ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN THỊ CẨM VÂN

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

CÁC MÔ HÌNH HỌC SÂU KẾT HỢP ĐA PHƯƠNG THỨC TIÊN TIẾN NHẬN DIỆN CẢM XÚC TRONG HỘI THOẠI VỚI THÔNG TIN KHÔNG ĐẦY ĐỦ VÀ MẤT CÂN BẰNG

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội – 2026

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

1. **Cam Van Thi Nguyen**, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. "Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction". *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore. Association for Computational Linguistics https://doi.org/10.18653/v1/2023.emnlp-main.937 (**Scopus**, **CORE Rank A* Conference**)
2. **Cam-Van Thi Nguyen**, Cao-Bach Nguyen, Duc-Trong Le, and Quang-Thuy Ha. 2024. "Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition". *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024*), pages 4259–4265, Torino, Italia. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.380/ (**Scopus**, **CORE Rank B Conference**).
3. **Cam-Van Thi Nguyen**, Hai-Dang Kieu, Quang-Thuy Ha, Xuan-Hieu Phan, Duc-Trong Le. "Mi-CGA: Cross-modal Graph Attention Network for robust emotion recognition in the presence of incomplete modalities", *Neurocomputing*, Volume 623, 2025, 129342, ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2025.129342 (**SCIE Q1 Journal, Impact Factor: 6.5**)
4. **Cam-Van Thi Nguyen**, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. 2024. "Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition". *In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24).* Association for Computing Machinery, New York, NY, USA, 9330–9339. https://doi.org/10.1145/3664647.3681648 (**Scopus**, **CORE Rank A* Conference**)
5. Phuong-Anh Nguyen, The-Son Le, Duc-Trong Le, **Cam-Van Thi Nguyen***. 2026. "Leveraging Self-Paced Curriculum Learning for Enhanced Modality Balance in Multimodal Conversational Emotion Recognition." *Neural Computing and Applications* https://doi.org/10.1007/s00521-026-12160-6 (**Scopus Q1 Journal, in press**)

Danh mục này gồm 05 công trình

Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction

Cam-Van Thi Nguyen[1] , **Anh-Tuan Mai**[1,2], **The-Son Le**[1]
Hai-Dang Kieu[1], **Duc-Trong Le**[1]
[1]VNU University of Engineering and Technology, Hanoi, Vietnam
[2]FPT Software AI Center
{vanntc, 20020269, 21020089, dangkh_uet, trongld}@vnu.edu.vn

Abstract

Emotion recognition is a crucial task for human conversation understanding. It becomes more challenging with the notion of multimodal data, e.g., language, voice, and facial expressions. As a typical solution, the global- and the local context information are exploited to predict the emotional label for every single sentence, i.e., utterance, in the dialogue. Specifically, the global representation could be captured via modeling of cross-modal interactions at the conversation level. The local one is often inferred using the temporal information of speakers or emotional shifts, which neglects vital factors at the utterance level. Additionally, most existing approaches take fused features of multiple modalities in an unified input without leveraging modality-specific representations. Motivating from these problems, we propose the Relational Temporal Graph Neural Network with Auxiliary Cross-Modality Interaction (CORECT), an novel neural network framework that effectively captures conversation-level cross-modality interactions and utterance-level temporal dependencies with the modality-specific manner for conversation understanding. Extensive experiments demonstrate the effectiveness of CORECT via its state-of-the-art results on the IEMOCAP and CMU-MOSEI datasets for the multimodal ERC task.

1 Introduction

Our social interactions and relationships are all influenced by emotions. Given the transcript of a conversation and speaker information for each constituent utterance, the task of Emotion Recognition in Conversations (ERC) aims to identify the emotion expressed in each utterance from a predefined set of emotions (Poria et al., 2019). The multimodal nature of human communication, which involves verbal/textual, facial expressions, vocal/acoustic, bodily/postural, and symbolic/pictorial expressions, adds complexity to the task of Emotion Recognition in Conversations (ERC) (Wang et al., 2022).

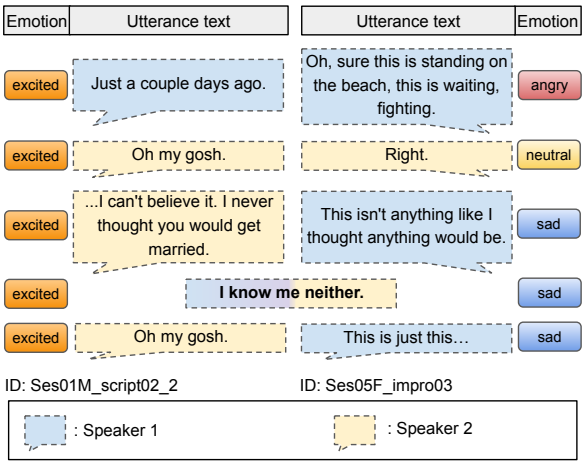


Figure 1: Examples of temporal effects on conversations

Multimodal ERC, which aims to automatically detect the a speaker's emotional state during a conversation using information from text content, facial expressions, and audio signals, has garnered significant attention and research in recent years and has been applied to many real-world scenarios (Sharma and Dhall, 2021; Joshi et al., 2022).

Massive methods have been developed to model conversation's context. These approaches can be categorized into two main groups: graph-based methods (Ghosal et al., 2019; Zhang et al., 2019; Shen et al., 2021b) and recurrence-based methods (Hazarika et al., 2018a; Ghosal et al., 2020; Majumder et al., 2019; Hu et al., 2021). In addition, there have been advancements in multimodal models that leverage the dependencies and complementarities of multiple modalities to improve the ERC performance (Poria et al., 2017; Hazarika et al., 2018b; Zadeh et al., 2018). One limitation of these methods is their heavy reliance on nearby utterances when updating the state of the query utterance, which can restrict their overall performance. Recently, Graph Neural Network (GNN)-based methods have been proposed for the multimodal ERC task due to their ability to capture

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15154–15167
December 6-10, 2023

long-distance contextual information through their relational modeling capabilities. However, those models rely on fused inputs being treated as a single node in the graph (Ghosal et al., 2019; Joshi et al., 2022), which limits their ability to capture modality-specific representations and ultimately hampers their overall performance.

The temporal aspect of conversations is crucial, as past and future utterances can significantly influence the query utterance as Figure 1. The sentence "*I know me neither*" appears with opposing labels on different dialogues, which could be caused by sequential effects from previous or future steps. There are only a few methods that take into account the temporal aspect of conversations. MMGCN (Wei et al., 2019) represents modality-specific features as graph nodes but overlooks the temporal factor. DAG-ERC (Shen et al., 2021b) incorporates temporal information, but focuses solely on text modality. Recently, COGMEN (Joshi et al., 2022) proposes to learn contextual, inter-speaker, and intra-speaker relations, but neglects modality-specific features and partially utilizes cross-modal information by fusing all modalities' representations at the input stage.

The aforementioned limitations motivate us to propose a **CO**nversation understanding model using **RE**lational **T**emporal Graph Neural Network with Auxiliary **C**ross-Modality Interaction (**CORECT**). It comprises two key components: the *(i) Relational Temporal Graph Convolutional Network (RT-GCN)*; and the *(ii) Pairwise Cross-modal Feature Interaction (P-CM)*. The *RT-GCN* module is based on RGCNs (Schlichtkrull et al., 2018) and GraphTransformer (Yun et al., 2019) while the *P-CM* is built upon (Tsai et al., 2019). Overall, our main contributions are as follows:

- We propose the CORECT framework for Multimodal ERC, which concurrently exploit the utterance-level local context feature from multimodal interactions with temporal dependencies via RT-GCN, and the cross-modal global context feature at the conversation level by P-CM. These features are aggregated to enhance the performance of the utterance-level emotional recognition.
- We conduct extensive experiments to show that CORECT consistently outperforms the previous SOTA baselines on the two publicly real-life datasets, including IEMOCAP and CMU-MOSEI, for the multimodal ERC task.
- We conduct ablation studies to investigate the effect of various components and modalities on CORECT for conversation understanding.

2 Related Works

This section presents a literature review on Multimodal Emotion Recognition (ERC) and the application of Graph Neural Networks for ERC.

2.1 Multimodal Emotion Recognition in Conversation

The complexity of conversations, with multiple speakers, dynamic interactions, and contextual dependencies, presents challenges for the ERC task. There are efforts to model the conversation context in ERC, with a primary focus on the textual modality. Several notable approaches include CMN (Hazarika et al., 2018b), DialogueGCN (Ghosal et al., 2019), COSMIC (Ghosal et al., 2020), DialogueXL (Shen et al., 2021a), DialogueCRN (Hu et al., 2021), DAG-ERC (Shen et al., 2021b).

Multimodal machine learning has gained popularity due to its ability to address the limitations of unimodal approaches in capturing complex real-world phenomena (Baltrušaitis et al., 2018). It is recognized that human perception and understanding are influenced by the integration of multiple sensory inputs. There have been several notable approaches that aim to harness the power of multiple modalities in various applications (Poria et al., 2017; Zadeh et al., 2018; Majumder et al., 2019), etc. CMN (Hazarika et al., 2018b) combines features from different modalities by concatenating them directly and utilizes the Gated Recurrent Unit (GRU) to model contextual information. ICON (Hazarika et al., 2018a) extracts multimodal conversation features and employs global memories to model emotional influences hierarchically, resulting in improved performance for utterance-video emotion recognition. ConGCN (Zhang et al., 2019) models utterances and speakers as nodes in a graph, capturing context dependencies and speaker dependencies as edges. However, ConGCN focuses only on textual and acoustic features and does not consider other modalities. MMGCN (Wei et al., 2019), on the other hand, is a graph convolutional network (GCN)-based model that effectively captures both long-distance contextual information and multimodal interactive information.

More recently, Lian et al. Lian et al. (2022) propose a novel framework that combines semi-

supervised learning with multimodal interactions. However, it currently addresses only two modalities, i.e., text and audio, with visual information reserved for future work. Shi and Huang (2023) introduces MultiEMO, an attention-based multimodal fusion framework that effectively integrates information from textual, audio and visual modalities. However, neither of these models addresses the temporal aspect in conversations.

2.2 Graph Neural Networks

In the past few years, there has been a growing interest in representing non-Euclidean data as graphs. However, the complexity of graph data has presented challenges for traditional neural network models. From initial research on graph neural networks (GNNs)(Gori et al., 2005; Scarselli et al., 2008), generalizing the operations of deep neural networks were paid attention, such as convolution (Kipf and Welling, 2017), recurrence (Nicolicioiu et al., 2019), and attention (Velickovic et al., 2018), to graph structures. When faced with intricate interdependencies between modalities, GNN is a more efficient approach to exploit the potential of multimodal datasets. The strength of GNNs lies in its ability to capture and model intra-modal and inter-modal interactions. This flexibility makes them an appealing choice for multimodal learning tasks.

There have been extensive studies using the capability of GNNs to model the conversations. DialogueGCN (Ghosal et al., 2019) models conversation using a directed graph with utterances as nodes and dependencies as edges, fitting it into a GCN structure. MMGCN (Wei et al., 2019) adopts an undirected graph to effectively fuse multimodal information and capture long-distance contextual and inter-modal interactions. Lian et al. (2020) proposed a GNN-based architecture for ERC that utilizes both text and speech modalities. DialogueCRN (Hu et al., 2021) incorporates multiturn reasoning modules to extract and integrate emotional clues, enabling a comprehensive understanding of the conversational context from a cognitive perspective. MTAG (Yang et al., 2021) is capable of both fusion and alignment of asynchronously distributed multimodal sequential data. COGMEN (Joshi et al., 2022) uses GNN-based architecture to model complex dependencies, including local and global information in a conversation. Chen et al. (2023) presents Multivariate Multi-frequency Multimodal Graph Neural Network, M^{3}Net for short, to explore the relationships between modalities and context. However, it primarily focuses on modality-level interactions and does not consider the temporal aspect within the graph.

3 Methodology

Figure 2 illustrates the architecture of CORECT to tackle the multimodal ERC task. It consists of main components namely Relational Temporal Graph Convolution Network (RT-GCN) and Pairwise Cross-modal Feature Interaction. For a given utterance in a dialogue, the former is to learn the local-context representation via leveraging various topological relations between utterances and modalities, while the latter infers the cross-modal global-context representation from the whole dialogue.

Given a multi-speaker conversation C consisting of N utterances $[u_1, u_2, \ldots, u_N]$, let us denote S as the respective set of speakers. Each utterance u_i is associated with three modalities, including audio (a), visual (v), and textual (l), that can be represented as u_i^a, u_i^v, u_i^l respectively. Using local- and global context representations, the ERC task aims to predict the label for every utterance $u_i \in C$ from a set of M predefined emotional labels $Y = [y_1, y_2, \ldots, y_M]$.

3.1 Utterance-level Feature Extraction

Here, we perform pre-processing procedures to extract utterance-level features to facilitate the learning of CORECT in the next section.

3.1.1 Unimodal Encoder

Given an utterance u_i, each data modality manifests a view of its nature. To capture this value, we employ dedicated unimodal encoders, which generate utterance-level features, namely $x_i^a \in \mathbb{R}^{d_a}$, $x_i^v \in \mathbb{R}^{d_v}$, $x_i^l \in \mathbb{R}^{d_l}$ for the acoustic, visual, and lexical modalities respectively, and d_a, d_v, d_l are the dimensions of the extracted features for each modality.

For textual modality, we utilize a Transformer (Vaswani et al., 2017) as the unimodal encoder to extract the semantic feature x_i^l from u_i^l as follows:

$$x_i^l = \textbf{Transformer}(u_i^l, \mathbf{W}_{trans}^l) \quad (1)$$

where $\mathbf{W}_{trans}^l$ is the parameter of Transformer to be learned.

For acoustic and visual modalities, we employ a fully-connected network as the unimodal encoder

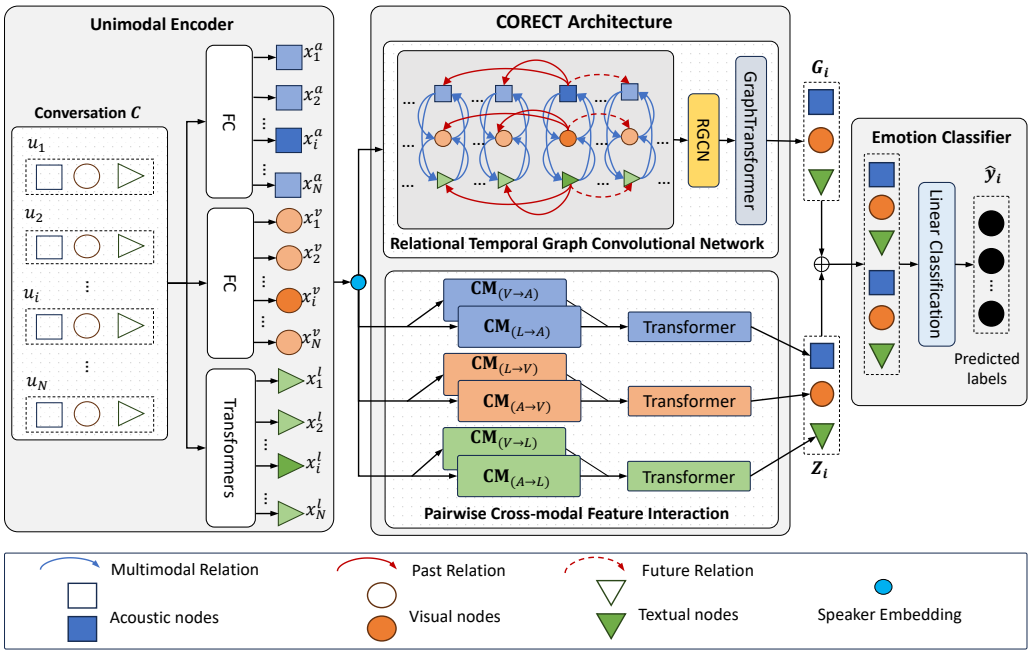


Figure 2: Framework illustration of CORECT for the multimodal emotion recognition in conversations

to extract context features for each modality type via the following procedure:

$$x_i^\tau = \mathbf{FC}(u_i^\tau; \mathbf{W}_{fc}^\tau), \tau \in \{a, v\} \quad (2)$$

where **FC** is the fully connected network, $\mathbf{W}_{fc}^\tau \in \mathbb{R}^{d_\tau \times d_{in}^\tau}$ are trainable parameters; d_{in}^τ is the input dimension of modality τ

3.1.2 Speaker Embedding

Inspired by MMGCN (Wei et al., 2019), we leverage the significance of speaker information. Let us define **Embedding** as a procedure that takes the identity of speakers and produce the respective latent representations. The embedding of multi-speaker could be inferred as:

$$\mathcal{S}_{emb} = \mathbf{Embedding}(S, \mathcal{N}_S) \quad (3)$$

where $\mathcal{S}_{emb} \in \mathbb{R}^{N \times \mathcal{N}_S}$ and $\mathcal{N}_S$ is the total number of participants in the conversation. The extracted utterance-level feature could be enhanced by adding the corresponding speaker embedding:

$$\mathbf{X}_\tau = \eta \mathcal{S}_{emb} + \mathcal{X}_\tau, \tau \in \{a, v, l\} \quad (4)$$

where $\mathcal{X}_\tau \in \mathbb{R}^{N \times d_\tau}$ refers to the global-context representation from the whole dialogue obtained from the respective unimodal encoder; $\mathbf{X}_\tau$ represents the enhanced representation with the inclusion of the speaker embedding; $\eta \in [0, 1]$ indicates the contribution ratio.

3.2 Relational Temporal Graph Convolutional Network (RT-GCN)

RT-GCN is proposed to capture local context information for each utterance in the conversation via exploiting the multimodal graph between utterances and their modalities.

3.2.1 Multimodal Graph Construction

Let us denote $\mathcal{G}(\mathcal{V}, \mathcal{R}, \mathcal{E})$ as the multimodal graph built from conversations, where $\{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ refers to the set of utterance nodes with the three modality types ($|\mathcal{V}| = 3 \times N$), the set of edges and their relation types. Figure 3 provides an illustrative example of the relations represented on the constructed graph.

Nodes. Each utterance u_i generates three nodes u_i^a, u_i^v, and u_i^l, which x_i^a, x_i^v, and x_i^l are the respective audio, visual, and lexical feature vectors.

Edges. The edge $(u_i^\tau, u_j^\tau, r_{ij}) \in \mathcal{E}, \tau \in \{a, v, l\}$ represents the interaction between u_i^τ and u_j^τ with the relation type $r_{ij} \in \mathcal{R}$. In the scope of paper, we consider two groups of relations: $\mathcal{R}_{multi}$ and $\mathcal{R}_{temp}$. Specifically, $\mathcal{R}_{multi}$ represents the intra connections between the three modalities within the same utterance, reflecting multimodal interactions. On the other hand, $\mathcal{R}_{temp}$ captures the inter connections between utterances of the same modality within a specified time window. This temporal relationship includes past/previous utterances de-

noted as $\mathcal{P}$ and next/future utterances denoted as $\mathcal{F}$. As a result, there are 15 edge types created with the definitions of the two groups.

Multimodal Relation. Emotions in dialogues cannot be solely conveyed through lexical, acoustic, or visual modalities in isolation. The interactions between utterances across different modalities play a crucial role. For example, given an utterance in a graph, its visual node has different interactive magnitude with acoustic- and textual nodes. Additionally, each node has a self-aware connection to reinforce its own information. Therefore, we can formalize 9 edge types of $\mathcal{R}_{multi}$ to capture the multimodal interactions within the dialogue as:

$$\mathcal{R}_{multi} = \begin{cases} \{(u_i^a, u_i^v), (u_i^v, u_i^a), (u_i^a, u_i^a)\} \\ \{(u_i^v, u_i^l), (u_i^l, u_i^v), (u_i^v, u_i^v)\} \\ \{(u_i^l, u_i^a), (u_i^a, u_i^l), (u_i^l, u_i^l)\} \end{cases} \quad (5)$$

Temporal Relation. It is vital to have distinct treatment for interactions between nodes that occur in different temporal orders (Poria et al., 2017). To capture this temporal aspect, we set a window slide $[\mathcal{P}, \mathcal{F}]$ to control the number of past/previous and next/future utterances that are set has connection to current node u_i^τ. This window enables us to define the temporal context for each node and capture the relevant information from the dynamic surrounding utterances. Therefore, we have 6 edge types of $\mathcal{R}_{temp}$ as follows:

$$\mathcal{R}_{temp} = \begin{cases} \{(u_j \overset{\text{past}}{\rightarrow} u_i)^\tau | i - \mathcal{P} < j < i\} \\ \{(u_i \overset{\text{future}}{\leftarrow} u_j)^\tau | i < j < i + \mathcal{F}\} \end{cases} \quad (6)$$

where $\tau \in \{a, v, l\}$; $i, j \in \overline{1, N}$; $\overset{\text{future}}{\leftarrow}$ and $\overset{\text{past}}{\rightarrow}$ indicate the past and future relation respectively.

3.2.2 Graph Learning

With the objective of leveraging the nuances and variations of heterogeneous interactions between utterances and modalities in the multimodal graph, we seek to employ Relational Graph Convolutional Networks (RGCN) (Schlichtkrull et al., 2018). For each relation type $r \in \mathcal{R}$, node representation is inferred via a mapping function $f(\mathbf{H}, \mathbf{W}_r)$, where $\mathbf{W}_r$ is the weighted matrix. Aggregating all 15 edge types, the final node representation could be computed by $\sum_r^{\mathcal{R}} f(\mathbf{H}, \mathbf{W}_r)$.

To be more specific, the representation for the i-th utterance is inferred as follows:

$$g_i^\tau = \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \mathbf{W}_r \cdot x_i^\tau + \mathbf{W}_0 \cdot x_i^\tau \quad (7)$$

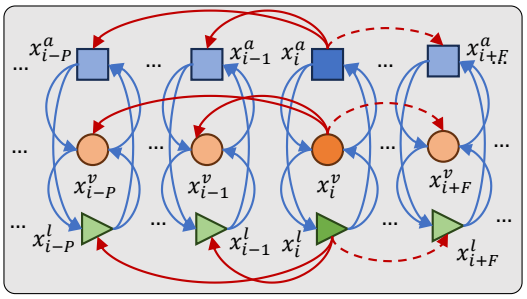


Figure 3: An example construction of a graph illustrating the relationship among utterance nodes representing audio (square), visual (circle), and text (triangle) modalities with window size $[\mathcal{P}, \mathcal{F}] = [2, 1]$ for query utterance i-th. The solid blue, solid red, and dashed red arrows indicate cross-modal-, past temporal- and future temporal connections respectively.

where $\mathcal{N}_r(i)$ is the set of the node i's neighbors with the relation $r \in \mathcal{R}$, $\mathbf{W}_0, \mathbf{W}_r \in \mathbb{R}^{d_{h1} \times d_\tau}$ are learnable parameters ($h1$ is the dimension of the hidden layer used by R-GCN), and $x_i^\tau \in \mathbb{R}^{d_\tau \times 1}$ denotes the feature vector of node u_i^τ; $\tau \in \{a, v, l\}$.

To extract rich representations from node features, we utilize a Graph Transformer model (Yun et al., 2019), where each layer comprises a self-attention mechanism followed by feed-forward neural networks. The self-attention mechanism allows vertices to exploit information from neighborhoods as well as capturing local and global patterns in the graph. Given g_i^τ is the representation of i^{th} utterance with modality $\tau \in \{a, v, l\}$ obtained from RGCNs, its representation is transformed into:

$$o_i^\tau = ||_{c=1}^{C} [\mathbf{W}_1 g_i^\tau + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}^\tau \mathbf{W}_2 g_j^\tau] \quad (8)$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_{h_2} \times d_{h_1}}$ are learned parameters ($h2$ is the dimension of the hidden layer used by Graph Transformer); $\mathcal{N}(i)$ is the set of nodes that has connections to node i; $||$ is the *concatenation* for C head attention; and the attention coefficient of node j, i.e., $\alpha_{i,j}^\tau$, is calculated by the *softmax* activation function:

$$\alpha_{i,j}^\tau = softmax(\frac{(\mathbf{W}_3 g_i^\tau)^\top (\mathbf{W}_4 g_i^\tau)}{\sqrt{d}}) \quad (9)$$

$\mathbf{W}_3, \mathbf{W}_4 \in \mathbb{R}^{d_\alpha \times d_{h_1}}$ are learned parameters.

After the aggregation throughout the whole graph, we obtain new representation vectors:

$$\mathbf{G}^\tau = \{o_1^\tau, o_2^\tau, \ldots, o_N^\tau\} \quad (10)$$

where $\tau \in \{a, v, l\}$ indicates the corresponding audio, visual, or textual modality.

3.3 Pairwise Cross-modal Feature Interaction

The cross-modal heterogeneities often elevate the difficulty of analyzing human language. Exploiting cross-modality interactions may help to reveal the "unaligned" nature and long-term dependencies across modalities. Inspired by the idea (Tsai et al., 2019), we design the *Pairwise Cross-modal Feature Interaction (P-CM)* method into our proposed framework for conversation understanding. A more detailed illustration of the *P-CM* module is presented in Appendix A.1.2

Given two modalities, e.g., audio a and textual l, let us denote $\mathbf{X}^a \in \mathbb{R}^{N\times d_a}$, $\mathbf{X}^l \in \mathbb{R}^{N\times d_l}$ as the respective modality-sensitive representations of the whole conversation using unimodal encoders. Based on the transformer architecture (Vaswani et al., 2017), we define the Queries as $Q^a = \mathbf{X}^a W_{Q^a}$, Keys as $K^l = \mathbf{X}^l W_{K^l}$, and Values as $V^l = \mathbf{X}^l W_{V^l}$. The enriched representation of $\mathbf{X}^a$ once performing cross-modal attention on the modality a by the modality l, referred to as $\mathbf{CM}^{l\to a} \in \mathbb{R}^{N\times d_V}$, is computed as:

$$\mathbf{CM}^{l\to a} = \sigma\left(\frac{\mathbf{X}^a\mathbf{W}_{Q^a}(\mathbf{W}_{K^l})^\top(\mathbf{X}^l)^\top}{\sqrt{d_k}}\right)\mathbf{X}_l\mathbf{W}_{V^l} \tag{11}$$

where σ is the *softmax* function; $\mathbf{W}_{Q^a} \in \mathbb{R}^{d_a\times d_K}$, $\mathbf{W}_{K^l} \in \mathbb{R}^{d_l\times d_K}$, and $\mathbf{W}_{V^l} \in \mathbb{R}^{d_l\times d_V}$ are learned parameters. The value of d_Q, d_K, d_V is the dimension of queues, keys and values respectively. $\sqrt{d_k}$ is a scaling factor and $d_{(.)}$ is the feature dimension.

To model the cross-modal interactions on unaligned multimodal sequences, e.g., audio, visual, and lexical, we utilize D cross-modal transformer layers. Suppose that $\mathbf{Z}^a_{[i]}$ is the modality-sensitive global-context representation of the whole conversation for the modality l at the $i-$th layer; $\mathbf{Z}^a_{[0]} = \mathbf{X}^a$. The enriched representation of $\mathbf{Z}^{[i]}_l$ denoted as $\mathbf{Z}^{[i]}_{l\to a}$ by applying cross-modal attention of the modality l on the modality a is computed as the following procedure:

$$\begin{aligned}
\mathbf{Z}^{l\to a}_{[0]} &= \mathbf{Z}^a_{[0]} \\
\overline{\mathbf{Z}}^{l\to a}_{[i]} &= \mathbf{CM}^{l\to a}_{[i]}(LN(\mathbf{Z}^{l\to a}_{[i-1]}), LN(\mathbf{Z}^{l\to a}_0)) \\
&\quad + LN(\mathbf{Z}^{l\to a}_{[i-1]}) \\
\mathbf{Z}^{l\to a}_{[i]} &= (LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))^{FFN} + LN(\overline{\mathbf{Z}}^{l\to a}_{[i]})
\end{aligned} \tag{12}$$

where $\overline{\mathbf{Z}}$ is the intermediate representation; *LN* is a layer normalization (Ba et al., 2016), which helps to stabilize the learning process and enhance the convergence of the model. $LN(\overline{\mathbf{Z}}^{a\to v}_{[i]}))^{FFN}$ expresses the transformation by the position-wise feed-forward block as:

$$\begin{aligned}
LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))^{FFN} = \max(0, LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))\mathbf{\Omega}_1 \\
+ \mathbf{b}_1)\mathbf{\Omega}_2 + \mathbf{b}_2
\end{aligned} \tag{13}$$

where $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are linear projection matrices; $\mathbf{b}_1$ and $\mathbf{b}_2$ are biases.

Likewise, we can easily compute the cross-modal representation $\mathbf{Z}^{[i]}_{a\to l}$, indicating that information from the modality a is transferred to the modality l. Finally, we concatenate all representations at the last layer, i.e., the $D-$th layer, to get the final cross-modal global-context representation $\mathbf{Z}^{[D]}_{a\rightleftarrows l}$. For other modality pairs, $\mathbf{Z}^{[D]}_{v\rightleftarrows l}$ and $\mathbf{Z}^{[D]}_{v\rightleftarrows a}$ could be obtained by the similar process.

3.4 Multimodal Emotion Classification

The local- and global context representation resulted in by the RT-GCN and P-CM modules are fused together to create the final representation of the conversation:

$$\begin{aligned}
\mathbf{H} &= Fusion([\mathbf{G}, \mathbf{Z}]) \\
&= [\{o^\tau_1, o^\tau_2, \ldots, o^\tau_N\}, \{\mathbf{Z}^{[D]}_{a\rightleftarrows v}, \mathbf{Z}^{[D]}_{v\rightleftarrows l}, \mathbf{Z}^{[D]}_{l\rightleftarrows a}\}]
\end{aligned} \tag{14}$$

where $\tau \in \{a, v, l\}$; $Fusion$ represents the concatenation method. $\mathbf{H}$ is then fed to a fully connected layer to predict the emotion label y^i for the utterance u_i:

$$v_i = ReLU(\mathbf{\Phi}_0 h_i + b_0) \tag{15}$$

$$p_i = \text{softmax}(\mathbf{\Phi}_1 v_i + b_1) \tag{16}$$

$$\hat{y}^i = \text{argmax}(p_i) \tag{17}$$

where $\mathbf{\Phi}_0, \mathbf{\Phi}_1$ are learned parameters.

Dataset	Dialogues			Utterances		
	train	valid	test	train	valid	test
IEMOCAP (6-way)	108	12	31	5,146	664	1,623
IEMOCAP (4-way)	108	12	31	3,200	400	943
MOSEI	2,249	300	646	16,327	1,871	4,662

Table 1: Statistics for IEMOCAP, MOSEI datasets

4 Experiments

This section investigate the efficacy of CORECT for the ERC task through extensive experiments in comparing with state-of-the-art (SOTA) baselines.

Methods	IEMOCAP (6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)
bc-LSTM (Poria et al., 2017)	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10
CMN (Hazarika et al., 2018b)	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON (Hazarika et al., 2018a)	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN (Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN (Ghosal et al., 2019)	47.10	**80.88**	58.71	66.08	70.97	61.21	65.54	65.04
MMGCN (Wei et al., 2019)	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
DialogueCRN (Hu et al., 2021)	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN (Joshi et al., 2022)	55.76	80.17	63.21	61.69	**74.91**	63.90	67.04	67.27
CORECT (Ours)	**59.30**	80.53	**66.94**	**69.59**	72.69	**68.50**	**69.93** (↑ 2.89)	**70.02** (↑ 2.75)

Table 2: The results on IEMOCAP (6-way) multimodal (A+V+T) setting. The results in **bold** indicate the highest performance, while the underlined results represent the second highest performance. The ↑ illustrates the improvement compared to the previous state-of-the-art model.

4.1 Experimental Setup

Dataset. We investigate two public real-life datasets for the multimodal ERC task including IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Bagher Zadeh et al., 2018). The dataset statistics are given in Table 1.

IEMOCAP contains 12 hours of videos of two-way conversations from 10 speakers. Each dialogue is divided into utterances. There are in total 7433 utterances and 151 dialogues. The 6-way dataset contains six emotion labels, i.e., *happy, sad, neutral, angry, excited*, and *frustrated*, assigned to the utterances. As a simplified version, ambiguous pairs such as *(happy, exited)* and *(sad, frustrated)* are merged to form the 4-way dataset.

CMU-MOSEI provides annotations for 7 sentiments ranging from highly negative (-3) to highly positive (+3), and 6 emotion labels including *happiness, sadness, disgust, fear, surprise*, and *anger*.

Evaluation Metrics. We use *weighted F1-score* (w-F1) and *Accuracy* (Acc.) as evaluation metrics. The w-F1 is computed $\sum_{k=1}^{K} freq_k \times F1_k$, where $freq_k$ is the relative frequency of class k. The accuracy is defined as the percentage of correct predictions in the test set.

