-modal emotional cues. Furthermore, since each modality contributes uniquely to detecting instantaneous emotional shifts, we integrate learnable modality-specific weights. These weights are dynamically adjusted during training to prioritize the most informative modalities.

3.4.1 Hypergraph Construction

We employ hypergraph theory to model complex relational structures in multi-modal conversations. Given a conversation sequence with N utterance segments, we construct a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent multi-modal interaction patterns. The node set $\mathcal{V} = \{v_i^x | i \in [1, N], x \in \{e, a, v\}\}$ contains all unimodal segments, where v_i^e, v_i^a, v_i^v correspond to EEG, audio, and video modalities of the i-th segment respectively, totaling $3N$ nodes.

Based on the principles of multi-modal synergy and temporal continuity in emotional expression, we design two types of hyperedge connection patterns to capture different levels of relational dependencies. Intra-segment hyperedges $e_i^s = \{v_i^e, v_i^a, v_i^v\}$ connect all modality nodes within the same segment, modeling the synergistic effects of instantaneous multi-modal emotional expression and forming the intra-segment hyperedge set $\mathcal{E}_s$ with $|\mathcal{E}_s| = N$. Inter-segment hyperedges $e_x^t = \{v_i^x | i \in [1, N]\}$ connect nodes of the same modality across different time segments, capturing temporal evolution patterns of emotional states and constituting the inter-segment hyperedge set $\mathcal{E}_t$ with $|\mathcal{E}_t| = 3$. This design enables the hypergraph to simultaneously model synergistic relationships between modalities and evolutionary relationships across time, totaling $N + 3$ hyperedges.

To effectively model the differentiated contributions of nodes in different hyperedge types, we introduce an adaptive weight allocation mechanism with a node weight function $\omega : \mathcal{V} \times \mathcal{E} \to \mathbb{R}^+$ to quantify the influence intensity of node v_i^x within a specific hyperedge e_j.

We design a hierarchical weighting strategy where intra-segment hyperedges $e_i^s \in \mathcal{E}_s$ use node weights α_s^x to reflect the synergistic contributions of different modalities in instantaneous emotional expression, while inter-segment hyperedges

$e_x^t \in \mathcal{E}_t$ use weights α_t^x to capture the continuity strength of the same modality in temporal evolution. This constructs the weighted incidence matrix $\hat{\mathbf{H}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$:

$$\hat{\mathbf{H}}_{ij} = \begin{cases} \alpha_s^x, & \text{if } v_i^x \in e_j^s \text{ and } e_j^s \in \mathcal{E}_s; \\ \alpha_t^x, & \text{if } v_i^x \in e_j^t \text{ and } e_j^t \in \mathcal{E}_t; \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Additionally, we define the hyperedge weight diagonal matrix $\mathbf{W}_e = \text{diag}(\beta_s^1, ..., \beta_s^N, \beta_t^1, \beta_t^2, \beta_t^3)$, where β_s^i and β_t^j represent the importance weights of intra-segment and inter-segment hyperedges respectively, enabling adaptive learning of optimal weight distributions for each modality under different relational patterns.

3.4.2 Multi-modal Hypergraph Fusion

Based on the successful application of hypergraph convolution [34], we adapt it to multi-modal EEG emotion recognition tasks and optimize the information propagation mechanism by integrating our proposed hierarchical weighting strategy. This fusion process gradually refines higher-order multi-modal and contextual relationships through iterative node-hyperedge information exchange. Specifically, we implement information propagation through L-layer iterative updates:

$$\begin{aligned} V^{(1)} &= \sigma\left(\mathbf{D}_{\mathcal{V}}^{-1}\mathbf{I}\mathbf{W}_e\mathbf{D}_{\mathcal{E}}^{-1}\hat{\mathbf{H}}^T V^{(0)}\right) \\ &\cdots \\ V^{(L)} &= \sigma\left(\mathbf{D}_{\mathcal{V}}^{-1}\mathbf{I}\mathbf{W}_e\mathbf{D}_{\mathcal{E}}^{-1}\hat{\mathbf{H}}^T V^{(L-1)}\right) \end{aligned} \tag{19}$$

where $V^{(l)} = \{v_{i,(l)}^x | i \in [1, N], x \in \{e, a, v\}\} \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the input feature at layer l, σ is the non-linear activation function, $\mathbf{D}_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{D}_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ are the node degree matrix and hyperedge degree matrix respectively, used for feature normalization. Finally, we concatenate the representations of three modalities to obtain the fused feature:

$$f_i = [v_{i,(L)}^e; v_{i,(L)}^a; v_{i,(L)}^v] \tag{20}$$

3.5 Emotion Classification

To achieve final emotion prediction, we employ a multi-layer perceptron classifier to process the fused multi-modal features f_i. This classifier adopts a two-layer fully connected structure that can effectively learn complex emotional patterns. Specifically, the classification process includes non-linear transformation in the hidden layer, probability computation in the output layer, and final prediction decision:

$$\begin{aligned} h_i &= \text{ReLU}(\mathbf{W}_c f_i + \mathbf{b}_c) \\ P_i &= \text{softmax}(\mathbf{W}_o h_i + \mathbf{b}_o) \\ \hat{y}_i &= \arg\max_c P_i[c] \end{aligned} \tag{21}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are the weight matrix and bias vector of the classification layer respectively, $\mathbf{W}_o$ and $\mathbf{b}_o$ are the parameters of the output layer, and $P_i \in \mathbb{R}^C$ represents the probability distribution over C emotion categories.

