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

Zijian Kang^{1,*}, Yueyang Li^{1,*}, Shengyu Gong¹, Weiming Zeng^{1,†}, Hongjie Yan², Lingbin Bian³, Wai Ting Siok³, and Nizhuan Wang^{3,†} *l

¹Lab of Digital Image and Intelligent Computation, Shanghai Maritime University, Shanghai 201306, China ²Affiliated Lianyungang Hospital of Xuzhou Medical University, Lianyungang 222002, China ³Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

ABSTRACT

Emotional Recognition in Conversation (ERC) is an important method for diagnosing health conditions such as autism or depression, as well as understanding emotions in individuals who struggle to express their feelings. Current ERC methods primarily rely on complete semantic textual information, including audio and visual data, but face challenges in integrating physiological signals such as electroencephalogram (EEG). This paper proposes a novel Hypergraph Multi-Modal Learning Framework (Hyper-MML), designed to effectively identify emotions in conversation by integrating EEG with audio and video information to capture complex emotional dynamics. Experimental results demonstrate that Hyper-MML significantly outperforms traditional methods in emotion recognition. This is achieved through a Multi-modal Hypergraph Fusion Module (MHFM), which actively models higher-order relationships between multi-modal signals, as validated on the EAV dataset. Our proposed Hyper-MML serves as an effective communication tool for healthcare professionals, enabling better engagement with patients who have difficulty expressing their emotions.

Keywords Emotion Recognition in Conversation · Hypergraph Learning · Multi-modal Feature Fusion · EEG · Audio · Video.

1 Introduction

Emotion Recognition in Conversation (ERC) holds significant potential for diagnosing health and mental conditions such as autism and depression. Recent research suggests 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 research in ERC primarily focuses on textual analysis, supplemented by visual cues like facial expressions and acoustic cues such as intonation and loudness [2]. These methods typically rely on complete, uninterrupted texts or dialogue transcripts, integrating multi-modal data to provide context for individual utterances. Context modeling in ERC generally incorporates three key elements: 1) the content of preceding exchanges, 2) the timing of conversational turns, and 3) speaker-specific details such as identity and evolving emotional states [4, 5]. However, disruptions in the textual flow – such as fragmented sentences or missing dialogue segments – can break the semantic structure and distort logical relationships between utterances. This incoherence reduces the accuracy of emotion recognition, limiting its practical applications in diagnosing and treating conditions like autism and depression. To address these limitations, psychotherapists need alternative indicators that remain robust even when conversational data is imperfect.

Physiological signals – particularly electroencephalogram (EEG) data – provide a direct window into neural activity and emotional states, surpassing text-based methods in objectivity and immediacy [3, 18]. While textual analysis

^{**:} Zijian Kang and Yueyang Li are co-first authors. †: Weiming Zeng and Nizhuan Wang are corresponding authors. This work was supported by the National Natural Science Foundation of China (No.31870979), the Hong Kong Polytechnic University Faculty Reserve Fund (Project ID: P0053738), and the Hong Kong Polytechnic University Start-up Fund (Project ID: P0053210).

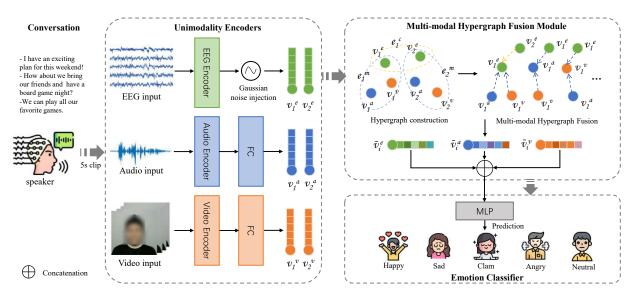


Figure 1: Overall framework of the proposed Hyper-MML.

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. 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 language-independent 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 [14]. This integrative framework allows psychologists to correlate physiological responses (e.g., brainwave patterns) with behavioural cues (e.g. voice tone, facial expressions), constructing a comprehensive emotional profile. These insights support the development of tailored treatment strategies that better address individual patient needs.

