Dynamic Interactive Bimodal Hypergraph Networks for Emotion Recognition in Conversations

Xuping Chen, Wuzhen Shi*

Shenzhen Key Laboratory of Digital Creative Technology
Guangdong Province Engineering Laboratory for Digital Creative Technology
College of Electronics and Information Engineering, Shenzhen University
2310433006@email.szu.edu.cn, wzhshi@szu.edu.cn

Abstract

The advancement in multimodal research has increased focus on Emotion Recognition in Conversations (ERC), targeting accurately identifying emotional changes. Methods based on graph convolution can better capture the dynamic changes of emotions and improve the accuracy and robustness of emotion recognition. However, existing methods do not distinguish the interaction patterns of a conversation, which results in limiting their ability to model contextual emotional relationships. In this paper, we propose a Dynamic Interactive Bimodal HyperGraph Convolutional Networks (DIB-HGCN), which creatively constructs two types of sub-hypergraphs, i.e., the monologic sub-hypergraph and the dialogic sub-hypergraph, for modeling emotion relationships of different interaction patterns. The monologic sub-hypergraph is used to explore the contextual consistent emotions during the speaker's monologue interactions, while the dialogic sub-hypergraph focuses on capturing the emotional transfers in the dialogic interactions. Meanwhile, the single window partitioning mechanism fails to accommodate the distinct emotional velocity variations across the two interaction patterns. Therefore, we set up dynamic windows in the monologic interactions to fully utilize the information of sentence nodes with consistent emotions, and we add fragment windows to the dialogic interactions to prevent information interference caused by frequent emotional transfers. The experimental results show that our proposed method outperforms existing methods on two benchmark multimodal ERC datasets.

Introduction

Emotional communication plays a pivotal role in human interactions, serving as both the fabric of interpersonal connections and a vital avenue for fostering deeper mutual understanding (Zhao et al. 2023). Therefore, the accurate identification of emotions within conversations has emerged as a topic of significant concern across various sectors of society (Gao et al. 2022). The development of multimodal technologies happens to provide an approach to overcome the limitations of ERC, thereby enabling a more comprehensive understanding of the speaker's emotional changes (Kalateh et al. 2024). It allows for a more comprehensive integration of emotional information from different modalities, which can significantly enhance the overall user experience of human-computer interaction and other related applications(Ghosal et al. 2019).

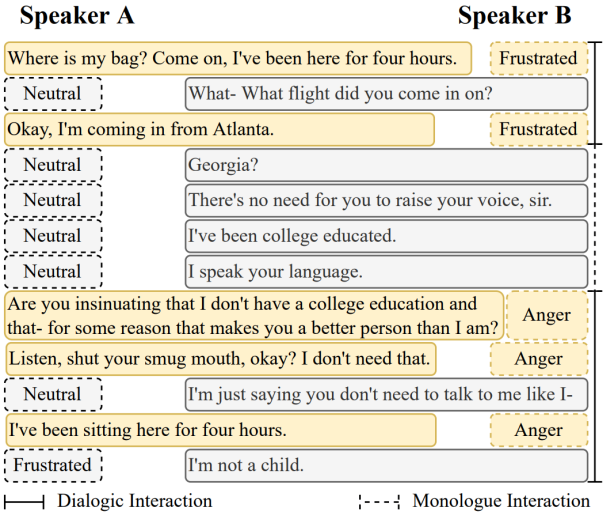

Figure 1: Examples of emotional consistency in monologue interaction patterns and emotional transfers in dialogic interaction patterns in conversations.

Compared to traditional emotion recognition of isolated utterance, ERC is characterized by its complexity and subtlety, requiring a consideration of the distinct roles of each participant. Consequently, recent efforts have sought to incorporate conversational participant information into the recognition process. For example, some researchers have integrated speaker information into conversational modeling by leveraging the time series modeling capabilities of RNNs (Hazarika et al. 2018a,b). Furthermore, conversations are more complex than mere time series data, which has directed researchers' attention to graph convolutional networks. Some researchers explore graph construction with speaker embeddings to capture speaker information effectively (Hu et al. 2022; Li et al. 2024). CauAIN (Zhao, Zhao, and Lu 2022), from the perspective of emotional cause detection, models the dependency relationships within and be-

*Corresponding author: Wuzhen Shi

tween speakers. The above methods have achieved good results, fully demonstrating the importance of conversational participant information. However, they haven't distinguished the interaction patterns, impairing their capability to capture the nuanced contextual emotional connections.

Fig.1 illustrates the distinction between dialogic and monologue interactions. Dialogic interaction involves communication among multiple parties with reciprocal responses. In contrast, monologue interaction is characterized by a continuous, one-sided expression of opinions without direct interaction or response from others. Fig.1 also depicts a scenario where a service staff member provides information with stable emotions after a passenger loses their backpack during the monologue interaction. However, the interaction pattern turns into the dialogic pattern when informing the passenger of the loss, leading to an emotional shift in the staff. The passenger's anger and verbal abuse in response to the news prompted the staff's emotions to transition from neutral to frustrated. Monologue interactions exhibit contextual emotional consistency, whereas, in dialogic interactions, the emotions are significantly influenced by the interplay among participants. To address this issue, we propose a bimodal hypergraph that adapts to various interaction patterns, comprising dialogic sub-hypergraphs and monologic sub-hypergraphs.