During training stage, we employ categorical cross-entropy loss combined with L2 regularization:

$$\mathcal{L} = -\frac{1}{\sum_{s=1}^N |D_s|}\sum_{s=1}^N\sum_{i=1}^{|D_s|}\log P_{s,i}[y_{s,i}] + \lambda\|\theta\|_2 \tag{22}$$

where N represents the number of dialogues, $|D_s|$ denotes the number of utterance segments in the s-th dialogue, $P_{s,i}$ is the probability distribution, $y_{s,i}$ is the ground truth label, λ is the regularization weight, and θ represents the trainable parameters of the model.

Table 1: Subject-wise performance of Hyper-MML on EAV dataset.

Subject	Acc	F1	Subject	Acc	F1	Subject	Acc	F1
1	67.50	67.15	15	86.67	86.63	29	69.17	68.15
2	91.67	91.71	16	70.83	70.17	30	80.83	81.30
3	85.83	86.06	17	99.17	99.17	31	85.00	84.60
4	83.33	82.17	18	76.67	76.44	32	65.00	63.96
5	70.83	69.62	19	71.67	71.94	33	83.33	82.99
6	80.00	79.58	20	92.50	92.46	34	70.83	69.79
7	81.67	81.44	21	86.67	86.05	35	67.50	67.55
8	74.17	74.80	22	82.50	82.51	36	74.17	73.97
9	83.33	83.18	23	77.50	76.66	37	62.50	61.96
10	70.00	66.16	24	85.83	85.66	38	85.83	85.98
11	75.00	74.46	25	74.17	74.68	39	76.67	75.24
12	76.67	75.81	26	70.83	70.17	40	70.83	70.89
13	84.17	84.12	27	85.83	85.27	41	76.67	76.27
14	70.00	68.71	28	83.33	83.99	42	78.33	77.98
						Average	78.21	77.80

4 Experiments

4.1 EEG-based ERC Dataset

EAV[9]: The recently released multi-modal dialogue emotion dataset, EAV, includes EEG data from 30 channels, audio recordings, and facial expression videos from 42 subjects. This dataset represents the first publicly available collection that integrates EEG, audio, and video in a conversational context. Each subject engaged in 200 interactions within prompt-based dialogue scenarios, eliciting five distinct emotions: Neutral, Anger, Happy, Sad and Calm. Each interaction consisted of 20 seconds of listening followed by 20 seconds of speaking. For our evaluation, we focused exclusively on the speaking data of the subjects and followed the authors' preprocessing methods, segmenting the 20-second speech data stream into 5-second intervals. This approach aims to simulate the interruptions in conversation flow that individuals might encounter due to health conditions or hardware/software issues. By doing so, it disrupts the complete semantic structure and reflects scenarios in which the text modality may be missing or incomplete.

AFFEC[10]: The Advancing Face-to-Face Emotion Communication (AFFEC) dataset is a multimodal dataset designed to capture the dynamic complexities of face-to-face emotional interactions by integrating various modalities, including EEG, eye-tracking, GSR, facial actions, and Big Five personality assessment. The AFFEC dataset includes a total of 72 participants, covering educational levels from high school to doctoral degrees to ensure participant diversity. The dataset contains 84 simulated dialogues targeting six different emotions (anger, disgust, fear, happiness, neutrality, and sadness), totaling over 5,000 trials with continuous emotion annotation, which is subsequently discretized into three-level categories (High, Medium, Low) for both arousal and valence dimensions in the actual emotion recognition tasks. For our evaluation, we utilized three modalities: EEG, eye-tracking, and GSR. To address the challenge of limited sample size per subject for subject-wise analysis, we implemented a temporal windowing strategy that segments each trial into overlapping time windows, effectively expanding the dataset by a factor of five while preserving the temporal dynamics of emotional states. This data augmentation approach ensures sufficient training samples for robust subject-specific model learning while maintaining the integrity of physiological signal patterns.

4.2 Implementation Details and Evaluation Metrics

Our experiments were conducted on a Windows 11 system equipped with an NVIDIA RTX 3090 GPU. The framework was implemented using Python 3.8 and PyTorch 1.7.1. We employed the Adam optimizer for parameter updates with a learning rate of 0.0001. Training configurations included 40 epochs, batch size of 16, and dropout rate of 0.5 to prevent overfitting.