In multi-modal dialogue recognition tasks, a standard approach involves using graph neural networks (GNNs) 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 natively 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 signle interaction step. This capability makes hypergraphs a more precise and efficient framework for multi-modal dialogue analysis [7].

In this study, we propose a novel Hypergraph Multi-Modal Learning framework (Hyper-MML) centered on EEG signals, which has been validated as a state-of-the-art (SOTA) method on the EAV dataset [8]. Our framework advances ERC through two 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 (e.g., shifts between frustration, surprise, or relief) in conversations. Unlike traditional text-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) Multi-modal Hypergraph Fusion Module (MHFM): A specialized module that enhances cross-modal interaction (EEG-audio-video) 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.

2 Method

2.1 Problem Definition

The EEG-based ERC aims to infer the emotional state of each incomplete semantic segment of utterances. For each complete utterance, we segment it into N segments $\{s_1, s_2, ..., s_N\}$, where each segment involves three sources of segment-aligned data corresponding to three modalities: EEG (s_i^e) , audio (s_i^a) , and video (s_i^v) , represented as follows:

$$s_i = \{s_i^e, s_i^a, s_i^v\} \tag{1}$$

The objective of EEG-based ERC task is to predict the emotional category of each fragment s_i from a predefined set of C emotional categories, i.e., Happy, Sad, Calm, Angry and Neutral. Figure 1 illustrates the proposed Hyper-MML framework based on EEG-audio-video triplets. In general, the Hyper-MML consists of three key modules: Unimodality Encoders, Multi-modal Hypergraph Fusion Module and Emotion Classifier.

2.2 Unimodality Encoders

Effectively capturing contextual information between utterance segments is challenging due to incomplete semantic data. To address this, we propose extracting short-context-window embeddings within each of three modalities (EEG, audio, video) at the segment level. For EEG signals, which reflect instantaneous neural activity, our focus lies in extracting temporal features and dynamic patterns to capture rapid emotional shifts.

1) Acoustic and Visual Embedding: For audio and video modalities, we used established fully connected networks [9, 10] as encoders. The short-context-aware feature encoding for each segment can be formulated as follows:

$$v_i^a = \mathbf{W}_1 f_i^a + b_i^a, \ v_i^v = \mathbf{W}_2 f_i^v + b_i^v \tag{2}$$

where f_i^a, f_i^v are the context-independent raw feature of segment i from the audio and video modalities, respectively. The raw audio features f_i^a are extracted using the openSMILE toolkit with the IS10 configuration [9] from the audios, while the raw facial expressions features f_i^v are extracted using a pre-trained MA-NET [10] from the videos. The unimodality encoder for each audio and video outputs the short-context-aware raw feature embedding v_i^a, v_i^v accordingly.

2) **EEG Embedding**: For EEG signals, to capture effective subject-specific information, we use a specialized EEG encoder NESTA that jointly learns subject-specific channel transformations and adaptively captures spectral patterns while preserving key temporal-spectral information within the EEG signals [11].

$$v_i^e = \text{NESTA}(s_i^a) \tag{3}$$

2.3 Multi-modal Hypergraph Fusion Module (MHFM)

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 MHFM 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.

1) Hypergraph Construction: Generally, a conversation with N utterance segments can be reformulated as a hypergraph $\mathcal{H} = (\mathcal{V}_{\mathcal{H}}, \mathcal{E}_{\mathcal{H}})$, in which each node $v \in \mathcal{V}_{\mathcal{H}}$ corresponds to a unimodal segment, and every hyperedge $e \in \mathcal{E}_{\mathcal{H}}$ encodes multimodal or contextual dependencies. Let $\mathbf{I} \in \mathbb{R}^{|\mathcal{V}_{\mathcal{H}}| \times |\mathcal{E}_{\mathcal{H}}|}$ represent the incident matrix, in which a nonzero entry $\mathbf{I}_{ve} = 1$ indicates that the hyperedge e is incident with the node v; otherwise $\mathbf{I}_{ve} = 0$. Each segment is represented by three nodes, i.e., v_i^e, v_i^a, v_i^v , in the hypergraph, corresponding to EEG, audio and video modalities, respectively.