Baseline Models. CORECT is compared against SOTA baselines specific to each dataset. For IEMOCAP, we consider two model groups namely: i) *RNN-based models* include bc-LSTM (Poria et al., 2017), CMN (Hazarika et al., 2018b), ICON (Hazarika et al., 2018a), DialogueRNN (Majumder et al., 2019); ii) *Graph-based methods* are DialogueGCN (Ghosal et al., 2019), MMGCN (Wei et al., 2019), DialougueCRN (Hu et al., 2021), CHFusion (Majumder et al., 2018), and COGMEN (Joshi et al., 2022). For CMU-MOSEI, we investigate multimodal models including MultilogueNet (Shenoy and Sardana, 2020), TBJE (Delbrouck et al., 2020), and COGMEN (Joshi et al., 2022).

Modality Settings	IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)
bc-LSTM (Poria et al., 2017)	75.20	75.13
CHFusion (Majumder et al., 2018)	76.59	76.80
COGMEN (Joshi et al., 2022)	82.29	82.15
CORECT (Ours)	**84.73** (↑ 2.44)	**84.64** (↑ 2.49)

Table 3: The results on the IEMOCAP (4-way) dataset in the multimodal (A+V+T) setting. The ↑ indicates the improvement compared to the previous SOTA model.

Implementation Details. Due to the space limit, the implementation details for feature extraction and interaction are described in Appendix A.1.

4.2 Comparison With Baselines

We further qualitatively analyze CORECT and the baselines on the IEMOCAP (4-way), IEMOCAP (6-way) and MOSEI datasets.

IEMOCAP: In the case of IEMOCAP (6-way) dataset (Table 2), CORECT performs better than the previous baselines in terms of F1 score for individual labels, excepts the *Sad* and the *Excited* labels. The reason could be the ambiguity between similar emotions, such as *Happy* & *Excited*, as well as *Sad* & *Frustrated* (see more details in Figure 6 in Appendix A.2). Nevertheless, the accuracy and weighted F1 score of CORECT are 2.89% and 2.75% higher than all baseline models on average. Likewise, we observe the similar phenomena on the IEMOCAP (4-way) dataset with a 2.49% improvement over the previous state-of-the-art models as Table 3. These results affirm the efficiency of CORECT for the multimodal ERC task.

Methods	Sentiment Classification Accuracy (%)		Emotion Classification (Binary, 1 vs. all) weighted F1-score (%)					
	2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilouge-Net (Shenoy and Sardana, 2020)	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE (Delbrouck et al., 2020)	82.40	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN (Joshi et al., 2022)	82.95	45.22	70.88	70.91	74.20	*87.79*	81.83	*86.05*
CORECT (Ours)	**83.66**	**46.31**	**71.35**	**72.86**	**76.77**	**87.90**	**84.26**	**86.48**

Table 4: Results on CMU-MOSEI dataset compared with previous works. The **bolded** results indicate the best performance, while the underlined results represent the second best performance.

Sub-Modules	IEMOCAP (6-way)		IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
-w/o RT-GCN	66.61	66.55 (↓ 3.47)	80.69	80.54 (↓ 4.10)
-w/o P-CM	66.54	66.64 (↓ 3.38)	82.18	82.16 (↓ 2.48)
-w/o $\mathcal{R}_{multi}$	66.54	66.82 (↓ 3.20)	82.61	82.53 (↓ 2.11)
-w/o $\mathcal{R}_{temp}$	67.04	67.34 (↓ 2.68)	82.08	82.07 (↓ 2.57)
CORECT	**69.93**	**70.02**	**84.73**	**84.64**

Table 5: The performance of CORECT in different strategies under the fully multimodal (A+V+T) setting. **Bolded** results represent the best performance, while underlined results depict the second best. The ↓ represents the decrease in performance when a specific module is ablated compared to our CORECT model.

CMU-MOSEI: Table 4 presents a comparison of the CORECT model on the CMU-MOSEI dataset with current SOTA models in two settings: Sentiment Classification (2-class and 7-class) and Emotion Classification. Apparently, CORECT consistently outperforms other models with sustainable improvements. One notable observation is the *italicized* results for the *Fear* and *Surprise* labels, where all the baselines have the same performance of 87.79 and 86.05 respectively. During the experimental process, when reproducing these baseline's results, we found that the binary classifiers were unable to distinguish any samples for the *Fear* and *Surprise* labels. However, with the help of technical components, i.e., *RT-GCN* and *P-CM*, our model shows significant improvement even in the presence of severe label imbalance in the dataset. Due to space limitations in the paper, we provide additional experiments on the CMU-MOSEI dataset for all possible combinations of modalities in Table 7 (Appendix A.2).

4.3 Ablation study

Effect of Main Components. The impact of main components in our CORECT model is presented via Table 5. The model performance on the 6-way IEMOCAP dataset is remarkably degraded when the *RT-GCN* or *P-CM* module is not adopted with the decrease by 3.47% and 3.38% respectively. Similar phenomena is observed on the 4-way IEMOCAP dataset. Therefore, we can deduce that the effect of *RT-GCN* in the CORECT model is more significant than that of *P-CM*.

For different relation types, ablating either $\mathcal{R}_{multi}$ or $\mathcal{R}_{temp}$ results in a significant decrease in the performance. However, the number of labels may affect on the multimodal graph construction, thus it is no easy to distinguish the importance of $\mathcal{R}_{multi}$ and $\mathcal{R}_{temp}$ for the multimodal ERC task.

Table 8 (Appendix A.2) presents the ablation results for uni- and bi-modal combinations. In the unimodal settings, specifically for each individual modality (A, V, T), it's important to highlight that both *P-CM* module and multimodal relations $\mathcal{R}_{multi}$ are non-existent. However, in bimodal combinations, the advantage of leveraging cross-modality information between audio and text (A+T) stands out, with a significant performance boost of over 2.75% compared to text and visual (T+V) modalities and a substantial 14.54% compared to visual and audio (V+A) modalities.

Additionally, our experiments have shown a slight drop in overall model performance (e.g., 68.32% in IEMOCAP 6-way, drop of 1.70%) when excluding Speaker Embedding $\mathcal{S}_{emb}$ from CORECT.

Effect of the Past and Future Utterance Nodes. We conduct an analysis to investigate the influence of past nodes ($\mathcal{P}$) and future nodes ($\mathcal{F}$) on the model's performance. Unlike previous studies

Modality Settings	IEMOCAP (6-way)		IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
A	52.31	51.49	67.02	65.48
T	67.22	67.26	82.82	82.65
V	38.63	37.67	49.73	47.97
A+T	68.27	68.36	83.14	83.13
T+V	65.50	65.61	81.76	81.75
V+A	54.16	53.82	69.03	68.21
CORECT (A+T+V)	**69.93**	**70.02**	**84.73**	**84.64**

Table 6: The performance of CORECT under various modality settings.

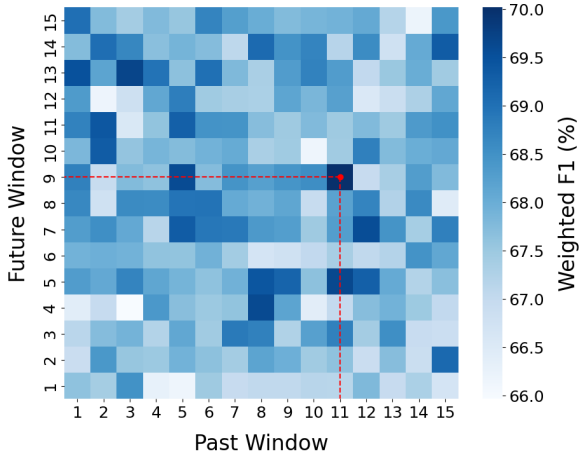


Figure 4: The effects of $\mathcal{P}$ and $\mathcal{F}$ nodes in the past and future of CORECT model on the IEMOCAP (6-way) The red-dash line implies our best setting for $\mathcal{P}$ and $\mathcal{F}$.

(Joshi et al., 2022; Li et al., 2023) that treated $\mathcal{P}$ and $\mathcal{F}$ pairs equally, we explore various combinations of $\mathcal{P}$ and $\mathcal{F}$ settings to determine their effects. Figure 4 indicates that the number of past or future nodes can have different impacts on the performance. From the empirical analysis, the setting $[\mathcal{P}, \mathcal{F}]$ of $[11, 9]$ results in the best performance. This finding shows that the contextual information from the past has a stronger influence on the multimodal ERC task compared to the future context.

Effect of Modality. Table 6 presents the performance of the CORECT model in different modality combinations on both the IEMOCAP and CMU-MOSEI datasets.

For IEMOCAP (Table 2 and Table 3), the textual modality performs the best among the unimodal settings, while the visual modality yields the lowest results. This can be attributed to the presence of noise caused by factors, e.g., camera position, environmental conditions. In the bi-modal settings, combining the textual and acoustic modalities achieves the best performance, while combining the visual and acoustic modalities produces the worst result. A similar trend is observed in the CMU-MOSEI dataset (Table 4), where fusing all modalities together leads to a better result compared to using individual or paired modalities.

5 Conclusion

In this work, we propose CORECT, an novel network architecture for multimodal ERC. It consists of two main components including RT-GCN and P-CM. The former helps to learn local-context representations by leveraging modality-level topological relations while the latter supports to infer cross-modal global-context representations from the entire dialogue. Extensive experiments on two popular benchmark datasets, i.e., IEMOCAP and CMU-MOSEI, demonstrate the effectiveness of CORECT, which achieves the new state-of-the-art record for multimodal conversational emotion recognition. Furthermore, we also provide ablation studies to investigate the contribution of various components in CORECT. Interestingly, by analyzing the temporal aspect of conversations, we have validated that capturing the long-term dependencies, e.g., past relation, improves the performance of the multimodal emotion recognition in conversations task.

Limitations

Hyper-parameter tuning is a vital part of optimizing machine learning models. Not an exception, the learning of CORECT is affected by hyper-parameters such as the number of attention head in P-CM module, the size of Future and Past Window. Due to time constraints and limited computational resources, it was not possible to tune or exploring all possible combinations of these hyper-parameters, which might lead to local-minima convergences. In future, one solution for this limitation is to employ automated hyper-parameter optimization algorithms, to systematically explore the hyperparameter space and improve the robustness of the model. As another solution, we may upgrade CORECT with learning mechanisms to automatically leverage important information, e.g., attention mechanism on future and past utterances.

Acknowledgements

Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.TS143.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. 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.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.

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

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 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)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 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)*, pages 7042–7052, Online. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

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

Zheng Lian, Bin Liu, and Jianhua Tao. 2022. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*.

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *INTERSPEECH*, pages 2347–2351.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133.

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

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Andrei Nicolicioiu, Iulia Duta, and Marius Leordeanu. 2019. Recurrent space-time graph neural networks. *Advances in neural information processing systems*, 32.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Garima Sharma and Abhinav Dhall. 2021. A survey on automatic multimodal emotion recognition in the wild. *Advances in data science: Methodologies and applications*, pages 35–64.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. 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)*, pages 1551–1560, Online. Association for Computational Linguistics.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA. Association for Computational Linguistics.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021, Online. Association for Computational Linguistics.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. 2019. Graph transformer networks. *CoRR*, abs/1911.06455.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

A Appendix

A.1 Implementation Details

A.1.1 Multimodal Raw Feature Extraction

The multimodal feature extraction process involves extracting features from the acoustic, lexical, and visual modalities for each utterance.

For IEMOCAP, the audio features, with a size of 100, are obtained using the OpenSmile Toolkit (Eyben et al., 2010); visual features, with a size of 512, are extracted using OpenFace (Baltrusaitis et al., 2018); textual features, with a size of 768, are derived using sBERT (Reimers and Gurevych, 2019).

For MOSEI, the audio features are extracted using librosa (McFee et al., 2015) with 80 filter banks, resulting in a feature vector size of 80.The visual features, with a size of 35, are obtained from (Bagher Zadeh et al., 2018). The textual features, with a size of 768, are obtained using sBERT (Reimers and Gurevych, 2019).

A.1.2 Pairwise Cross-modal Feature Interaction

Figure 5 illustrates details of Pairwise Cross-modal Feature Interaction (P-CM).

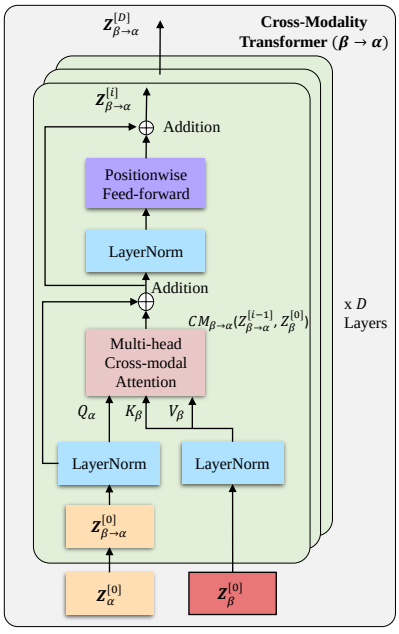
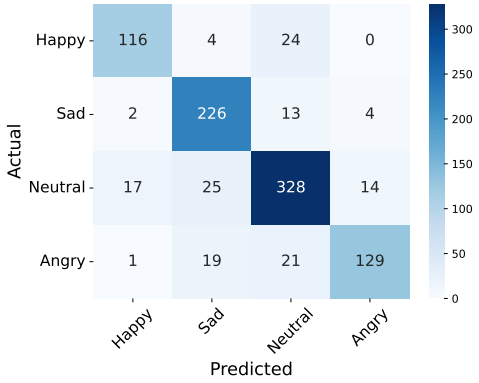


Figure 5: Illustration of the P-CM module between modality β and α.

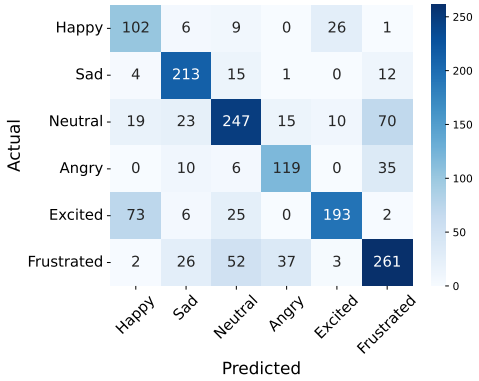
A.2 Additional Experiment Result

Table 8 showcases the results on the IEMOCAP dataset (both 6-way and 4-way) for all the modality combinations of the CORECT model, while Table 7 presents an ablation study conducted on the CMU-MOSEI dataset, considering various modality combination settings.

Figure 6 shows the confusion matrix for prediction on IEMOCAP (4-way) and IEMOCAP (6-way), respectively.



(a) Confusion matrix on the IEMOCAP (4-way).



(b) Confusion matrix on the IEMOCAP (6-way).

Figure 6: Visualization the confusion matrices of CORECT under multimodal (A+V+T) setting. Most of False predictions observed on IEMOCAP (6-way) came from the ambiguity between pair of labels: *Happy* and *Excited*, *Neural* and *Frustrate*.

A.3 Reproducibility

CORECT is implemented using Pytorch [1], and run experiments on Google Colab Pro. We choose Adam as the optimizer and set the dropout rate to 0.5. The numbers of multi-head attentions used in Graph Transformer and P-CM are selected as 7 and 2, respectively. For IEMOCAP dataset, the learning rate is 0.0003; Window size $[\mathcal{P}, \mathcal{F}]$ is tested on various settings in the range of [1,15]. For CMU-MOSEI dataset, the learning rate is 0.0006; Window size $[\mathcal{P}, \mathcal{F}]$ is picked between [5,4] due to the property of short dialogue in CMU-MOSEI. Refer-

[1] https://pytorch.org/

Datasets	Modality Settings	Sentiment Class Accuracy (%)		Emotion Class weighted F1-score (%)					
		2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilogue-Net (Shenoy and Sardana, 2020)	A+T+V	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE (Delbrouck et al., 2020)	A+T	82.4	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN (Joshi et al., 2022)	A+T+V	82.95	45.22	70.88	70.91	74.20	*87.79*	81.83	*86.05*
CORECT (Ours)	T	84.13	45.80	67.82	72.12	75.55	*87.79*	84.63	*86.05*
	A+T	**84.28**	44.89	67.49	71.53	75.39	*87.79*	**84.69**	*86.05*
	A+T+V	83.66	**46.31**	**71.35**	**72.86**	**76.77**	**87.90**	84.26	**86.48**

Table 7: Ablation study on CMU-MOSEI dataset. The multimodal implementation (A+V+T) consistently outperformed the baseline models in most of the modality combinations. For the 2-class sentiment and the *Disgust* class emotion, our approach reaches a competitive performance.

Dataset		A		T		V		A+T		T+V		V+A		A+V+T	
		Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	F1	Acc	W-F1	Acc	F1
IEMOCAP (6-way)	w/o RT-GCN	35.12	30.01	64.7	64.34	30.99	26.88	67.10	66.92	65.37	65.50	52.13	51.80	66.61	66.55
	w/o P-CM	-	-	-	-	-	-	65.87	65.89	65.00	65.07	53.54	52.86	66.54	66.64
	w/o $\mathcal{R}_{multi}$	-	-	-	-	-	-	66.30	66.27	64.76	64.78	53.67	53.48	66.54	66.82
	w/o $\mathcal{R}_{temp}$	41.53	39.49	63.65	63.72	27.66	27.37	67.34	67.33	65.43	65.29	50.65	49.67	67.04	67.34
	CORECT	**52.31**	**51.49**	**67.22**	**67.26**	**38.63**	**37.67**	**68.27**	**68.36**	**65.50**	**65.61**	**54.16**	**53.82**	**69.93**	**70.02**
IEMOCAP (4-way)	w/o RT-GCN	55.25	52.18	80.38	80.25	34.04	31.33	81.87	81.18	80.17	80.04	58.96	58.57	80.69	80.54
	w/o P-CM	-	-	-	-	-	-	80.91	80.94	80.38	80.04	69.25	69.00	82.18	82.16
	w/o $\mathcal{R}_{multi}$	-	-	-	-	-	-	81.76	81.78	80.38	80.47	69.14	68.84	82.61	82.53
	w/o $\mathcal{R}_{temp}$	56.84	54.88	80.70	80.70	41.04	39.75	82.08	81.99	81.34	81.36	57.16	56.62	82.08	82.07
	CORECT	**67.02**	**65.48**	**82.82**	**82.85**	**49.73**	**47.97**	**83.14**	**83.13**	**81.76**	**81.75**	**69.03**	**68.21**	**84.73**	**84.64**

Table 8: Ablation study on IEMOCAP dataset. It shows the consistency of our proposal method since any ablation experiments (on both modal settings and modules) results in a reduction of overall performance. On unimodal setting of $\{A, V, T\}$, *P-CM* module and multimodal relation are not exist. Therefore there are no ablation of *P-CM* and $\mathcal{R}_{multi}$ on these unimodal setting, denotes by "-".

ring to the training log on the IEMOCAP (6-way) dataset using Google Colab Pro, each mini-batch (size of 10 dialouges) takes approximately 0.4s. The similar ratio is observed on the MOSEI dataset.

Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition

Cam-Van Thi Nguyen, Cao-Bach Nguyen, Quang-Thuy Ha, Duc-Trong Le
VNU University of Engineering and Technology, Hanoi, Vietnam
{vanntc, 19020218, thuyhq, trongld}@vnu.edu.vn

Abstract

Emotion recognition in conversation (ERC) is a crucial task in natural language processing and affective computing. This paper proposes MultiDAG+CL, a novel approach for Multimodal Emotion Recognition in Conversation (ERC) that employs Directed Acyclic Graph (DAG) to integrate textual, acoustic, and visual features within a unified framework. The model is enhanced by Curriculum Learning (CL) to address challenges related to emotional shifts and data imbalance. Curriculum learning facilitates the learning process by gradually presenting training samples in a meaningful order, thereby improving the model's performance in handling emotional variations and data imbalance. Experimental results on the IEMOCAP and MELD datasets demonstrate that the MultiDAG+CL models outperform baseline models. We release the code for MultiDAG+CL and experiments: https://github.com/vanntc711/MultiDAG-CL.

Keywords: Multimodal Emotion Recognition, Curriculum Learning, Directed Acyclic Graph

1. Introduction

Online social networks' growing popularity has sparked interest in capturing emotions in conversations. Emotion Recognition in Conversation (ERC) has emerged as a critical task in various domains such as chatbots (Ghosh et al., 2017), healthcare (Li et al., 2019), and social media analysis (Polzin and Waibel, 2000). In the field of ERC, researchs can be broadly categorized into unimodal and multimodal approaches. Unimodal approaches usually focus on using text as the main modality for emotion recognition. Several models have been proposed in the past to tackle unimodal ERC task. DialogueRNN (Majumder et al., 2019) introduces a recurrent network to track speaker states and context during the conversation. DialogueGCN (Ghosal et al., 2019) utilizes graph structures to combine contextual dependencies.

Multimodal Emotion Recognition in Conversation (Multimodal ERC) classifies emotions in conversation turns using text, audio, and visual cues. By incorporating multiple modalities, it provides a comprehensive representation of emotional expressions, including tone of voice, facial expressions, and body language, resulting in improved accuracy and robustness in emotion recognition compared to traditional unimodal ERC approaches. Several models have been proposed to address the task of multimodal ERC. The MFN (Zadeh et al., 2018) synchronizes multimodal sequences using a multi-view gated memory. ICON (Hazarika et al., 2018) provides conversational features from modalities through multi-hop memories. The bc-LSTM (Poria et al., 2017) leverages an utterance-level LSTM to capture multimodal features. MMGCN (Hu et al., 2021) uses a graph-based fusion module to capture intra- and inter-modality contextual features. CTNet (Lian et al., 2021) utilizes a transformer-based structure to model interactions among multimodal features. CORECT (Nguyen et al., 2023) leverages relational temporal GNNs with cross-modality interaction support, effectively capturing conversation-level interactions and utterance-level temporal relations.

A Directed Acyclic Graph (DAG) is a directed graph without any directed cycles, comprising vertices and edges, where each edge is directed from one vertex to another, ensuring no closed loops. Building upon this concept, Yu et al. (2019) introduced Directed Acyclic Graph Neural Network (DAG-GNN). Additionally, Shen et al. (2021) presented DAG-ERC, a model combining graph-based and recurrence-based neural architectures to capture information flow in long-distance conversations. However, DAG-ERC's focus has been primarily on unimodal text data, with limited exploration in other modalities. Curriculum Learning (CL), inspired by human learning, progressively introduces more complex concepts starting from a simple initial state. It establishes a sequence of curricula where the best curriculum with the simplest examples is used to train the classifier in each learning round (Bengio et al., 2009; Soviany et al., 2022). CL incorporates two key factors: a *difficulty measurer* to assess the difficulty level of training examples, and a *training scheduler* to determine the order of example presentation during training. The difficulty measurer assesses the difficulty level of training examples, while the training scheduler determines the order in which examples are presented to the model during training. For the ERC task, Yang et al. (2022) proposes a hybrid CL framework specifically for the textual modality only.

LREC-COLING 2024, pages 4259–4265
20-25 May, 2024.

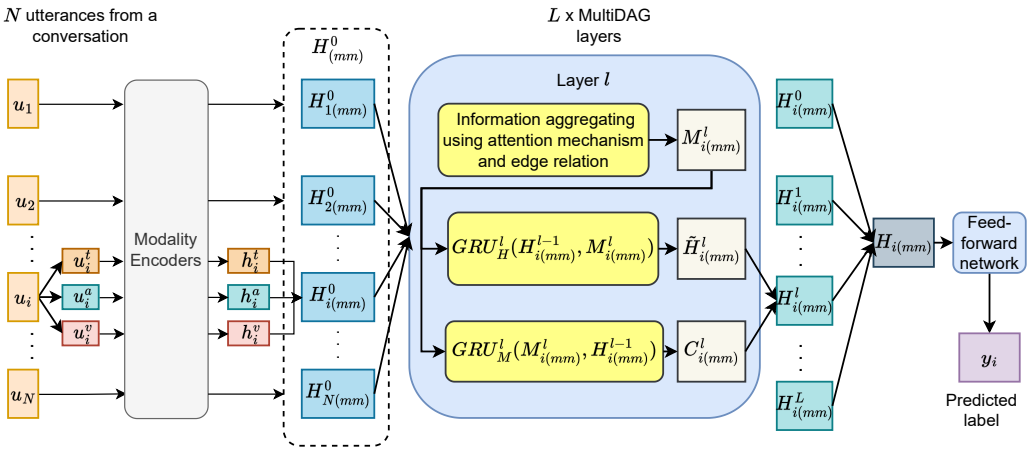


Figure 1: Overall structure of MultiDAG.

In this paper, we proposes **MultiDAG+CL**, a multimodal model inspired by DAG-ERC (Shen et al., 2021), designed to overcome the limitations of text-based approaches. It integrates multimodal features using DAG-GNN, enabling a comprehensive understanding of emotions in conversations. Leveraging Curriculum Learning, our model, **MultiDAG+CL**, addresses emotional shift issues and imbalanced data, significantly enhancing ERC model performance on IEMOCAP and MELD datasets. Notably, we are the first to integrate multimodal ERC models with Curriculum Learning strategy.

2. Methodology

Consider a conversation C having utterances $\{u_1, u_2, \ldots, u_N\}$ where N is the number of utterances. An utterance is a coherent piece of information conveyed by a single participant p_m at a specific moment, where $m \geq 2$. The task of Emotion Recognition in Conversation (ERC) is to predict emotion label of each utterance u_i with predefined emotion label set $E = \{y_1, y_2, \ldots, y_r\}$. Following the multimodal approach, we represent an utterance in terms of three different modalities: audio (a), visual (v), and textual (l). The raw feature representation of utterance u_i is $u_i = \{u_i^a, u_i^v, u_i^l\}$.

MultiDAG+CL consists of two core components: *MultiDAG* and *Curriculum Learning-CL*. The *MultiDAG* component represents the model that combines multimodal features without CL integration. The *-CL* component is where Curriculum Learning is incorporated to enhance model performance.

2.1. Multimodal ERC with Directed Acyclic Graph - MultiDAG

2.1.1. Modality Encoder

We use modality-specific encoders to generate context-aware utterance feature encoding. For the textual modality, a bidirectional LSTM network captures sequential textual context information, while a Fully Connected Network is used for the acoustic and visual modalities as follows:

$$h_i^a = Enc_A(u_i^a); h_i^v = Enc_V(u_i^v); h_i^l = Enc_L(u_i^l) \quad (1)$$

where Enc_A, Enc_V, Enc_L are modality encoder for audio, visual, textual modalities, respectively. These encoders generate the context-aware raw feature encodings h_i^a, h_i^v, h_i^l accordingly. The multimodal feature vector for an utterance $u_{i(mm)}$ corresponding to available modalities is:

$$H_{i(mm)}^0 = h_i^a \oplus h_i^v \oplus h_i^l \quad (2)$$

2.1.2. MultiDAG Construction

Each utterance in a conversation receives information exclusively from past utterances. This one-way information flow is effectively represented by a Directed Acyclic Graph (DAG), where information moves from predecessors to successors. This characteristic allows the DAG to gather information for a query utterance not only from neighboring utterances but also from more distant ones. Following the multimodal representation input, we initialize the Directed Acyclic Graph Gated Neural Network (DAG-GNN) (Yu et al., 2019). The integration of both remote and local information is executed in a manner analogous to the approach undertaken in DAG-ERC by Shen et al. (2021). The comprehensive architecture of MultiDAG is visually represented in Figure 1.

At each layer l of the MultiDAG, the hidden state of the utterances is continuously computed from the first utterance to the last utterance. For each utterance $u_{i(mm)}$, the attention weight between $u_{i(mm)}$ and the preceding nodes is calculated by using the hidden state of $u_{i(mm)}$ at layer $l-1$ to attend to the hidden states of the nodes at layer l:

$$a_{ij(mm)}^l = S_{j \in \mathcal{N}_{i(mm)}}(W_\alpha^l [H_{j(mm)}^l \| H_{i(mm)}^{l-1}]) \quad (3)$$

Here, S denotes the Softmax function; $N_{i(mm)}$ represents the set of preceding nodes leading to u_i, W_a^l is a trainable weight matrix, $H_{i(mm)}^{l-1}$ is the hidden state of $u_{i(mm)}$ at layer $l-1$, and $||$ denotes concatenation.

The attention weight is further utilized in combination with edge relationships to aggregate information.

$$M_{i(mm)}^l = \sum_{j \in \mathcal{N}_{i(mm)}} a_{ij(mm)} W_{r_{ij}}^l H_{j(mm)}^l \quad (4)$$

where $W_{r_{ij}}^l \in \{W_0^l, W_1^l\}$ are trainable parameters. The $0/1$ value represents the edge relationship, distinguishing different or same speakers.

The aggregated information $M_{i(mm)}^l$ interacts with the previous layer's hidden state of $u_{i(mm)}$, $H_{i(mm)}^{l-1}$, through a GRU to generate the final hidden state $\tilde{H}_{i(mm)}^l$ at the current layer:

$$\tilde{H}_{i(mm)}^l = GRU_H^l(H_{i(mm)}^{l-1}, M_{i(mm)}^l) \quad (5)$$

where $H_{i(mm)}^{l-1}$, $M_{i(mm)}^l$, and $\tilde{H}_{i(mm)}^l$ represent the input, hidden state, and output of the GRU network, respectively. This step is the node information unit. Another GRU serves as the context information unit, modeling the flow of information from the historical context through a layer. In this unit, the roles of H_i^{l-1} and M_i^l in the GRU are exchanged, where $H_{i(mm)}^{l-1}$ controls the propagation of $M_{i(mm)}^l$:

$$C_{i(mm)}^l = GRU_M^l(M_{i(mm)}^l, H_{i(mm)}^{l-1}) \quad (6)$$

The hidden states of u_i from all layers are concatenated together to create final representation:

$$H_{i(mm)} = \|_{l=0}^{L}(\tilde{H}_{i(mm)}^l + C_{i(mm)}^l) \quad (7)$$

This representation is then passed through a Feed-Forward Network to perform emotion prediction. The objective function used to train the model is the cross-entropy loss function.

2.2. Curriculum Learning - CL

We design a **Difficulty Measure Function (DMF)** based on the frequency of emotional shift in conversations, and simultaneously construct a **Training Scheduler** to implement the training process according to the predefined learning curriculum.

2.2.1. Difficulty Measure Function (DMF)

When designing the difficulty measurement function for a conversation, it is essential to determine what makes a conversation easier or more difficult than others. Taking inspiration from Yang et al. (2022), we constructed a function to calculate the difficulty of a conversation based on the frequency of emotional shift. Here, an emotional shift is defined as occurring when the emotion expressed in two consecutive utterances by the same speaker is different. Specifically, $e(u_i) \neq e(u_k)$, $p(u_i) = p(u_k)$, $\nexists j : i < j < k$, $p(u_i) = p(u_j) = p(u_k)$. Here, $e(u_i)$ and $e(u_k)$ is the emotions of two consecutive utterances u_i and u_k, respectively. The more frequent the emotional shift occur in a conversation, the more it is considered difficult. Therefore, the difficulty of i-th conversation c_i is as follows:

$$\mathcal{DIF}(c_i) = \frac{N_{shift}(c_i) + N_{sp}(c_i)}{N_u(c_i) + N_{sp}(c_i)} \quad (8)$$

where $N_{shift}(c_i)$ and $N_u(c_i)$ represent the number of emotional shift in conversation c_i and the total number of utterances in c_i, respectively. $N_{sp}(c_i)$ is the number of speakers appearing in conversation c_i and acts as a smoothing factor. The algorithm for calculating the difficulty of the conversation is fully described in the Algorithm 1.