In conversations, emotional transfers are intricately interwoven with contextual cues. Consequently, ERC must holistically analyze the interplay of utterances to pinpoint the emotion of each sentence. Many previous studies have investigated how to model conversational information as graphs to capture intricate relationships. Some researchers construct graphs based on the global information of a conversation. MMGCN (Hu et al. 2021) takes different modal features of each utterance in conversations as nodes and connects nodes of the same mode and nodes of different modes in the same utterance. M^3NET (Chen et al. 2023a) constructs the hypergraph networks to analyze complex multivariate relationships in ERC. Some researchers construct graphs based on contextual information. GraphCFC and MA-CMU-SGRNet (Zhang et al. 2024) use a fixed contextual window to associate sentences with relevant contexts. DCGCN (Yang et al. 2024) dynamically adjusts the fragment window based on the proposed concept of discourse density. Research based on global modeling suggests that it is necessary to fully utilize the information of each node in a conversation, while research based on contextual modeling suggests that distant nodes can interfere with emotion recognition. Emotional transfer rates in conversations differ, making the above single-window partitioning mechanism unable to satisfy conversations with various interaction patterns. We suggest that global information is more pertinent to monologue interactions with gradual emotional shifts, while local contextual information is better suited to dialogic interactions characterized by rapid emotional changes. Therefore, in monologic interactions, we dynamically modulate the window size to optimize sentence information exploitation. Conversely, in dialogic interactions, we utilize a fragment window to minimize disruption from large differential emotions in distant utterances.

In summary, our contributions mainly include:

- We construct a bimodal hypergraph to adapt to the complexity of ERC. We capture the consistent emotions of the same speaker in the monologic interactions and the emotional transfers in the dialogic interactions by distinguishing the interaction patterns of the conversation.
- We set up dynamic windows for monologue interactions while setting up fragment windows for dialogic interactions. To better achieve deep integration of utterance context, we dynamically adjust the window size of the monologic sub-hypergraph based on continuous rounds to model its global information, while setting fragment windows for local modeling of dialogic sub-hypergraph to avoid interference.
- We conduct extensive experiments to validate the state-of-the-art performance of our method and perform ablation studies on our proposed modules to verify their effectiveness.

Related Work

Speaker-specific Modeling in Conversations

Conventional methods primarily rely on static speaker-specific modeling. ConGCN (Zhang et al. 2019) and CoG-BART (Tang et al. 2022) have performed this approach by incorporating speaker information into each utterance, thereby defining the connections between utterances. In conversations, emotions are intricately intertwined, with their influence being communicated through verbal cues, intonation, and facial gestures (Mariooryad and Busso 2013), which places a higher demand on speaker-specific modeling. Therefore, more recent methods, such as SEGD (Chen et al. 2023b), which utilizes speaker state encoders, are moving towards dynamic speaker-specific modeling to explore dependencies both within and between speakers.

Static speaker-specific modeling emphasizes conversation context but neglects dynamic speaker interactions, whereas dynamic modeling takes into account intra- and inter-speaker variation but struggles with rapid emotion changes, complicating accurate emotion identification in dialogic interaction. These limitations inspire us not to model specific speakers, but to directly model interaction patterns explicitly. Concerning these challenges, our work distinguishes monologic and dialogic interactions in conversations, focusing on emotional consistency in the former and capturing emotional transfers in the latter.

Context Modeling in Conversations

RNN-based Methods. LSTM (Hochreiter and Schmidhuber 1997) is commonly used for sequential data processing, suitable for modeling conversational sequences. bc-LSTM (Poria et al. 2017) and MFN (Zadeh et al. 2018) utilize different LSTMs for each modality. Furthermore, DialogueCRN (Hu, Wei, and Huai 2021) initiates a comprehensive understanding of conversational contexts from the perspective of emotional cognition, employing multiple LSTMs to integrate contextual relationships. Transitioning to GRUs,

which offer a streamlined alternative for capturing contextual dependencies, models like CMN (Hazarika et al. 2018b) and ICON (Hazarika et al. 2018a) utilize GRUs (Chung et al. 2014) for contextual modeling by simulating speakers' memories and interdependencies. DialogueRNN (Majumder et al. 2019) enhances sequential modeling with GRUs for global, speaker, and emotional state encoding.

Transformer-based Methods. The invention of the Transformer has sparked a new wave in artificial intelligence, including the ERC and other fields. KET (Zhong, Wang, and Miao 2019) utilizes a hierarchical self-attention mechanism, complemented by external commonsense knowledge, to dissect the emotional substrates of conversational data. DialogXL (Shen et al. 2021) introduces a dialog-aware self-attention framework, which concurrently maintains the integrity of the historical context. Ctnet (Lian, Liu, and Tao 2021) adopts the Transformer architecture to capitalize on its robust capacity for contextual modeling, thereby facilitating the acquisition of comprehensive global contextual information within conversational sequences.

GNN-based Methods. GNN effectively models the complex node information in ERC. GNN-based methods are classified into undirected, directed graph, and hypergraph approaches based on edge characteristics. Undirected graph-based methods typically construct graphs by globally establishing pairwise connections between nodes within the same modality. GraphCFC, on the other hand, leverages the directionality of edges to delineate nodes and contextual speaker relationships. DCGCN refines this approach by dynamically adjusting the contextual window according to its graph construction methodology. HGNN (Feng et al. 2019) represents one of the pioneering convolutional methods for hypergraphs, introducing a general hypergraph neural network framework and demonstrating its efficacy in modeling complex high-order data dependencies. M^{3}NET stands out as the first to introduce the hypergraph concept into ERC.