To capture relationships that extend beyond pairwise interactions in multi-modal emotion recognition based on utterance segments, the complex multi-modal relationships of each utterance segment are constructed as hyperedges. Each node $v_i^x(x \in \{e, a, v\})$ is connected to other modalities of the same utterance segment $\{v_i^y|y \neq x, y \in \{e, a, v\}\}$, forming a multi-modal hyperedge e^m . Additionally, we connect EEG signals $\{v_i^e|i \in [1, N]\}$ from different segments of the same utterance to create short-context hyperedges e^c . The hyperedges $\mathcal{E}_{\mathcal{H}}$ are therefore divided into two subsets: the multi-modal hyperedge set \mathcal{E}_m and the contextual hyperedge set \mathcal{E}_c . This approach enables the constructed hypergraph to capture higher-order mutual information and contextual information between multi-modal data, thereby transcending the limitations of pairwise interactions.

Inspired by the hypergraph study of edge-dependent vertex weights [12], we set different node weights for multimodal hyperedges and context-dependent nodes, aiming to distinguish the contributions of modality nodes to different relational patterns. Therefore, the edge-dependent vertex weights can be represented by a weighted incidence matrix:

$$\hat{\mathbf{H}}_{ij} = \begin{cases} \gamma_m(e_j), & \text{if } v_i \in e_j \text{ and } e_j \in \mathcal{E}_m; \\ \gamma_c(e_j), & \text{if } v_i \in e_j \text{ and } e_j \in \mathcal{E}_c; \\ 0, & \text{otherwise.} \end{cases}$$

$$\tag{4}$$

in which $\gamma_m(e_j)$ is the multi-modal edge-dependent vertex weights, while $\gamma_c(e_j)$ is the context edge-dependent vertex weights. Analogously, the hyperedge weight matrix can be defined as follow:

$$\mathbf{W}_{e} = \operatorname{diag}(w_{m}(e_{1}), ..., w_{m}(e_{|\mathcal{E}_{m}|}), w_{c}(e_{1}), ..., w_{c}(e_{|\mathcal{E}_{c}|}))$$
(5)

where $w_m(e_i)$ and $w_c(e_i)$ is the multi-modal hyperedge weight and the context hyperedge weight, respectively.

2) Hypergraph Feature Fusion: Inspired by M^3Net [6], the core of MHFM involves a hypergraph convolution operation that propagates multivariate embeddings across the hypergraph. In this operation, we dynamically adjust the importance of each modality through learnable weights, enabling the model to prioritize the most relevant features for emotion recognition tasks. This flexibility allows MHFM to adaptively capture the complex interactions and contextual relationships inherent in multi-modal data.

Mathematically, the module updates the embeddings based on the aggregated features of neighboring nodes, enhancing the representation of each modality while maintaining the integrity of the hypergraph structure. The formula for MHFM is as follows:

$$V^{(l+1)} = \sigma \left(\mathbf{D}_{\mathcal{H}}^{-1} \mathbf{I} \mathbf{W}_e \mathbf{B}_{\mathcal{H}}^{-1} \hat{\mathbf{H}}^T V^{(l)} \right)$$
 (6)

in which $V^{(l)} = \{v^x_{i,(l)} | i \in [1,N], x \in \{e,a,v\}\} \in \mathbb{R}^{\mathcal{V}_{\mathcal{H}} \times \mathcal{E}_{\mathcal{H}}}$ is the input at layer l. σ is the non-linear activation function. $\mathbf{D}_{\mathcal{H}} \in \mathbb{R}^{|\mathcal{V}_{\mathcal{H}}| \times |\mathcal{V}_{\mathcal{H}}|}$ is the node degree matrix used to normalize node features. $\mathbf{B}_{\mathcal{H}} \in \mathbb{R}^{|\mathcal{E}_{\mathcal{H}}| \times |\mathcal{E}_{\mathcal{H}}|}$ and hyperedge degree matrix that reflects the connectivity of hyperedges. After performing L iterations, we get the outputs of the last iteration $v_{i,(L)}^x$ as the multivariate representations:

$$\bar{v}_i^e = v_{i,(L)}^e, \quad \bar{v}_i^a = v_{i,(L)}^a, \quad \bar{v}_i^v = v_{i,(L)}^v$$
 (7)