Algorithm 1 CL Training with DMF

Input: $\mathcal{D}$ - training dataset; M - training model
k - number of buckets in baby step scheduler
$\mathcal{DIF}$ - difficulty measure function
t - number of epochs; n - number of utterances
e - the emotion label of the utterances
$p(u_i)$ - the speaker's corresponding utterance u_i

S - Set containing the emotion sequence of speakers; $S[p][i]$ represents the emotion in the i-th utterance of speaker p
Output: M^* - the optimal model
$S = \emptyset$, $N_{es} = 0$
for $i = 1$ to n **do**
  $S[p[i]] \leftarrow S[p[i]] \cup \{e[i]\}$
end for
N_{sp} = length(S)
for $p \in S$ **do**
  for $i = 1$ to length($S[p]$) $- 1$ **do**
    if $S[p[i]] \neq S[p[i+1]]$ **then**
      $N_{shift} \leftarrow N_{shift} + 1$
    end if
  end for
end for
$\mathcal{DIF} = \frac{N_{shift}+N_{sp}}{n+N_{sp}}$
$\mathcal{D}\prime = sort(\mathcal{D}, \mathcal{DIF})$
$\mathcal{D}\prime = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^k\}$ where $\mathcal{DIF}(d_a) < \mathcal{DIF}(d_b), d_a \in \mathcal{D}^i, d_b \in \mathcal{D}^j, \forall i < j$
$\mathcal{D}^{train} = \emptyset$
for $i = 1$ to t **do**
  if $i \leq k$ **then**
    $\mathcal{D}^{train} = \mathcal{D}^{train} \cup D^i$
  end if
  TRAIN($M, \mathcal{D}^{train}$)
end for
return $M*$

Model	IEMOCAP								MELD	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
bc-LSTM (Poria et al., 2017)	33.82	78.76	56.75	64.35	60.25	60.75	60.51	60.42	59.62	57.29
MFN (Zadeh et al., 2018)	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	60.80	57.80
ICON (Hazarika et al., 2018)	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	58.20	56.30
DialogueRNN (Majumder et al., 2019)	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	60.31	57.66
DialogueGCN (Ghosal et al., 2019)	**51.57**	80.48	57.69	53.95	72.81	57.33	63.22	62.89	58.62	56.36
DAG-ERC (Shen et al., 2021)	47.59	79.83	69.36	66.67	66.79	**68.66**	67.53	68.03	61.04	63.66
MMGCN (Hu et al., 2021)	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	60.42	58.31
CTNet (Lian et al., 2021)	51.3	79.9	65.8	67.2	**78.7**	58.8	68.0	67.5	62.0	60.5
DAG-ERC+HCL (Yang et al., 2022)	-	-	-	-	-	-	-	68.73	-	63.89
COGMEN (Joshi et al., 2022)	-	-	-	-	-	-	68.2	67.6	-	-
MultiDAG (Ours)	49.65	79.83	66.40	67.59	71.78	67.90	68.30	68.45	64.29	63.87
MultiDAG+CL (Ours)	45.26	**81.40**	**69.53**	**70.33**	71.61	66.94	**69.11**	**69.08**	**64.41**	**64.00**

Table 1: Performance of approaches on IEMOCAP and MELD datasets. **Bold** indicates the highest performance, and underlining denotes the second-highest. "-" represents missing values due to unavailability in original papers.

2.2.2. Training Scheduler

The training scheduler is used to organize and schedule the training process by arranging conversations. Specifically, the dataset $\mathcal{D}$ is divided into multiple different bins $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$, where conversation with similar difficulty are grouped into the same bin. The training process starts with the easiest bin. After training for a certain number of epochs, the next bin is mixed into the current training dataset. Finally, once all bins have been mixed and used, additional epochs of training are performed.

3. Experimental Setup

3.1. Datasets and Baselines

We evaluate our approach on the following two ERC datasets: IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). The detailed statistics of the datasets are reported in Table 2. For the data processing, we use the same split as the work in (Hu et al., 2021). We compare our method against several state-of-the-art baselines, including unimodal and multimodal learning approaches. (Due to the space limit, they are brief described in Section 1). The evaluation metrics used are Accuracy (Acc.) and weighted average F1-score (w-F1).

Datasets	Conversation			Utterances			Avg.
	Train	Valid	Test	Train	Valid	Test	utterances
IEMOCAP	120		31	5810		1623	66.8
MELD	1038	114	280	9989	1109	2610	9.57

Table 2: Dataset statistics

3.2. Implementation Details

We perform hyperparameter tuning for our proposed model on each dataset using hold-out validation with separate validation sets. For the IEMOCAP dataset, the hyperparameter configuration includes a learning rate of 0.0005, a dropout rate of 0.4, 30 epochs of training, and 4 layers of MultiDAG+CL. For the MELD dataset, the hyperparameter configuration for the MultiDAG+CL model is as follows: a learning rate of 0.00001, a dropout rate of 0.1, 60 epochs of training, and 2 layers of Multi-DAG.

4. Results and Analysis

4.1. Comparision with Baselines

We conducted a comprehensive comparison of our proposed approach with SOTA multimodal ERC methods, and the results are summarized in Table 1. Due to space constraints, we only report Acc. and w-F1 for the MELD dataset. Our approach, *MultiDAG+CL*, which combines the *MultiDAG* model with a curriculum learning strategy, achieves SOTA performance on both the IEMOCAP and MELD datasets. *MultiDAG+CL* outperforms previous SOTAs by 1.05% (DAG-ERC on IEMOCAP) and 0.34% (DAG-ERC on MELD), respectively. Specifically, our models achieve improvements in individual emotion recognition tasks in most cases, especially for the *Sad*, *Neutral* and *Angry* emotions. In the meantime, we find *Happy*, *Sad*, and *Angry* emotions can be confused with the *Neutral* emotion in some cases (as shown in Fig. 2). Such phenomenon is related to imbalanced class distribution.

4.2. Effect of Modality

Table 3 compares the performance of MultiDAG and MultiDAG+CL under various multimodal settings on both benchmark datasets. In IEMOCAP, the textual modality performs best among the unimodal settings, while the visual modality shows the lowest results due to noise from factors like camera position and environmental conditions. In bimodal settings, the combination of textual and acoustic modalities performs the best, while the combina-

tion of visual and acoustic modalities yields the lowest result. Similar observations are made in the MELD dataset.

Modality	MultiDAG		MultiDAG+CL	
	IEMOCAP	MELD	IEMOCAP	MELD
T	68.17	63.66	67.12	63.47
A	49.37	40.27	50.58	40.17
V	33.79	31.27	36.69	31.27
T + A	68.42	63.61	68.45	63.56
T + V	67.56	63.69	67.40	63.62
A + V	52.40	40.54	51.86	39.99
T + V + A	**68.45**	**63.87**	**69.08**	**64.00**

Table 3: Results of MultiDAG and MultiDAG+CL under different modality settings. T, A, V represent the text, audio, visual modality, resepectively.

4.3. Effect of Curriculum Learning

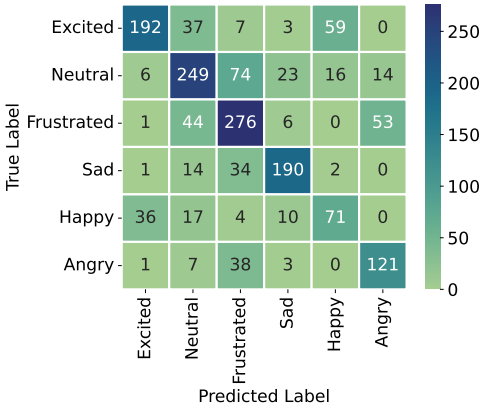
The MultiDAG+CL model demonstrates notable performance improvement by incorporating curriculum learning for both the IEMOCAP and MELD datasets. The effectiveness of curriculum learning relies on factors like the difficulty measure design and training strategy, including the number of buckets in the training set. We perform experiments to select the optimal number of buckets in the CL training scheduler. The results shown in the Table 4, indicate that for the IEMOCAP dataset, the optimal number of buckets is 5, while for the MELD dataset, it is 12. These findings suggest that the CL strategy is effective in improving the performance of the MultiDAG model on both datasets, with the specific number of buckets tailored to each dataset's representations. In summary, our proposed MultiDAG+CL model with curriculum learning, significantly contribute to the achieved results.

IEMOCAP		MELD	
Number of buckets	w-F1	Number of buckets	w-F1
4	68.05	5	63.94
5	**69.08**	8	63.83
7	68.84	10	63.89
10	68.38	12	**64.00**
15	68.36	14	63.96

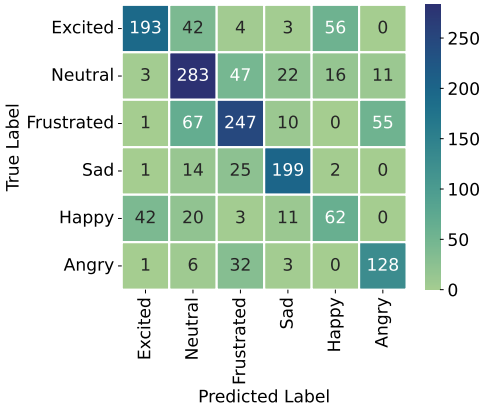
Table 4: Results of MultiDAG+CL for different number of buckets in CL training scheduler.

4.4. Performance for Emotion-shift

From the confusion matrices of the MultiDAG and MultiDAG+CL models (Figure 2), it can be observed that the prediction accuracy for the "Happy", "Neutral", "Sad", and "Angry" labels is improved when CL is incorporated into the model. Particularly, the misclassification rate of the "Neutral" label as "Disgust" decreases significantly from 19.3% in the MultiDAG model to only 12.3% in MultiDAG+CL. However, the prediction accuracy for the "Disgust" and "Happy" labels decreases.



(a) MultiDAG



(b) MultiDAG+CL

Figure 2: The confusion matrices on the IEMOCAP.

5. Conclusion

This paper proposes MultiDAG+CL, a novel approach for Multimodal Emotion Recognition in Conversation that leverages Directed Acyclic Graphs to integrate textual, acoustic, and visual features within a unified framework. The incorporation of Curriculum Learning (CL) addresses challenges related to emotional shifts and data imbalance, enhancing the model's performance. Through extensive experiments, we evaluate the performance of both the proposed MultiDAG+CL models. Future work includes exploring alternative training schedulers for Curriculum Learning and incorporating a learning-based approach to model emotion label similarity.

Acknowledgements

Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS147.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 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)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 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)*, pages 5666–5675, Online. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng. 2019. Towards discriminative representation learning for speech emotion recognition. In *IJCAI*, pages 5060–5066.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000.

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

Cam Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore. Association for Computational Linguistics.

Thomas S Polzin and Alexander Waibel. 2000. Emotion-sensitive human-computer interfaces. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network

for conversational emotion recognition. 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)*, pages 1551–1560, Online. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11595–11603.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Neurocomputing 623 (2025) 129342

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Mi-CGA: Cross-modal Graph Attention Network for robust emotion recognition in the presence of incomplete modalities

Cam-Van Thi Nguyen [a], Hai-Dang Kieu [a,b], Quang-Thuy Ha [a], Xuan-Hieu Phan [a], Duc-Trong Le [a,*]

[a] University of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam
[b] College of Engineering and Computer Science, VinUniversity, Hanoi, Viet Nam

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Modality incompleteness
Graph Attention Network
Cross-modality attention

ABSTRACT

Multimodal Emotion Recognition in Conversation (Multimodal ERC) is crucial for understanding human communication across various applications. However, the challenge of missing modalities impedes the development of robust models. Existing approaches often overlook scenarios where multiple modalities are absent simultaneously and fail to explore deep semantic interactions between modalities. Additionally, learning high-dimensional interactive features from limited samples is challenging due to missing data. This paper proposes Mi-CGA, a framework tailored for incomplete multimodal learning in conversational contexts. Mi-CGA comprises two main components: Incomplete Multimodal Representation (IMR) and Cross-modal Graph Attention Network (CGA-Net). IMR simulates incomplete modalities, while CGA-Net extracts rich information from conversational graphs. CGA-Net consists of three key modules: Modality Feature Estimation reconstructs missing data, Multi-head Graph Attention Network enhances utterance-level representation, and Cross-modal Attention Network improves conversation-level representation. Experimental results on three benchmark datasets (IEMOCAP, CMU-MOSI, and CMU-MOSEI) consistently demonstrate that Mi-CGA outperforms several representative baseline models, marking a significant advancement for the Multimodal ERC task. Source code for Mi-CGA is available at https://github.com/dangkh/Mi-CGA.

1. Introduction

Our interaction with the world involves multiple modalities such as sight, touch, hearing, smell, and taste, representing diverse sensory experiences [1]. Beyond sensory experiences, modalities extend to gestures, speech, written language, and gaze, facilitating information conveyance [2]. Human communication includes thoughts, sentiment, and emotions conveyed through language, voice, and facial expressions. The rise of automated Multimodal Emotion Recognition in Conversation (Multimodal ERC or MERC) and Multimodal Sentiment Analysis (MSA) tasks reflects the growing interest in enabling natural human–computer interactions (HCI) through machine learning advancements.

Compared to unimodal data, multimodal data offers the potential to capture diverse aspects of emotion and provides supplementary information, thereby significantly enhancing the performance of Emotion Recognition in Conversations (ERC) tasks [3,4]. However, in the realm of multimodal learning, the assumption of complete training data, where all modalities are consistently available in all training instances, is untenable in real-world scenarios. The prevalence of missing modalities constitutes a pervasive challenge, stemming from various factors, as illustrated in Fig. 1. Consequently, three fundamental technical challenges must be addressed under these circumstances. Firstly, certain previous approaches assume that only one modality can be missing and often overlook the practical scenario of multiple missing modalities simultaneously. Additionally, some methods miss the opportunity to leverage even limited remaining information to enhance the analysis of available modalities. Thus, more flexible methods are required to address varying rates of missing data [5,6]. Secondly, existing approaches often lack deep semantic interactions between modalities at the feature level. Additionally, the utilization of GNN-based models to represent and model these relationships remains underexplored. Lastly, the inconsistency in dimensions and varying feature sets is a common challenge when dealing with multimodal data containing different combinations of severely incomplete modalities [7]. It poses difficulties when attempting to apply complete multimodal fusion models [8–10],

* Corresponding author.
E-mail addresses: vanntc@vnu.edu.vn (C.-V.T. Nguyen), dangkh_uet@vnu.edu.vn (H.-D. Kieu), thuyhq@vnu.edu.vn (Q.-T. Ha), hieupx@vnu.edu.vn (X.-H. Phan), trongld@vnu.edu.vn (D.-T. Le).

https://doi.org/10.1016/j.neucom.2025.129342
Received 31 January 2024; Received in revised form 8 June 2024; Accepted 2 January 2025
Available online 7 January 2025

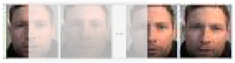
Modality	Demonstration	Possible Reasons
Text	…They act like they are too cool to talk to me…	• Unfamiliar terms • Automated speech recognition fault
Audio		• Background noise • Sensor failure
Video		• Face undetected • Fast motions

Fig. 1. Illustration of uncertain missing modalities in Multimodal Emotion Recognition Task.

which treat each independent multimodal instance within a unified architecture.

To tackle the aforementioned challenges, this study proposes **Mi-CGA**, an end-to-end graph neural network designed for handling incomplete modalities in the Multimodal ERC task. It has two main components namely Incomplete Multimodal Representation (IMR) and the "Cross-modal Graph Attention Network (CGA-Net)". The IMR component is to simulate the incomplete modality problems in conversations, which will be resolved by the CGA-Net component to maintain or improving the Multimodal ERC task. Specifically, the primary component CGA-Net comprises three crucial modules. As the first module, the "Modality Feature Estimation (FE)" module reconstructs missing features, ensuring a comprehensive data representation. The second module named "Graph Attention Network (GAT)" captures multimodal representations by prioritizing and weighting information based on attention scores, to enhance utterance-level representation. Equally important, the "Cross-modal Attention Network (CMA)" module facilitates inter-modalities information exchange, improving the model's ability to capture cross-modal relationships and enhancing its performance in Multimodal ERC task. The overall architecture of **Mi-CGA** is illustrated in Fig. 2. Our main contributions are summarized as follows:

- We propose Mi-CGA, a novel framework for robust Multimodal ERC in diverse incomplete modality scenarios, leveraging Graph Neural Networks and Cross-modal Attention mechanisms efficiently.
- We design the CGA-Net as a pivotal model component that optimizes information extraction from conversational graphs. Within the CGA-Net, the "Modality Feature Estimation (FE)" component addressing uncertain missing modalities using advanced Graph Neural Network techniques. Additionally, CGA-Net captures inter-modality information and enhances multimodal capabilities through cross-modal Attention mechanisms.
- Extensive experiments on benchmark datasets, including IEMOCAP, CMU-MOSI, and CMU-MOSEI, validate Mi-CGA's superior performance.

The paper is structured as follows: Section 2 provides a comprehensive overview of related works. Detailed insights into the proposed model are presented in Section 3. Experiment settings are thoroughly explained in Section 4, and the experimental evaluation, along with results, is extensively covered in Section 5. Finally, Section 6 wraps up the study by summarizing findings, discussing limitations and potential directions for future research.

2. Related work

In recent years, the field of Emotion Recognition in Conversations (ERC) has witnessed the emergence of a diverse range of effective models catering to both *unimodal* [11–14] and *multimodal* [15–18] data. It is noteworthy that all these models are designed under *the assumption of complete data availability*, implying the presence of all modalities for analysis. These models may encounter challenges and fail when confronted with scenarios involving missing modalities.

Substantial research efforts have been directed towards addressing the challenge of missing modalities in Multimodal Machine Learning. Simple strategies involve removing incomplete data samples or applying conventional imputation techniques like zero/average imputation and matrix completion [31–34]. However, these methods risk introducing extra noise into the original data, potentially causing performance degradation and may necessitate intricate auxiliary models like deep generative models. Recent approaches can be categorized into three groups: (1) *Data Augmentation*: Randomly removing inputs to simulate missing modalities, (2) *Generative Methods*: Predicting missing modalities based on available data, and (3) *Joint Multimodal Representations*: Learning representations that capture information from various modalities. The related works on missing/incomplete modalities, particularly those on MSA and Multimodal ERC, are presented chronologically in Table 1.

Data Augmentation: Parthasarathy and Sundaram [34] introduced a data augmentation strategy where visual inputs are randomly ablated during training to mimic real-world scenarios with missing modalities in audio–visual multimodal emotion recognition, improving recognition performance. Meanwhile, Wang et al. [25] proposed PANet and M2R2, which use iterative data augmentation to train emotion recognition models, focusing on learning a common representation in the presence of missing modalities at the utterance level.

Generative Methods: Earlier contributions to addressing the challenge of missing modalities in multimodal data include the CRA network by Tran et al. [19], which captures intricate relationships between different modalities. Zhang et al. [20] introduced a cross-partial multi-view network to impute missing views by learning latent multi-view representations with adversarial strategies. More recently, DiCMoR [29] reduces the distribution gap by transferring distributions from available to missing modalities, while IMDer [30] uses a score-based diffusion model to map input Gaussian noise into the desired distribution space of missing modalities.

Join Multimodal Representation Methods: Recent advancements include MMIN [5], who integrated CRA [19] to recover missing modalities, while Yuan et al. [23] introduced a Transformer-based feature reconstruction network aimed at guiding feature extraction for missing modalities. MTMSA [27] used a modality translation module to convert visual and auditory modalities into textual ones. GCNet [28] captured temporal and speaker-related information in incomplete conversational data.

While some models excel in tasks such as Multimodal Sentiment Analysis (MSA), they may not be directly applicable to Multimodal Emotion Recognition (multimodal ERC) tasks [21,23,26,29,30]. Additionally, certain approaches overlook random feature missing rates within modalities or only consider the complete absence of entire modalities [5,6]. Moreover, despite the recent effectiveness of GNNs-based models in various tasks, their utilization in missing data reconstruction remains relatively limited [28,35]. In summary, our model, Mi-CGA, employs a *joint learning approach*, utilizing textual, visual, and audio modalities concurrently to enhance Multimodal ERC even in the presence of incomplete data. It effectively tackles varying missing data rates and leverages the strengths of GNN structures suited for conversational data.

3. Methodology

3.1. Problem definition and notations

Let us consider a conversation C of a sequence of N utterances $\{U_1, U_2, \ldots, U_N\}$. Each utterance U_i in C associates with three data modalities including audio (a), visual (v), and text (t), i.e., $U_i = \{u_i^{(a)}, u_i^{(v)}, u_i^{(t)}\}$. Suppose that there exists an overall random missing rate

Table 1
A chronological summary of related works on missing modalities.

Model	Approach	Main technique	Studied problem	Modality missing	Datasets	Advantages	Disadvantages
CRA [19]	Generative	Autoencoder	Imputation	Uncertain missing	GRSS, RGB-D, MTPIE, HFSD	Offers a data imputation method that harnesses the advantages of both autoencoder networks and residual learning	No clear proof for this model's suitability in MERC.
CPM-Nets [20]	Generative	GANs	Multi-view Learning	Arbitrary view-missing	Hand-written, Animal, CUB, ADNI, etc.	Simultaneously leverages all samples and views, and is adaptable to arbitrary view-missing patterns.	Cannot be utilized for MERC and only use the visual modality.
MCTN0 [21]	Join learning	RNNs	MSA	Uncertain missing	CMU-MOSI, ICT-MMMO, Youtube	Offers a way to learn joint representations with input coming just from the source modality.	No clear proof for this model's suitability in MERC.
MeLIM [6]	Generative	Generative network	Metric learning	Uncertain missing	ADNI	Integrate metric learning with the data generating process to address the missing modality problem in patient similarity analysis.	Consider only incomplete pairwise modalities. Healthcare domain application.
HGMF [22]	Join learning	Graph-based transductive learning	Multimodal analysis	Uncertain missing	ModelNet40, NT, IEMOCAP	Take advantage of a heterogeneous hypernode graph structure to capture interactions from incomplete modalities	Conduct binary classification task on only 3 emotion labels. Not compatible with MERC
TFR-Net [23]	Join learning	Transformer	MSA	Uncertain missing	CMU-MOSI, CMU-MOSEI	Enhances models' robustness to random missing in non-aligned modality sequences.	No clear proof for this model's suitability in MERC.
MMIN [5]	Join learning	CRA	MERC	Uncertain missing	IEMOCAP, MSP-IMPROV	Predicts the presence of any missing modality based on the available modalities, considering various scenarios of missing conditions.	Consider only the complete absence of modalities.
SMIL [7]	Generative	Bayesian, Meta Learning	Inference Missing Modality	Severely Missing	CMU-MOSI, MM-IMDb, avMNIST	Proposes multimodal learning with severely missing modality that leverages Bayesian Meta Learning.	Only considering incomplete pairwise modalities.
TATE [24]	Join learning	Transformer	MSA	Uncertain missing	IEMOCAP, CMU-MOSI	Designs a tag encoding module that addresses scenarios when there is a single modality or multiple ones are missing.	No clear proof for this model's suitability in MERC.
M2R2 [25]	Data augmentation	Bidirectional GRU, CPM-Nets	MERC	Uncertain missing	IEMOCAP, MELD	Train an ERC model through iterative data augmentation, enhancing its performance by learning a shared representation.	Missing modalities at the utterance level.
MM-Align [26]	Join learning	Optimal Transport, Meta Learning	MSA	Severely missing	CMU-MOSI, CMU-MOSEI	Teaches the alignment dynamics between temporal modality for the inference in the event of lacking modality sequences by applying optimal transport.	Only considering incomplete pairwise modalities.
MTMSA [27]	Join learning	Transformer, Modality Translation	MSA	Uncertain missing	CMU-MOSI, IEMOCAP	Employ a modality translation module to translate the visual and auditory modalities into the textual modality.	No clear proof for this model's suitability in MERC.
GCNet [28]	Join learning	GNNs	MERC	Uncertain missing	IEMOCAP, CMU-MOSI, CMU-MOSEI	Designs the framework to capture temporal and speaker information in the incomplete conversational data.	Reconstruction loss based on MSE might easily lead to overfitting of the model.
DiCMoR [29]	Generative	Distribution transfer	MSA	Uncertain missing	CMU-MOSI, CMU-MOSEI	Transferring distributions from available modalities to missing modalities reduces the distribution gap between them.	No clear proof for this model's suitability in MERC.
IMDer [30]	Generative	Score-based diffusion	MSA	Uncertain missing	CMU-MOSI, CMU-MOSEI	Uses a score-based diffusion model to map input Gaussian noise into the distribution space of missing modalities	No clear proof for this model's suitability in MERC.

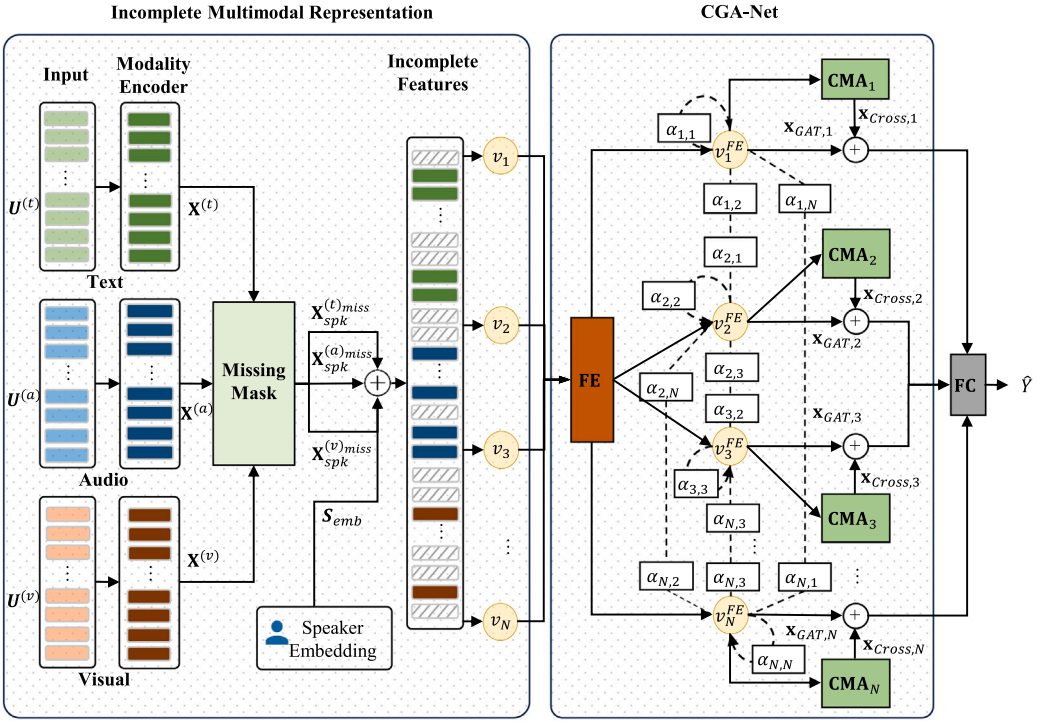


Fig. 2. This figure illustrates the Mi-CGA model's process, starting with Incomplete Multimodal Representation (IMR), where independent representations of three modalities undergo various missing scenarios and are concatenated to form Incomplete Features for each utterance. The incomplete representation then passes through the CGA-Net module, comprising FE module for missing feature reconstruction, Multi-head Graph Attention Network (MulGAT) for enhancing utterance-level representation, and Cross-modal Attention Network (CMA) for integrating information across modalities. Finally, classification predicts the emotion.

of $\rho \in [0,1]$ on multimodal information, let us denote $x_i^{(a)miss} \in \mathbb{R}^{d_a}$, $x_i^{(v)miss} \in \mathbb{R}^{d_v}$, and $x_i^{(t)miss} \in \mathbb{R}^{d_t}$ are incomplete representation of $u_i^{(a)}, u_i^{(v)}$ and $u_i^{(t)}$ with latent dimensions d_a, d_t, d_v.

Considering the multimodal emotion recognition task in the presence of incomplete modalities, Mi-CGA takes the tuple $(x_i^{(a)miss}, x_i^{(v)miss}, x_i^{(t)miss})$ as input and seeks to predict the corresponding emotion label $\hat{y}_i$ of U_i from a predefined emotion label set $E = \{y_1, y_2, \ldots, y_{|E|}\}$.

3.2. Incomplete Multimodal Representation(IMR)

3.2.1. Unimodal encoder

The Unimodal Encoder generates utterance-level embeddings for each modality. For text, we employ a bi-directional Long Short-Term Memory network (biLSTM) to capture sequential contextual information. For visual and acoustic modalities, we adopt a Fully Connected Network to independently process and encode contextual features, following the approach outlined in [36].

To express the multimodal context-aware feature encoding for each utterance, we can adopt the following procedure:

$$x_i^{(a)} = W_e^a u_i^{(a)} + b_i^{(a)} \tag{1}$$

$$x_i^{(v)} = W_e^v u_i^{(v)} + b_i^{(v)} \tag{2}$$

$$x_i^{(t)} = [\overrightarrow{LSTM}(u_i^{(t)}, x_{i-1}^{(t)}), \overleftarrow{LSTM}(u_i^{(t)}, x_{i+1}^{(t)})] \tag{3}$$

where $x_i^{(a)} \in \mathbb{R}^{d_a}, x_i^{(v)} \in \mathbb{R}^{d_v}, x_i^{(t)} \in \mathbb{R}^{d_t}$ are context-aware representations for audio, visual and text modalities respectively, and d_a, d_v, d_t are latent dimensions; W_e^v, W_e^a are parameters to be learned.

For a given conversation C, the multimodal sequences denoted as $\mathbf{X} = [\mathbf{X}^{(a)}, \mathbf{X}^{(t)}, \mathbf{X}^{(v)}]$ could be constructed as:

$$\mathbf{X}^{(a)} = [x_1^{(a)}, x_2^{(a)}, \ldots, x_N^{(a)}] \tag{4}$$

$$\mathbf{X}^{(v)} = [x_1^{(v)}, x_2^{(v)}, \ldots, x_N^{(v)}] \tag{5}$$

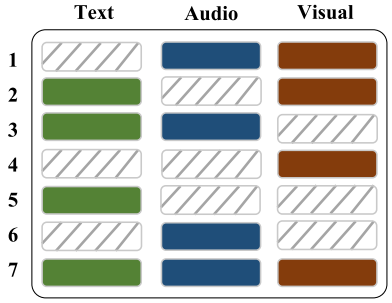


Fig. 3. Seven missing patterns for $M = 3$. Each row illustrates a missing pattern, in which a rectangle with diagonal lines implies the missing modality.

$$\mathbf{X}^{(t)} = [x_1^{(t)}, x_2^{(t)}, \ldots, x_N^{(t)}] \tag{6}$$

where $\mathbf{X}^{(a)} \in \mathbb{R}^{N \times d_a}, \mathbf{X}^{(v)} \in \mathbb{R}^{N \times d_v}, \mathbf{X}^{(t)} \in \mathbb{R}^{N \times d_t}$.

3.2.2. Missing mask generation

Similar to GCNet [28], we simulate the presence of incomplete modalities across the entire conversation with an overall missing rate ρ. It is important to ensure each training sample retains at least one available modality so that ρ is bounded in $[0, \frac{M-1}{M}]$, where M is the number of modalities. Fig. 3 shows a trimodal dataset ($M = 3$) with seven distinct missing patterns.

The missing mask matrix $\mathcal{M}$ is generated for the entire conversation C with a given ρ. Each utterance, having M modalities, simulates missing modalities by randomly selecting which ones will be missing based on a probability of ρ. This consistent missing rate is maintained throughout training, validation, and testing phases.

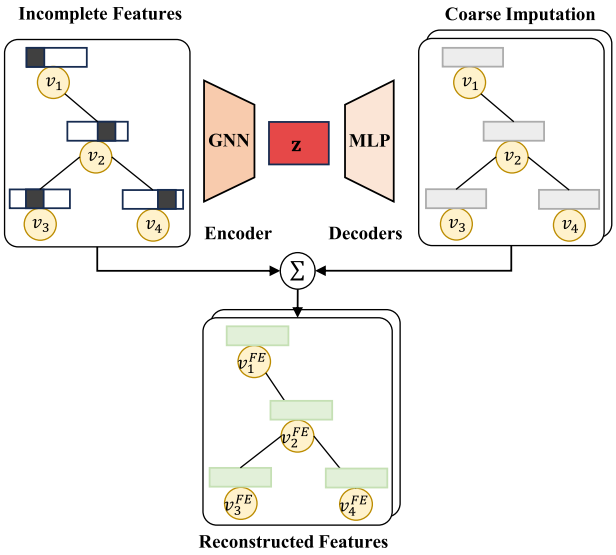


Fig. 4. Multimodal Feature Estimation Module (FE).

Let us denote $\mathbf{X}^{miss} = [\mathbf{X}^{(a)_{miss}}, \mathbf{X}^{(t)_{miss}}, \mathbf{X}^{(v)_{miss}}] \in \mathbb{R}^{N\times d}$ as the conversation-level representation in the presence of incomplete modalities with the missing rate ρ. It is inferred as follows:

$$\mathbf{X}^{miss} = \mathbf{X} \odot \mathcal{M} \tag{7}$$

where $\mathcal{M} \in \{0,1\}^{N\times d}$, $\mathbf{X} = [\mathbf{X}^{(a)}, \mathbf{X}^{(t)}, \mathbf{X}^{(v)}] \in \mathbb{R}^{N\times d}$ denotes the feature matrix representing all utterances in the input conversation, $d = d_a + d_t + d_v$ and $\odot$ is the element-wise multiplication.

3.2.3. Enhancing conversation-level representation with speaker embedding

Recent studies [36,37] validate the significance of the speaker information in improving utterance representations. Inspired by this finding, we employ a procedure **S-Emb** to generate latent representations based on the identities of speakers. Given a conversation C and its respective speaker set S, the embedding $S_{emb} \in \mathbb{R}^{N\times|S|}$ of the participants is:

$$S_{emb} = \mathbf{S-Emb}(S) \tag{8}$$

The enhanced conversation-level representation $\mathbf{X}^{miss}_{spk} \in \mathbb{R}^{N\times\bar{d}}$ ($\bar{d} = d + |S|$) by incorporating the corresponding speaker embeddings is as follows:

$$\mathbf{X}^{miss}_{spk} = \eta S_{emb} \oplus \mathbf{X}^{miss} \tag{9}$$

where $\eta \in [0,1]$ denotes the contribution ratio, $\oplus$ is a concatenation operation.