In addition to the above methods, RGCN (Chen et al. 2019) integrates residual design into graph convolution for complex contextual features, MA-CMU-SGRNet combines GNN and Transformer for local and global semantics, and PIRNet (Lian, Liu, and Tao 2022) uses a multi-stage approach to blend personal influence with context. RNN excels in sequence but lacks long-term memory. The transformer captures global dependencies but may miss sequence information. GNN with hypergraphs models complex relations effectively. GNN-based methods excel at relational modeling, where hypergraphs provide an effective way to capture complex multivariate higher-order relationships. Inspired by the above work, we extend the existing hypergraph approach to fine-grained conversational scenarios to better infer emotional changes during conversations.

Methodology

Fig.2 shows the overall structure of DIB-HGCN, which aims to accurately capture the contextual consistent emotion and the emotional transfers in conversations under different interaction patterns. After obtaining the multimodal data of a conversation, we encode each utterance into a representation, mapping each to its speaker. Then, based on the speaker's information, we determine the interaction pattern and divide it into multiple segments. Next, we propose a bimodal hypergraph to model utterance information separately in different interaction patterns through sub-hypergraphs. We construct dynamic windows and fragment windows to better adapt to different interaction patterns. Furthermore, we use hypergraph attention convolution to update node information layer by layer. Meanwhile, we also parallelly construct an undirected graph of node information and use undirected graph attention convolution to update node information. We concatenate these features and then obtain the final representation of emotions through the output layer. More details will be introduced in subsequent sections.

Problem Formulation

We denote a conversation by $U = [u_1, u_2, ..., u_N]$, where N denotes the number of utterances in a conversation. Correspondingly, the speaker of each utterance can be represented as $U_{sp} = [sp_1, sp_2, ..., sp_N](M \geq 2)$, where M represents the number of speakers. The unimodal input features for utterances in a conversation are denoted as $U^m = [u_1^m, u_2^m, ..., u_N^m], m \in \{a, v, t\}$, where a, v, t represent acoustic, visual, and textual modalities, respectively. Define the set of emotion labels $Y = [y_1, y_2, ..., y_k]$, where k denotes the number of emotional categories. The goal of the ERC task is to predict the emotion labels for each utterance $u_i, \forall i \in [1, N]$ based on the available contextual multimodal data from a predefined emotion set Y.

Unimodal Encoding

Prior studies (Yao and Shi 2024; Li et al. 2024; Chen et al. 2023a) demonstrate that RNN-based methods excel in textual feature extraction but lack enhanced performance in other modalities. Therefore, for textual modality features, we employ BiGRUs for encoding the initial representation of the i^{th} utterance x_i^t and an embedding layer to compute the speaker embedding, adding to obtain the unimodal textual representation v^t:

$$\begin{aligned} x_i^t &= \mathrm{BiGRU}(u_i^t, x_{i(+,-)1}^t) \\ S_{emb} &= \mathrm{Embedding}(\mathrm{S}, \mathrm{D}) \\ v^t &= S_{emb} + x^t \end{aligned} \tag{1}$$

where S denotes the speakers set and D is the speakers count.

For acoustic and visual modalities, we use a fully connected network to obtain the representations $\{v_i^a, v_i^t\}$:

$$\begin{aligned} v_i^a &= w_1 u_i^a + b_1 \\ v_i^v &= w_2 u_i^v + b_2 \end{aligned} \tag{2}$$

where w_1, w_2, b_1 and b_2 are trainable parameters of the fully connected layers.

Bimodal Hypergraph-based Feature Extraction

Bimodal Hypergraph Construction

In ERC, the multiple attributes in conversations belong to high-order information, such as multiple speaker and multimodal information (Chen et al. 2023a). We use bimodal