Model performance was evaluated using two standard metrics: accuracy (Acc) and F1-score (F1). Accuracy measures the proportion of correctly classified samples, while F1-score provides a balanced assessment considering both precision and recall, particularly important for imbalanced emotion datasets.

Table 2: Performance comparison between our method and other competing methods on EAV dataset across five emotion categories (Neutral, Anger, Happy, Sad, Calm). The best result in each column is presented in **bold**.

Methods	EAV						
	Emotion Categories (F1)					Overall	
	Neutral	Anger	Happy	Sad	Calm	Acc	F1
bc-LSTM[18]	58.16	64.57	47.13	64.48	52.14	57.21	57.30
DialogueRNN[11]	63.75	67.12	44.16	65.19	65.50	61.20	61.14
DialogueGCN[12]	68.32	70.18	56.17	80.14	68.94	68.73	68.75
MMGCN[3]	72.88	**74.12**	55.38	81.28	73.01	71.31	71.33
MM-DFN[35]	71.89	70.34	59.15	74.42	71.92	69.74	69.54
GraphMFT[36]	71.71	72.91	60.15	82.66	72.51	71.55	72.99
M3NET[37]	73.45	71.94	79.36	81.64	70.72	75.14	75.42
HAUCL[29]	75.21	72.80	82.44	78.17	70.97	75.91	75.92
AGF-IB[38]	73.53	71.76	81.87	80.91	72.34	76.19	76.08
AMERL[22]	75.18	71.88	72.80	79.25	**72.62**	74.82	74.35
Hyper-MML	**76.39**	73.05	**82.89**	**84.55**	72.14	**78.21**	**77.80**

Table 3: Performance comparison between our method and other competing methods on AFFEC dataset for arousal and valence classification tasks. The best result in each column is presented in in **bold**.

Methods	AFFEC task			
	Perceived-Arousal	Perceived-Valence	Felt-Arousal	Felt-Valence
bc-LSTM[18]	35.21	31.45	42.67	38.92
DialogueRNN[11]	38.94	33.78	45.83	41.26
DialogueGCN[12]	43.25	37.89	50.74	46.18
MMGCN[3]	46.83	40.56	53.27	49.71
MM-DFN[35]	45.92	39.84	52.45	48.89
GraphMFT[36]	47.68	41.87	54.69	51.02
M3NET[37]	49.34	43.21	56.85	52.78
HAUCL[29]	50.87	44.67	58.42	54.31
AGF-IB[38]	51.23	45.12	58.91	54.87
AMERL[22]	50.45	44.38	57.96	53.94
Hyper-MML	**53.76**	**47.89**	**60.23**	**57.32**

4.3 Competing Methods

In order to validate the effectiveness of our proposed Hyper-MML for multi-modal ERC task, we conducted extensive comparative experiments. However, there is currently insufficient exploration in the field of dialogue emotion recognition based on EEG. Aside from one multi-modal dialogue emotion recognition model utilizing the EEG modality, known as AMERL, there are virtually no other direct benchmark models available in the literature. Therefore, we selected several emotion recognition benchmark models that are widely applied in the field of Natural Language Processing (NLP), which typically integrate text, audio, and facial expression modalities for multi-modal emotion analysis. For these benchmark models, we replaced the text modality embeddings with EEG modality embeddings to assess the potential of EEG signals in emotion recognition. We believe that EEG signals can effectively capture emotional information and complement audio and facial expression modalities, thereby providing a more comprehensive perspective for emotion recognition. Below, we present a brief introduction to some of the comparative models employed in our study:

bc-LSTM[18]: A bidirectional LSTM variant that processes utterances in both temporal directions, capturing contextual dependencies for improved sentiment classification in video sequences.

DialogueRNN[11]: This RNN-based architecture models participant states and contextual dynamics using gated recurrent units, effectively tracking emotional evolution between speakers and listeners.

DialogueGCN[12]: The first graph convolutional approach for conversational emotion recognition, addressing context propagation limitations in RNN methods by modeling self-dependency and inter-speaker relationships.

MMGCN[3]: A deep graph convolutional framework that fuses multimodal information while capturing long-range contextual dependencies and inter-speaker interactions in dialogue sequences.

MM-DFN[35]: This dynamic fusion network integrates text, audio, and video through interactive multiview memory modules, adapting feature importance based on conversational context.

Table 4: Subject-wise performance of Hyper-MML on AFFEC dataset.