3.3. Cross-modal Graph Attention Network (CGA-net)

We build a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ to represent the multimodal data, in which each conversation is considered a fully connected graph and each utterance is a node $v_i \in \mathcal{V}$, v_i contain information of three modalities $v_i = [v_i^a, v_i^v, v_i^l]$. For edges construction, we assume that each utterance has the connections to others in the same dialogue. The connection of nodes is represented in form of an adjacency matrix $\mathbf{A}$.

3.3.1. Modality Feature Estimation (FE)

Interconnected nodes associating with neighboring utterances and sharing mutual modal information may exhibit an underlying resemblance, which could be useful to help reconstruct missing features. Motivated from this intuition, we propose a multimodal feature estimation module consisting of a GNN-based encoder f and a MLP-based decoder g as Fig. 4. The encoder f is to generate an embedding vector for each node in line with the prevailing topological relationships while the decoder g is to approximate the missing features. The coarse reconstructed representation $\mathbf{X}^{coarse} \in \mathbb{R}^{N\times\bar{d}}$ of the conversation C is as follows:

$$\mathbf{X}^{coarse} = g_\theta\big(f_\phi(\mathbf{X}^{miss}_{spk}, \tilde{\mathbf{A}})\big) \tag{10}$$

where ϕ,θ are learnable parameters of f,g. Specifically, we build the encoder network f_ϕ upon GGCN[1]) [38], $\tilde{\mathbf{A}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}$ is the symmetric normalization of $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, and $\mathbf{I} \in \mathbb{R}^{N\times N}$ being the identity matrix. For the decoder g_θ, we adopt a linear transformation as $g_\theta(z) = \mathbf{W}_\theta z + b_\theta$, where W_θ and b_θ are learnable weights and biases. Using Eq. (10), the missing positions corresponding to the generated missing mask are populated with values from the reconstructed matrix at those same positions.

Although the imputed features using neighbors' information might contain sufficient information, the reconstructed values and existing values in $\mathbf{X}^{coarse}$ maybe very distinguishable which could cause poor performance in the final classification. To overcome this problem, a smooth step is employed by adding a normalization layer to generate the final imputed conversation-level representation $\mathbf{X}^{FE} \in \mathbb{R}^{N\times\bar{d}}$ as follows:

$$\mathbf{X}^{FE} = Norm(\mathbf{X}^{coarse}, \mathbf{X}^{miss}_{spk}) \tag{11}$$

$$= (1-\lambda)\mathbf{X}^{coarse} + \lambda\mathbf{X}^{miss}_{spk} \tag{12}$$

where λ is the controlling hyperparameter.

To amplify the significance of the FE layer, we integrate a reconstruction objective function into the overall objective function, which will be presented in Section 3.5.

3.3.2. Multi-head Graph Attention Network (mulgat)

GNNs [39] involve information aggregation from neighbors and state updating to refine node states. A key challenge is accurately capturing complex relationships within the graph, especially when reconstructing missing features. Graph Attention Networks (GATs) [40] address this by using attention-based aggregation, assigning different importance to neighbors, which effectively focuses on relevant parts of the graph.

Inspired by GAT, we introduce a single-head graph attention sublayer, **S-GAT** for short, to help improve the utterance-level representation within a conversation as:

$$\hat{\mathbf{X}}^{FE} = \mathbf{S-GAT}(\mathbf{X}^{FE}, \Theta) \tag{13}$$

where $\hat{\mathbf{X}}^{FE} \in \mathbb{R}^{N\times\bar{d}}$ is the enhanced representation of $\mathbf{X}^{FE}$, and Θ is the parameter set of $\mathbf{S-GAT}$ to be learned.

For each node $v_i \in \mathcal{G}$, let us consider its respective feature vector $x_i \in \mathbb{R}^{\bar{d}}$, and neighbor set $\mathcal{N}_{(v_i)} = \{v_j \in \mathcal{V} | (v_j, v_i) \in \mathcal{E}\}$. The fundamental processes of **S-GAT** can be described as follows: *Aggregation*: The attention mechanism is employed to infer the aggregated representation $\mathbf{x}_{agg,i} \in \mathbb{R}^{\bar{d}}$ for each node v_i as:

$$\mathbf{x}_{agg,i} = \sum_{v_j\in\mathcal{N}_{(v_i)}} \alpha_{ij} W_{agg}\mathbf{x}_j \tag{14}$$

where α_{ij} signifies the attention coefficient or edge weight between node v_i and its neighboring node v_j; $W_{agg} \in \mathbb{R}^{\bar{d}\times\bar{d}}$ is a learnable parameter. Followed by GATv2s [41], we compute attention coefficients as the following:

$$\alpha_{ij} = \frac{\exp(e_{i,j})}{\sum_{v_j\in\mathcal{N}_{(v_i)}}\exp(e_{i,j})} \tag{15}$$

where $e_{i,j}$ measures the significance of the features of the neighbor v_j to the node v_i, which is computed as:

$$e_{i,j} = \Phi_{att}^{\top}\sigma\big(\Theta_{att}[\mathbf{x}_i \oplus \mathbf{x}_j]\big) \tag{16}$$

[1] It is no difficult to replace GCN with other graph neural networks

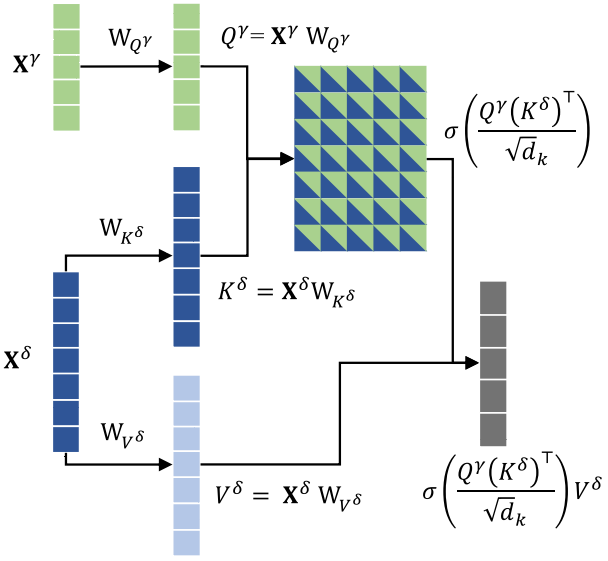


Fig. 5. Crossmodal attention between sequence $\mathbf{X}^{\delta}$ and $\mathbf{X}^{\gamma}$.

in which $\Phi_{att}, \Theta_{att} \in \mathbb{R}^{2\times\tilde{d}}$ are parameters to be learned; σ denotes a non-linear activation function, e.g., LeakyReLU; and $\oplus$ represents the concatenation operation.

State updating: The enhanced representation $\mathbf{x}_{com,i} \in \mathbb{R}^{\tilde{d}}$ of $\mathbf{x}_i$ by the weighted average of the aggregated representations of neighboring nodes as:

$$\mathbf{x}_{com,i} = \sigma\Big(\sum_{v_j \in \mathcal{N}_{(v_i)}} \alpha_{ij} W_{com} \mathbf{x}_{agg,j}\Big) \tag{17}$$

where α_{ij} is attention coefficient computed as Eq. (15); W_{com} is a parameter to be learned; and σ is a non-linear activation function, e.g., LeakyReLU. As a result, $\hat{\mathbf{X}}^{FE}$ is built via concatenating all $\mathbf{x}_{com,i}, i \in [1, N]$ of nodes, i.e., utterances, with their respective orders within the conversation C.

To stabilize the learning process of self-attention, we employ the multi-head attention mechanism (**MulGAT**) based on the concatenate strategy to derive the utterance-level representation in $\mathbf{X}^{FE}$. Formally, the enhanced representation $\mathbf{X}_{GAT}$ using **MulGAT** is inferred as follows:

$$\mathbf{X}_{GAT} = \mathbf{MulGAT}(\mathbf{X}^{FE}, \Psi) = \Psi\left[\hat{\mathbf{X}}_1^{FE}, \ldots, \hat{\mathbf{X}}_H^{FE}\right] \tag{18}$$

where H is the number of attention heads, $\hat{\mathbf{X}}_i^{FE}$ is the output from the ith $\mathbf{S}-\mathbf{GAT}$, and the bracket expresses the concatenation of single-head outputs, Ψ is a trainable parameter.

3.3.3. Cross-modal Attention Network (CMA)

Inspired by the decoder transformer in Neural Machine Translation [42], offering a latent adaptation across modalities is an effective strategy to fuse cross-modal information. Given two input modalities δ and γ, it is straightforward to extract the two respective uni-modal sequences denoted as $\mathbf{X}^{\delta} \in \mathbb{R}^{N\times d_{\delta}}$ and $\mathbf{X}^{\gamma} \in \mathbb{R}^{N\times d_{\gamma}}$. To capture interactions between every modality pair, Fig. 5 shows how we apply the cross-modal attention mechanism (CMA) to measure the transmitting information from modality δ to modality γ, which is denoted as "$\delta \to \gamma$". Specifically, the cross-modal attention $\mathbf{H}^{\gamma}$ from modality δ to modality γ is expressed through the CMA module as follows:

$$\mathbf{H}^{\gamma} = \mathbf{CMA}_{\delta\to\gamma}(\mathbf{X}^{\delta}, \mathbf{X}^{\gamma}) \tag{19}$$

$$= \text{Softmax}\left(\frac{Q^{\gamma}(K^{\delta})^{\top}}{\sqrt{d_k}}\right) V^{\delta} \tag{20}$$

$$= \text{Softmax}\left(\frac{\mathbf{X}^{\gamma} W_{Q^{\gamma}} (\mathbf{X}^{\delta})^{\top} (W_{K^{\delta}})^{\top}}{\sqrt{d_k}}\right) \mathbf{X}^{\delta} W_{V^{\delta}} \tag{21}$$

where we employ scaled dot-product attention from [42] with the *Queries* as $Q^{\gamma} = \mathbf{X}^{\gamma} W_{Q^{\gamma}}$, *Keys* as $K^{\delta} = \mathbf{X}^{\delta} W_{K^{\delta}}$, and *Values* as $V^{\delta} = \mathbf{X}^{\delta} W_{V^{\delta}}$; $W_{Q^{\gamma}} \in \mathbb{R}^{d_{\gamma}\times d_Q}$, $W_{K^{\delta}} \in \mathbb{R}^{d_{\delta}\times d_K}$, and $W_{V^{\delta}} \in \mathbb{R}^{d_{\delta}\times d_V}$ are trainable weights with d_Q, d_K, d_V representing the respective dimensions of queues, keys, and values; and $d_{(.)}$ denotes the feature dimension, $\sqrt{d_k}$ is a scaling factor; $\mathbf{H}^{\gamma} \in \mathbb{R}^{N\times d_V}$.

Likewise, we can also obtain the cross-modal attention $\mathbf{H}^{\delta} := \mathbf{CMA}_{\gamma\to\delta}(\mathbf{X}^{\gamma}, \mathbf{X}^{\delta}) \in \mathbb{R}^{N\times d_V}$. Finally, these cross-modal attention representations are concatenated to procedure the cross-modal attention representation for the entire dialogue denoted as $\mathbf{X}^{\delta\rightleftarrows\gamma} \in \mathbb{R}^{2\times d_V}$.

$$\mathbf{X}^{\delta\rightleftarrows\gamma} = [\mathbf{H}^{\gamma}, \mathbf{H}^{\delta}] \tag{22}$$

Revisiting the incomplete multimodal approach, we can compute the cross-modal attention representations for the conversation C as the following:

$$\begin{aligned}
\mathbf{X}_{Cross}^{a\rightleftarrows t} &= \mathbf{CMA}_{a\rightleftarrows t}(\mathbf{X}^{(a)FE}, \mathbf{X}^{(t)FE}) \\
\mathbf{X}_{Cross}^{t\rightleftarrows v} &= \mathbf{CMA}_{t\rightleftarrows v}(\mathbf{X}^{(t)FE}, \mathbf{X}^{(v)FE}) \\
\mathbf{X}_{Cross}^{v\rightleftarrows a} &= \mathbf{CMA}_{v\rightleftarrows a}(\mathbf{X}^{(v)FE}, \mathbf{X}^{(a)FE})
\end{aligned} \tag{23}$$

where $\mathbf{X}^{(a)FE} \in \mathbb{R}^{d_a}$, $\mathbf{X}^{(t)FE} \in \mathbb{R}^{d_t}$, and $\mathbf{X}^{(v)FE} \in \mathbb{R}^{d_v}$ present the feature representations of audio, textual, and visual modalities inferred from the FE module in Section 3.3.1. The aggregated cross-modal attention representation X_{Cross} of a given conversation C is computed as:

$$\mathbf{X}_{Cross} = [\mathbf{X}_{Cross}^{a\rightleftarrows t}, \mathbf{X}_{Cross}^{t\rightleftarrows v}, \mathbf{X}_{Cross}^{v\rightleftarrows a}] \tag{24}$$

where the bracket manifests the concatenation operation.

3.4. Emotion classification

To utilize both topology-dependent and cross-modal attention mechanism, we then concatenate all attention components at the final stage to generate the final feature representation as follows:

$$\mathbf{X}_{\text{Final}} = [\mathbf{X}_{GAT}, \mathbf{X}_{Cross}] \tag{25}$$

For each utterance $U_i \in C$, feature vector $\hat{x}_i \in \mathbf{X}_{\text{Final}}$ is fed through a fully-connected layer to predict the emotion label $\hat{y}_i$ of the utterance u_i:

$$\begin{aligned}
l_i &= \text{RELU}(W_l \hat{x}_i + b_l) \\
s_i &= \text{Softmax}(W_s l_i + b_s) \\
\hat{y}_i &= \underset{k}{\text{argmin}}(s_i[k])
\end{aligned} \tag{26}$$

where $W_l \in \mathbb{R}^{\tilde{d}\times\tilde{d}}, b_l \in \mathbb{R}^{\tilde{d}}, W_s \in \mathbb{R}^{|E|\times\tilde{d}}, b_s \in \mathbb{R}^{|E|}$ are parameters to be learned.

3.5. Model training

Here, we propose a dual-loss objective function to concurrently minimize both classification and reconstruction loss as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rct} \tag{27}$$

where $\mathcal{L}_{cls}$ is the classification objective function with $y_i \in \{0,1\}^{|E|}$ is the ground-truth one-hot label, we have:

$$\mathcal{L}_{cls} = -\frac{1}{L}\sum_{i=1}^{L} y_i log(\hat{y}_i) \tag{28}$$

Equally important, $\mathcal{L}_{rct}$ is referred to the reconstruction objective function. Here, we aim to exploit two approaches for computing $\mathcal{L}_{rct}$ including Mean Square Error (MSE) versus Kullback–Leibler divergence (KL-divergence). The $\mathcal{L}_{MSE}$ is computed as follows:

$$\mathcal{L}_{MSE} = \frac{1}{\| X \|}\sum^{\|X\|}(\mathbf{X} - \mathbf{X}^{FE})^2 \tag{29}$$

where $\mathbf{X}, \mathbf{X}^{FE}$ denote original feature matrix and imputed feature matrix respectively; $\| X \|$ denoted for size of feature vector X. As another

Algorithm 1 Mi-CGA Training Procedure

Input: The training set $D = \{(x_i^{(t)}, x_i^{(a)}, x_i^{(v)}), y_i\}_{i=1}^{N}$, $m \in \{t, a, v\}$
Output: Prediction emotion label $\hat{y}$

Phase 1. Incomplete Multimodal Representation (IMR)
Encode unimodal to multimodal feature $\mathbf{X}$ as Eq. (1)–(6)
Create a missing mask matrix $\mathcal{M}$ with a missing ratio ρ.
Create a conversation-level representation $\mathbf{X}^{miss} = \mathbf{X} \odot \mathcal{M}$ with missing ratio ρ as Eq. (7).
Enhance conversation-level representation with speaker embedding $\mathbf{X}_{spk}^{miss}$ as Eq. (8)–(9)
for each training epoch **do**
  for batch in dataLoader **do**
    Phase 2. Cross-modal Graph Attention Network (CGA-Net)
    Create reconstructed representation $\mathbf{X}^{coarse} = g_\theta\left(f_\phi(\mathbf{X}_{spk}^{miss}, \tilde{\mathbf{A}})\right)$ using encoder f and decoder g as Eq. (10)
    Generate the final feature estimated representation $\mathbf{X}^{FE}$ with smoothing factor λ as Eqs. (11)–(12)
    Improve utterance-level representation $\mathbf{X}^{FE}$ to $\hat{\mathbf{X}}^{FE}$ using sublayer **S-GAT** as Eqs. (13)–(17)
    Create enhanced representation $\mathbf{X}_{GAT}$ using multi-head attention mechanism **MulGAT** as Eq. (18)
    Compute the cross-modal attention representation $\mathbf{X}_{Cross}$ using Cross-modal Attention (CMA) as Eqs. (23)–(24)
    Phase 3. Emotion Classification
    Generate the final feature representation at the final stage by the concatenation operation: $\mathbf{X}_{\text{Final}} = [\mathbf{X}_{GAT}, \mathbf{X}_{Cross}]$
    Calculate the predicted emotion label $\hat{y}$ as Eq. (26)
    Phase 4. Model Optimization
    Create dual-loss objective function to concurrently minimize both classification and reconstruction $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rct}$ as Eqs. (27)–(30)
    Update all network parameters using BP algorithm
  end for
end for

option, $\mathcal{L}_{KL}$ is estimated as:

$$\mathcal{L}_{KL} = D_{KL}(p \parallel \hat{p}) = \sum_{i \in \hat{p}} p log \frac{p}{\hat{p}_i} + (1-p) log \frac{1-p}{1-\hat{p}_i} \tag{30}$$

where p is sparsity parameter and $\hat{p} = \text{Sigmoid}(\mathbf{X}^{FE})$ is expected as the fraction of activation of neurons. Similar to the idea of Sparse Auto-Encoder [43], instead of reconstructing whole features that could lead to over-fitting, we assume that reconstructed $\hat{p}$ should only recover the fraction of activation of neurons. Therefore, we desire to minimize the difference between two distributions is measured by the KL-divergence.

The entire calculation process of the Mi-CGA model is described in Algorithm 1.

4. Experimental setup

4.1. Datasets

We consider three real-life dataset namely IEMOCAP [44], CMU-MOSI [45], and CMU-MOSEI [46]. Consistent with prior research [5, 28], our primary focus is on the Multimodal ERC task.

To ensure a fair comparison, we adopt two prevalent label processing methods in IEMOCAP dataset: the four-class (4-way) [11,47] and the six-class (6-way) [48,49]. For CMU-MOSI and CMU-MOSEI, we address the positive/negative classification problem, where scores < 0 are negative and scores > 0 are positive. The datasets are split into train and test sets with an 8:2 ratio. The data distribution is shown in Table 2.

IEMOCAP [44] is a multimodal ERC dataset of 10,000 videos capturing actors' emotional conversations, labeled with one of six emotions: *happy, sad, neutral, angry, excited, frustrated*. For clarity, pairs like (*happy, excited*) and (*sad, frustrated*) are combined, creating a 4-way dataset.

Table 2
Data Statistics.

Datasets	Dialogues			Utterances		
	Train	Valid	Test	Train	Valid	Test
IEMOCAP (6-way)	120		31	5,810		1,623
IEMOCAP (4-way)	120		31	4,290		1,241
CMU-MOSI	52	10	31	1,284	229	686
CMU-MOSEI	2,248	300	676	16,326	1,871	4,659

CMU-MOSI [45] is a multimodal sentiment analysis dataset with 2199 short video snippets from 93 YouTube movie reviews, each labeled with a sentiment score from −3 to +3.

CMU-MOSEI [46] is an extended version of MOSI, featuring 22,856 annotated clips with sentiment and emotion labels, offering more diverse samples, speakers, and topics.

4.2. Feature extraction

We analyze three modalities: acoustic, lexical, and visual. To ensure a fair comparison, we use the multimodal feature extraction process from GCNet [28].[2] The extraction process is as follows:

Text Modality: We use the pre-trained DeBERTa-large model [50],[3] which improves upon BERT [51] and RoBERTa, to extract 1024-dimensional lexical features for each utterance.

Audio Modality: Acoustic feature extraction relies on the pre-trained wav2vec model [52]. Specifically, we use the pre-trained *wav2vec-large*[4] to extract 512-dimensional acoustic representations for each utterance.

Visual Modality: We use the pre-trained MA-Net [53][5] for visual feature extraction. Aligned faces are obtained through MTCNN [54], and average encoding is applied to condense features into 1024-dimensional representations for each utterance.

4.3. Baselines

To comprehensively assess Mi-CGA's performance, we conducted a thorough comparison with various baselines and state-of-the-art (SOTA) models in incomplete multimodal ERC. These include CPM-Net [20], AE [55], CRA [19], MMIN [5], GCNet [28], DiCMoR [29], and IMDer [30]. Brief descriptions of these baselines are provided in the Related Works section.

4.4. Evaluation strategy

For each testing dialogue, Mi-CGA and baseline models are required to generate the predicted emotion label for every single utterance within the conversation. Similar to related baselines, the accuracy (Acc.) and weighted-F1 score (w-F1.) are used in our experiment as the evaluation metrics.

4.5. Implementation details

We utilize Adam optimizer with a learning rate of 0.003 and weight decay of $1e^{-5}$ with a number of epochs is 200. All experiments are conducted on a machine with NVIDIA RTX 3060Ti with 12 GB of memory. For the structure of Mi-CGA, we stack 2 layers of GAT along with 4-head attentions. The coefficient λ in Eqs. (11)–(12) and p is set to default as 0.5 and 0.2 respectively.

[2] https://github.com/zeroQiaoba/GCNet
[3] https://huggingface.co/microsoft/deberta-large
[4] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec
[5] https://github.com/zengqunzhao/MA-Net

Table 3
Comparison with existing works for various missing rates. We report the weighted average F1 (w-F1) scores in percentages, where a higher w-F1 indicates superior performance. The best result is indicated in **bold**, while the second-highest one is marked with the underline. The row Δ quantifies the improvements of Mi-CGA over the second-highest model.

Dataset	Models	Missing rates								
		0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	Average
IEMOCAP (4-way)	CPM-Net	58.00	55.29	53.65	52.52	51.01	49.09	47.38	44.76	51.46
	AE	74.82	71.36	67.40	62.02	57.24	50.56	43.04	39.86	58.29
	CRA	76.26	71.28	67.34	62.24	57.04	49.86	43.22	38.56	58.23
	MMIN	74.94	71.84	69.36	66.34	63.30	60.54	57.52	55.44	64.91
	GCNet	78.36	77.48	77.34	76.22	75.14	73.80	71.88	71.38	75.20
	Mi-CGA	**83.42**	**82.83**	**82.27**	**81.50**	**83.17**	**80.08**	**79.96**	**79.35**	**81.50**
	Δ	5.06	5.35	4.93	5.28	8.03	6.28	8.08	7.97	6.30
IEMOCAP (6-way)	CPM-Net	41.05	37.33	36.22	35.73	35.11	33.64	32.26	31.25	35.32
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	56.96	53.94	51.46	48.42	45.60	42.82	40.18	37.84	47.15
	GCNet	58.64	58.50	57.64	57.08	56.12	54.40	53.60	53.46	56.18
	Mi-CGA	**66.04**	**65.83**	**64.07**	**63.08**	**61.72**	**59.96**	**59.52**	**59.18**	**62.65**
	Δ	7.36	7.33	6.43	6.00	5.60	5.56	5.92	5.72	6.47
CMU-MOSI	CPM-Net	71.90	68.91	71.12	70.59	64.95	65.88	64.02	61.79	67.77
	AE	56.76	52.82	48.66	42.26	35.18	29.12	25.08	23.18	39.13
	CRA	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN	85.20	81.91	78.22	74.60	70.14	67.72	64.04	61.53	72.92
	GCNet	85.01	82.54	80.17	78.54	76.48	73.45	69.46	68.35	76.75
	DiCMoR	85.60	83.90	82.00	80.20	77.70	76.40	73.00	70.08	78.70
	IMDer	85.60	84.80	**83.40**	81.00	78.50	75.90	74.00	71.20	79.30
	Mi-CGA	**87.21**	**85.02**	83.28	**81.83**	**79.56**	**78.62**	**75.63**	**73.05**	**80.05**
	Δ	1.61	0.22	−0.12	0.83	1.06	2.22	1.63	1.85	0.75
CMU-MOSEI	CPM-Net	78.47	74.79	74.48	73.81	72.39	70.43	68.73	67.07	72.52
	AE	86.66	84.37	82.58	80.57	78.80	76.43	74.26	72.81	79.56
	CRA	86.48	84.19	82.25	80.12	78.55	75.85	74.07	72.46	79.25
	MMIN	85.78	83.77	81.85	79.77	77.63	75.36	72.95	71.18	78.54
	GCNet	87.12	**86.50**	85.50	84.53	83.55	82.44	80.27	80.20	83.76
	DiCMoR	85.10	83.50	81.50	79.30	77.40	75.80	73.70	72.20	78.60
	IMDer	85.10	84.60	82.40	80.70	78.10	77.40	75.50	74.60	79.80
	Mi-CGA	**87.61**	86.21	**85.80**	**84.81**	**84.26**	**84.82**	**82.85**	**81.56**	**83.92**
	Δ	0.49	−0.29	0.30	0.28	0.71	2.38	2.58	1.36	0.16

5. Results and analysis

This section provides a comprehensive analysis and discussion of the experimental results. We conduct sufficient experiments to verify the following four research questions:

RQ1. **Comparison with SOTA baseline models** (Section 5.1): Does Mi-CGA demonstrate superior performance in terms of weighted average F1 (w-F1) compared to baseline models in Multimodal ERC?

RQ2. **Importance of the Modalities** (Section 5.2): How does the utilization of different modalities impact the performance and effectiveness of Mi-CGA with varying missing modality ratios?

RQ3. **Effects of MulGAT and CMA in CGA-Net** (Section 5.3): To what extent do the various modules in CGA-Net contribute to the overall performance?

RQ4. **Advantage of FE Module** (Section 5.4): What are advantages of the Feature Estimation (FE) module compared to naive imputation methods?

5.1. Performance comparison with SOTA baseline models (RQ1)

Table 3 shows the comparison results of Mi-CGA against the SOTA techniques in Multimodal ERC with feature incompleteness. Additionally, we provide the average performance across all missing rate ratios to offer a comprehensive assessment of our model's effectiveness.

Our proposed model, Mi-CGA, consistently outperforms competing approaches across multiple datasets. On the IEMOCAP (4-way) dataset, it shows a performance gain of 6.30% over GCNet in average w-F1 score. For IEMOCAP (6-way), Mi-CGA sets a new SOTA record with 62.65% accuracy, improving by 6.47% over the prior best-performing model (GCNet). Similar improvements are observed on the CMU-MOSI and CMU-MOSEI datasets, validating the robustness and effectiveness of our approach in incomplete multimodal learning for Multimodal ERC.

Delving further into the results associated with varying data incompleteness ratios, ranging from complete data ($\rho = 0.0$) to severe missing data ($\rho = 0.7$), reveals that our model demonstrates effectiveness in scenarios with both complete and incomplete modalities. Specifically, in the context of modality-complete data (i.e., $\rho = 0.0$), our model Mi-CGA consistently demonstrates significant improvements across all datasets, ranging from 0.49% (CMU-MOSEI) to 7.36% (IEMOCAP 6-way) compared to currently advanced approaches. This phenomenon is also observed in severely modality-incomplete data (i.e., $\rho = 0.7$), with improvements ranging from 1.36% (CMU-MOSEI) to 7.97% (IEMOCAP(4-way)) compared to other baseline models.

Moreover, the experimental results presented in Table 3 clearly demonstrate that Mi-CGA exhibits significantly smaller performance degradation compared to the baseline models as the missing rate increases. In IEMOCAP (6-way) dataset, the performance of the baseline models drop significantly, with declines ranging from 6.98% (GCNet) to 37.70% (CRA). In contrast, our Mi-CGA exhibits a much more modest decline, experiencing only a 4.07% decrease in performance. This trend holds across the remaining datasets, where our model also achieves comparable results with other currently advanced models.

Table 4
Results of modality ablation experiments on IEMOCAP dataset. Here, "A", "V", and "T" represent the audio, visual, and textual modalities, respectively. The best results are **bolded**, the second-highest result is denoted by the underline, and "-" signifies that no results were observed at that missing rate.

Modalities	IEMOCAP (6-way)							
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
A	51.33	–	–	–	–	–	–	–
V	46.02	–	–	–	–	–	–	–
T	63.76	–	–	–	–	–	–	–
A+V	54.96	53.98	51.69	50.82	52.23	51.71	–	–
A+T	65.67	64.92	63.25	61.13	60.53	**60.41**	–	–
V+T	65.39	65.37	63.28	60.02	59.91	59.63	–	–
A+V+T	**66.04**	**65.83**	**64.07**	**63.08**	**61.72**	59.96	**59.52**	**59.18**
Modalities	IEMOCAP (4-way)							
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
A	73.77	–	–	–	–	–	–	–
V	63.68	–	–	–	–	–	–	–
T	80.70	–	–	–	–	–	–	–
A+V	75.22	75.66	74.87	75.56	74.01	78.26	–	–
A+T	81.50	81.40	79.50	79.90	78.61	77.93	–	–
V+T	82.91	82.44	80.77	80.79	80.34	79.60	–	–
A+V+T	**83.42**	**82.83**	**82.27**	**81.72**	**82.37**	**80.08**	**79.96**	**79.35**

5.2. Importance of the modalities (RQ2)

In this experimental setup, three scenarios were considered: (**S1**) Exclusive use of a single modality for analysis, ensuring no missing modalities given the assumption detailed in Section 3.2 (i.e., $\rho = 0.0$); (**S2**) Utilization of any two modalities for emotion recognition (e.g., A+V, A+T, and V+T), with varying missing modality ratios from $\rho = 0.0$ to $\rho = 0.5$; (**S3**) Simultaneous employment of all three modalities (A+V+T), with missing modality ratios ranging from $\rho = 0.0$ to $\rho = 0.7$.

The results in Table 4 highlight the significance of modalities across three scenarios. On the IEMOCAP (6-way) dataset (**S1**), the text modality outperforms audio and visual, with w-F1 scores 12.43% and 17.74% higher, respectively. Similar trends are observed in the IEMOCAP (4-way) dataset, with the text modality surpassing audio and visual by

6.93% and 17.02%, respectively, affirming its prominent role in multimodal ERC.

In scenario **S2** on the IEMOCAP (6-way) dataset, combining the text modality with others leads to better performance compared to combinations without it. Specifically, omitting the text modality results in an average decrease of about 9.76% in the weighted F1 (w-F1) score compared to combinations lacking audio or visual. When analyzing the results in Table 4 for the IEMOCAP (6-way) dataset, especially as the missing rate increases from 0.0 to 0.5, we observe that the A+V combination experiences a smaller decline compared to A+T and V+T. The performance drop for A+V is 3.24%, while for A+T and V+T, it is 5.51% on average. This underscores the importance of the text modality, particularly in cases of severe missing data, where incomplete text leads to a significant overall performance drop. This trend is similarly observed in the IEMOCAP (4-way) dataset.

For the scenario **S3**, the best results emerge when all three modalities are employed concurrently, underscoring the synergistic contributions of these modalities in the multimodal ERC task across both datasets.

5.3. Effects of MulGAT and CMA in CGA-net (RQ3)

In these experiments, we created two model variants by selectively removing specific modules from Mi-CGA. The goal was to assess the effectiveness of these modules by evaluating the performance of the resulting model variants. The following model variants were generated: (1) $\mathbf{CGA}_{-\mathbf{G}}$, which omits the Multi-head Graph Attention Network from CGA-Net; (2) $\mathbf{CGA}_{-\mathbf{C}}$, which excludes the Cross-modal Attention Network from CGA-Net.

The module ablation experiment results, as presented in Table 5, offer insights into the performance of the different model variations. For the IEMOCAP (4-way) dataset, we observe that $\mathbf{CGA}_{-\mathbf{G}}$ experiences an average reduction of 0.91% across all missing rates, while $\mathbf{CGA}_{-\mathbf{C}}$ shows an average decrease of 0.38% across all missing rates. These values are slightly higher on the IEMOCAP (6-way) dataset, standing at 2.03% and 0.45%, respectively. Importantly, $\mathbf{CGA}_{-\mathbf{G}}$ consistently displays a lower overall decrease.