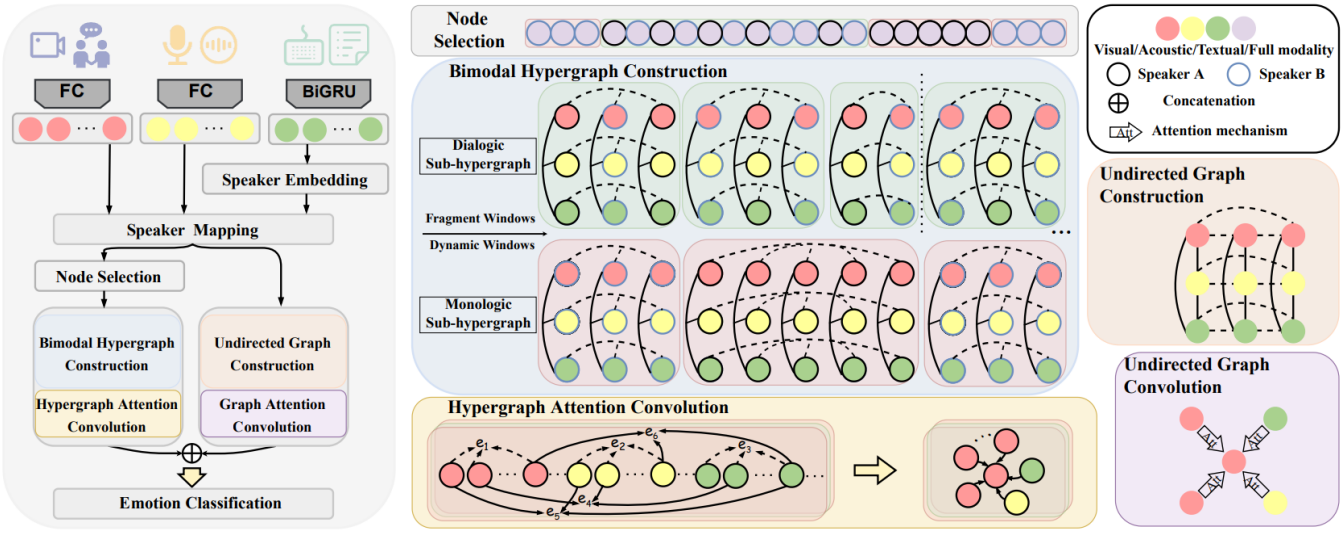

Figure 2: Overall Framework of DIB-HGCN.

hypergraphs to capture high-order information while distinguishing the interaction patterns of conversations.

Bimodal Hypergraph Partitioning: We construct a hypergraph $G = (V, E)$ to represent each conversation where each node $\forall v \in \mathcal{V}_{\mathcal{H}}, (|\mathcal{V}_{\mathcal{H}}| = 3N)$ corresponds to a unimodal utterance, denoted with $\{v_i^t, v_i^a, v_i^v\}$. Every hyperedge $\forall e \in \mathcal{E}_{\mathcal{H}}$ encodes multimodal or contextual dependencies. Each conversation is further divided into dialogic sub-hypergraphs $G_{\mathbf{d}}$ and monologic sub-hypergraphs $G_{\mathbf{m}}$:

$$\begin{aligned} G_{\mathbf{d}} &= \cup_0^p G_{\mathbf{d}}^p \\ G_{\mathbf{m}} &= \cup_0^q G_{\mathbf{m}}^q \\ G &= G_{\mathbf{d}} \cup G_{\mathbf{m}} \end{aligned} \tag{3}$$

Node Selection: Each modality of an utterance is represented as a node in a hypergraph, i.e.,v_i^m represents the nodes of a certain modality for the i^{th} utterance. Meanwhile, we define a function S that maps the index of nodes to the corresponding speakers, facilitating the node selection.

In monologue interactions, the conversation consists of one person continuously outputting ideas, so we consider that the nodes in each monologic sub-hypergraph need to satisfy the same speaker sp for 3 consecutive utterances, $sp \in [2, M]$. Mathematically defined as:

$$\begin{aligned} &\forall v_i, v_{i+1}, ...v_{i+d_q} \in G_{\mathbf{m}}^q, \\ \text{s.t.} \quad & S(v_n) == S(v_{n(+,-)1}) == sp \\ \text{or} \quad & S(v_n) == S(v_{n+1}) == S(v_{n+2}) == sp \\ \text{or} \quad & S(v_n) == S(v_{n-1}) == S(v_{n-2}) == sp \end{aligned} \tag{4}$$

where $n \in [i, i + d_q]$, $G_{\mathbf{m}}^q$ denotes the q^{th} monologic sub-hypergraph, d_q denotes its context distance.

During dialogic interactions, both parties participate in dialogue together to promote emotional transfers. We hope that at least one of the nodes in the time step before and after the current node is different from its speaker's information. Note that for start or end nodes, only check if adjacent nodes have different speakers. If both patterns apply, classify it as a monologue sub-hypergraph first. The nodes in the dialogic sub-hypergraph need to meet the following requirements:

$$\begin{aligned} &\forall v_i, v_{i+1}, ...v_{i+d_p} \in G_{\mathbf{d}}^p, \\ \text{s.t.} \quad & S(v_n) \neq S(v_{n+1}) \\ \text{or} \quad & S(v_n) \neq S(v_{n-1}) \end{aligned} \tag{5}$$

where $G_{\mathbf{d}}^p$ denotes the p^{th} dialogic sub-hypergraph, n denotes the n^{th} utterance node in sub-hypergraph, $n \in (i, i + d_p)$, d_p denotes its context distance.

Window Settings: We separately set up windows for the sub-hypergraphs. In monologue interactions, we form dynamic window sets $D' = \{d_1, d_2, \ldots, d_q\}$ of monologic sub-hypergraphs, using the contextual distance dynamics as a window to be used for hyperedge connections.

Since emotions transfer frequently during dialogic interactions, we apply fragment windows of j size to capture localized interaction emotions when d exceeds the segmentation threshold $\mathrm{S_{eg}}$ to avoid introducing noise. Note that when d is not divisible by j, the remaining nodes automatically form a new subgraph.

Hyperedge: Each node in the window $v_l^m (m \in \{a, v, t\}), l \in [i, i + d]$ is firstly connected to all other utterances in the same modality in the same window $\{v_h^m | (h \in [i, i + d], h \neq l)\}$ with one contextual hyperedge. Moreover, each node v_l^m is connected to other modalities of the same utterances $\{v_l^z | (z \in \{a, v, t\}, z \neq m)\}$, with one multimodal hyperedge. The nodes in each window are connected through the hyperedges we defined above and thus obtain a node incidence matrix of hypergraph $\mathbf{H}$, in which $\mathbf{H}_{ve} = 1$ indicates that the hyperedge e is incident with the node v; otherwise $\mathbf{H}_{ve} = 0$. Note that $\mathbf{H}_{ve}$ must be 0 for nodes and hyperedges that are in different windows.