Finally, we concatenate the embeddings of three modalities to obtain the emotional embedding of the utterance segment as follow:

$$e_i = \bar{v}_i^e \oplus \bar{v}_i^a \oplus \bar{v}_i^v \tag{8}$$

3) Emotion Classification: The emotion classifier takes as input the concatenated multivariate representations to perform emotion prediction. Referring to prior works [5], we finally feed e_i into a multilayer perceptron (MLP) with fully connected layers to predict the emotion label \hat{y}_i for the segment:

$$l_i = \text{ReLU}(\mathbf{W}_l e_i + b_l), \tag{9}$$

$$\mathcal{P}_i = \text{softmax}(\mathbf{W}_{smax}l_i + b_{smax}),\tag{10}$$

$$\mathcal{P}_{i} = \operatorname{softmax}(\mathbf{W}_{smax}l_{i} + b_{smax}), \tag{10}$$
$$\hat{y}_{i} = \arg \max(\mathcal{P}_{i}[\tau]) \tag{11}$$

in which l_i is the output of the hidden layer after applying the ReLU activation function, W_l and b_l are the weight matrix and bias vector for the input layer, respectively. \mathcal{P}_i denotes the probability distribution over the emotion classes, with W_{smax} and b_{smax} representing the weight matrix and bias vector for the output layer. Finally, \hat{y}_i is the predicted emotion label for the utterance segment. We use categorical cross-entropy along with L2-regularization as the loss function during training, following the work [15]:

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log P_{i,j}[y_{i,j}] + \lambda \|\theta\|_2$$
(12)

where N represents the number of dialogues, c(i) denotes the number of utterance segments within dialogue i. Additionally, $P_{i,j}$ refers to the probability distribution of class labels, while $y_{i,j}$ indicates the ground-truth label for segment j in the dialogue i. The parameter λ is used as the weight for L2-regularization, and θ represents the parameters that can be trained in the model.

Table 1: Accuracy and F1-score compared with baseline. * indicates significant improvement over AMERL (p < 0.05).

	Method					Method			
	AMERL		Ours			AMERL		Ours	
subject	ACC(%)	F1(%)	ACC(%)	F1(%)	subject	ACC(%)	F1(%)	ACC(%)	F1(%)
1	66.60	-	67.50	67.15	22	76.37	-	81.67	81.71
2	76.27	-	85.00	84.89	23	64.63	-	78.33	77.77
3	75.43	-	85.83	86.06	24	85.13	-	85.83	85.66
4	81.83	-	83.33	82.17	25	67.73	-	74.17	74.68
5	59.47	-	70.83	69.62	26	68.57	-	70.83	70.17
6	69.73	-	80.00	79.58	27	83.87	-	85.83	85.27
7	80.43	-	81.67	81.44	28	81.17	-	83.33	83.99
8	68.97	-	74.17	74.80	29	62.53	-	69.17	68.15
9	76.43	-	83.33	83.18	30	56.10	-	80.83	81.30
10	69.37	-	70.00	66.16	31	64.30	-	84.17	84.60
11	57.03	-	75.00	74.46	32	63.83	-	65.00	63.96
12	55.17	-	76.67	75.81	33	80.10	-	83.33	82.99
13	75.27	-	84.17	84.12	34	68.20	-	70.83	69.79
14	57.97	-	70.00	68.71	35	62.97	-	67.50	67.55
15	65.53	-	85.83	85.76	36	74.23	-	74.17	73.97
16	57.83	-	70.83	70.17	37	61.83	-	62.50	61.96
17	89.63	-	99.17	99.17	38	76.27	-	85.83	85.98
18	74.97	-	76.67	76.44	39	71.50	-	76.67	75.24
19	68.10	-	71.67	71.94	40	66.90	-	70.83	70.89
20	86.50	-	92.50	92.46	41	72.33	-	76.67	76.27
21	78.10	-	87.50	87.04	42	77.10	-	78.33	77.98
A	19 68.10 - 71.67 71.94 20 86.50 - 92.50 92.46				ts	70.86	-	76.65*	76.80

3 Experiments and Results

3.1 Dataset and Experimental Setting

The recently released multi-modal dialogue emotion dataset, EAV [8], 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. In addition, the proposed model is implemented using PyTorch and torch-geometric packages. The networks are trained with 1 NVIDIA GeForce RTX 3090. We use accuracy and F1-score as the metrics to measure the performance. Paired t-test is performed to test the significance of performance improvement with a default significance level of 0.05. Models are trained using Adam [17] with a batch size of 16.