Similar trends are evident on both CMU-MOSI and CMU-MOSEI datasets. However, on the CMU-MOSEI dataset at missing rates of 0.4 and 0.6, $\mathbf{CGA}_{-\mathbf{G}}$ achieves the best results, although the improvement over Mi-CGA is not significantly substantial. As the data missing rates increase, the efficacy of cross-modality learning diminishes, as the aggregation of more information from extensively missing modalities may introduce heightened noise into the model.

5.4. Advantage of the multimodal FE module (RQ4)

Dealing with missing data poses a fundamental challenge in machine learning. Zero imputation, a simple method, involves replacing missing values with zeros. In graph contexts, a common approach is to impute missing information by borrowing from neighboring nodes [56]. Here, we compare our Feature Estimation (FE) approach against two baseline strategies: setting the global mean (**MeanImp**) as missing values and assigning missing features to 0 (**ZeroImp**). Fig. 6 shows the performance comparison between our FE and these imputation strategies.

In the CMU-MOSI dataset, our FE module maintains stable performance with increasing missing data levels. However, the ZeroImp strategy shows a notable performance drop from 73.11% to 53.32% (a decrease of 19.79%). Similarly, the MeanImp strategy also sees a decrease of about 19.21% as missing data levels rise. Filling missing values with zeros or global means cannot recover lost information, leading to significant performance degradation when learning from the remaining non-zero data points. Using neighboring node averages for imputation proves ineffective as the missing rate rises, as it still relies on zero-dominated averages. Overall, our FE approach consistently outperforms the naive imputation strategies and provides results that are significantly more competitive and reliable.

Table 5
Effectiveness of MulGAT and CMA. "$\mathbf{CGA}_{-G}$" and "$\mathbf{CGA}_{-C}$" denote the absence of the MulGAT and CMA modules in CGA-Net respectively. The best results are indicated in **bold**.

ρ	Module	IEMOCAP (4-way)	IEMOCAP (6-way)	CMU-MOSI	CMU-MOSEI
0.0	$\mathbf{CGA}_{-G}$	81.59	63.59	80.63	87.58
	$\mathbf{CGA}_{-C}$	83.21	65.52	78.51	**87.62**
	Mi-CGA	**83.42**	**66.04**	**87.21**	87.61
0.1	$\mathbf{CGA}_{-G}$	81.58	63.84	75.29	86.33
	$\mathbf{CGA}_{-C}$	82.74	65.26	76.15	86.37
	Mi-CGA	**82.83**	**65.83**	**85.02**	**86.50**
0.2	$\mathbf{CGA}_{-G}$	81.70	61.60	76.16	85.23
	$\mathbf{CGA}_{-C}$	82.20	63.90	76.65	85.25
	Mi-CGA	**82.27**	**64.07**	**83.28**	**85.93**
0.3	$\mathbf{CGA}_{-G}$	80.93	60.14	71.68	84.22
	$\mathbf{CGA}_{-C}$	81.47	61.79	74.02	**84.24**
	Mi-CGA	**81.50**	**63.08**	**81.09**	84.13
0.4	$\mathbf{CGA}_{-G}$	82.86	59.45	66.43	**83.31**
	$\mathbf{CGA}_{-C}$	82.95	61.57	69.64	83.00
	Mi-CGA	**83.17**	**61.72**	**79.32**	83.09
0.5	$\mathbf{CGA}_{-G}$	79.50	57.68	67.83	83.58
	$\mathbf{CGA}_{-C}$	79.61	59.88	71.37	83.66
	Mi-CGA	**80.08**	**59.96**	**79.83**	**83.69**
0.6	$\mathbf{CGA}_{-G}$	78.86	57.66	60.68	**80.78**
	$\mathbf{CGA}_{-C}$	78.99	58.63	64.19	80.48
	Mi-CGA	**79.96**	**59.52**	**75.10**	80.74
0.7	$\mathbf{CGA}_{-G}$	78.32	59.03	58.75	78.89
	$\mathbf{CGA}_{-C}$	78.35	59.01	61.41	79.33
	Mi-CGA	**79.35**	**59.18**	**69.57**	**79.66**

5.5. Ablation study

In this section, we perform ablation studies on the IEMOCAP dataset for both the (4-way) and (6-way) settings. The goal is to investigate the impact of the smooth factor in the Feature Estimation (FE) component and assess the effects of different loss functions in Mi-CGA. This can be expressed through the following questions:

Abl1. How does the smooth factor λ in the Feature Estimation (FE) module control the impact of estimated features?

Abl2. How do the investigations of different losses in Mi-CGA impact the model's performance?

5.5.1. Effects of smooth factor λ in FE (Abl1)

In this section, we examine how the smoothing factor (λ) influences the final estimated features ($\mathbf{X}^{FE}$) by bridging the gap between initial missing features ($\mathbf{X}^{miss}_{spk}$) and raw features estimated from neighboring nodes ($\mathbf{X}^{coarse}$). We vary λ from 0 (no smoothing) to larger values like 0.9, indicating more influence from the coarse estimated features. Table 6 shows the results on the IEMOCAP dataset for different λ settings.

In both the IEMOCAP (4-way) and IEMOCAP (6-way) datasets, we observe that the optimal value for the parameter λ, which maximizes the overall performance of the Mi-CGA model, is consistently 0.5 across various missing rates. This value signifies that the final estimated features leverage information equally from neighboring nodes and the original node features that were initially missing. It supplements information from neighboring nodes without entirely discarding the original node's information, indicating that our Feature Estimation (FE) module selectively augments missing features, striking a balance that enhances the model's performance.

5.5.2. Effect of different losses (Abl2)

To examine the impacts of different loss functions, we conduct experiments by substituting various loss functions and evaluating their effect on performance. Specifically, we compare the Mean Squared

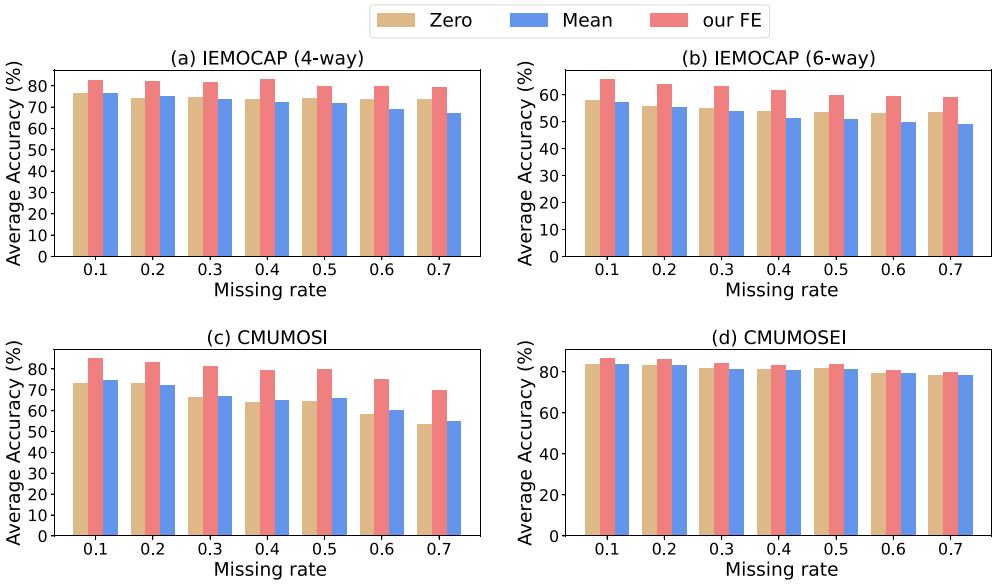


Fig. 6. Illustration of the robustness in performance of our proposed feature estimation against basis approaches in the different rate of missing in modalities. The performance of Mi-CGA goes along with the feature estimation module is represented in red bar that indicates a notable improvement in all datasets and various missing rates.

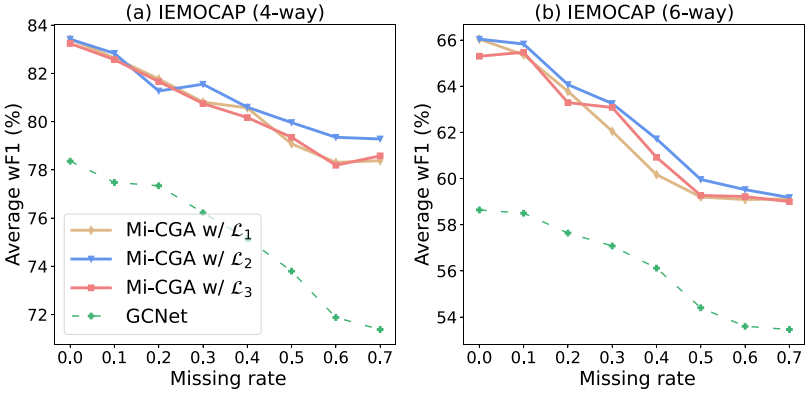


Fig. 7. Illustration of our Mi-CGA performance with different types of objective function on IEMOCAP datasets.

Table 6
An investigation of the impact of the smoothing factor λ. The best results are **bolded**, the second-highest result is denoted by the underline.

Settings	IEMOCAP (4-way)						
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
w/o smooth	81.68	82.10	80.50	**83.40**	**80.20**	74.51	79.50
$\lambda = 0.1$	81.61	82.19	**81.57**	82.68	79.88	75.05	79.73
$\lambda = 0.2$	82.12	81.98	80.44	82.25	79.77	75.92	78.51
$\lambda = 0.5$	**82.83**	**82.27**	81.50	83.17	80.08	**77.86**	**80.79**
$\lambda = 0.9$	82.58	81.87	80.73	81.38	78.79	76.75	79.61
Settings	IEMOCAP (6-way)						
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
w/o smooth	64.86	62.04	63.31	59.45	57.14	60.40	58.44
$\lambda = 0.1$	63.50	62.38	63.28	59.10	57.72	60.11	56.04
$\lambda = 0.2$	63.62	62.88	63.69	60.14	57.91	59.99	56.51
$\lambda = 0.5$	**65.83**	**64.07**	63.08	**61.72**	**59.96**	**61.32**	**59.18**
$\lambda = 0.9$	65.76	63.60	**63.93**	61.10	58.03	58.92	56.09

Error (MSE) loss and the Kullback–Leibler (KL) divergence loss, as described in Section 3.5. Additionally, we assess our model's performance against GCNet [28], a previous SOTA model that utilizes the MSE loss.

We assess our model with three reconstruction loss ($\mathcal{L}_{rct}$) settings: (1) $\mathcal{L}_1$: employing Mean Squared Error (MSE) loss ($\mathcal{L}_{cls} + \mathcal{L}_{MSE}$); (2) $\mathcal{L}_2$: utilizing KL-divergence loss ($\mathcal{L}_{cls} + \mathcal{L}_{KL}$); (3) $\mathcal{L}_3$: using only the classification loss $\mathcal{L}_{cls}$ without any reconstruction loss. Fig. 7 shows the performance comparison among different loss settings. KL-divergence demonstrates superior performance compared to the other two loss settings across both the IEMOCAP (4-way) and IEMOCAP (6-way) datasets, highlighting Mi-CGA's effectiveness in optimizing the specified loss function. This could be because the MSE loss strictly emphasizes exact reconstruction, while KL-divergence is more flexible in regularizing the similarity between value distributions.

We varied the sparse parameter p in the KL divergence to assess its impact on the effectiveness of KL as a reconstruction loss function in our model. Results from Fig. 8 show that incorporating $\mathcal{L}_{KL}$ as the reconstruction loss in Eq. (27) enhances the performance of our model across all missing cases in both IEMOCAP (4-way) and IEMOCAP (6-way) datasets compared to models without the reconstruction step. However, performance differences across different p coefficients are minimal.

6. Conclusion

In conclusion, this paper introduces Mi-CGA, a framework designed to address challenges caused by incomplete modalities in Multimodal Emotion Recognition in Conversation (ERC). The core of Mi-CGA is the Cross-modal Graph Attention Network (CGA-Net), which uses Graph

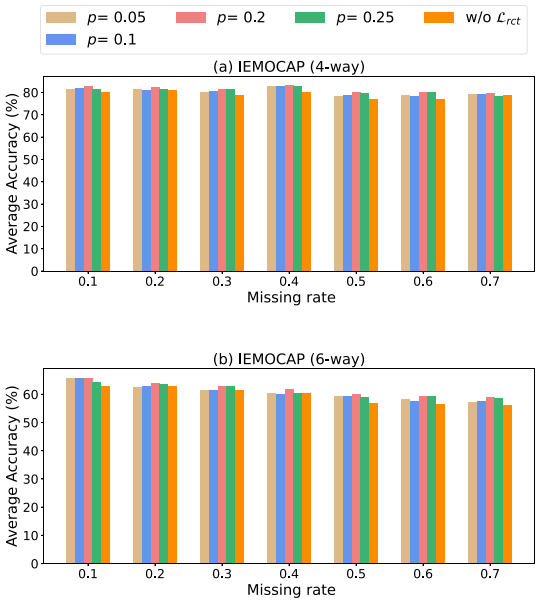


Fig. 8. A comparison of the impact of the p value in $\mathcal{L}_{rct}$. The default setting in Mi-CGA is $p = 0.2$. The results for different p values show minimal variation, but they consistently outperform the scenario without $\mathcal{L}_{rct}$.

Neural Networks (GNNs) and a Cross-modal Attention mechanism to extract detailed information from conversational graphs. Extensive experiments on benchmark datasets (IEMOCAP, CMU-MOSI, and CMU-MOSEI) show the model's effectiveness in improving Multimodal ERC despite incomplete modalities. A key challenge is the randomly generated Missing Mask, which makes it impossible to predict specific missing modalities or positions, only the overall percentage of missing data. While this simulates real-world data loss, it can negatively affect performance, especially when a modality is almost entirely masked. This impacts CGA-Net's ability to integrate information across modalities (CMA module) and from neighboring nodes (MulGAT module).

Future work can enhance Mi-CGA by: (1) Developing strategies to generate Missing Masks that balance information loss across modalities, optimizing which modalities and regions to mask; (2) Exploring hyperparameter optimization algorithms to systematically improve the model's robustness and reduce the computational burden, addressing the current limitations in real-time applications.

CRediT authorship contribution statement

Cam-Van Thi Nguyen: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Hai-Dang Kieu:** Writing – review & editing, Validation, Software, Methodology. **Quang-Thuy Ha:** Writing – review & editing, Supervision. **Xuan-Hieu Phan:** Writing – review & editing, Supervision. **Duc-Trong Le:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Duc-Trong Le reports financial support was provided by Vietnam National University Hanoi. Cam-Van Nguyen Thi reports financial support was provided by Vingroup Big Data Institute. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been done under the research project QG.23.37 "Research approaches on improving the reliability of deep learning classifiers" of Vietnam National University, Hanoi. Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS147.

Data availability

Data will be made available on request.

References

[1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[2] P. Grifoni, Multimodal human computer interaction and pervasive services, IGI Global, 2009.
[3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[4] K. Ezzameli, H. Mahersia, Emotion recognition from unimodal to multimodal analysis: A review, Inf. Fusion (2023) 101847.
[5] J. Zhao, R. Li, Q. Jin, Missing modality imagination network for emotion recognition with uncertain missing modalities, 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), 2021, pp. 2608–2618.
[6] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, A. Zhang, Metric learning on healthcare data with incomplete modalities., in: IJCAI, vol. 3534, 2019, p. 3540.
[7] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, X. Peng, Smil: Multimodal learning with severely missing modality, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 2302–2310.
[8] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114.
[9] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.
[10] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Bagher Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2247–2256.
[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Cputational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.
[12] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, 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, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 154–164.
[13] W. Shen, J. Chen, X. Quan, Z. Xie, Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 15, 2021, pp. 13789–13797.
[14] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, 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), Association for Computational Linguistics, Online, 2021, pp. 1551–1560.
[15] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2122–2132.
[16] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.
[17] Z. Lian, B. Liu, J. Tao, CTNet: Conversational transformer network for emotion recognition, IEEE/ACM Trans. Audio Speech, Lang. Process. 29 (2021) 985–1000.

[18] C.V.T. Nguyen, T. Mai, S. The, D. Kieu, D.-T. Le, Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15154–15167.
[19] L. Tran, X. Liu, J. Zhou, R. Jin, Missing modalities imputation via cascaded residual autoencoder, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1405–1414.
[20] C. Zhang, Z. Han, H. Fu, J.T. Zhou, Q. Hu, et al., CPM-nets: Cross partial multi-view networks, Adv. Neural Inf. Process. Syst. 32 (2019).
[21] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6892–6899.
[22] J. Chen, A. Zhang, Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1295–1305.
[23] Z. Yuan, W. Li, H. Xu, W. Yu, Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4400–4407.
[24] J. Zeng, T. Liu, J. Zhou, Tag-assisted multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1545–1554.
[25] N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, J. Zhang, M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation, IEEE Trans. Artif. Intell. (2022).
[26] W. Han, H. Chen, M.-Y. Kan, S. Poria, MM-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10498–10511.
[27] Z. Liu, B. Zhou, D. Chu, Y. Sun, L. Meng, Modality translation-based multimodal sentiment analysis under uncertain missing modalities, Inf. Fusion (2023) 101973.
[28] Z. Lian, L. Chen, L. Sun, B. Liu, J. Tao, GCNet: Graph completion network for incomplete multimodal learning in conversation, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[29] Y. Wang, Z. Cui, Y. Li, Distribution-consistent modal recovering for incomplete multimodal learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22025–22034.
[30] Y. Wang, Y. Li, Z. Cui, Incomplete multimodality-diffused emotion recognition, Adv. Neural Inf. Process. Syst. 36 (2024).
[31] L. Cai, Z. Wang, H. Gao, D. Shen, S. Ji, Deep adversarial learning for multi-modality missing data completion, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1158–1166.
[32] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (2010) 2287–2322.
[33] C. Zhang, Y. Cui, Z. Han, J.T. Zhou, H. Fu, Q. Hu, Deep partial multi-view learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2402–2415.
[34] S. Parthasarathy, S. Sundaram, Training strategies to handle missing modalities for audio-visual expression recognition, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020, pp. 400–404.
[35] N.V. Doan, T.D. Nguyen, D.T. Nguyen, C.-V.T. Nguyen, H.-D. Kieu, D.-T. Le, GAT-FP: Addressing imperfect multimodal learning using graph attention networks and feature propagation, in: Proceedings of the 12th International Symposium on Information and Communication Technology, 2023, pp. 327–334.
[36] J. Hu, Y. Liu, J. Zhao, Q. Jin, 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), Association for Computational Linguistics, Online, 2021, pp. 5666–5675.
[37] T. Kim, P. Vossen, EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa, 2021, CoRR, abs/2108.12009.
[38] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.
[39] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81.
[40] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, 2017, CoRR, abs/1710.10903.
[41] S. Brody, U. Alon, E. Yahav, How attentive are graph attention networks?, 2021, arXiv preprint arXiv:2105.14491.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[43] A. Ng, et al., Sparse autoencoder, in: CS294A Lecture notes, vol. 72, (no. 2011) 2011, pp. 1–19.
[44] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[45] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (2016) 82–88.
[46] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
[47] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting, vol. 2018, NIH Public Access, 2018, p. 2122.
[48] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6818–6825.
[49] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 164–172.
[50] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021.
[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[52] S. Schneider, A. Baevski, R. Collobert, M. Auli, Wav2vec: Unsupervised pre-training for speech recognition, in: Interspeech 2019, ISCA, 2019.
[53] Z. Zhao, Q. Liu, S. Wang, Learning deep global multi-scale and local attention features for facial expression recognition in the wild, IEEE Trans. Image Process. 30 (2021) 6544–6556.
[54] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[55] P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings, 2012, pp. 37–49.
[56] E. Rossi, H. Kenlay, M.I. Gorinova, B.P. Chamberlain, X. Dong, M.M. Bronstein, On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features, in: Learning on Graphs Conference, PMLR, 2022, pp. 11:1–11:16.

M.Sc Cam-Van Thi Nguyen received the B.S. degree (Hons.) in Information Technology (2017) and the M.S. degree in Information Systems (2019) from the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam (VNU-UET). She is currently working towards the Ph.D degree from VNU-UET. From 2017 to present, she has been working as a Researcher at Data Science and Knowledge Technology Laboratory (DS&KTLab), VNU-UET. Her current research interests include conversational AI, affective computing, deep learning, multimodal sentiment analysis and emotion recognition.

M.Sc Hai-Dang Kieu received the B.S. degree and M.S. degree in Computer Science from the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam (VNU-UET), in 2017 and 2021. He had worked as researcher at Bournemouth University, UK from 2019-2020. Now, he is pursuing Ph.D degree in a joint program at the VinUniversity and the University of Technology Sydney. He also works as a research assistant and teaching assistant at VinUniversity and VNU-UET. His topic of interests include: generative model, diffusion model, deep learning, multimodality learning.

Assoc. Prof. Ha Quang Thuy works for the Data Science and Knowledge Technology Laboratory (DS&KT Lab) and the Department of Information Systems, Faculty of Information Technology, VNU-University of Engineering and Technology (VNU-UET), Vietnam National University. He received his PhD degree at the Hanoi University of Science (HUS), VNU in 1997. His research interests are in Rough Sets, Description Logics, Lifelong Machine Learning, and Data Mining (Text - Web - Social Media Mining and Process Mining).

Assoc. Prof. Xuan-Hieu Phan received the B.S. and M.S. degrees in information technology and computer science from the College of Technology (Coltech, now UET), Vietnam National University in Hanoi (VNUH), in 2001 and 2003, respectively, and the Ph.D. degree in computer and information science from the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2006. From 2006 to 2008, he was a JSPS Postdoctoral Fellow at the Graduate School of Information Science (GSIS), Tohoku University. From 2008 to 2010, he was also a Research Fellow at the Centre for Health Informatics (CHI), University of New South Wales (UNSW), Sydney, NSW, Australia. He is currently an Associate Professor at VNUH-UET. His research interests include natural language processing, machine learning, information retrieval, web and text mining, and business intelligence.

Dr. Duc-Trong Le received the B.S. degree in Information Technology from University of Engineering and Technology, Vietnam National University, Hanoi (VNU-UET) in 2011 and the PhD degree in Information Systems from Singapore Management University in 2019. Currently, he is the Deputy Head of the Computer Science department, Faculty of Information Technology, VNU-UET. His research interest focuses on social/web mining, recommendation systems, reliable machine learning and multimodal learning. He also serves as Program Committee at top-tier AI/ML conferences, e.g., AAAI, IJCAI, and Reviewer for Q1 journals namely TKDE, TKDD, TITS, NEUNET, NEUCOM.

Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition

Cam-Van Thi Nguyen
VNU University of Engineering and Technology
Hanoi, Vietnam
vanntc@vnu.edu.vn

The-Son Le
VNU University of Engineering and Technology
Hanoi, Vietnam
21020089@vnu.edu.vn

Anh-Tuan Mai
VNU University of Engineering and Technology
Hanoi, Vietnam
20020269@vnu.edu.vn

Duc-Trong Le
VNU University of Engineering and Technology
Hanoi, Vietnam
trongld@vnu.edu.vn

Abstract

Multimodal Emotion Recognition in Conversations (ERC) is a typical multimodal learning task in exploiting various data modalities concurrently. Prior studies on effective multimodal ERC encounter challenges in addressing modality imbalances and optimizing learning across modalities. Dealing with these problems, we present a novel framework named **Ada2I**, which consists of two inseparable modules namely **Ada**ptive Feature Weighting (AFW) and **Ada**ptive Modality Weighting (AMW) for *feature-level* and *modality-level* balancing respectively via leveraging both **I**nter- and **I**ntra-modal interactions. Additionally, we introduce a refined disparity ratio as part of our training optimization strategy, a simple yet effective measure to assess the overall discrepancy of the model's learning process when handling multiple modalities simultaneously. Experimental results validate the effectiveness of **Ada2I** with state-of-the-art performance compared to baselines on three benchmark datasets, particularly in addressing modality imbalances.

CCS Concepts

• **Information systems** → **Sentiment analysis**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

Keywords

Multimodal Emotion Recognition, Imbalance Modality, Adaptive Feature Weighting, Adaptive Modality Weighting, Disparity ratio

ACM Reference Format:
Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. 2024. Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681648

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681648

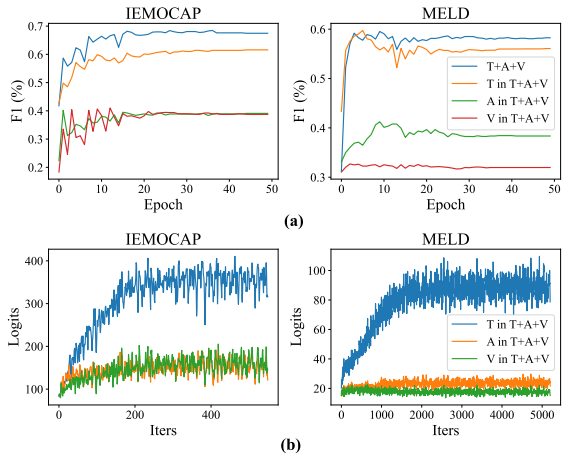


Figure 1: (a) Weighted F1 scores for the multimodal setting (T+A+V) compared with each unimodal encoder, and (b) batch-average unimodal-logit scores.

1 Introduction

Multimodal learning is an approach to building models that can process and integrate information from multiple heterogeneous data modalities [2, 20, 21], including image, text, audio, video, and table. Since numerous tasks in the real world involve multiple modalities, multimodal learning has become increasingly important and attracted widespread attention as an effective way to accomplish these tasks. In recent years, the field of Emotion Recognition in Conversations (ERC) has witnessed a surge in effective models [8, 27, 30]. Moving beyond unimodal recognition, the utilization of multimodal data offers a multidimensional perspective for more nuanced emotion discernment [9, 19, 24]. Consequently, the incorporation of multimodal data is a natural evolution for enhancing emotion recognition in conversations. However, the widespread adoption of multimodal learning has revealed underlying challenges, with a primary focus on modality imbalances. These imbalances entail disparities in the contributions of individual modalities to the final decision-making process.

As illustrated in Figure. 1, the text modality quickly addresses the overall model performance and the joint logit scores, whereas

the visual and audio modalities remain under-optimized throughout the training process. In addressing modality imbalance, diverse terminologies have emerged to characterize this phenomenon and explore its underlying causes. Terms such as "greedy nature" [39], "modality collapse" [15], and "modality imbalance" [6, 22] have been employed in various studies. These terms are associated with factors such as the "suppression of dominant modalities" [26], "different convergence rates" [36], "diminishing modal marginal utility" [37], or "modality competition" [14]. In essence, two primary perspectives emerge regarding this problem [37]: firstly, modalities exhibit varying levels of dominance, with models often overly reliant on a dominant modality with the highest convergence speed, thereby impeding the full utilization of other modalities with slower convergence speeds. Secondly, modal encoder optimization varies, necessitating the adoption of multiple strategies. Some approaches [7, 26] attempt to modulate the learning rates of different modalities based on the fusion modality. However, these approaches often overlook the impact of *intra-modal data enhancement* [46]. For instance, right from the initial representations through the modal encoder, the outputs can lead to misleading final results, resulting in its weakened position across all modalities. Hence, from the outset, it is crucial to enhance representations for each modality, regardless of whether they are weak or strong, as it can affect the imbalance in learning across modalities.

Moreover, current methodologies primarily focus on interactions between pairs of modalities [6, 22, 26, 40], resulting in complex computations and inadequate treatment across all modalities. These methods are commonly applied in tasks such as audio-visual learning [26, 40] and multimodal affective computing [46], often using datasets related to sarcasm detection, sentiment analysis, or humor detection. However, there is a lack of methods explicitly tailored for multimodal ERC tasks, especially for well-known multimodal datasets like IEMOCAP [3], MELD [28], and CMU-MOSEI [1]. Additionally, in recent prominent studies [17, 33], while overall performance for multimodal ERC tasks has notably increased, a closer examination of the *"importance of modality"* reveals that pairwise modalities consistently fail to achieve satisfactory performance, creating a significant gap compared to leveraging all three modalities simultaneously. Therefore, it is crucial to simultaneously leverage learning from all modalities while also significantly enhancing the capabilities of weaker modalities to improve the overall learning performance of multimodal ERC models in practical applications.

In this paper, we propose a novel framework named **Ada2I** that addresses imbalances in learning across audio, text, and visual modalities for multimodal ERC. It consists of two primary modules including **Ada**ptive Feature Weighting (AFW) and **Ada**ptive Modality Weighting (AMW) for *feature-level* and *modality-level* balancing respectively in the consideration of **I**nter- and **I**ntra-modal interactions. Focusing on *feature-level balancing* using Adaptive Feature Weighting (AFW), we apply tensor contraction to infer feature-aware attention weights for each modality, which aims to produce a feature-level balanced representation for each conversation. As an important component of AFW, Attention Mapping Network controls the balancing via maximizing the alignment between unimodal features and their corresponding attention coefficients. For *modality-level balancing* using Adaptive Modality Weighting (AMW), we further exploit feature-level balanced representations from the preceding AFW module to generate modality-level balanced ones through modality-wise normalization of features and learning weights before being used to enhance the emotion recognition. Additionally, we utilize the concept of disparity ratio, although with modifications compared to the study by Peng et al. [26], called OGM-GE, as a value to supervise the training process and evaluate the model. Specifically, while OGM-GE [26] introduced gradient modulation for pairs of modalities, we refine it to handle all three modalities simultaneously—textual, visual, and audio—in the Multimodal emotion recognition in conversation task. This adjustment reduces model complexity and overall processing time, leading to enhanced efficiency. To summarize, our contributions are as follows:

- We propose an end-to-end framework named **Ada2I** that addresses the issue of imbalance learning across modalities comprehensively for the multimodal ERC task. It not only considers modality-level imbalances but also leverages feature-level representations to contribute to the balancing step in the learning process.
- With two modules intricately designed yet inseparable, Adaptive Feature Weighting (AFW) is crafted to enhance the representation of each conversation at the feature level, while Adaptive Modality Weighting (AMW) is proposed to optimize the modality-level learning weights during training. Additionally, we redefine the disparity ratio, a simple yet effective measure, to assess the overall discrepancy of the model's learning process when simultaneously handling multiple modalities, rather than just two as in the original approach from Peng et al. [26].
- Our empirical experiments illustrate the effectiveness and enhancements of **Ada2I** in comparison to existing state-of-the-art approaches dealing with modality imbalance across three prevalent multimodal ERC datasets including IEMOCAP [3], MELD [28], and CMU-MOSEI [1].

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal Emotion Recognition (ERC) has emerged as a focal point within the affective computing community, garnering significant attention in recent years. The integration of multimodal data provides a multidimensional perspective, enabling a more nuanced understanding of emotions. Moreover, researchers have increasingly turned to multimodal fusion techniques, combining text, audio, and visual cues to enhance multimodal ERC performance [9, 10, 16, 18, 24, 25]. ICON [9] employs two Gated Recurrent Units (GRUs) to capture speaker information, supplemented by global GRUs to track changes in emotional states throughout conversations. Similarly, MMGCN [38] utilizes Graph Convolutional Networks (GCNs) to capture contextual information, effectively leveraging multimodal dependencies and speaker information. On the other hand, Multilogue-Net [31] introduces a solution utilizing a context-aware RNN and employing pairwise attention as a fusion mechanism. TBJE [4], adopts a transformer-based architecture with modular co-attention to jointly encode multiple modalities. Additionally, COGMEN [16] is a multimodal context-based graph neural network that integrates both local (speaker information) and global (contextual information) aspects of conversation. Moreover,

CORECT [24] employs relational temporal Graph Neural Networks (GNNs) with cross-modality interaction support, effectively capturing conversation-level interactions and utterance-level temporal relations. GraphMFT [18] utilizes multiple enhanced graph attention networks to capture intra-modal contextual information and inter-modal complementary information. More recently, DF-ERC [17] emphasizes both feature disentanglement and fusion while taking into account both multimodalities and conversational contexts. Moreover, AdaIGN [33] employs the Gumbel Softmax trick to adaptively select nodes and edges, enhancing intra- and cross-modal interactions. *While these methods primarily focus on designing model structures, they overlook the challenges posed by modality imbalance during multimodal learning.*

2.2 Imbalanced multimodal learning

Despite the suggestion by [13] that integrating multiple modalities could enhance the accuracy of latent space estimations, thereby improving the efficacy of multimodal models, our investigation within the multimodal ERC task reveals a phenomenon contradicting this notion. The problem of modality imbalance persists as a significant challenge in multimodal learning frameworks involving low-quality data [43], particularly in tasks such as multimodal ERC. Conventional methods often prioritize one modality over others, assuming that certain types of sensory data are more relevant for a given task. For example, textual cues may receive greater emphasis, while visual or audio cues alone might be prioritized [16, 24, 38]. Current methodologies for addressing imbalanced multimodal learning primarily focus on tasks such as audio-visual learning with a focus on optimizing pairwise modality learning [6, 26, 40], sentiment analysis, and sarcasm detection [46]. However, these approaches often have task-specific limitations and framework restrictions, limiting their broader applicability. For instance, Wang et al. [36] identified that different modalities overfit and generalize at different rates, leading to suboptimal solutions when jointly trained using a unified optimization strategy. Peng et al. [26] proposed OGM-ME method where the better-performing modality dominates the gradient update, suppressing the learning process of other modalities. MMCosine [40] employs normalization techniques on features and weights to promote balanced and improved fine-grained learning across multiple modalities. Notably, there is a lack of specific approaches tailored for multimodal ERC apart from the work by Wang et al. [37]. Recently, Wang et al. [37] observed a phenomenon referred to as "diminishing modal marginal utility" and proposed fine-grained adaptive gradient modulation, which was applied to ERC, while I^{2}MCL considers both data difficulty and modality balance for multimodal learning based on curriculum learning for affective computing, though not specifically for emotion recognition. To comprehensively address the challenge of modality imbalance in multimodal ERC, we propose an end-to-end model that ensures balance among text, audio, and visual modalities during training.