Hypergraph Attention Convolution

Referring to previous work (Chen et al. 2023a), we measure the contribution of each node to the hyperedges through

training parameters $\mathcal{P}_e(v)$ and update the correlation matrix based on the weights to avoid network complexity:

$$\mathcal{H}_{v,e} = \begin{cases} \mathcal{P}_e(v), & \text{if hyperedge } e \text{ is incident with node } v; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Similarly, we obtain the weights of each hyperedge through training parameters $\mathcal{W}(e)$ to form the hyperedge weight matrix. Mathematically:

$$\mathcal{W} = \begin{bmatrix} \mathcal{W}(e_1) & 0 & 0 & \cdots & 0 \\ 0 & \mathcal{W}(e_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathcal{W}(e_n) \end{bmatrix} \tag{7}$$

Finally, we can use hypergraph attention convolution to update the node features layer by layer:

$$\begin{aligned} \mathbf{V}_h^{(1)} &= \sigma(\mathbf{D}_{\mathcal{H}}^{-1}\mathbf{H}\mathcal{W}\mathbf{B}^{-1}\mathcal{H}^{\top}\mathbf{V}_h^{(0)}) \\ &... \\ \mathbf{V}_h^{(L)} &= \sigma(\mathbf{D}_{\mathcal{H}}^{-1}\mathbf{H}\mathcal{W}\mathbf{B}^{-1}\mathcal{H}^{\top}\mathbf{V}_h^{(L-1)}) \end{aligned} \tag{8}$$

where $\mathbf{V}_h^{(L)} = \{v_{i,(L)}^m | i \in [1, N], m \in \{t, a, v\}\}$ is the output at layer L. σ is a non-linear activation function. $\mathbf{D}_{\mathcal{H}} \in \mathbb{R}^{|\mathcal{V}_{\mathcal{H}}| \times |\mathcal{V}_{\mathcal{H}}|}$ is the hyperedge degree matrix and $\mathbf{B} \in \mathbb{R}^{|\mathcal{E}_{\mathcal{H}}| \times |\mathcal{E}_{\mathcal{H}}|}$ is the node degree matrix. By this means, the high-order relationships are gradually refined.

Undirected Graph-based Feature Extraction

Undirected Graph Construction

In the previous section, we obtain local information associations using multiple sub-hypergraphs. In this section, we refer to (Hu et al. 2021) to construct an undirected graph $G_{un} = (V_{un}, E_{un})$ adaptively capturing global information using an attention mechanism, where the node features used are consistent with the hypergraph, denoted as $\{v_i^t, v_i^a, v_i^v\}$. Unlike hypergraphs, utterance nodes in an undirected graph are connected in pairs. Specifically, each node v_i^m is connected to other utterance nodes of the same modality $\{v_j^m | j \in [1, N], j \neq i\}$ and different modalities of the same utterance $\{v_i^z | z \in \{a, v, t\}, z \neq m\}$.

Undirected Graph Attention Convolution

We refer to previous work (Chen et al. 2023a) to extract high- and low-frequency features using designed filters for attention weights. Mathematically:

$$\alpha_{(i,j)} = \frac{\tanh(w_3(v_{i,(k)} \oplus v_{j,(k)}))}{\sqrt{|N_j|}\sqrt{|N_i|}} \tag{9}$$

where N_j is the neighboring nodes of node i, $\oplus$ is the concatenation operation and $w_3 \in \mathbb{R}^{2D_h \times 1}$ is a trainable weight matrix. $\tanh(\cdot)$ is the hyperbolic tangent function. By this means, the coefficient $\alpha_{i,j}$ can readily model the varying importance of different frequency constituents.

The information of each node is updated using the information of neighboring nodes through K iterations:

$$\begin{aligned} v_{i,(1)}^m &= v_{i,(1)}^m + \sum_{j \in \mathcal{N}_i} \alpha_{(i,j)} v_{j,(0)}^m \\ &... \\ v_{i,(K)}^m &= v_{i,(K-1)}^m + \sum_{j \in \mathcal{N}_i} \alpha_{(i,j)} v_{j,(K-1)}^m \end{aligned} \tag{10}$$

Finally, we obtain the representation of each modality $\{v_{i,(K)}^t, v_{i,(K)}^a, v_{i,(K)}^v\}$ and concatenate them to obtain the final feature representation of the undirected graph:

$$\mathbf{V}_{un}^{(K)} = v_{i,(K)}^t \oplus v_{i,(K)}^a \oplus v_{i,(K)}^v \tag{11}$$