Subject	Acc	F1	Subject	Acc	F1	Subject	Acc	F1
1	62.12	58.34	25	45.45	41.23	49	68.94	65.78
2	73.48	71.92	26	59.09	55.67	50	47.73	43.89
3	56.06	52.41	27	71.97	69.85	51	65.15	62.03
4	68.18	65.29	28	53.03	49.17	52	60.61	57.24
5	44.70	40.83	29	66.67	63.51	53	72.73	70.18
6	58.33	54.76	30	49.24	45.67	54	55.30	51.42
7	63.64	60.29	31	70.45	67.83	55	67.42	64.15
8	51.52	47.89	32	46.97	43.21	56	61.36	58.07
9	69.70	66.94	33	64.39	61.12	57	73.48	71.26
10	57.58	53.85	34	52.27	48.54	58	48.48	44.73
11	65.91	62.67	35	71.21	68.45	59	62.88	59.41
12	50.76	46.92	36	45.45	41.67	60	56.82	53.18
13	67.42	64.28	37	59.85	56.23	61	69.70	66.87
14	54.55	50.84	38	73.48	71.15	62	51.52	47.64
15	71.21	68.73	39	47.73	43.95	63	64.39	61.28
16	46.21	42.18	40	61.36	58.19	64	58.33	54.89
17	60.61	57.12	41	55.30	51.73	65	72.73	70.04
18	68.94	65.87	42	68.18	65.42	66	46.97	43.35
19	53.79	50.16	43	72.73	70.11	67	60.61	57.48
20	65.15	61.89	44	50.00	46.25	68	54.55	51.07
21	49.24	45.38	45	63.64	60.47	69	67.42	64.52
22	71.97	69.12	46	57.58	53.96	70	52.27	48.81
23	45.45	41.74	47	70.45	67.59	71	65.91	62.84
24	59.09	55.43	48	48.48	44.61	72	59.85	56.37
						Average	60.23	56.78

GraphMFT[36]: A graph attention-based technique that simultaneously captures intra-modal context and inter-modal complementarity through enhanced attention mechanisms.

M3NET[37]: This approach models multivariate relationships and multi-frequency signals using graph neural networks, capturing complex utterance interdependencies in conversational scenarios.

HAUCL[29]: A hypergraph-based framework combining variational autoencoders with contrastive learning to dynamically adjust connections while reducing contextual redundancy and over-smoothing.

AGF-IB[38]: This method eliminates inter-modal heterogeneity through information bottleneck theory and adversarial learning, while employing graph contrastive learning for semantic information capture.

AMERL[22]: A multimodal framework specifically designed for EEG integration, using dynamic attention mechanisms to adaptively weight EEG, video, image, and audio features for robust emotion recognition.

5 Results and Discussion

5.1 Comparison with Baselines

Table 1 presents the subject-wise performance of our proposed Hyper-MML framework across all 42 participants in the EAV dataset. The results demonstrate considerable inter-subject variability, with accuracy ranging from 62.50% to 99.17%, reflecting the inherent individual differences in EEG signal characteristics. Despite this variability, our method achieves an overall average accuracy of 78.21% and F1-score of 77.80%, indicating robust performance in EEG-based conversational emotion recognition.

Table 2 provides a comprehensive comparison between our Hyper-MML framework and other baseline methods on the EAV dataset. The baseline methods span from traditional RNN-based approaches (bc-LSTM, DialogueRNN) to advanced graph-based methods (DialogueGCN, MMGCN, GraphMFT) and recent hypergraph techniques (M3NET, HAUCL, AGF-IB). Our approach demonstrates superior performance across multiple evaluation metrics, achieving the highest overall accuracy of 78.21% and F1-score of 77.80%, representing significant improvements of 2.30% in accuracy and 1.88% in F1-score compared to the previous best-performing method AGF-IB. The performance gains are particularly pronounced in the recognition of specific emotional categories: our method achieves the highest F1-scores

Table 5: Comparison of EEG encoding methods on EAV and AFFEC datasets. AFFEC results from Felt-Arousal task. The best result in each column is presented in **bold**.

Encoder	EAV		AFFEC	
	Acc	F1	Acc	F1
MLP	61.80	61.56	55.97	55.67
EEGNet[15]	69.18	69.94	58.08	54.84
TSConv[39]	73.22	72.12	58.12	54.67
ATMS[40]	76.12	75.83	59.12	55.08
NESTA[30]	76.40	76.37	59.80	55.82
ABEMA	**78.21**	**77.80**	**60.23**	**56.78**

for Sad (84.55%) and Happy (82.89%) emotions, with improvements of 3.64% and 0.45% respectively over the previous best results.

While hypergraph-based methods generally outperform traditional approaches, our method achieves superior results through more effective utilization of the hypergraph structure rather than increased model complexity. Our superior performance stems from two key factors rather than architectural complexity. EEG integration provides direct neural correlates, offering more reliable indicators than reconstructed features in competing methods. Additionally, AHFM explicitly models both intra-segment multimodal synergy and inter-segment temporal continuity, while existing methods focus primarily on complex fusion without optimizing hypergraph structure.