3 Methodology

3.1 Preliminary

3.1.1 Tensor Ring Decomposition. A tensor of order K (K dimensions) $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_K}$ can be represented as a sequence of core tensors of order 3: $\mathcal{G}_j \in \mathbb{R}^{d_j \times r_j \times r_{j+1}}$, where the last core tensor has the form $\mathcal{G}_K \in \mathbb{R}^{d_K \times r_K \times r_1}$. The dimensions $r_1, r_2, ..., r_k$ are called tensor ranks. In that case, $\mathcal{T}$ is represented in the form of a tensor ring $Tr\{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_k\}$ as follows: $\mathcal{T} = \mathcal{G}_1 \times_3^2 \mathcal{G}_2 \times_4^2 ... \times_{k+2}^2 \mathcal{G}_k$. In which $\times_n^m$ denotes the tensor contraction operation with mode-$\binom{m}{n}$. For example, with $\mathcal{G}_1 \in \mathcal{R}^{d_1 \times r_1 \times r_2}$, $\mathcal{G}_2 \in \mathcal{R}^{d_2 \times r_2 \times r_3}$ and $\mathcal{G}_3 \in \mathcal{R}^{d_3 \times r_3 \times r_1}$, $\mathcal{T}$ is represented as $\mathcal{G}_1 \times_3^2 \mathcal{G}_2 \times_4^2 \mathcal{G}_3 \in \mathcal{R}^{d_1 \times d_2 \times d_3}$.

3.1.2 Problem Definition. In the context of a conversation C with N utterances $\{u_1, u_2, \ldots, u_N\}$, the task of Emotion Recognition in Conversations (ERC) is to predict the emotion label for each utterance in the conversation from a predefined emotion category set $\mathcal{E}$. Each utterance is associated with M modalities, i.e. textual (t), audio (a), and visual (v) modalities, represented as:

$$u_i = \{u_i^t, u_i^a, u_i^v\}, i \in \{1, \ldots, N\} \tag{1}$$

where $u_i \in \mathbb{R}^{M \times d}$, d signifies the dimension of modal features. For each modality m, we derive multimodal features $\{\mathbf{X}^m\}_{m \in \{t,a,v\}} \in \mathbb{R}^{d_m \times N}$ for the conversation C. Here, $\{d_m\}_{m \in \{t,a,v\}}$ is the feature dimension of each modality.

In the following sub-section, we outline our proposed model Ada2I, including its main sub-modules: *(1) Modality Encoder, (2) Adaptive Feature Weighting* and *(3) Adaptive Modality Weighting*. We also refine the disparity ratio metric as part of our *Training Optimization Strategy*. Figure 2 illustrates architecture of Ada2I.

3.2 Modality Encoder

Given a conversation C, a **Transformer** [34] network is utilized as the encoder to generate a unimodal representation $\mathbf{Z}^m \in \mathbb{R}^{N \times d_m}$ respecting to the modality m as:

$$\mathbf{Z}^m = \phi(\theta^{(m)}, \mathbf{X}^m), m \in \{t, a, v\} \tag{2}$$

where the function $\phi(\theta^{(m)})$ is the Transformer network with learnable parameter $\theta^{(m)}$.

3.3 Adaptive Feature Weighting (AFW)

3.3.1 Tensor-based Multimodal Interaction Representation. Motivated by the tensor-ring decomposition method introduced by [44], we extend the traditional attention mechanism by replacing the query (**Q**) and key (**K**) representations with tensor-ring decomposition-based counterparts. This modification results in query tensor-ring representation $\mathcal{G}_Q$ and key tensor-ring representation $\mathcal{G}_K$, which facilitate the acquisition of more compact modality representations. Additionally, inspired by [32], we integrate a tensor-based multi-way interaction transformer architecture into our model. This enhancement allows the model to capture multi-way interactions among modalities, thereby enhancing its capability to discern intricate multimodal relationships.

We employ a tensor-ring-based generation function to retrieve the multi-interaction multimodal query tensor $\mathcal{Q}$ and key tensor $\mathcal{K}$ from the input modality presentations $\mathbf{Z}^m$. Specifically, we compute $\mathcal{Q}$ and $\mathcal{K}$ as follows:

$$\begin{cases} \mathcal{Q} = \mathbf{Tr}\{\mathcal{G}_Q^{(t)}, \mathcal{G}_Q^{(a)}, \mathcal{G}_Q^{(v)}\} \in \mathbb{R}^{d_t \times d_a \times d_v} \\ \mathcal{K} = \mathbf{Tr}\{\mathcal{G}_K^{(t)}, \mathcal{G}_K^{(a)}, \mathcal{G}_K^{(v)}\} \in \mathbb{R}^{d_t \times d_a \times d_v} \end{cases} \tag{3}$$

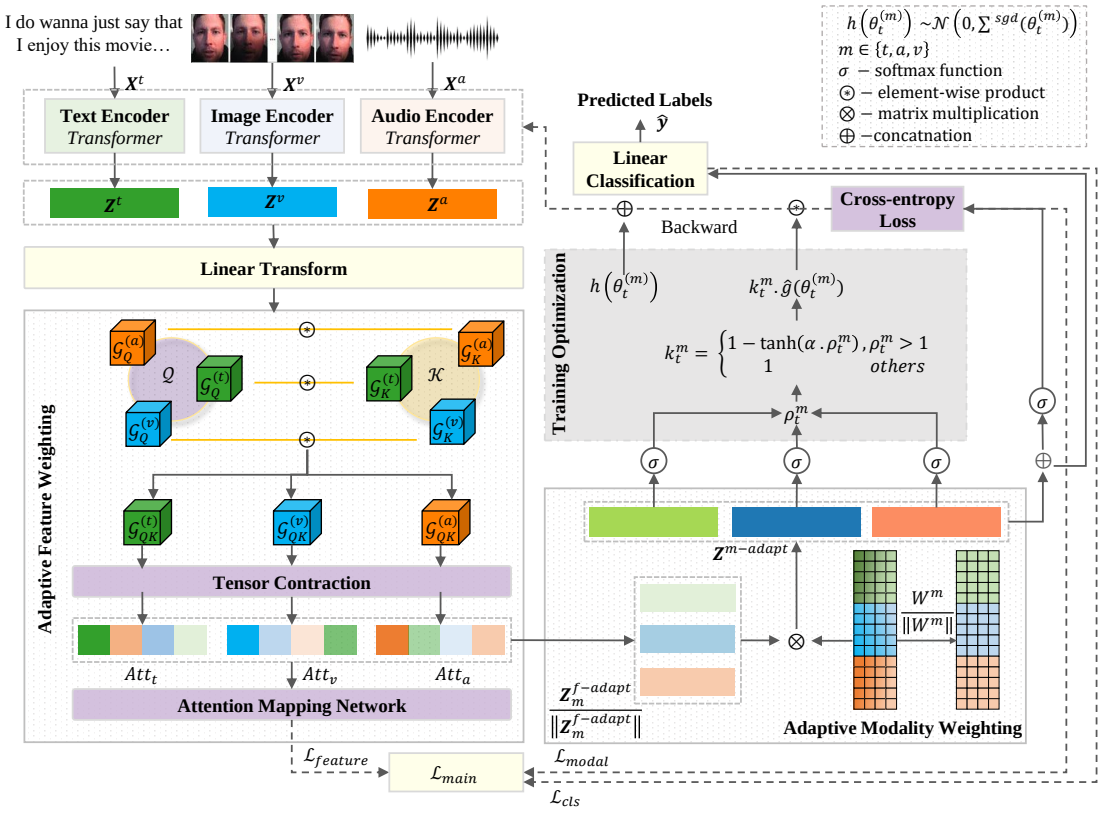


Figure 2: Illustration of Ada2I framework

Here, **Tr{.}** represents the tensor-ring decomposition function, which naturally provides the low-rank core tensor representations $\mathcal{G}_Q^m$ and $\mathcal{G}_K^m$ for each modality.

To perform multimodal attention in the tensor space, we need to compute the attention coefficient matrix, Θ, from the tensorized input. To achive this, we can first compute the Tensor-ring Key representation and Tensor-ring Query representation of input data, $\mathcal{G}_Q^m \in \mathbb{R}^{d_m \times r_s \times r_w}$ and $\mathcal{G}_K^m \in \mathbb{R}^{d_m \times r_s \times r_w}$, where $m \in \{t, a, v\}$, the index $s, w \in \{1, 2, 3\}$, and $s \neq w$. The attention coefficient matrix Θ of modality m is formulated as follows:

$$\Theta^m = \text{softmax}\left(\frac{1}{\sqrt{d_k}} \mathcal{G}_Q^m \odot \mathcal{G}_K^m\right) \tag{4}$$

where $\odot$ denotes the element-wise product, $\sqrt{d_k}$ is a scaling factor.

More specifically, the modality m core tensor $\mathcal{G}_K$ and $\mathcal{G}_Q$ are computed using a Linear Transform (Figure 3), as expressed below:

$$\begin{cases} \mathcal{G}_Q^m = reshape\big((\mathbf{Z}^m W_{Q_m}^{(1)}) \otimes_1 (\mathbf{Z}^m W_{Q_m}^{(2)})\big) \\ \mathcal{G}_K^m = reshape\big((\mathbf{Z}^m W_{K_m}^{(1)}) \otimes_1 (\mathbf{Z}^m W_{K_m}^{(2)})\big) \end{cases} \tag{5}$$

where $m \in \{t, a, v\}$, $W_{Q_m}^{(1)} \in \mathbb{R}^{d_m \times r_s}$, $W_{Q_m}^{(2)} \in \mathbb{R}^{d_m \times r_w}$, $W_{K_m}^{(1)} \in \mathbb{R}^{d_m \times r_s}$, $W_{K_m}^{(2)} \in \mathbb{R}^{d_m \times r_w}$ are the linear transformation matrix; $\otimes_1$ denotes the mode-1 Khatri-Rao product.

3.3.2 Adaptive Feature Weighting (AFW). This module addresses the varying impact of each modality on inter-modality and intra-modality interactions using attention mechanism. First, we calculate the attention pooling matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{r_s \times r_w}$ by averaging $\Theta^{(m)}$ across the modality dimension d_m, $m \in \{t, a, v\}$. Inspired by MMT [32], the *feature*-aware attention matrix $Att_m \in \mathbb{R}^{N \times d_m}$ for a given modality m is computed as follows:

$$Att_m = \text{Linear}\left(\Theta^m \times_3^1 \mathbf{A}^{(t)} \times_3^1 \mathbf{A}^{(a)} \times_3^1 \mathbf{A}^{(v)}\right) \tag{6}$$

where $\times_3^1$ is the $mode-(_3^1)$ tensor contraction. The *feature*-aware balanced representation $\mathbf{Z}_m^{f-adapt} \in \mathbb{R}^{N \times d_m}$ of the conversation C for a given modality m is computed as:

$$\mathbf{Z}_m^{f-adapt} = Att_m \mathbf{Z}^m + \beta \mathbf{Z}^m \tag{7}$$

where $\beta \in [0, 1]$ is a balancing parameter to regulate the contribution of the original unimodal feature vector $\mathbf{Z}^m$.

3.4 Adaptive Modality Weighting (AMW)

Our key focus is to achieve balanced contributions from each modality during the training. Similar to [40], we observe the imbalance problem in multimodal ERC through experiments analyzing the modality-wise weight in norm of each label during training. Apparently, the dominant unimodal encoder, e.g., text, tends to have its weight in norm increase much faster than the weaker modalities,

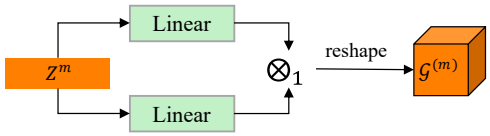


Figure 3: Linear Transform block to compute core tensor.

i.e., audio and visual, leading to divergent unimodal logit scores and distorting the joint fusion representation. Inspired by [35, 45], we propose to incorporate modality-wise L2 normalization to properly weight features, mitigating imbalances arising from differing data distributions and noise levels across modalities. This dynamic adjustment prevents any single modality from dominating the fusion process, thus enhancing overall performance. Therefore, the modality-level balanced representation $\mathbf{Z}^{m-adapt}$ of the given conversation is calculated as follows:

$$\mathbf{Z}^{m-adapt} = \sum_{m}^{\{t,a,v\}} \frac{W^m \mathbf{Z}_m^{f-adapt}}{\|W^m\| \|\mathbf{Z}_m^{f-adapt}\|} + b \tag{8}$$

where $W^m \in \mathbb{R}^{d_m \times |\mathcal{E}|}$ symbolizes the output matrix of the model pertaining to modality m, and $\mathcal{E}$ is the set of emotion classes.

For **emotion recognition**, we feed $\mathbf{Z}^{m-adapt}$, into the mulilayer preceptron (MLP) with ReLU activation function to compute the output $\hat{y}_i \in \mathbb{R}^{N \times |\mathcal{E}|}$.

$$\hat{y}_i = \text{MLP}(\mathbf{Z}^{m-adapt}) \tag{9}$$

The output $\hat{y}_i$ is utilized to predict emotion labels.

3.5 Learning

First, we investigate the standard cross-entropy loss for this downstream task, i.e., mutilmodal ERC as:

$$\mathcal{L}_{cls} = -\frac{1}{B} \sum_{i}^{B} y_i log \hat{y}_i \tag{10}$$

where B is the batch size.

Second, in order to align between the original unimodal representation of modality m and its respective *feature*-aware attention weights as Eq (6), we employ Attention Mapping Network as follows:

$$\hat{Att}_m = \Phi_m(\mathbf{Z}_m, \psi^{(m)}), m \in \{t, a, v\} \tag{11}$$

where $\Phi_m(\cdot)$ is a feed-forward neural network with the parameter $\psi^{(m)}$, $\hat{Att}_m \in \mathbb{R}^{N \times d_m}$ is the *feature*-aware self-attention weights of the modality m. To enhance feature-level balance across all modalities, we introduce a L1-norm loss $\mathcal{L}_{feature}$ as:

$$\mathcal{L}_{feature} = \frac{1}{B} \sum_{i}^{B} \left(\sum_{m}^{\{t,a,v\}} |Att_m^i - \hat{Att}_m^i| \right) \tag{12}$$

Additionally, we also consider the modality-level balance loss $\mathcal{L}_{modal}$, which is computed as:

$$\mathcal{L}_{modal} = -\frac{1}{B} \sum_{i}^{B} \log \frac{e^{\mathbf{Z}_i^{m-adapt}}}{\sum_{j=1}^{|\mathcal{E}|} e^{\mathbf{Z}_j^{m-adapt}}} \tag{13}$$

where $\mathbf{Z}_j^{m-adapt}$ represents the output of the j-th class for the i-th sample. Finally, we combine the all loss functions into a joint objective function, which is used to optimize all trainable parameters in an end-to-end manner:

$$\mathcal{L}_{main} = \mathcal{L}_{modal} + \mathcal{L}_{feature} + \mathcal{L}_{cls} \tag{14}$$

Recent studies have brought attention to the challenge of handling imbalanced optimization in joint learning models, particularly when dealing with multiple modalities. Peng et al. [26] introduce the OGM-GE method to address optimization imbalances encountered during the simultaneous training of dual-modal systems, i.e., visual and audio. However, directly applying the OGM-GE method to our framework is not practical as it only deals with two modalities. In contrast, our framework caters to more than two modalities across different domains, specifically tailored for the multimodal ERC task. Therefore, leanrable parameter of encoder layer is optimized during training process as the following strategy:

$$\theta_{t+1}^{(m)} = \theta_t^{(m)} - \eta . \hat{g}(\theta_t^{(m)}) \tag{15}$$

where $\hat{g}(\theta^m t) = \frac{1}{o} \sum x \in B_t \nabla_{\theta_t^m} \ell(x, \theta_t^{(i)})$ represents an unbiased estimation of the full gradient $\nabla \theta_t^m \ell(x, \theta_t^{(i)})$ using a random mini-batch B_t chosen at the t-th step with size o. The term $\nabla \theta_t^m \ell(x, \theta_t^{(i)})$ denotes the gradient with respect to B_t.

We adjust the balance of modalities through gradient parameter adjustments. For each output at step t, we compute the discrepancy ratio for each modality using the softmax of the cosine similarity between the output weights and the corresponding feature vectors:

$$s_t^m = \sum_{j=1}^{L} \sum_{k=1}^{\mathcal{E}} \mathbb{I}_{k=y_j} \text{softmax}(cos\langle W_k^m, \mathbf{Z}_k^m \rangle + \frac{b_k}{M}) jk \tag{16}$$

where $\mathbb{I}_{k=y_j}$ equals 1 if $k = y_j$ and 0 otherwise, and softmax(.) estimates the unimodal performance of the multimodal model, M denotes the count of modalities. Specifically, for the multimodal ERC task under consideration, we delineate three modalities: text (t), audio (a), and visual (v). The discrepancy ratio is calculated as:

$$\rho_t^m = \frac{s_t^m}{\min_{m \in \{t,a,v\}} (s_t^j)} \tag{17}$$

The learnable parameters are updated according to:

$$\theta_{t+1}^{(m)} = \theta_t^{(m)} - \eta . \hat{g}(\theta_t^{(m)}) . k_t^m \tag{18}$$

where the modulation coefficient k_t^m is determined by $1 - \tanh(\alpha \cdot \rho_t^m)$ if $\rho_t^m > 1$, and 1 otherwise. Here, α is a hyperparameter controlling the degree of modulation. Additionally, to enhance the adaptability of the modulation process, Gaussian noise $h(\theta_t^{(i)})$ sampled from a distribution $\mathcal{N}(0, \sum^{sgd}(\theta_t^{(i)}))$ is introduced after parameter updates:

$$\theta_{t+1}^{(i)} = \theta_t^{(i)} - \eta \cdot \hat{g}(\theta_t^{(i)}) \cdot k_t^i + \eta \cdot h(\theta_t^{(i)}) \tag{19}$$

Training Optimization Strategy The training process of Ada2I is illustrated in Algorithm 1.

Algorithm 1 Ada2I Training Procedure

Input: The training set $\mathcal{D} = \{(x_i^t, x_i^a, x_i^v), y_i\}_{i=1}^N, m \in \{t, a, v\}$
Output: Prediction emotion label $\hat{y}$
  for each training epoch **do**
    for minibatch $\mathcal{B} = \{(x_i^t, x_i^a, x_i^v), y_i\}_{i=1}^N\}$ sampled from $\mathcal{D}$ **do**
      #Refer to Subsection 3.2
      Encode unimodal feature $\mathbf{X}^m$ to $\mathbf{Z}^m$ as Eq (2)
      #Refer to Subsection 3.3
      Multimodal feature representation as Eq (3)
      Calculate coefficient matrix Θ^m as Eq (4)
      Calculate modality-aware attention Att_m as Eq (6)
      Compute fused feature $\mathbf{Z}_m^{f-adapt}$ with β using Eq (7)
      #Refer to Subsection 3.4
      Compute logit output $\mathbf{Z}^{m-adapt}$ with modality-wise L2 normalization as Eq (8)
      Produce prediction of multimodal data $\hat{y}_i$ as Eq (9)
      #Refer to Subsection 3.5
      Use cross-entropy loss to calculate $\mathcal{L}_{cls}$ as Eq (10)
      Use L_1 to calculate $\mathcal{L}_{feature}$ as Eq (12)
      Use cross-entropy to calculate $\mathcal{L}_{modal}$ as Eq (13)
      Add $\mathcal{L}_{feature}$, $\mathcal{L}_{modal}$ and $\mathcal{L}_{cls}$ to compute $\mathcal{L}_{main}$ as Eq (14)
      Compute discrepancy ratio $\rho_t^m = \frac{s_t^m}{\min_{m \in \{t,a,v\}}(s_t^j)}$
      Compute modulation coefficient k_t^m
      Update using $\theta_{t+1}^{(i)} = \theta_t^{(i)} - \eta \cdot \hat{g}(\theta_t^{(i)}) \cdot k_t^i + \eta \cdot h(\theta_t^{(i)})$
    end for
  end for

3.6 Datasets

Datasets: We consider three benchmark datasets for multimodal ERC namely: IEMOCAP [3], MELD [28], and CMU-MOSEI [1]. The dataset statistics are illustrated in Table 1.

Table 1: Data Statistics

Datasets	Dialogues			Utterances		
	train	valid	test	train	valid	test
IEMOCAP	120		31	5,810		1,623
MELD	1,039	114	280	9,989	1,109	2,610
CMU-MOSEI	2,248	300	676	16,326	1,871	4,659

IEMOCAP. This dataset comprises 12 hours of video recordings of dyadic conversations involving 10 speakers. It includes 151 dialogues, segmented into 7,433 utterances, each annotated with one of six emotion labels: happy, sad, neutral, angry, excited, or frustrated.

MELD. This dataset is based on the TV series Friends, includes 13,709 video clips featuring multi-party conversations, each labeled with one of Ekman's six universal emotions: joy, sadness, fear, anger, surprise, and disgust.

CMU-MOSEI:. This dataset is a prominent resource for sentiment and emotion analysis, comprises 3,228 YouTube videos divided into 23,453 segments, featuring contributions from 1,000 speakers covering 250 topics. It includes six emotion categories: happy, sad, angry, scared, disgusted, and surprised, with sentiment intensity ranging from -3 to 3.

Table 2: Hyper-parameter settings

Parameter/Module	IEMOCAP	MELD	CMU-MOSEI
Text Feature Extraction	sBERT[1]		
Audio Feature Extraction	Wave2vec-Large [29], OpenSmile [5]		
Visual Feature Extraction	MTCNN [42], MA-Net[2], DenseNet [12]		
Text embedding dim. d_t	768	768	768
Audio embedding dim. d_a	512	300	512
Visual embedding dim. d_v	1024	342	1024
hidden dim	300	200	500
tensor rank	11	6	10
η	0.037	0.4	0.4
β	0.01	0.55	0.2
learning rate	1.7e-4	1.2e-4	1.9e-4
batch size	10	10	32
epoch	50	50	30

4 Experimential Setup

4.1 Baselines and Evaluation Metrics

Baselines: Ada2I is compared against several state-of-the-art (SOTA) baseline approaches for evaluating performance in multimodal ERC, particularly addressing modality imbalance problems. For the IEMOCAP and MELD datasets, we consider baseline models such as DialogueRNN [23], DialogueGCN [8], MMGCN [38], BiDDIN [41], and MM-DFN [11]. We report the best results obtained from [37], which enhanced these models to address modality imbalance. Additionally, we consider other SOTA models for multimodal ERC that do not explicitly address modality imbalance, including COGMEN [16], CORECT [24], GraphMFT [18], DF-ERC [17], and AdaIGN [33].

For the CMU-MOSEI dataset, we evaluated various baseline models for sentiment classification tasks, which include both 2-class sentiment, featuring only positive and negative sentiment, and 7-class sentiment, ranging from highly negative (-3) to highly positive (+3). These baseline models include Multilouge-Net [31], TBJE [4], COGMEN [16], CORECT [24], OGM-GE [26], and I^2MCL [46]. Notably, OGM-GE and I^2MCL specifically address the issue of imbalanced modalities in multimodal ERC, whereas the others do not.

Evaluation Metrics: Similar to prior studies [23, 37, 38], we evaluate the effectiveness of emotion recognition using Accuracy (Acc) and Weighted F1 Score (W-F1) as our primary evalucation metrics.

4.2 Experimental Settings

We derive multimodal features for each utterance from acoustic, lexical, and visual modalities using a combination of models and pre-trained models, as outlined in Table 2.

We employ PyTorch[3] for training our architecture and Comet[4] for logging all experiments, leveraging its Bayesian optimizer for hyperparameter tuning. Additional parameters can be found in Table 2.

[1]https://www.sbert.net/
[2]https://github.com/zengqunzhao/MA-Net
[3]https://pytorch.org/
[4]https://comet.ml

Table 3: Comparison of results in the multimodal setting of Ada2I with the modality-balanced baseline model enhanced by FAGM [37] (denoted by †). The best performance is indicated in bold, and the second-best performance is underlined.

Methods	IEMOCAP								MELD							
	T+A+V		T+A		T+V		A+V		T+A+V		T+A		T+V		A+V	
	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc
DialogueRNN†	61.31	61.61	61.90	61.98	60.19	59.95	48.31	50.71	56.42	58.05	56.46	58.01	55.67	57.39	40.46	45.39
DialogueGCN†	62.76	63.22	64.36	64.39	61.25	62.23	49.20	49.85	54.61	58.96	54.80	57.28	55.26	57.10	10.02	44.44
BiDDIN†	58.81	58.84	58.88	58.16	59.04	58.96	46.36	46.77	57.47	59.18	56.56	58.05	56.93	58.10	44.39	48.62
MM-DFN†	64.92	64.57	63.91	64.20	61.02	60.60	54.48	55.03	55.75	60.8	57.10	60.00	57.73	60.65	42.05	48.66
MMGCN†	64.53	64.51	63.25	63.40	61.02	61.06	54.14	54.90	58.48	61.15	57.59	60.69	57.14	59.46	43.49	48.43
Ada2I (Ours)	**68.97**	**68.76**	**66.91**	**67.28**	**65.48**	**65.43**	**55.16**	**55.64**	**60.38**	**63.03**	**60.08**	**62.64**	**58.62**	**61.95**	**55.16**	**55.64**
Δ(%)	↑*4.05*	↑*4.19*	↑*2.55*	↑*2.89*	↑*4.23*	↑*3.20*	↑*0.68*	↑*0.61*	↑*1.90*	↑*1.88*	↑*2.49*	↑*1.95*	↑*0.89*	↑*1.30*	↑*10.77*	↑*6.98*

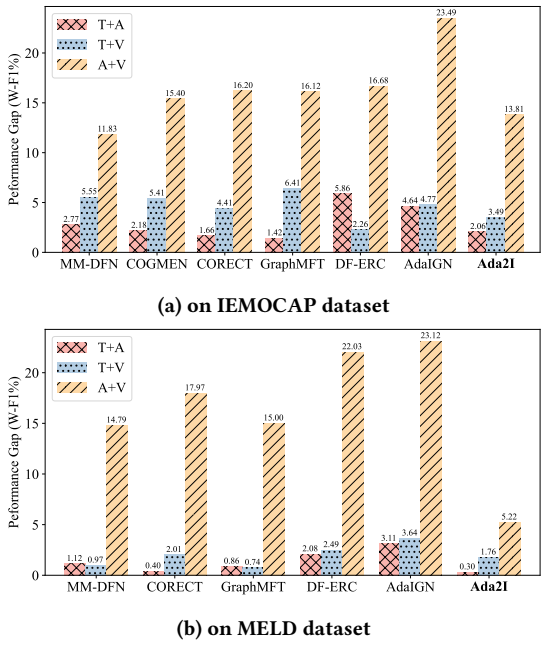
5 Results and Discussion

5.1 Performance Comparison against Baselines

IEMOCAP and MELD dataset: As depicted in Table 3, our model Ada2I performs better than the previous SOTA baselines in the context of balanced modality consideration on all modality combinations on both datasets. Indeed, in the AV modality pair on the MELD dataset, traditionally deemed the weakest, we observe a substantial performance boost in Multimodal ERC. Specifically, there is a noteworthy enhancement of 10.77% on WF1 and 6.98% on Accuracy compared to the previous SOTA model. This progress effectively reduces the performance discrepancy compared to modality pairs where text plays a dominant role.

We also compare Ada2I with SOTA baseline models for multimodal ERC, particularly those focusing solely on multimodal fusion and architectural design without addressing modality imbalance. Figure 4b demonstrates that our proposed Ada2I significantly reduces the performance gap in WF1 between learning from all three modalities simultaneously (T+A+V) and pair-wise modality combinations on the MELD dataset. Most notably, with the weaker modality pair (audio+visual) consistently lagging behind in performance compared to the full modality combination (i.e., with AdaIGN, this gap is 23.12%), Ada2I boosts the model and shortens the gap to only 5.22%. Similarly, with the text+audio (T+A) and text+visual (T+V) pairs, this gap is also substantially reduced, indicating that the model has learned in a more balanced manner, leveraging additional useful information from non-dominant modalities. The significant improvement is similarly observed on the IEMOCAP dataset in Figure 4a.

CMU-MOSEI dataset: Table 4 shows that Ada2I outperforms all baseline models. Specifically, when compared to OGM-GE and I^{2}MCL, two models proposed for addressing modality imbalance during training, Ada2I demonstrates superior performance across all modality combinations. When compared to other baseline models that do not consider modality balancing, Ada2I also demonstrates significant balancing capabilities, reducing the performance gap between modality pairs. For instance, in the CORECT model, the gap between T+A+V and A+V is 15.09% for 2-class sentiment, and this figure increases to 21.76% for 7-class sentiment. However, with Ada2I, these gaps are significantly reduced to 10.32% and 13.07%, respectively, underscoring the effectiveness of Ada2I in addressing modality imbalances.



(a) on IEMOCAP dataset

(b) on MELD dataset

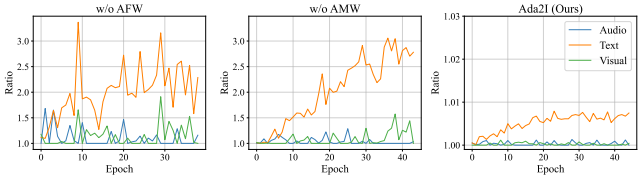
Figure 4: Performance gap visualizations between the multimodal setting (T+A+V) and pair-wise modality combinations are evaluated using the W-F1 metric across the IEMOCAP and MELD datasets.

5.2 Ablation Study

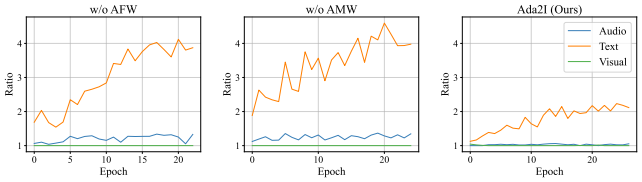
5.2.1 Balancing Interpretation. We conduct ablation studies with the two main modules of the model, AMW and AFW, to assess their impact on the Ada2I model. Additionally, through the Discrepancy Ratio, we interpret the model's balancing by observing its changes. A smaller Discrepancy Ratio indicates a more balanced optimization process. Figure 5 shows that the discrepancy ratios ρ^t, ρ^v, and ρ^a significantly decrease when both AMW and AFW are combined within Ada2I, with all ratios approaching approximately 1 on the IEMOCAP dataset. In contrast, when one of the modules is ablated, the ratios for audio (ρ^a) and visual (ρ^v) are approximately 1.5, while for text, it increases to around 3. Similarly, on the MELD dataset, our proposed model Ada2I has reduced this discrepancy

Table 4: Results on the CMU-MOSEI dataset with accuracy (Acc.) as the metric. The best performance is in bold. Cells with "-" indicate missing results, and † denotes results reproduced from the code provided in the original paper.

Methods	2-class				7-class			
	T+A+V	T+A	T+V	A+V	T+A+V	+TA	T+V	A+V
Multilouge-Net [31]	82.10	80.18	80.06	**75.16**	44.83	-	-	-
TBJE [4]	81.50	82.40	-	-	44.40	45.50	-	-
COGMEN† [16]	82.95	85.00	82.99	65.95	43.90	44.31	42.68	24.27
CORECT† [24]	83.98	84.28	82.83	68.89	46.31	44.89	43.76	24.55
I^2MCL [46]	81.05	-	-	-	-	-	-	-
OGM-GE† [26]	84.58	84.03	83.67	71.53	45.43	43.68	44.44	31.53
Ada2I (Ours)	**85.25**	**85.08**	**85.21**	74.93	**47.71**	**47.35**	**47.37**	**34.64**
Δ(%)	↑0.67	↑0.08	↑1.54	↓0.23	↑2.28	↑1.85	↑2.93	↑3.11



(a) IEMOCAP



(b) MELD

Figure 5: The change of the discrepancy ratio ρ^t, ρ^a, ρ^v on the IEMOCAP and MELD datasets during training, along with various ablation tests including without AMW and without AFW, are compared to the Ada2I model.

ratio of text from over 4 (w/o AFW) to approximately half, reaching around 2, while for audio and visual, it brings them close to the 1 mark. In summary, the combined design of both modules AMW and AFW enhances balanced learning across modalities during training, highlighting the significance and inseparability of feature-level and modality-level balancing.