Emotion Classification and Loss Function

The emotion classifier combines the final feature representations of hypergraphs and undirected graphs as input for emotion classification.:

$$\begin{aligned} \hat{E} &= \text{Relu}(\mathbf{V}_h^{(L)} \oplus \mathbf{V}_{un}^{(K)}) \\ \mathcal{P} &= \text{softmax}(w_4\hat{E} + b_4) \\ \hat{y} &= \underset{\tau}{\text{argmax}}(\mathcal{P}[\tau]) \end{aligned} \tag{12}$$

where w_4 is trainable weight, $\mathcal{P} \in \mathbb{R}^C$ and $\hat{y}$ is the predicted label. We use categorical cross-entropy along with $L2$-regularization to calculate the loss of the network:

$$L = -\frac{1}{\sum_{r=1}^{R} N_r} \sum_{i=1}^{R} \sum_{j=1}^{N_i} log(\mathcal{P}_{i,j}[y_{i,j}]) + \eta \left\| \Theta \right\|_2 \tag{13}$$

where R is the number of conversation samples in the training set and N_r is the number of utterances in conversation sample r. $\mathcal{P}_{i,j}$ and $y_{i,j}$ denote the probabilistic distribution of class labels and the ground-truth label for utterance class j in conversation i, respectively. Θ refers to all trainable parameters. η is the L_2-regularization weight.

Experiments

4.1 Dataset

We chose IEMOCAP and MELD for our experiments. They have comprehensive multimodal data as well as a recognized benchmark status in ERC. The details are as follows:

IEMOCAP (Busso et al. 2008) comprises 151 videos featuring two-person conversations, including 7433 utterances. It is annotated with six distinct emotion categories: happiness, sadness, neutrality, anger, excitement, and frustration.

MELD (Poria et al. 2019) consists of video recordings from multi-person conversations extracted from the "Friends" television series, involving between three to nine participants per conversation. It encompasses 1433 conversations, 13708 utterances, and 304 unique speakers. The dataset is labeled with seven emotions: happiness, sadness, neutrality, anger, disgust, fear, and surprise.

Methods	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
A-DMN	64.6	64.3	-	60.45
COGMEN	68.2	67.6	-	-
CORECT	69.93	70.02	-	-
MM-DFN$^{\triangledown}$	68.95	68.39	62.49	59.46
SCMFN$^{\triangledown}$	71.23	71.21	67.01	66.25
M^{3}NET$^{\triangledown}$	70.73	70.66	67.51	65.97
DCGCN	68.31	-	66.25	-
MA-CMU-SGR	72.4	71.6	-	62.3
DIB-HGCN(Ours)	**72.58**	**72.46**	**68.01**	**66.61**

Table 1: Overall performance of all models on both IEMOCAP and MELD. $\triangledown$ from our reimplementation using open-source codes. Bold font denotes the best performances. Both Acc. and F1 are weighted averages.

Baselines

In addition to DIB-HGCN, we also utilize other models for ERC. We replicate the baseline method MM-DFN, SCMFN and M^{3}NET using the same parameters and software as authors, and the M^{3}NET results showed deviations, which may be caused by hardware differences and random variations.

A-DMN (Xing, Mai, and Hu 2020) uses RNNs and episodic memory for self and inter-speaker modeling. **COGMEN** (Joshi et al. 2022) uses a contextualized GNN to model speaker interactions and contextual information. **CORECT** (Nguyen et al. 2023) models conversation by leveraging relational temporal dynamics and cross-modality interactions. **MM-DFN** fuses multimodal features to reduce redundancy. **SCMFN** (Yao and Shi 2024) builds a speaker-centric graph and performs UDA-based cross-modal fusion. **M^{3}Net** captures multivariate information with hypergraphs in ERC. **DCGCN** integrates context dynamically to enhance emotion recognition. **MA-CMU-SGRNet** refines semantic graphs for emotional understanding.

Implementation Details

Experiments are conducted on a machine with NVIDIA GTX 4090 GPU and implemented by CUDA 12.1, Python 3.8, PyTorch 1.7.1, and torch-geometric 1.7.2. We adopt the previous approach, utilizing RoBERTa (Liu et al. 2019) for textual, OpenSmile (Schuller et al. 2011) for acoustic, and DenseNet (Huang et al. 2017) for visual feature extraction, respectively. After normalization, we reduce the dimensionality of these unimodal features to a final size of 200, ensuring consistency across modalities.

Due to differing numbers of emotion categories and speakers between the two datasets, separate parameters and models are trained for each. For IEMOCAP, we set the batch size to 16 and the epoch to 80. For MELD, we adjust the batch size to 32 and epoch to 30. To achieve reproducibility performance, we iterate through a range of random seeds, identifying 1722 for IEMOCAP and 67137 for MELD. Other parameter settings are detailed in section 5.

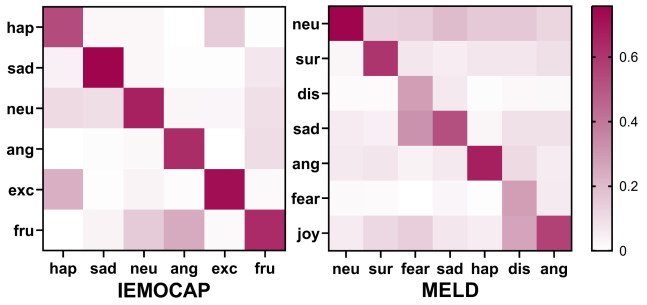

Figure 3: The confusion matrices of DIB-HGCN.

Evaluation Metrics

We evaluate the performance using weighted average accuracy and F1-score as the primary metric, consistent with prior studies (Hu et al. 2022; Zhang et al. 2024). The higher the evaluation index, the better the model performance.