To validate the generalizability of our approach, we further evaluated Hyper-MML on the AFFEC dataset, which presents a more challenging task of arousal and valence classification in three-level categories (Low/Medium/High). Table 3 demonstrates that our method consistently outperforms all baseline approaches across four distinct classification tasks, achieving the highest F1-scores of 53.76% for Perceived-Arousal, 47.89% for Perceived-Valence, 60.23% for Felt-Arousal, and 57.32% for Felt-Valence, representing improvements of 2.53%, 2.77%, 1.32%, and 2.45% respectively over the previous best-performing method AGF-IB. Results confirm established patterns: felt emotions outperform perceived emotions, and arousal tasks exceed valence tasks in accuracy. Table 4 presents the subject-wise performance across all 72 participants for Felt-Valence task on the AFFEC dataset, showing an overall average accuracy of 60.23% and F1-score of 56.78% with considerable inter-subject variability.

5.2 Effectiveness of ABEMA

As shown in Table 5, our proposed ABEMA encoder demonstrates superior performance compared to five established EEG processing methods across both datasets. On the EAV dataset, ABEMA achieves the highest accuracy of 78.21% and F1-score of 77.80%, outperforming the second-best method NESTA by 1.81% in accuracy and 1.43% in F1-score respectively. The comparative results clearly illustrate the evolution from traditional MLP approaches with 61.80% accuracy to advanced deep learning methods like EEGNetV4 and TSConv achieving 69.18% and 73.22% accuracy respectively, with attention-based methods ATMS and NESTA showing further improvements. ABEMA's hierarchical mutual-cross attention mechanism and adaptive frequency band fusion strategy achieve the best results by effectively modeling both intra-band and inter-band relationships in EEG signals.

We also report results from the Felt-Arousal classification task on the AFFEC dataset. AFFEC results maintain this advantage with 60.23% accuracy, exceeding NESTA by 0.43%. This consistent advantage across different task formulations demonstrates that ABEMA's subject-specific adaptation layer and hierarchical attention mechanism effectively address fundamental challenges in EEG-based emotion recognition, including inter-subject variability and complex frequency-domain relationships.

5.3 Modality Contribution Analysis

Table 6 presents a comprehensive analysis of individual modality contributions in our multimodal framework. Among the three modalities, EEG demonstrates the strongest individual performance with 71.83% accuracy and 72.04% F1-score, highlighting the critical role of physiological signals in capturing objective emotional states. Audio modality achieves competitive results with 69.47% accuracy and 69.88% F1-score, reflecting the rich emotional information embedded in vocal characteristics such as tone, pitch, and prosody. Visual modality shows the lowest individual performance at 68.36% accuracy and 68.54% F1-score, which aligns with the challenges of extracting subtle emotional cues from facial expressions in conversational contexts where speakers may not exhibit pronounced visual emotional indicators.

Table 6: Modality contribution analysis on EAV dataset comparing individual and combined modality performance. The best result in each column is presented in **bold**.

Modalities	EAV	
	Acc	F1
EEG	71.83	72.04
Audio	69.47	69.88
Video	68.36	68.54
EEG+Audio+Video	**78.21**	**77.80**

Table 7: Ablation study results on EAV dataset examining key components of ABEMA and AHFM modules. The best result in each column is presented in **bold**.

Component Settings		Acc	F1
ABEMA	w/o Intra-band Attention	72.11	72.28
	w/o Inter-band Attention	72.29	72.83
	w/o Both Attention	70.24	70.67
AHFM	w/o Node Weights (α_s^x, α_t^x)	75.92	75.49
	w/o Hyperedge Weights (β_s^i, β_t^j)	75.22	75.31
	w/o Both Weights	72.46	72.87
Hyper-MML		**78.2**1	**77.80**

The integration of all three modalities (EEG+Audio+Video) achieves substantial performance improvements, reaching 78.21% accuracy and 77.80% F1-score, representing gains of 6.38% in accuracy and 5.76% in F1-score compared to the best individual modality (EEG). This significant enhancement demonstrates the complementary nature of different modalities in capturing the multifaceted aspects of emotional expression. The synergistic effect suggests that while EEG provides reliable physiological indicators of emotional states, audio and visual cues contribute additional contextual information that enhances overall recognition accuracy. The superior performance of the combined approach validates our hypergraph-based fusion strategy's effectiveness in modeling complex higher-order relationships among modalities, enabling the framework to leverage the unique strengths of each modality while compensating for their individual limitations.

5.4 Ablation Study

To validate the effectiveness of key components in our Hyper-MML framework, we perform comprehensive ablation experiments on the the key components of Hyper-MML in Table 7.

Performance Evaluation of ABEMA: The ablation study on ABEMA reveals the critical importance of the hierarchical mutual-cross attention mechanism. Removing intra-band attention results in 72.11% accuracy and 72.28% F1-score, while removing inter-band attention leads to 72.29% accuracy and 72.83% F1-score, indicating that inter-band relationships are slightly more crucial than intra-band dependencies. The most significant performance degradation occurs when both attention mechanisms are removed, dropping to 70.24% accuracy and 70.67% F1-score, representing decreases of 7.97% and 7.13% respectively. This substantial decline confirms that the hierarchical attention design effectively captures complex frequency-domain relationships in EEG signals, enabling optimal integration of complementary information across different frequency bands.