5.2.2 Effect of Weight Normalization. As mentioned earlier, the unimodal weights also directly influence the encoder updating process. The imbalanced weight components induce gradients and subsequently lead to the inconsistent convergence of unimodalities. Here, we provide a clearer visualization of these unimodal weights before imbalance processing (Only Encoder) and in the Ada2I model in Figure 6 for the IEMOCAP dataset. It is evident that with Only Encoder, the text encoder (dominant modality) weight in norm grows much faster than audio and visual. After balancing, our model exhibits a more balanced optimization process.

5.2.3 Effect of Module. Table 5 provides an ablation on the modules. AFW and AMW are two closely linked and crucial modules in

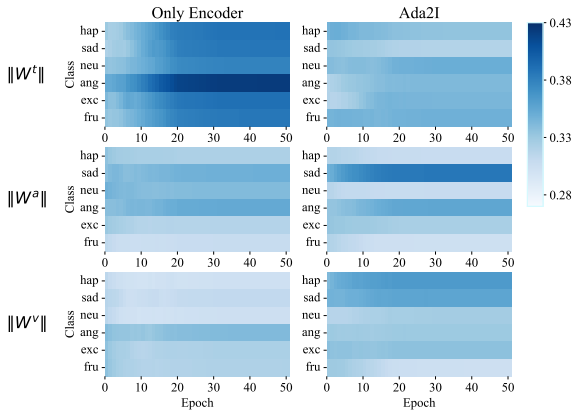


Figure 6: Modality-wise weights of each label normalized for the IEMOCAP dataset

Ada2I, ensuring model stability. Furthermore, Ada2I with training optimization balances the training across three modalities (text, audio, visual), preventing the text modality from dominating the others.

Table 5: Ablation studies of Ada2I on AFW, AMW, and training strategy. The symbol ↓ denotes the reduction in performance of the variants compared to Ada2I.

Modules	IEMOCAP		MELD	
	W-F1	Acc	W-F1	Acc
w/o AFW	66.24$_{(\downarrow 2.73)}$	65.99$_{(\downarrow 2.77)}$	59.65$_{(\downarrow 0.73)}$	62.45$_{(\downarrow 0.58)}$
w/o AMW	66.11$_{(\downarrow 2.86)}$	65.87$_{(\downarrow 2.89)}$	58.87$_{(\downarrow 1.51)}$	61.13$_{(\downarrow 1.90)}$
w/o traning optimization	67.95$_{(\downarrow 1.02)}$	68.08$_{(\downarrow 0.68)}$	58.13$_{(\downarrow 2.25)}$	59.92$_{(\downarrow 3.11)}$
Ada2I (Ours)	**68.97**	**68.76**	**60.38**	**63.03**

6 Conclusion

In this work, we present **Ada2I**, a framework designed to address modality imbalances and optimize learning in multimodal ERC. We identify and analyze existing issues in current ERC models that overlook the imbalance problem. From there, we propose a solution comprising integral modules: Adaptive Feature Weighting (AFW) and Adaptive Modality Weighting (AMW). The former enhances intra-modal representations for feature-level balancing, while the latter optimizes inter-modal learning weights with the balancing at modality level. Furthermore, we introduce a refined disparity ratio to optimize training, offering a straightforward yet effective measure to evaluate the model's overall discrepancy when handling multiple modalities simultaneously. Extensive experiments on the IEMOCAP, MELD, and CMU-MOSEI datasets validate its effectiveness, showcasing SOTA performance. In the future, we anticipate enhancing the efficiency of the framework and maximizing the utilization of emotional cues.

Acknowledgments

Cam-Van Thi Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS147.

References

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[4] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria (Eds.). Association for Computational Linguistics, Seattle, USA, 1–7. https://doi.org/10.18653/v1/2020.challengehml-1.1

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

[6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20029–20038.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[8] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 154–164. https://doi.org/10.18653/v1/D19-1015

[9] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2594–2604. https://doi.org/10.18653/v1/D18-1280

[10] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 2122–2132. https://doi.org/10.18653/v1/N18-1193

[11] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 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)*. IEEE, 7037–7041.

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems* 34 (2021), 10944–10956.

[14] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*. PMLR, 9226–9259.

[15] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *International Conference on Machine Learning*. PMLR, 9938–9964.

[16] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4148–4164.

[17] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5923–5934.

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

[19] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.

[20] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[21] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv preprint arXiv:2209.03430* (2022).

[22] Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Zitong Yu, Wenzhong Tang, and Alex Kot. 2024. Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing. *arXiv preprint arXiv:2402.19298* (2024).

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

[24] Cam-Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 15154–15167. https://doi.org/10.18653/v1/2023.emnlp-main.937

[25] Cam-Van Thi Nguyen, Cao-Bach Nguyen, Duc-Trong Le, and Quang-Thuy Ha. 2024. Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 4259–4265. https://aclanthology.org/2024.lrec-main.380

[26] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8238–8247.

[27] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 873–883. https://doi.org/10.18653/v1/P17-1081

[28] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 527–536. https://doi.org/10.18653/v1/P19-1050

[29] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).

[30] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1551–1560. https://doi.org/10.18653/v1/2021.acl-long.123

[31] Aman Shenoy and Ashish Sardana. 2020. Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria (Eds.). Association for Computational Linguistics, Seattle, USA, 19–28. https://doi.org/10.18653/v1/2020.challengehml-1.3

[32] Jiajia Tang, Kang Li, Ming Hou, Xuanyu Jin, Wanzeng Kong, Yu Ding, and Qibin Zhao. 2022. MMT: Multi-Way Multi-Modal Transformer for Multimodal Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, LD Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization*, Vol. 7. 3458–3465.

[33] Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19089–19097.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[35] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1041–1049.

[36] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12695–12705.

[37] Yunxiao Wang, Meng Liu, Zhe Li, Yupeng Hu, Xin Luo, and Liqiang Nie. 2023. Unlocking the Power of Multimodal Learning for Emotion Recognition in Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5947–5955.

[38] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[39] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*. PMLR, 24043–24055.

[40] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. 2023. MMCosine: Multi-Modal Cosine Loss Towards Balanced Audio-Visual Fine-Grained Learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[41] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*. 503–511.

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Qiao Yu. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[43] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947* (2024).

[44] Qibin Zhao, Guoxu Zhou, Shengli Xie, Liqing Zhang, and Andrzej Cichocki. 2016. Tensor ring decomposition. *arXiv preprint arXiv:1606.05535* (2016).

[45] Yutong Zheng, Dipan K Pal, and Marios Savvides. 2018. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5089–5097.

[46] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. 2023. Intra-and Inter-Modal Curriculum for Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3724–3735.

ORIGINAL ARTICLE

Leveraging self-paced curriculum learning for enhanced modality balance in multimodal conversational emotion recognition

Phuong-Anh Nguyen · The-Son Le · Duc-Trong Le · Cam-Van Thi Nguyen

Received: 17 March 2025 / Accepted: 27 April 2026

Abstract
Multimodal Emotion Recognition in Conversations (MERC) is a crucial task for understanding human interactions, where multimodal approaches integrating language, facial expressions, and vocal tone have led to significant advancements. However, a persistent challenge in multimodal architectures, including MERC, is modality misalignment and imbalanced learning. These issues often hinder models from effectively utilizing multimodal information, leading to suboptimal performance despite the availability of multiple modalities. To address this, we design a framework for MERC with a proposed plug-and-play module that builds upon Self-Paced Curriculum Learning (**SPCL**). As in Curriculum Learning (CL), an effective *Difficulty Measurer* is essential for structuring a meaningful *Learning Scheduler*. In this work, we propose a dual-level Difficulty Measurer tailored for MERC, addressing both intra- and inter-conversational dynamics. Unlike conventional approaches that assess difficulty only at the utterance level, our dual-level design incorporates a conversation-level difficulty score. The utterance-level score captures fine-grained modality-specific challenges, while the conversation-level score models broader dialogue structures, including emotional dependencies and modality coherence within the conversation. This holistic evaluation enables our Learning Scheduler to dynamically guide training from easier to more challenging instances. By integrating SPCL into existing MERC architectures, our method effectively mitigates modality imbalance and enhances model robustness. Extensive experiments on the IEMOCAP and MELD datasets confirm consistent improvements: on IEMOCAP, SPCL achieves gains ranging from approximately +1.2% to +6.6% in weighted F1-score over baseline models across different architectures and modality settings, while on MELD, it delivers even more pronounced improvements, with gains reaching up to +10.4% over baseline models. These gains underscore the practical value of SPCL for real-world MERC applications, as it substantially improves emotion recognition accuracy while maintaining compatibility as a plug-and-play module across diverse model architectures.

Keywords Self-paced curriculum learning · Modality balance · Multimodal emotion recognition

Published online: 02 June 2026

1 Introduction

Multimodal learning has gained significant attention in artificial intelligence due to its ability to integrate information from diverse sources, such as text (T), audio (A), and visual (V) data [1, 2]. By leveraging complementary modalities, multimodal models enhance understanding and improve predictive performance [3, 4]. Among its diverse applications, Emotion Recognition in Conversations (ERC) has emerged as a critical task spanning artificial intelligence, cognitive science, and social sciences. The primary objective of ERC is to detect the emotional undertones accompanying each utterance in a conversation. While language plays a crucial role, emotions are often conveyed through a combination of verbal and non-verbal cues, such as facial expressions, tonal variations, and bodily gestures [5]. Given the inherently multimodal nature of human communication, integrating multimodal data into ERC has naturally evolved into Multimodal Emotion Recognition in Conversations (MERC). By jointly modeling text, audio, and visual modalities, MERC seeks to enhance emotion recognition beyond the constraints of unimodal approaches [6–8].

However, the effectiveness of multimodal integration is often limited by modality imbalance, where certain modalities contribute disproportionately to learning, leading to suboptimal model performance. Prior research has examined this phenomenon from various perspectives, describing it as the dominance of specific modalities [9], discrepancies in convergence rates [10, 11], or diminishing marginal utility of modalities [12]. Zhang et al. [13] categorize these challenges into Property Discrepancy and Quality Discrepancy. Property Discrepancy arises because different modalities exhibit distinct learning behaviors due to their heterogeneous nature. For instance, audio features often require fewer training iterations before overfitting, whereas visual features typically converge more slowly. This inconsistency complicates optimization, making it difficult to balance learning across modalities. Quality Discrepancy refers to the uneven distribution of task-relevant information among modalities. While all modalities aim to represent the same underlying context, some provide stronger discriminative signals than others. Multimodal models, inherently greedy in their learning dynamics, tend to prioritize the most informative modality while underutilizing weaker ones, further exacerbating modality imbalance.

In MERC, text is often the dominant modality, providing explicit emotional cues, whereas audio and visual features are more subtle and context-dependent. Consequently, models may become overly reliant on textual data, failing to fully integrate information from other modalities. This imbalance ultimately leads to suboptimal fusion strategies and reduced overall performance. As shown in Fig. 1, we conducted an experiment comparing the performance of tri-modal models (ATV) with their bi-modal versions (TA and AV). While there were instances where the tri-modal approach outperformed its bi-modal counterparts, the results were inconsistent. In half of the cases, incorporating all three modalities led to a performance reduction, with accuracy dropping between 0.05% and 1.18%, and w-F1 dropping between 0.11% and 1.18%. These findings highlight the need to further investigate modality imbalance, a crucial factor affecting MER performance. To address modality imbalance in MERC, previous studies have explored several strategies, including pre-trained unimodal networks [14, 15], auxiliary learning objectives [16, 17], and optimization-based techniques [9, 18, 19]. Pre-trained unimodal

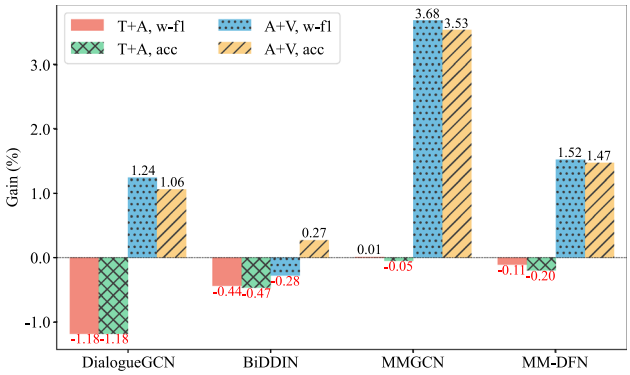


Fig. 1 Performance gain of training baseline models on full modality (ATV) over training on bi-modality (AT or TV). Evaluation uses W-F1 and Acc metrics across IEMOCAP dataset

networks leverage separately trained feature extractors for each modality before integrating them into a multimodal framework, improving individual modality representations but often incurring high computational costs and requiring large-scale labeled data [20, 21]. Auxiliary learning objectives introduce additional constraints or tasks, such as contrastive or self-supervised learning, to enhance modality-specific representations [17, 22], yet they may struggle to generalize across diverse conversational contexts. Optimization-based methods, which directly adjust gradient updates to prevent modality dominance, have shown promise in balancing multimodal contributions [9, 12, 23, 24], but their reliance on complex gradient manipulation makes them difficult to implement and tune.

Building upon these persistent challenges, we propose a novel approach based on Self-Paced Curriculum Learning to address modality imbalance in MERC. As an extension of Curriculum Learning, Self-Paced Curriculum Learning enhances this process by dynamically selecting training samples based on the model's learning progress, ensuring a more adaptive and balanced training trajectory. Specifically, we introduce a framework specifically designed for MERC, incorporating our proposed **SPCL** module to tackle modality imbalance. Our proposed SPCL module comprises two key components: (1) a **Difficulty Measurer**, which assesses the complexity of each training sample through a dual-level design that captures utterance-level recognition performance and conversation-level modality discrepancy to ensure balanced learning across modalities; and (2) a **Learning Scheduler**, which dynamically refines sample selection based on the model's evolving competence, progressively guiding training from easier to more challenging instances.

While modality imbalance has been explored in general multimodal learning, its unique challenges in MERC, especially those linked to conversational context and dialogue structures, have not been fully addressed. Our method fills this gap by offering an SPCL strategy that is carefully tailored for MERC. With this design, SPCL improves model stability, reduces problems from modality imbalance, and achieves better performance on MERC tasks.

Overall, our key contributions are as follows:

- We introduce a novel framework tailored for MERC, integrating our proposed module, SPCL, based on Self-Paced Curriculum Learning (SPCL), to address modality imbalance through adaptive sample selection.
- We design a Difficulty Measurer with a dual-level assessment mechanism, capturing both utterance-level recognition performance and conversation-level modality discrepancy. This ensures a more balanced multimodal representation learning process. Additionally, we design a Learning Scheduler that dynamically selects training samples based on model competence, facilitating a progressive learning process that enhances model robustness.
- We conduct extensive experiments on benchmark MERC datasets, IEMOCAP [25] and MELD [26], integrating SPCL as a plug-in module into four baseline models. Results demonstrate significant performance improvements, validating the effectiveness of our approach.

The remainder of this paper is organized as follows: Sect. 2 provides a comprehensive review of related work, highlighting existing approaches to modality imbalance in MERC. Section 3 introduces our proposed method, detailing the architecture and functionality of the SPCL module. Section 4 describes the experimental setup, including datasets, baseline models, and evaluation metrics. Section 5 presents our results, with an in-depth analysis of performance improvements, ablation studies, and discussions on the impact of each component. Finally, Sect. 6 concludes the paper and outlines potential directions for future research.

2 Related work

2.1 Multimodal conversational emotion recognition

Multimodal Emotion Recognition in Conversation (MERC) aims to identify emotions by leveraging multiple modalities, including textual, auditory, and visual data. The interplay between these modalities enhances emotion recognition, yet the complexity of conversational dynamics-such as speaker interactions and evolving emotional states-poses significant challenges. To tackle these issues, various methodologies have been developed, ranging from sequential modeling to advanced fusion and graph-based approaches [7, 8, 27, 28].

Early MERC models primarily relied on recurrent neural networks (RNNs) and transformers to model dialogue context. DialogueGCN [27] structures conversations as graphs, capturing speaker relationships through edge connections. COSMIC [6] integrates commonsense knowledge into an RNN-based framework to enhance context understanding, while DialogueCRN [29] employs cognitive-inspired mechanisms to model emotion flow across utterances. More recently, DialogXL [30] has demonstrated the effectiveness of transformers in processing long-range dependencies within conversations. A critical aspect of MERC is how modalities are integrated. Initial approaches, such as bc-LSTM [31] and CMN [32], focused on concatenating modality features but lacked sophisticated interaction mechanisms. Subsequent methods have improved fusion techniques by leveraging hierarchical structures [33], feature disentanglement [34], and attention-based co-learning [35]. Transformer-based models like TBJE [36] apply modular co-attention to refine cross-modal representations, while AdaIGN [37] selectively adjusts node and edge relationships within the fused feature space. Beyond sequential and fusion-based models, graph neural networks (GNNs) have emerged as a powerful tool for encoding conversational structures. MMGCN [7] captures multimodal dependencies through graph convolution, considering both speaker identity and context flow. COGMEN [38] extends this by incorporating conversational graphs that dynamically evolve over time. Meanwhile, CORECT [8] models relational interactions between utterances, allowing for refined representation learning across dialogues.

Despite their strengths, these methods primarily focus on architectural innovations and often overlook the critical issue of modality imbalance. Addressing this challenge requires not only optimizing fusion strategies but also devising techniques to balance the contributions of different modalities, ensuring robust performance even when certain modalities dominate or underperform.

2.2 Imbalanced multimodal learning

Multimodal learning, a rapidly growing field in artificial intelligence, focuses on leveraging and integrating data from multiple modalities, such as text, images, and audio, to improve model performance and enable richer understanding [2]. A critical challenge in multimodal learning is effectively integrating information from different modalities to enable complementary interactions. Traditional fusion strategies, such as early fusion, intermediate fusion, and late fusion, aim to combine multimodal information effectively. However, these strategies exhibit limited ability to resolve modality competition [39] and imbalanced multimodal learning, where dominant modalities overshadow others. This imbalance often results in modality inhibition [40, 41], where weaker modalities fail to contribute meaningfully to the final decision. As a consequence, when a dominant modality is missing or corrupted [42], the overall performance of these models often degrades significantly. Addressing this challenge requires balancing the contributions of all modalities while ensuring that weaker modalities are not entirely suppressed.

Recent studies [9, 39, 43] highlight that multimodal models often fail to outperform their best unimodal counterparts due to modality imbalance. To tackle this, various works have introduced optimization strategies aimed at rebalancing the learning process across modalities. One common approach is gradient modulation, where the model dynamically adjusts learning rates or gradients based on modality importance. For instance, Peng et al. [9] propose an on-the-fly gradient modulation strategy that monitors the contribution discrepancies of each modality

toward the learning objective, ensuring balanced optimization. Similarly, Fan et al. [18] introduce a prototypical modal rebalance (PMR) method that controls updating directions for each modality, allowing for more effective unimodal learning. FAGM [12] extends this by fine-tuning gradient updates at the parameter level, proportionally adjusting contributions from each modality to prevent over-reliance on dominant features.

Beyond gradient-based solutions, some methods enhance modality interaction to mitigate imbalance. RNA loss [44] introduces constraints in the loss function to align feature norms across modalities, ensuring more consistent feature representations and reducing modality discrepancies. OGM-GE [9] further improves balancing by dynamically adjusting gradients based on each modality's importance, preventing weaker modalities from being overshadowed. Other approaches focus on modality interaction enhancements. MLA [15] employs an alternating learning algorithm that iteratively updates different modalities, promoting stronger cross-modal dependencies. ReconBoost [45] leverages gradient boosting to dynamically adjust learning objectives, capturing underutilized information from weaker modalities. Knowledge distillation has also been explored as a way to improve weaker modalities, with UMT [46] transferring knowledge from well-trained unimodal teachers to guide multimodal representations. More recently, On-the-fly Prediction Modulation (OPM) [47] has been proposed as another approach to address modality imbalance. By monitoring the discriminative discrepancy between modalities during training, OPM dynamically drops features from the dominant modality with a certain probability, whereas OGM-GE [9] instead mitigates gradient contributions on-the-fly to balance learning across modalities. While they have demonstrated effectiveness in general multimodal tasks, they often introduce increased model complexity and additional training overhead.

Despite their contributions, these methods are often constrained by assumptions about network architecture, loss functions, or optimization methods, limiting their applicability in more general scenarios. Unlike these approaches, our method removes these restrictions by supporting arbitrary numbers of modalities, optimizers, and loss functions. Furthermore, our method is specifically designed for the MERC task under imbalanced scenarios, addressing both utterance-level and conversation-level modality imbalance. We not only consider the external disparity across modalities but also explore the intrinsic factors within conversations that contribute to imbalance.

2.3 Self-paced curriculum learning

Inspired by the structured progression of knowledge acquisition in human cognition, Curriculum Learning (CL) [48, 49] has emerged as a training paradigm that organizes the learning process by introducing samples in an incremental order of complexity-from simpler to more challenging examples. By steering the model toward an optimal parameter space, CL has demonstrated significant potential across diverse domains, including large language models [50], action recognition [51], affective computing [28, 52], and reinforcement learning [53]. A typical curriculum framework is composed of two fundamental components: a difficulty measurer, which quantifies the complexity of training samples, and a scheduler, which determines the timing and strategy for incorporating more complex samples into the training process. For the MERC task, Nguyen et al. [28] employ a Directed Acyclic Graph to integrate textual, acoustic, and visual features within a unified framework. Their model leverages CL to address challenges related to emotional shifts and data imbalance; however, it does not specifically focus on modality imbalance. In the broader context of multimodal imbalance learning, Qian et al. [54] propose a sample-level curriculum that dynamically assesses each sample's difficulty based on prediction deviation, consistency, and stability. They also introduce a modality-level curriculum to measure modality contributions from both global and local perspectives. Nevertheless, their method does not directly address the unique challenges posed by MERC tasks.

A key subset of CL, known as Self-Paced Learning (SPL), automates difficulty evaluation by using the model's current training loss as an indicator of sample complexity. Inspired by educational methodologies-where learners control their study pace by selecting topics, determining study methods, and managing learning duration [55–57]-SPL offers a dynamic and adaptive training process. Unlike traditional curriculum learning, which follows

predefined criteria for structuring the learning process, SPL dynamically adjusts the training curriculum based on the model's learning progress. By leveraging a loss-driven difficulty measurer, SPL adapts to different tasks and data distributions, ensuring a more tailored and efficient training process. Moreover, SPL seamlessly integrates curriculum design into the learning objective, making it a flexible tool for enhancing various machine learning frameworks [58].

In this work, we apply SPL to multimodal learning under imbalanced scenarios, particularly in MERC task. Our approach addresses modality discrepancy problem by introducing a Learning Scheduler strategy that designs an adaptive curriculum, dynamically selecting appropriate samples at each training step based on the model's response. This ensures a balanced learning process, enabling the model to effectively handle modality imbalance while improving overall robustness.

3 Methodology

In this section, we introduce the detailed architecture of our framework with our proposed Self-Paced Curriculum Learning-based (SPCL) module, designed to mitigate modality imbalance in MERC. Figure 2 illustrates the overall pipeline of our approach, showcasing how the SPCL module seamlessly integrates with existing MER models. In the following subsections, we provide an end-to-end overview of our framework, from obtaining unimodal emotion predictions to formulating the learning objective.We provide a detailed analysis of SPCL, including the formulation, implementation, and influence of the Difficulty Measurer and Learning Scheduler on the training process.

3.1 Problem definition

Given a predefined emotion category set $\mathcal{C}$ and a dataset $\mathcal{D} = \{U_1, U_2, \ldots, U_{|\mathcal{D}|}\}$, conversation U_i consists of N_i utterances and their corresponding labels $\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{iN_i}, y_{iN_i})\}$. The *Emotion Recognition in*

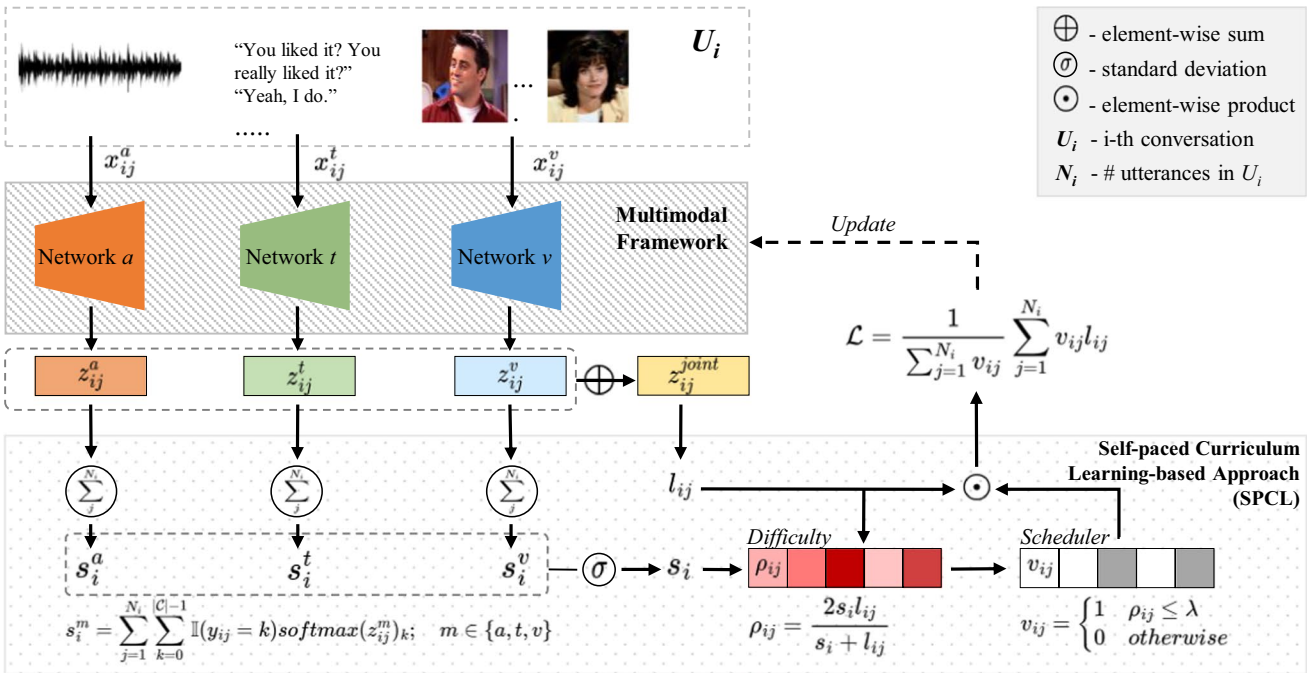


Fig. 2 Our framework pipeline with integrated SPCL module

Conversation (ERC) task is to predict an emotion label from $\mathcal{C}$ for each utterance x_{ij}. In the context of multimodal conversation, every utterance is represented through *M* modalities. Specifically, the modalities include audio (*a*), textual (*t*) and visual (*v*) modal. Thus, the input can be written as:

$$x_{ij} = \{x_{ij}^a, x_{ij}^t, x_{ij}^v\}$$

where $x_{ij}^m \in \mathbb{R}^{d_m}$ with d_m is the dimension the *m* modality.

In the following subsections, we introduce our framework, which includes 2 main sub-modules: *(1) Modality Prediction, (2) Self-paced Curriculum Learning-based (SPCL) module*.

3.2 Modality prediction

For each utterance x_{ij} in conversation U_i, an emotion prediction network is utilized to generate uni-modal prediction logit:

$$z_{ij}^m = \phi_m(x_{ij}^m; \theta^m), m \in \{a, t, v\} \tag{1}$$

where the function $\phi_m(\cdot) : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{|C|}$ is the uni-modal network with learnable parameter θ^m, and $z_{ij}^m \in \mathbb{R}^{|\mathcal{C}|}$ is the logit prediction corresponding to modal m.

In order to retrieve the ultimate prediction for utterance x_{ij}, i.e. cross-modal logit, we perform a simple fusion step on the uni-modal logits from above. Specifically, the cross-modal logit is the sum of uni-modal logits. This fusion step can be written as:

$$z_{ij}^{joint} = \sum_{m}^{\{a,t,v\}} z_{ij}^m \tag{2}$$

where $z_{ij}^{joint} \in \mathbb{R}^{|\mathcal{C}|}$ is the cross-modal logit prediction for utterance x_{ij}, and z_{ij}^m are the uni-modal logits from (Equation 1).

We evaluate the model's overall performance on utterance x_{ij} via the cross-modal logit as follow:

$$l_{ij} = -\log(softmax(z_{ij}^{joint})_{y_{ij}}) \tag{3}$$

where $l_{ij} \in \mathbb{R}$ is the loss with regard to utterance x_{ij}.

3.3 Self-paced Curriculum Learning-based Approach (SPCL)

Self-paced Curriculum Learning is employed to alleviate the imbalance between modals during training; at the same time, elevate MERC task performance. Adapting existing works using curriculum learning, we also design the learning curricula via two components: (1) *Difficulty Measurer* to determine the difficulty of all samples in the dataset, and (2) *Learning Scheduler* to control the learning pace.

3.3.1 Difficulty measurer

Traditional Difficulty Measurers treat each utterance as an independent sample, computing an utterance-level difficulty score to guide the sample selection process. In line with this approach, we adopt the utterance loss from Sect. 3.2 as the *utterance-level score*, under the assumption that a higher loss indicates greater misalignment among the unimodal representations of the utterance. However, this method is limited by its inability to capture the broader dialogue dynamics. Specifically, an utterance-level score alone fails to capture broader conversational

characteristics, such as emotional dependencies between utterances and the overall emotion distribution within a dialogue.

To address this limitation, we propose incorporating a *conversation-level* difficulty score alongside the utterance-level score to provide a more comprehensive assessment of sample difficulty. This conversation-level score, which is unique to each dialogue, is shared among all utterances within the same conversation, thereby ensuring a more holistic evaluation of multimodal interactions and emotional coherence.

To obtain the conversation-level score, firstly, we retrieve uni-modal scores for conversation U_i using uni-modal logits as follow:

$$s_i^m = \sum_{j=1}^{N_i} \sum_{k=0}^{|\mathcal{C}|-1} \mathbb{I}(y_{ij} = k) softmax(z_{ij}^m)_k \tag{4}$$

where s_i^m is a scalar that acts as the score of U_i w.r.t modal m, z_{ij}^m is the logit from Eq. 1, $\mathbb{I}(y_{ij} = k)$ is the indicator which equals 1 if $y_{ij} = k$ and 0 if otherwise, $softmax(\cdot)_k$ indicates the k^{th} value of softmax.

Next, we derive the cross-modal score of U_i by computing the standard deviation of uni-modal scores, which serves as a quantifiable measure of inter-modal variation. The standard deviation function evaluates the extent to which individual uni-modal scores s_i^m deviate from their mean, thereby offering a systematic assessment of modality misalignment. A higher standard deviation signifies pronounced discrepancies among modalities, indicating that certain modalities exert greater influence while others contribute minimally, ultimately leading to an imbalanced representation. Conversely, a lower standard deviation suggests a more equitable contribution across modalities, facilitating more coherent multimodal information fusion. Furthermore, standard deviation is a robust statistical measure that normalizes variations across different datasets, ensuring a consistent and reliable evaluation of modality divergence. Consequently, we define this function's output as the *conversation-level score*, leveraging its capability to effectively capture and quantify modality misalignment. Formally, this score is computed as follows:

$$s_i = \sigma(s_i^a, s_i^t, s_i^v) \tag{5}$$

where s_i is the conversation-level score of U_i, s_i^m with $m \in \{a, t, v\}$ are from Eq. 4, and $\sigma(\cdot)$ is the standard deviation function.

Finally, the final *difficulty* of utterance x_{ij} is formulated from utterance-level score l_{ij} and conversation-level score s_i. To ensure that the difficulty fairly represents both ER task and modality discrepancy, we combine these two values the using harmonic mean:

$$\rho_{ij} = \frac{2 s_i l_{ij}}{s_i + l_{ij}} \tag{6}$$

where $\rho_{ij} \in \mathbb{R}$ is the difficulty of utterance x_{ij}. The use of the harmonic mean is particularly advantageous in this context, as it penalizes extreme values and ensures that neither l_{ij} (recognition difficulty) nor s_i (modality misalignment) dominates the difficulty calculation. Unlike the arithmetic mean, which can be disproportionately influenced by large values, the harmonic mean emphasizes situations where both components are relatively balanced, prevents overly sensitive to one factor while ignoring the other. This formulation adaptively scales difficulty, allowing the SPCL to prioritize utterances that are easy to classify or exhibit insignificant modality misalignment, i.e. utterances with low ρ_{ij}.