3.2 Comparison with Baseline

To evaluate Hyper-MML, we compared our framework with the attention-based multi-modal emotion recognition framework (AMERL), which utilizes attention mechanisms to dynamically adjust the contributions of different modalities and is currently the only framework that applies multi-modal physiological signals to emotion recognition in conversations [13]. Specifically, AMERL uses an attention mechanism to dynamically adjust the attention weights of various modalities, prioritizing the key features of each modality and adapting to different input sizes. We evaluated the emotion classification accuracy and F1-score of 42 participants, and it is evident that our proposed Hyper-MML significantly outperformed the previous method on the EAV dataset, achieving new state-of-the-art (SOTA) records. The corresponding results are presented in Table 1.

3.3 Ablation Studies

To better demonstrate the rationale and effectiveness of the proposed model, we conducted an ablation study on the key components of Hyper-MML, with the results presented in Table 2. First, we validated the effectiveness of EEG modality compared to the text modality under conditions of incomplete semantic integrity in the utterance segments.

Table 2: Results of ablation experiments. * indicates significant improvement. E:EEG, T:Text, A:Audio, V:Video

	Modality	Graph Type	Accuracy(%)	F1-score(%)
1	T+A+V	Hypergraph	68.83	67.73
2	E+A+V	Graph	71.93	72.93
Hyper-MML	E+A+V	Hypergraph	76.65*	76.80*

We use speech recognition on the acoustic signals to obtain the transcribed text of the utterance segments. Subsequently, we extracted the features of the raw text using a pre-trained RoBERTa large language model [16]. Our experiments showed that the performance of EEG-based multi-modal recognition model outperformed that of the text-based multi-modal recognition model in the emotion recognition task involving incomplete utterance segments. Additionally, we verified the effectiveness of our hypergraph fusion module by replacing it with a standard graph convolution module, which facilitates complex interactions between multiple modalities through several pairwise relationships. Under this configuration, we observed a decrease in average accuracy by 4.72% on the EAV dataset, and the F1-score similarly dropped by 3.97%. This demonstrates the effectiveness of hypergraph modeling to capture higher-order relationships between modalities and contextual elements.

4 Conclusion

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. By integrating EEG signals with audio and video data, our framework effectively captures the intricate emotional dynamics inherent in conversational interactions. The MHFM significantly enhances the model's ability to process and integrate various modalities, leading to a more nuanced understanding of emotional states. Our experiments on the EAV dataset demonstrate that the proposed framework not only improves classification accuracy but also sets new benchmarks in emotion recognition. Future research should explore the generalizability of the Hyper-MML framework across diverse datasets and real-world applications, such as mental health monitoring and human-computer interaction.

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] 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.
- [4] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4652–4661, 2022.
- [5] 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.
- [6] 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.
- [7] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergen: A new method for training graph convolutional networks on hypergraphs. *Advances in Neural Information Processing Systems*, 32, 2019.
- [8] 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.
- [9] 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.

- [10] 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.
- [11] 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.
- [12] 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.
- [13] Kang Yin, Hye-Bin Shin, Dan Li, and Seong-Whan Lee. Eeg-based multimodal representation learning for emotion recognition. *arXiv preprint arXiv:2411.00822*, 2024.
- [14] 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.
- [15] 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.
- [16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [18] Yueyang Li, Weiming Zeng, Wenhao Dong, Di Han, Lei Chen, Hongyu Chen, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. A tale of single-channel electroencephalogram: Devices, datasets, signal processing, applications, and future directions. *arXiv* preprint arXiv:2407.14850, 2024.