Performance Evaluation of AHFM: The AHFM ablation experiments demonstrate the effectiveness of the adaptive weighting strategy in hypergraph construction. Removing node weights results in 75.92% accuracy and 75.49% F1-score, while removing hyperedge weights leads to 75.22% accuracy and 75.31% F1-score, suggesting that node-level adaptive weighting has a slightly greater impact on performance. When both weight mechanisms are eliminated, performance drops to 72.46% accuracy and 72.87% F1-score, representing decreases of 5.75% and 4.93% respectively. These results validate that the adaptive weight allocation mechanism enables the model to dynamically prioritize the most informative modalities and relationship types, optimizing information propagation within the hypergraph structure and significantly enhancing multimodal fusion effectiveness.

5.5 Visualization

As shown in Figure 5, the confusion matrix reveals strong diagonal dominance with robust performance for neutral, sad, and anger emotions, while some confusion occurs between emotionally similar categories such as happy and calm, reflecting the inherent challenge of distinguishing positive emotions with varying arousal levels. For the AFFEC dataset,

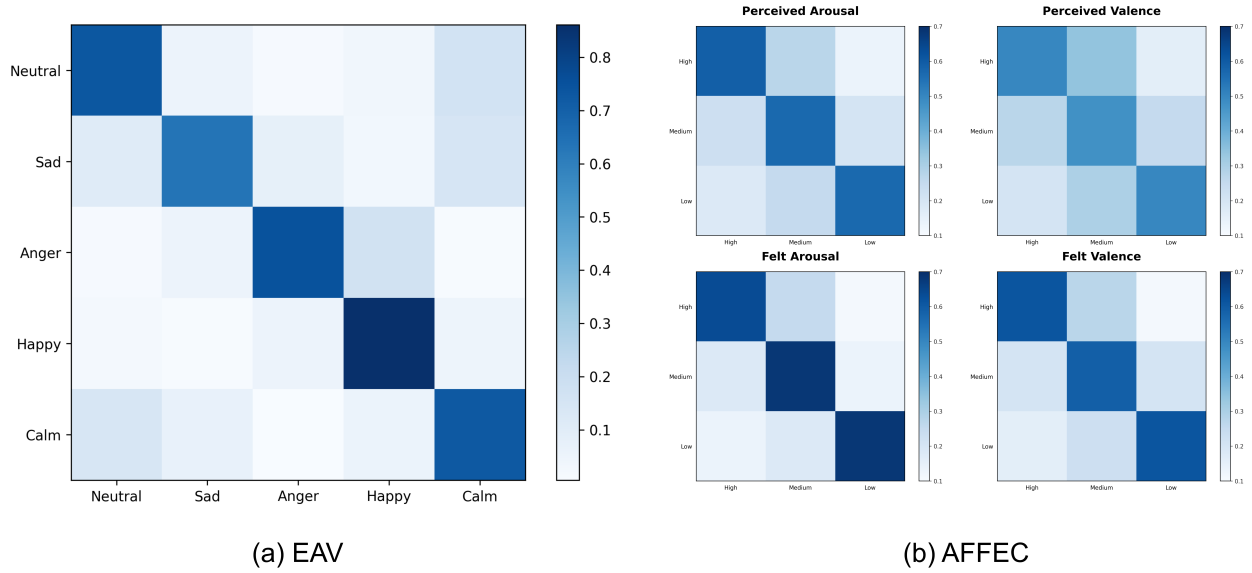

Figure 5: Confusion matrix on the EAV and AFFEC datasets. (a) EAV dataset shows confusion matrix for five emotion categories. (b) AFFEC dataset presents confusion matrices for four classification tasks: Perceived-Arousal, Perceived-Valence, Felt-Arousal, and Felt-Valence with three-level categories (High/Medium/Low).