Results and Analysis

Overall Performance

We compare our proposed DIB-HGCN with existing models. The overall performance of all models is presented in Tab.1. It can be seen that ours achieves SOTA results for all the metrics in all the datasets. It underlines the necessity of distinguishing interaction patterns in conversations.

Additionally, Fig.3 visualizes the performance of each emotion category across benchmark datasets via confusion matrices. DIB-HGCN enables the adaptive selection of contextual emotional information. This selection process is based on different emotional change rate of interaction patterns. Therefore, it bolsters the performance of emotion recognition. These results validate the effectiveness of our approach. The DIB-HGCN demonstrates robust classification accuracy for the majority of emotions in IEMOCAP. For MELD, the model struggles with fear and disgust, likely attributable to their infrequent representation in the dataset.

Ablation Study

We conduct ablation studies to evaluate the effectiveness of our proposed modules. In the first variant, we exclude the bimodal hypergraph. Instead, we directly construct an undirected graph with attention convolution. In the second variant, we solely construct a bimodal hypergraph. Moreover, as the third variant, we remove the sub-hypergraph construction, using only the hypergraph and undirected graph in parallel. We compare these variants as shown in Tab.2.

Methods	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
w/o Hypergraph branch	71.04	70.97	67.70	66.10
w/o Undirected graph branch	67.41	67.22	67.13	65.85
w/o Sub-hypergraphs	70.86	70.97	67.36	66.27

Table 2: Ablation studies of DIB-HGCN.

G_{d}	G_{m}	IEMOCAP		MELD	
		Acc.	F1	Acc.	F1
D	D	69.66	69.44	67.51	65.72
F	F	71.10	70.79	67.36	65.94
D	F	69.62	69.36	66.99	64.29
F	D	**72.58**	**72.46**	**68.01**	**66.61**

Table 3: Window settings in sub-hypergraphs. D denotes Dynamic windows and F denotes Fragment windows.

Effects of Hypergraph Branch: The removal of the hypergraph led to a decrease in performance metrics, with the IEMOCAP accuracy dropping by 1.54% and its F1-score decreasing by 2.49%, while the MELD accuracy and F1 score saw smaller reductions of 0.31% and 0.51%, respectively. This suggests that hypergraph enhances relationship capture.

Effects of Undirected Graph Branch: The removal of the undirected graph branch results in a marked degradation of performance, highlighting the critical importance of incorporating global information.

Effects of Window Settings: We investigate the rationality of setting windows in sub-hypergraphs. We set dynamic windows and fragment windows for two sub-hypergraphs respectively. According to Tab.3, using dynamic windows in dialogic sub-hypergraphs can lead to a significant decrease in performance, while using fragment windows in monologic sub-hypergraphs can also impair performance to some extent. These results validate the necessity to set the window base on the rate of emotional change.

Visualisation of Feature Distribution

We visualize the feature distributions of M^{3}NET and DIB-HGCN using u-map. Different colors represent different emotions. Fig.4 illustrates a notable degree of dispersion and substantial overlap among the feature distributions for different emotions as extracted by M^{3}NET. Conversely, the feature distributions yielded by ours exhibit a higher degree of differentiation and concentration. In the case of neutral and sad emotion, a more pronounced degree of discriminability is observed. These results corroborates the efficacy of DIB-HGCN in aggregating the features corresponding to each emotional category within a shared feature space.

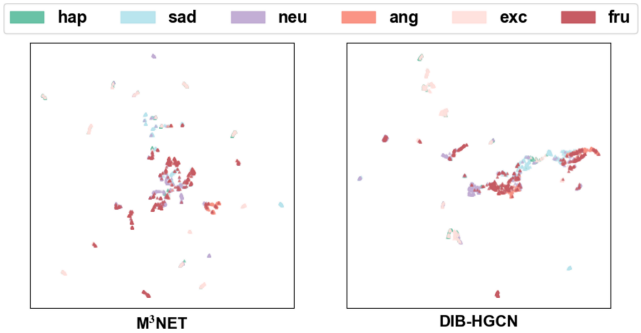

Figure 4: The visualization of the u-map representations.

Discussions on Parameter Settings

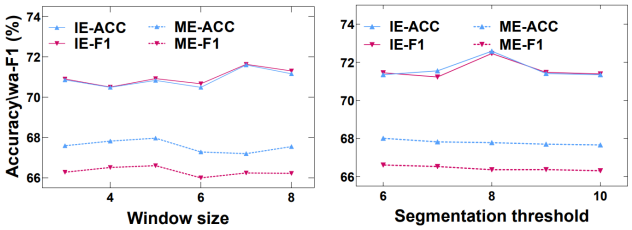

Figure 5: The effects of window size j and segmentation threshold $\mathrm{S_{eg}}$.

Effects of Window Size j and Segmentation Threshold $\mathrm{S_{eg}}$ We set parameters to constrain the fragment window. First, we only establish the window partitioning mechanism. To ensure sufficient context, we set the initial value to 3, while to prevent excessive information, the maximum is set to 8. On IEMOCAP, the j is set to 7, while on MELD, it is set to 5. Furthermore, We observe a decrease in evaluation metrics at a window size of 6 in Fig.5, prompting us to set the S_{eg} starting from 6. We gradually increase S_{eg} and find that IEMOCAP reaches its optimal result at 8, while MELD shows a decreasing trend after an increase at 6.