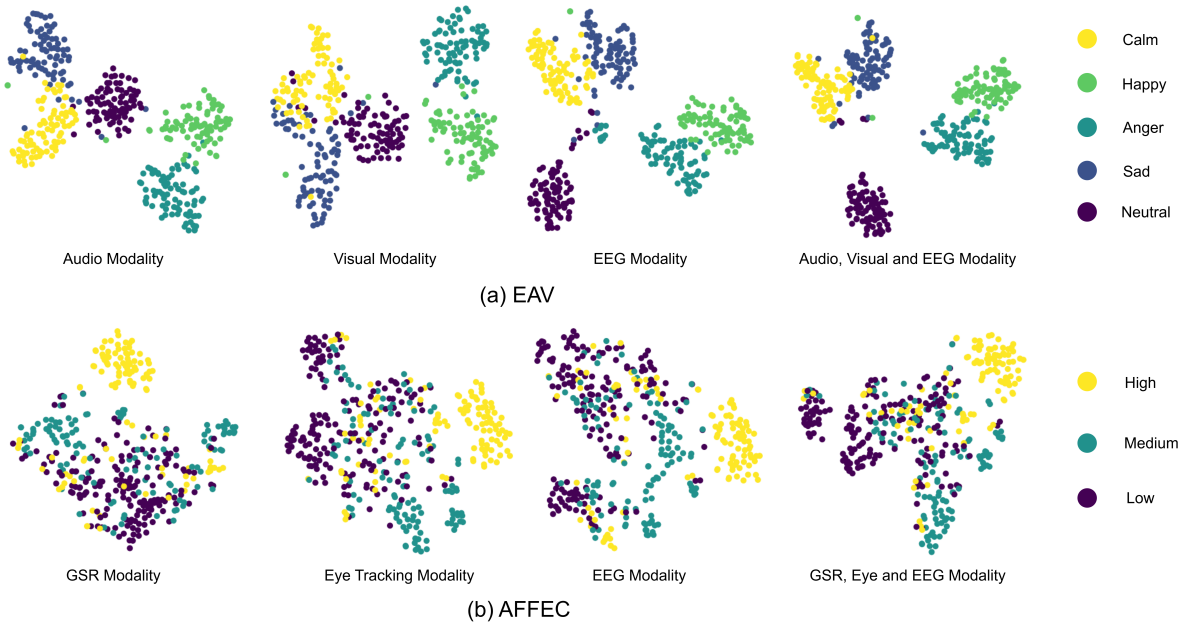

Figure 6: T-SNE visualization of learned features on the EAV and AFFEC datasets. (a) EAV dataset visualization across audio, visual, EEG modalities and their fusion. (b) AFFEC dataset visualization for the Felt-Arousal task showing GSR, Eye-tracking, EEG modalities and their multimodal fusion. Each dot represents a segment, and colors indicate emotion categories.

the Felt-Arousal and Felt-Valence tasks demonstrate superior performance compared to their Perceived counterparts, confirming that self-reported emotions are more reliably classified than observer-perceived emotions. The Medium category shows increased confusion with adjacent categories, which is partly attributed to the inherent class imbalance in the AFFEC dataset where Medium samples are less represented compared to High and Low categories.

Figure 6 shows the t-SNE visualization of learned features. For the EAV dataset, the individual modality analysis reveals distinct characteristics: Audio modality demonstrates moderate clustering with reasonable emotion separation, Visual modality shows relatively scattered distributions with less clear boundaries due to the challenges of facial expression analysis in conversational contexts, while EEG modality presents compact and well-separated clusters, validating the reliability of physiological signals in capturing objective emotional states. For the AFFEC dataset, we focus on the best-performing Felt-Arousal task with GSR, Eye-tracking, and EEG modalities, where EEG again demonstrates superior clustering performance despite the class imbalance challenges. Most significantly, the multimodal fusion visualization reveals dramatically enhanced feature discrimination with clearer inter-class boundaries and substantially reduced intra-class variance compared to individual modalities. This improved separability provides visual confirmation that our hypergraph-based AHFM successfully models higher-order relationships among modalities, leveraging their complementary strengths while compensating for individual limitations to achieve superior emotion recognition performance.

6 Conclusion and Future Work

In this study, we introduced the Hypergraph Multi-Modal Learning framework (Hyper-MML) for EEG-based emotion recognition in conversations, addressing the limitations of traditional methods that primarily rely on textual information and struggle with incomplete dialogue scenarios. By integrating EEG signals with audio and video data through our novel hypergraph architecture, our framework effectively captures the intricate emotional dynamics inherent in conversational interactions while providing objective physiological indicators independent of language coherence. The proposed ABEMA encoder with hierarchical mutual-cross attention successfully models complex frequency-domain relationships in EEG signals, while the AHFM significantly enhances the model's ability to process higher-order multimodal relationships, leading to a more nuanced understanding of emotional states. Our comprehensive experiments on both EAV and AFFEC datasets demonstrate that the proposed framework achieves state-of-the-art performance with substantial improvements over existing methods, validating the robustness and generalizability of our hypergraph-based approach. Future research should explore the integration of additional physiological signals to further enhance recognition accuracy, investigate real-time processing capabilities for practical deployment, and validate the framework's effectiveness in real-world clinical applications, such as mental health monitoring and adaptive human-computer interaction.

Acknowledgments

This work was supported by The Hong Kong Polytechnic University Start-up Fund (Project ID: P0053210), The Hong Kong Polytechnic University Faculty Reserve Fund (Project ID: P0053738), an internal grant from The Hong Kong Polytechnic University (Project ID: P0048377), The Hong Kong Polytechnic University Departmental Collaborative Research Fund (Project ID: P0056428), The Hong Kong Polytechnic University Collaborative Research with World-leading Research Groups Fund (Project ID: P0058097) and Research Grants Council Collaborative Research Fund (Project ID: P0049774).

References

[1] Amaia Hervás. Autism and Depression: clinical presentation, evaluation and treatment. *Medicina (Argentina)*, 83(Suppl 2):37–42, 2023.

[2] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953, 2019.

[3] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*, 2021.

[4] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. EEG based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022.

[5] Yueyang Li, Weiming Zeng, Wenhao Dong, Di Han, Lei Chen, Hongyu Chen, Zijian Kang, Shengyu Gong, Hongjie Yan, Wai Ting Siok, et al. A tale of single-channel electroencephalogram: Devices, datasets, signal processing, applications, and future directions. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[6] Yimin Zhao and Jin Gu. Feature fusion based on mutual-cross-attention mechanism for eeg emotion recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 276–285. Springer, 2024.

[7] Hongyu Chen, Weiming Zeng, Chengcheng Chen, Luhui Cai, Fei Wang, Yuhu Shi, Lei Wang, Wei Zhang, Yueyang Li, Hongjie Yan, et al. Eeg emotion copilot: Optimizing lightweight llms for emotional eeg interpretation with assisted medical record generation. *Neural Networks*, page 107848, 2025.

[8] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Min-Ho Lee, Adai Shomanov, Balgyn Begim, Zhuldyz Kabidenova, Aruna Nyssanbay, Adnan Yazici, and Seong-Whan Lee. EAV: EEG-Audio-Video dataset for emotion recognition in conversational contexts. *Scientific data*, 11(1):1026, 2024.

[10] Meisam J Sekiavandi, Laurits Dixen, Jostein Fimland, Sree Keerthi Desu, Antonia-Bianca Zserai, Ye Sul Lee, Maria Barrett, and Paolo Burelli. Advancing face-to-face emotion communication: A multimodal dataset (affec). *arXiv preprint arXiv:2504.18969*, 2025.

[11] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825, 2019.

[12] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[13] Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873, 2024.

[14] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[15] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[16] Ming Jin, Changde Du, Huiguang He, Ting Cai, and Jinpeng Li. Pgcn: Pyramidal graph convolutional network for eeg emotion recognition. *IEEE Transactions on Multimedia*, 26:9070–9082, 2024.

[17] Cunbo Li, Tian Tang, Yue Pan, Lei Yang, Shuhan Zhang, Zhaojin Chen, Peiyang Li, Dongrui Gao, Huafu Chen, Fali Li, et al. An efficient graph learning system for emotion recognition inspired by the cognitive prior graph of eeg brain network. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7130–7144, 2024.

[18] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal residual lstm network. In *Proceedings of the 27th ACM international conference on multimedia*, pages 176–183, 2019.

[19] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3):839–847, 2018.

[20] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2022.

[21] Dongyuan Tian, Yucheng Wang, Peiliang Gong, Zhewen Xu, Zhenghua Chen, Xiaohui Wei, and Min Wu. Accnet: Adaptive cross-frequency coupling graph attention for eeg emotion recognition. *Neural Networks*, page 107853, 2025.

[22] Kang Yin, Hye-Bin Shin, Dan Li, and Seong-Whan Lee. Eeg-based multimodal representation learning for emotion recognition. In *2025 13th International Conference on Brain-Computer Interface (BCI)*, pages 1–4. IEEE, 2025.

[23] Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. A survey on hypergraph representation learning. *ACM Computing Surveys*, 56(1):1–38, 2023.

[24] Uthsav Chitra and Benjamin Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *International Conference on Machine Learning*, pages 1172–1181. PMLR, 2019.

[25] Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2022.

[26] Lianghao Xia, Chao Huang, and Chuxu Zhang. Self-supervised hypergraph transformer for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2100–2109, 2022.

[27] Yuze Liu, Ziming Zhao, Tiehua Zhang, Kang Wang, Xin Chen, Xiaowei Huang, Jun Yin, and Zhishu Shen. Exploiting spatial-temporal data for sleep stage classification via hypergraph learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5430–5434. IEEE, 2024.

[28] Ding Ruan, Shuyi Ji, Chenggang Yan, Junjie Zhu, Xibin Zhao, Yuedong Yang, Yue Gao, Changqing Zou, and Qionghai Dai. Exploring complex and heterogeneous correlations on hypergraph for the prediction of drug-target interactions. *Patterns*, 2(12), 2021.

[29] Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiehua Zhang. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 4341–4348, 2024.

[30] Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-mcrl: Neural multimodal contrastive representation learning for eeg-based visual decoding. *arXiv preprint arXiv:2412.17337*, 2024.

[31] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[32] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[33] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[34] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770, 2023.

[35] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE, 2022.

[36] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427, 2023.

[37] Ravikiran Mane, Neethu Robinson, A Prasad Vinod, Seong-Whan Lee, and Cuntai Guan. A multi-view cnn with novel variance layer for motor imagery brain computer interface. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 2950–2953. Ieee, 2020.

[38] Yuntao Shou, Tao Meng, Wei Ai, Fuchen Zhang, Nan Yin, and Keqin Li. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. *Information Fusion*, 112:102590, 2024.

[39] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

[40] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.