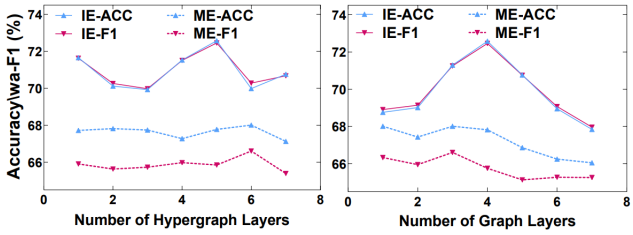

Figure 6: The effects of graph layers.

Effects of Graph Layers We use parallel layers of hypergraphs and undirected graphs to capture emotional information. By adjusting layer counts, Fig.6 shows that more layers can enhance performance up to a point, but beyond that, overfitting diminishes returns, suggesting that extra layers are redundant. IEMOCAP performance peaks at 5 hypergraphs and 4 undirected layers, with further layers adding complexity without advantage. MELD's performance is largely unaffected by layer stacking.

Conclusion

We construct an ERC model based on hypergraphs. DIB-HGCN is proposed to capture high-order relationships under different interaction patterns and dynamically fuse contextual information based on the rate of emotional change. Experimental results on benchmark ERC datasets demonstrate the effectiveness of the proposed method. Our current work is realized based on complete data. However, in real-world dialogue scenarios, there is a presence of missing modalities, either singular or multiple. For practical application, our subsequent efforts will be oriented toward operating on incomplete data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62101346, and in part by the Shenzhen Science and Technology Program under Grant JCYJ20240813141358076.

References

Busso, C.; Bulut, M.; Lee, C.; Kazemzadeh, E.; Provost, E.; Kim, S.; Chang, J.; Lee, S.; and Narayanan, S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.

Chen, F.; Shao, J.; Zhu, S.; and Shen, H. 2023a. Multivariate, Multi-Frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10761–10770.

Chen, J.; Hou, H.; Gao, J.; Ji, Y.; and Bai, T. 2019. RGCN: recurrent graph convolutional networks for target-dependent sentiment analysis. In *International Conference on Knowledge Science, Engineering and Management*, 667–675. Springer.

Chen, J.; Huang, P.; Huang, G.; Li, Q.; and Xu, Y. 2023b. Sdtn: Speaker dynamics tracking network for emotion recognition in conversation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Gao, Q.; Cao, B.; Guan, X.; Gu, T.; Bao, X.; Wu, J.; Liu, B.; and Cao, J. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Know.-Based Syst.*, 248(C).

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164. Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2594–2604. Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018b. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022. 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)*, 7037–7041.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7042–7052. Association for Computational Linguistics.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5666–5675. Association for Computational Linguistics.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Joshi, A.; Bhat, A.; Jain, A.; Singh, A. V.; and Modi, A. 2022. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN.

Kalateh, S.; Estrada-Jimenez, L. A.; Nikghadam-Hojjati, S.; and Barata, J. 2024. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access*, 12: 103976–104019.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2024. GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition. *IEEE Transactions on Multimedia*, 26: 77–89.

Lian, Z.; Liu, B.; and Tao, J. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 985–1000.

Lian, Z.; Liu, B.; and Tao, J. 2022. Pirnet: Personality-enhanced iterative refinement network for emotion recognition in conversation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2): 2863–2874.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. DialogueRNN: An

Attentive RNN for Emotion Detection in Conversations. AAAI'19/IAAI'19/EAAI'19. AAAI Press. ISBN 978-1-57735-809-1.

Mariooryad, S.; and Busso, C. 2013. Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on affective computing*, 4(2): 183–196.

Nguyen, C. V. T.; Mai, T.; The, S.; Kieu, D.; and Le, D.-T. 2023. Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15154–15167. Singapore: Association for Computational Linguistics.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–883. Association for Computational Linguistics.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9): 1062–1087. Sensing Emotion and Affect - Facing Realism in Speech Processing.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13789–13797.

Tang, S.; Wang, C.; Xu, K.; Huang, Z.; Xu, M.; and Peng, Y. 2022. An Emotion Evolution Network for Emotion Recognition in Conversation. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, 1231–1238. IEEE.

Xing, S.; Mai, S.; and Hu, H. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3): 1426–1439.

Yang, Z.; Li, X.; Cheng, Y.; Zhang, T.; and Wang, X. 2024. Emotion Recognition in Conversation Based on a Dynamic Complementary Graph Convolutional Network. *IEEE Transactions on Affective Computing*.

Yao, B.; and Shi, W. 2024. Speaker-Centric Multimodal Fusion Networks for Emotion Recognition in Conversations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8441–8445. IEEE.

Zadeh, A.; Liang, P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L. 2018. Memory Fusion Network for Multi-View Sequential Learning. AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *IJCAI*, 5415–5421. Macao.

Zhang, X.; Cui, W.; Hu, B.; and Li, Y. 2024. A Multi-Level Alignment and Cross-Modal Unified Semantic Graph Refinement Network for Conversational Emotion Recognition. *IEEE Transactions on Affective Computing*.

Zhao, S.; Hong, X.; Yang, J.; Zhao, Y.; and Ding, G. 2023. Toward Label-Efficient Emotion and Sentiment Analysis. *Proc. IEEE*, 111(10): 1159–1197.

Zhao, W.; Zhao, Y.; and Lu, X. 2022. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *IJCAI*, 4524–4530.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 165–176.