3.3.2 Learning scheduler

A learning scheduler, responsible for organizing and distributing training samples throughout the learning process, is employed to ensure such structured training. Here, we adopt a controlled approach by utilizing a

hard regularizer, ensuring a strictly progressive training schedule where training samples are either included or excluded based on their predefined difficulty level ρ_{ij}. Unlike soft regularization techniques, which gradually adjust sample importance through continuous weighting [59], a hard regularizer completely excludes difficult samples until the model is sufficiently trained to handle them. This strict progression helps prevent catastrophic forgetting and encourages a more stable knowledge accumulation process, as demonstrated in prior curriculum learning research [49].

Specifically, we define a mask value v_{ij} corresponding to utterance x_{ij}. Our v_{ij} is retrieved using a *hard regularizer* $g(\rho_{ij}, \lambda)$ that leads to a binary weighting:

$$v_{ij} = g(\rho_{ij}, \lambda) = \begin{cases} 1 & \rho_{ij} \leq \lambda, \\ 0 & otherwise \end{cases} \tag{7}$$

where ρ_{ij} is the difficulty, and $\lambda > 0$ is a threshold parameter that acts as the boundary splitting easy and hard samples.

To ensure adherence to the intended learning manner, the threshold parameter λ is initially set to a relatively small value. As training progresses, λ is gradually increased to regulate the difficulty level of the samples introduced to the model. Specifically:

$$\lambda^{(t)} = \begin{cases} \varepsilon & t = 0, \\ \alpha\lambda^{(t-1)} & t > 0 \end{cases} \tag{8}$$

where $\lambda^{(t)}$ is the difficulty threshold at epoch $t - th$, ε is a relatively small number, $\alpha > 1$ is the aging hyper-parameter used to monitor the learning pace. Here, ε and α are hand-selected via conducting experiments.

3.4 Multi-modal learning with SPCL

In a standard MERC task, the objective loss is computed as the sum of the negative log likelihood loss for each sample l_{ij}, as follows:

$$\mathcal{L} = \frac{1}{\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} 1} \sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} l_{ij} \tag{9}$$

with $\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} 1$ refers to the total number of utterances in the dataset.

At each training step, once the loss is obtained, the parameters of the multimodal framework, i.e. the learnable parameters θ^m of the uni-modal networks $\phi^m(\cdot)$, are updated using gradient-based optimization as following:

$$\theta^{m^{(t+1)}} \leftarrow \theta^{m^{(t)}} - \eta \frac{\partial \mathcal{L}}{\partial \theta^{m^{(t)}}}, \quad m \in \{a, t, v\} \tag{10}$$

However, with the integration of our SPCL module, we refine this process by selectively excluding difficult samples from the loss computation. Specifically, we mitigate the influence of these challenging samples by scaling l_{ij} with a binary mask v_{ij}, which functions as a gating mechanism. This mask ensures that only easy-to-moderate samples contribute to the loss during the initial training stages, facilitating a more stable and progressive learning trajectory. Our new SPCL loss function is formally defined as:

$$\mathcal{L}_{SPCL} = \frac{1}{\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} v_{ij}} \sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} v_{ij} l_{ij} \tag{11}$$

where $\mathcal{L}_{SPCL}$ is the new loss, $\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{N_i} v_{ij}$ is the number of easy samples.

Consequently, the uni-modal networks' parameters are updated based on this filtered loss, ensuring that the model first learns from well-aligned, lower-difficulty samples before gradually incorporating harder samples as training progresses.

$$\theta^{m^{(t+1)}} \leftarrow \theta^{m^{(t)}} - \eta \frac{\partial \mathcal{L}_{SPCL}}{\partial \theta^{m^{(t)}}}, \quad m \in \{a, t, v\} \tag{12}$$

Overall, the whole training process is described in Algorithm 1.

Input $\mathcal{D} = \{U_1, U_2, \ldots, U_{|\mathcal{D}|}\}$
Initialize $\lambda \leftarrow \varepsilon$, α
for epoch t **do**
  for mini-batch $\mathcal{B}$ **do**
    Retrieve z_{ij}^m, z_{ij}^{joint} (Equation 1), (Equation 2)
    Calculate utterance loss l_{ij} (Equation 3)
    Calculate uni-modal score s_i^m (Equation 4)
    Calculate conversation-level score s_i (Equation 5)
    Calculate difficulty ρ_{ij} (Equation 6)
    Retrieve mask v_{ij} (Equation 7)
    Calculate SPCL loss $\mathcal{L}_{SPCL}$ (Equation 11)
    Update model parameters with $\mathcal{L}_{SPCL}$ (Equation 12)
  end for
  Update threshold λ (Equation 8)
end for

Algorithm 1 Pseudo-code for training a multimodal ERC framework with SPCL integration

4 Experimental setup

4.1 Datasets

We conduct experiments on two benchmark datasets for ERC task that support multi-modal, namely: IEMOCAP [25], MELD [26]. Statistics of the two datasets are summarized in Table 1.

IEMOCAP [25]: A dataset of 12-hour video recordings involving 10 actors. This dataset includes 151 dialogues of binary speakers, split into a total of 7,433 utterances. Each utterance is annotated with one of the 6 emotion labels: happy, sad, neutral, angry, excited, or frustrated.

MELD [26]: A dataset deprived from the TV series "Friends". This dataset provides 1,433 multi-party dialogues, segmented into 13,709 utterances. These utterances are classified into: happy, sad, angry, scared, disgusted, and surprised, with sentiment intensity ranging from −3 to 3.

4.2 Baselines and evaluation metrics

To evaluate the robustness and stability of our proposed method, we incorporate it into 4 existing models for ERC task, namely: DialogueGCN [27], BiDDIN [60], MMGCN [7], MM-DFN [61]. In particular, these models are used as our emotion prediction network $\phi_m(\cdot)$ in Eq. 1.

Table 1 Statistics for IEMOCAP and MELD

Dataset	Dialogues			Utterances		
	Train	Valid	Test	Train	Valid	Test
IEMOCAP	120		31	5,810		1,623
MELD	1,039	114	280	9,989	1,109	2,610

- DialogueGCN [27]: Models intra- and inter-speaker dependencies using a bidirectional GRU for sequential encoding and a speaker-level graph encoder. Nodes exchange contextual information via similarity-based attention. Originally text-only, we extended it to multimodal by incorporating visual and audio data with late fusion.
- BiDDIN [60]: Captures intra- and inter-modal dependencies using a bidirectional GRU for modality-specific encoding and a graph-based encoder for cross-modal interactions. Nodes refine representations via message passing, with edge weights set by similarity-based attention. Emotion classification is performed as a node classification task in a multimodal graph.
- MMGCN [7]: A multimodal, speaker-aware model using a graph-based fusion framework. A deep GCN refines node embeddings, integrating intra-speaker, inter-speaker, and cross-modal relationships. Through message passing, long-distance contextual information is aggregated for emotion classification as a node classification task.
- MM-DFN [61]: Utilizes a graph-based fusion mechanism for intra- and inter-modal dependencies. A modality-specific encoder processes features separately, while a dynamic fusion module filters redundancy and preserves complementary signals. Emotion classification is framed as a node classification task in a multimodal graph.

Since our method requires directly computing uni-modal scores from uni-modal logits, all baselines follow a late-fusion structure. To provide a clearer comparison, we also evaluate our approach against existing frameworks and modules designed to address imbalanced multimodal learning, including RNA loss [44], OGM-GE [9], and FAGM [12]. Specifically, RNA loss introduces constraints in the loss function to align feature norms across modalities, ensuring more balanced representations. OGM-GE mitigates modality imbalance by dynamically adjusting gradient updates based on the discrepancy ratio. FAGM is a plug-in method that rebalances gradients at the parameter level by proportionally adjusting them based on modality dominance, preventing any single modality from overwhelming the learning process. While originally developed for bimodal settings, these methods are extended to trimodal models in our experiments. All modules and frameworks are integrated into the baselines and evaluated under the same training environment for a fair comparison.

In order to assess the performance of our model, we employ two key evaluation metrics: the Weighted F1-score (w-F1) and Accuracy (Acc.). These metrics provide insights into the effectiveness and overall correctness of the predictions made by our classifier.

4.3 Multimodal raw feature extraction

The multimodal feature extraction process involves extracting features from the acoustic, lexical, and visual modalities for each utterance. For both IEMOCAP and MELD datasets, audio features are obtained using the OpenSmile Toolkit [62]; visual features are extracted using OpenFace [63]; textual features are derived using sBERT [64]. The hyper-parameter settings used in the experiments are presented in (Table 2).

Table 2 Hyper-parameters settings

Parameter/Module	IEMOCAP	MELD
Text feature extraction	sBERT [64]	
Audio feature extraction	OpenSmile Toolkit [62]	
Visual feature extraction	OpenFace Toolkit [63]	
Text embedding dim. d_t	768	768
Audio embedding dim. d_a	512	300
Visual embedding dim. d_v	1024	342
ε	[0.6, 1.2]	
α	[1.05, 1.4]	
Learning-rate	[0.0001, 0.0003]	
Batch size	16	32
Epoch	50	50

4.4 Reproducibility

SPCL is implemented using Pytorch[1], and run experiments on Google Colab and Kaggle. We choose Adam as the optimizer. The batch size is 16 and 32 for IEMOCAP and MELD dataset, respectively. Since each combination of baseline and dataset have different converging rates, the hyper-parameters are tested on various settings. Particularly, learning-rate is selected within the range of [0.0001, 0.0003]; hyper-parameter ε, i.e. initial value of threshold λ, is picked from range of [0.6, 1.2]; raging hyper-parameter α is selected from range of [1.05, 1.4].

5 Result and discussion

We qualitatively analyze our proposed Self-paced Curriculum Learning-based Approach (SPCL) and the baselines on the IEMOCAP and MELD datasets. We also conducted extensive experiments to prove the utility of each individual components of the Difficulty Measurer in the ablation study section.

5.1 Comparision with baselines

5.1.1 Analysis of experimental results on IEMOCAP

Table 3 presents a comparative performance analysis of our proposed SPCL module against multiple baselines on the IEMOCAP dataset. The results demonstrate that integrating SPCL consistently improves weighted F1-score (w-F1) and accuracy (Acc) across all modality combinations (TAV, TA, TV, AV), outperforming existing methods. The performance gap with other imbalance-mitigation methods (Δ) and the improvement over the original baseline model without any balancing strategy (Δ_{Base}) highlight the effectiveness of SPCL.

Overall Performance Improvements: Across all baseline models, our method achieves state-of-the-art performance, yielding the highest accuracy and weighted F1 scores, with statistically significant improvements over the strongest existing approach, such as FAGM. Notably, our approach demonstrates substantial gains in TAV and AV settings, where modality imbalance poses a significant challenge. Compared to the baseline without any balancing strategy, our method consistently delivers marked performance enhancements. In the DialogueGCN (TAV) setting, the baseline achieves 60.43% w-F1 and 60.54% accuracy, whereas our method significantly improves these to 66.99% w-F1 (+6.56%) and 67.03% accuracy (+6.49%). Similarly, in MM-DFN (TAV), our method surpasses the baseline by 5.62% in w-F1 and 5.66% in accuracy. The improvements are also consistent in the AV setting, where our method achieves 58.30% w-F1 and 58.47% accuracy for DialogueGCN, representing gains of +10.41% w-F1 and +9.98% accuracy over the baseline.

While FAGM achieves competitive performance in some cases, other methods such as RNA loss and OGM-GE frequently result in performance degradation. For instance, in DialogueGCN (TAV), RNA loss reduces w-F1 from 60.43% to 58.43% and accuracy from 60.54% to 58.47%. OGM-GE further degrades performance to 57.16% w-F1 and 57.24% accuracy. This indicates the limitations of static regularization approaches in handling modality imbalance. OPM yields only modest and inconsistent improvements over the baseline. For instance, in MM-DFN (AV), it raises w-F1 from 53.30% to 54.02%, yet remains 4.28% below our method's 58.30%. This indicates that fixed reweighting strategies like OPM are insufficient for capturing dynamic modality contributions under imbalance.

These results suggest that the success of our method stems from its ability to dynamically adapt to the evolving learning difficulty of samples and the shifting contributions of different modalities. Unlike static or manually designed weighting schemes, our SPCL framework leverages real-time feedback from both utterance-level performance and conversation-level modality discrepancies. This dual-level perspective enables the model to

[1] https://pytorch.org/.

Table 3 Performance comparison of baseline models with our SPCL module and other plug-in methods on IEMOCAP. **Bold** and underlined denote the best and second-best results, respectively. Δ indicates the performance gap to the previous SOTA, while Δ_{Base} measures the improvement of SPCL over the original baseline. Values marked with † denote statistically significant improvements ($p < 0.05$) based on paired t-tests

Model	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [27]								
Baseline	60.43	60.54	61.61	61.72	59.19	59.48	47.89	48.49
+ RNA loss	58.43	58.47	57.42	57.73	56.23	56.62	47.40	49.29
+ OGM-GE	57.16	57.24	59.30	59.52	55.88	56.13	43.71	44.98
+ OPM	58.89	59.72	57.02	57.55	60.48	60.54	49.80	51.76
+ FAGM	62.76	63.22	64.36	64.39	61.25	62.23	49.20	49.85
+ **SPCL**	$66.99^{\dagger}_{\pm 1.03}$	$67.03^{\dagger}$	$65.32^{\dagger}_{\pm 0.99}$	$65.46^{\dagger}$	$64.47^{\dagger}_{\pm 0.21}$	$64.46^{\dagger}$	$57.89^{\dagger}_{\pm 1.00}$	$58.59^{\dagger}$
Δ	*4.23*	*3.81*	*0.96*	*1.07*	*3.22*	*2.23*	*8.09*	*6.83*
Δ_{Base}	*6.56*	*6.49*	*3.71*	*3.74*	*5.28*	*4.98*	*10.00*	*10.10*
BiDDIN [60]								
Baseline	58.29	58.20	58.73	58.67	58.57	57.93	45.35	46.03
+ RNA loss	58.63	58.55	58.02	57.92	57.29	57.24	42.54	44.82
+ OGM-GE	58.06	57.98	57.71	57.73	57.58	57.55	39.84	40.42
+ OPM	56.27	56.62	57.82	57.60	52.59	52.60	37.72	40.48
+ FAGM	58.81	58.84	58.88	58.16	59.04	58.96	**46.36**	46.77
+ **SPCL**	$59.90^{\dagger}_{\pm 0.13}$	$60.73^{\dagger}$	$60.24^{\dagger}_{\pm 1.11}$	$60.43^{\dagger}$	$61.10^{\dagger}_{\pm 0.82}$	$61.91^{\dagger}$	$46.34_{\pm 0.43}$	$49.11^{\dagger}$
Δ	*1.09*	*1.89*	*1.36*	*1.76*	*2.06*	*2.95*	*−0.02*	*2.34*
Δ_{Base}	*1.61*	*2.53*	*1.51*	*1.76*	*2.53*	*3.98*	*0.99*	*3.08*
MMGCN [7]								
Baseline	62.67	62.67	62.66	62.72	58.99	59.14	47.22	49.23
+ RNA loss	63.13	63.28	59.25	59.27	56.30	56.50	50.35	51.20
+ OGM-GE	62.42	62.69	62.33	62.42	58.83	59.03	51.90	53.54
+ OPM	64.60	64.10	62.30	62.70	59.70	59.60	50.60	52.00
+ FAGM	64.53	64.51	63.25	63.40	61.02	61.06	**54.14**	**54.90**
+ **SPCL**	$67.66^{\dagger}_{\pm 0.57}$	$67.71^{\dagger}$	$66.75^{\dagger}_{\pm 0.42}$	$66.51^{\dagger}$	$65.00^{\dagger}_{\pm 1.05}$	$65.09^{\dagger}$	$53.70_{\pm 0.71}$	$54.04_{\pm 0.94}$
Δ	*3.06*	*3.20*	*3.50*	*3.11*	*3.98*	*4.03*	*−0.44*	*−0.86*
Δ_{Base}	*5.00*	*5.04*	*4.09*	*3.79*	*6.01*	*5.95*	*6.48*	*4.81*
MM-DFN [61]								
Baseline	61.54	61.72	61.98	62.12	59.78	59.93	48.42	49.11
+ RNA loss	60.23	60.49	60.18	60.41	57.74	57.92	45.63	46.32
+ OGM-GE	59.92	60.13	60.57	60.69	58.33	58.49	44.98	45.51
+ OPM	63.30	62.91	64.43	64.45	64.06	63.89	53.55	53.79
+ FAGM	63.45	63.72	63.83	63.94	61.58	61.72	50.35	51.02
+ **SPCL**	$67.16^{\dagger}_{\pm 0.67}$	$67.08^{\dagger}$	$66.03^{\dagger}_{\pm 0.86}$	$66.09^{\dagger}$	$64.31^{\dagger}_{\pm 0.66}$	$64.70^{\dagger}$	$53.38^{\dagger}_{\pm 0.67}$	$53.47^{\dagger}$
Δ	*3.71*	*3.36*	*1.60*	*1.64*	*0.25*	*0.81*	*−0.17*	*−0.32*
Δ_{Base}	*5.62*	*5.36*	*4.05*	*3.97*	*4.53*	*4.77*	*4.96*	*4.36*

prioritize informative yet underrepresented modalities and to avoid overfitting to dominant signals. As a result, the training process becomes more balanced and effective, leading to superior generalization performance across various MERC settings.

Impact on Modality Combinations: Among different modality combinations, the TAV setting exhibits the most substantial improvements with SPCL, effectively addressing modality imbalance. Across models, SPCL outperforms FAGM, achieving w-F1 gains ranging from 1.17% to 3.70%, demonstrating the benefits of adaptive sample selection in enhancing multimodal alignment. The TA and TV settings also experience consistent improvements, particularly in MMGCN when integrating SPCL compared to integrating FAGM, where accuracy increases from 63.26% to 66.15% (+2.89%), and in MM-DFN, where it improves from 63.94% to 66.80%

(+2.86%). This suggests that SPCL effectively strengthens the interaction between textual and non-textual modalities.

The AV setting, which poses the greatest challenge due to the absence of textual features, exhibits the most pronounced improvements. In DialogueGCN, SPCL surpasses FAGM, improving w-F1 from 49.20% to 57.98% and accuracy from 49.85% to 58.49%, achieving gains of 8.78% and 8.64%, respectively. Similarly, in MM-DFN, SPCL enhances w-F1 from 50.04% to 58.30% and accuracy from 50.91% to 58.47%, with improvements of 8.26% and 7.56%. These findings highlight the robustness of SPCL in optimizing non-textual modality fusion, making it particularly effective in overcoming modality imbalance.

5.1.2 Analysis of experimental results on MELD

Table 4 presents a comparative performance analysis of our proposed SPCL module against multiple baselines on the MELD dataset. Similar to IEMOCAP, integrating SPCL consistently improves weighted F1-score (w-F1) and accuracy (Acc) across all modality combinations (TAV, TA, TV, AV), surpassing existing approaches.

Overall Performance Improvements: Our method consistently achieves the highest performance across all baseline models in the TAV setting, outperforming competitive approaches such as FAGM. For example, in MM-DFN, with SPCL integrated, our method improves the weighted F1-score from 57.55% (FAGM) to 59.17% (+1.62%) and from the baseline's 57.52%, yielding a total gain of +1.65%. Similarly, in MMGCN, SPCL increases the weighted F1-score from 58.48% (FAGM) to 59.11% (+0.63%) and over the baseline's 57.71%, achieving a total improvement of +1.40%. Across all evaluated models in the TAV setting, SPCL achieves an average weighted F1-score improvement of 0.85% over the second-best method and 2.25% over the baseline models, demonstrating consistent effectiveness in enhancing multimodal interactions.

While FAGM remains competitive, SPCL demonstrates a more adaptive learning strategy, particularly within transformer-based architectures. For instance, in MM-DFN on MELD, SPCL surpasses the second-best method OPM by 0.42% in the TAV setting (59.17% vs. 58.75%). Consistent with findings on IEMOCAP, static regularization techniques such as RNA loss and OGM-GE often fail to deliver consistent performance improvements. While RNA loss improves performance in certain cases (e.g., 56.65% weighted F1-score in DialogueGCN's TAV setting), it does not consistently achieve the best results across different models.

However, in DialogueGCN on MELD, our method does not consistently yield superior performance. In the TAV setting, SPCL achieves a weighted F1-score of 57.87%, which is only 0.14% higher than OGM-GE (57.73%). The limited effectiveness of our curriculum-based training on MELD may be attributed to the dataset's shorter and more fragmented conversational structure. As SPCL progressively introduces more complex samples, its training schedule might not align optimally with MELD's data distribution, thereby limiting its potential gains within this specific architecture.

Impact of Different Modality Combinations: The performance trends across modality combinations on MELD are largely consistent with those observed on IEMOCAP, further validating the effectiveness of our proposed approach. The TAV setting particularly benefits from SPCL, as its adaptive sample selection enhances multimodal balance and improves overall recognition performance. Additionally, the TA and TV settings exhibit notable improvements, demonstrating the capacity of SPCL to mitigate modality imbalance across diverse multimodal configurations.

Similar to IEMOCAP, SPCL consistently outperforms FAGM across models in the TA, TV, and AV settings. In the TA setting, SPCL achieves weighted F1-score improvements ranging from 1.39% to 1.75% over FAGM, with the most pronounced gains observed in BiDDIN (+2.08%) and MM-DFN (+2.41%) relative to the baseline. In the TV setting, SPCL maintains superior performance, particularly in BiDDIN (+3.54%) and MMGCN (+2.04%) over the baseline model. For the AV setting, while the improvements over competing methods are more moderate, SPCL attains the highest weighted F1-score in MM-DFN (42.42%) and MMGCN (44.34%), with notable gains in MM-DFN (+2.38%) compared to the baseline. These findings further reinforce the efficacy of SPCL in addressing modality imbalance and enhancing multimodal emotion recognition in conversations.

Table 4 Performance comparison of baseline models with our SPCL module and other plug-in methods on MELD. **Bold** and underlined denote the best and second-best results, respectively. Δ indicates the performance gap to the previous SOTA, while Δ_{Base} measures the improvement of SPCL over the original baseline. Values marked with † denote statistically significant improvements ($p < 0.05$) based on paired t-tests

Model	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [27]								
Baseline	53.11	55.08	51.99	54.22	54.22	56.07	43.54	44.54
+ RNA loss	56.65	58.47	54.21	58.35	53.78	58.12	**43.64**	**47.32**
+ OGM-GE	57.73	57.36	**56.38**	**58.81**	56.15	57.78	42.05	46.51
+ OPM	54.47	57.12	53.26	56.17	53.21	57.66	40.52	43.64
+ FAGM	54.61	58.96	54.80	57.28	55.26	57.10	40.02	44.44
+ **SPCL**	**57.87**$_{\pm 1.49}$	60.77†	58.04$_{\pm 0.56}$	60.84	**56.18**$_{\pm 1.38}$	58.61†	42.28$_{\pm 0.79}$	46.64
Δ	*0.14*	*1.81*	*1.66*	*2.03*	*0.03*	*0.49*	*−1.36*	*−0.68*
Δ_{Base}	*4.76*	*5.69*	*6.05*	*6.62*	*1.96*	*2.54*	*−1.26*	*2.10*
BiDDIN [60]								
Baseline	56.41	58.54	56.23	57.85	56.46	58.06	43.07	47.35
+ RNA loss	52.18	49.16	53.21	50.31	52.59	49.43	41.05	44.60
+ OGM-GE	55.27	53.41	51.96	47.74	52.18	48.58	43.03	46.97
+ OPM	53.87	57.62	54.73	58.58	56.25	59.77	40.69	47.39
+ FAGM	57.47	59.18	56.56	58.05	56.93	58.10	**44.39**	**48.62**
+ **SPCL**	57.60$^{\dagger}{}_{\pm 0.25}$	60.86†	58.08$^{\dagger}{}_{\pm 0.30}$	61.22†	58.10$^{\dagger}{}_{\pm 0.43}$	61.00†	42.30$_{\pm 0.23}$	48.15
Δ	*0.13*	*1.68*	*1.52*	*2.64*	*1.17*	*1.23*	*−2.09*	*−0.47*
Δ_{Base}	*1.19*	*2.32*	*1.85*	*3.37*	*1.64*	*2.94*	*−0.77*	*1.12*
MMGCN [7]								
Baseline	57.71	59.95	57.29	59.79	56.73	59.31	42.38	**49.12**
+ RNA loss	56.94	58.62	56.00	57.59	55.48	57.70	41.84	46.91
+ OGM-GE	57.59	59.92	56.80	59.77	56.20	59.08	42.20	48.81
+ OPM	55.78	57.24	56.27	59.77	55.29	59.23	42.72	47.20
+ FAGM	58.48	61.15	57.59	60.69	57.14	59.46	43.49	48.43
+ **SPCL**	59.11$^{\dagger}{}_{\pm 0.48}$	61.32†	58.93$^{\dagger}{}_{\pm 0.29}$	61.65†	58.14$^{\dagger}{}_{\pm 1.17}$	60.64†	43.79$^{\dagger}{}_{\pm 0.31}$	49.10
Δ	*0.63*	*0.17*	*1.34*	*0.96*	*1.00*	*1.18*	*0.30*	*−0.02*
Δ_{Base}	*1.40*	*1.37*	*1.64*	*1.86*	*1.41*	*1.33*	*1.41*	*−0.02*
MM-DFN [61]								
Baseline	57.52	59.90	57.11	59.47	57.46	59.68	40.04	43.91
+ RNA loss	56.02	58.20	54.13	55.59	54.13	55.59	36.39	47.54
+ OGM-GE	56.53	58.39	55.86	59.08	56.25	58.24	40.60	48.43
+ OPM	58.75	61.42	57.67	61.38	58.28	61.49	**42.51**	47.16
+ FAGM	57.55	60.80	57.10	60.00	57.73	60.65	42.05	**48.66**
+ **SPCL**	59.17$^{\dagger}{}_{\pm 0.30}$	61.91†	59.11$^{\dagger}{}_{\pm 0.32}$	62.31†	58.91$^{\dagger}{}_{\pm 0.17}$	61.94†	43.32$_{\pm 0.57}$	48.59†
Δ	*0.42*	*0.49*	*1.44*	*0.93*	*0.63*	*0.45*	*0.81*	*−0.07*
Δ_{Base}	*1.65*	*2.01*	*2.00*	*2.84*	*1.45*	*2.26*	*3.28*	*4.68*

5.2 Discussion and analysis

In this section, we provide further analysis and insights into the effectiveness of our proposed SPCL framework.

5.2.1 Impact of key components

We conduct an ablation study to evaluate the impact of the two key components in our Difficulty Measurer: the utterance-level score l_{ij} and the conversation-level score s_i. Specifically, we systematically remove each component from the difficulty formulation of ρ_{ij} in Eq. 6 and assess the resulting performance, as summarized in Table 5.

Table 5 Ablation study on IEMOCAP for our proposed SPCL module. The subscript ↓ or ↑ denotes the performance change compared to our SPCL module when a sub-module is ablated

Method	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [27]	60.43	60.54	61.61	61.72	59.19	59.48	47.89	48.49
+ SPCL (Ours)	**66.99**	**67.03**	**65.32**	65.46	**64.47**	**64.46**	**57.89**	**58.59**
w/o utt-score	$63.11_{\downarrow 3.88}$	63.24	$65.31_{\downarrow 0.01}$	65.72	$63.56_{\downarrow 0.91}$	63.72	$55.29_{\downarrow 2.60}$	56.13
w/o conv-score	$64.59_{\downarrow 2.40}$	64.94	$64.87_{\downarrow 0.45}$	**65.66**	$63.49_{\downarrow 0.98}$	63.86	$55.25_{\downarrow 2.64}$	56.06
BiDDIN [60]	58.29	58.20	58.73	58.67	58.57	57.93	45.35	46.03
+ SPCL (Ours)	**59.90**	**60.73**	59.40	60.24	**61.10**	**61.91**	**46.34**	**49.11**
w/o utt-score	$57.59_{\downarrow 2.31}$	59.18	$\mathbf{60.41}_{\uparrow 1.01}$	**60.59**	$60.67_{\downarrow 0.43}$	61.28	$45.02_{\downarrow 1.32}$	48.31
w/o conv-score	$58.61_{\downarrow 1.29}$	59.22	$59.14_{\downarrow 0.26}$	60.46	$59.48_{\downarrow 1.62}$	60.08	$45.41_{\downarrow 0.93}$	48.50
MMGCN [7]	62.67	62.67	62.66	62.72	58.99	59.14	47.22	49.23
+ SPCL (Ours)	**67.66**	**67.71**	65.62	65.84	66.01	65.91	**53.70**	54.04
w/o utt-score	$64.60_{\downarrow 3.06}$	65.05	$63.89_{\downarrow 1.73}$	64.01	$62.55_{\downarrow 3.46}$	62.75	$50.31_{\downarrow 3.39}$	52.18
w/o conv-score	$65.78_{\downarrow 1.88}$	65.85	$\mathbf{66.02}_{\uparrow 0.40}$	**65.98**	$\mathbf{66.11}_{\uparrow 0.10}$	**66.07**	$52.00_{\downarrow 1.70}$	**55.32**
MM-DFN [61]	61.84	61.84	61.95	62.04	60.32	60.37	50.96	52.87
+ SPCL (Ours)	**67.16**	**67.08**	**66.09**	**66.51**	**65.43**	**64.91**	**53.38**	**57.40**
w/o utt-score	$66.46_{\downarrow 0.70}$	66.39	$65.18_{\downarrow 0.91}$	65.45	$63.73_{\downarrow 1.70}$	63.94	$52.11_{\downarrow 1.27}$	52.53
w/o conv-score	$64.49_{\downarrow 2.67}$	64.54	$65.20_{\downarrow 0.89}$	65.45	$63.95_{\downarrow 1.48}$	64.29	$50.93_{\downarrow 2.45}$	53.13

Overall, across all baseline models and modality settings, removing either component leads to consistent performance degradation, confirming their complementary roles in SPCL. The performance drop is particularly pronounced in the TAV setting, which involves the full modality set and exhibits more complex inter-modal dynamics. For instance, in DialogueGCN (TAV), removing the utterance-level score causes a drop of 3.88% in weighted F1, while removing the conversation-level score results in a 2.67% decrease.

The utterance-level score proves to be especially critical, as its removal leads to substantial and consistent performance drops across multiple models and settings (e.g., -3.06% in MMGCN (TAV), -1.70% in MMDFN (TV)). This highlights its importance in capturing fine-grained, modality-specific discrepancies at the local level, thus guiding SPCL in effective pacing and intra-utterance balancing.

Conversely, the conversation-level score contributes to modeling broader, global-level patterns such as turn-wise modality shifts or long-range emotional dependencies. While its removal leads to smaller declines compared to the utterance-level score, it still yields meaningful gains when present (e.g., $+2.64\%$ in DialogueGCN (AV), $+1.62\%$ in BiDDIN (TV)).

In a few cases, using only one of the two difficulty scores still surpasses the baseline performance (e.g., BiDDIN (TV) without conv-score achieves $+0.91\%$ over the baseline), underscoring the independent utility of each score. However, the full SPCL module consistently achieves the best results across all cases, reaffirming that both levels of difficulty modeling are necessary for addressing the diverse imbalance patterns in MERC.

5.2.2 Curricula expanding rate and hyper-parameters tuning

We define the curriculum expanding rate as the ratio of easy samples to the total samples at each training epoch. This rate ranges between 0 and 1, where a value of 1 indicates training on the entire dataset. However, it is not guaranteed to increase consistently unless carefully tuned. The expanding rates of various baselines, when integrated with our module, are illustrated in Fig. 3.

This rate is directly influenced by the tuning of ε and α, which can be explained through the updating of λ in Eq. 8, and varies depending on the baseline architecture, as different models exhibit unique sensitivity to data distribution.

Our study on the curriculum expanding rate reveals that the best performance is achieved when the rate maintains a consistently increasing trend, as exemplified by DialogueGCN. This suggests that a gradual yet steady

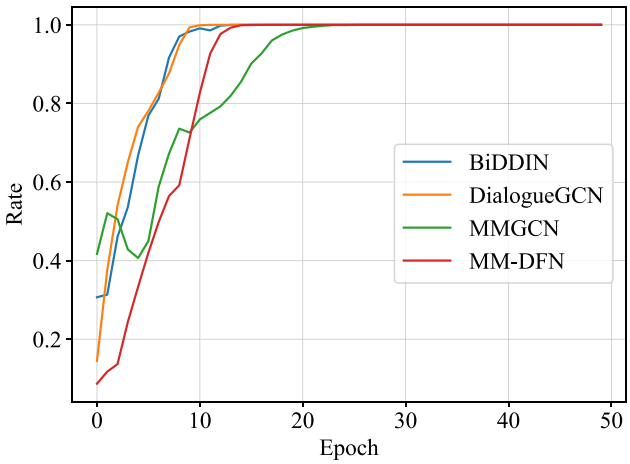
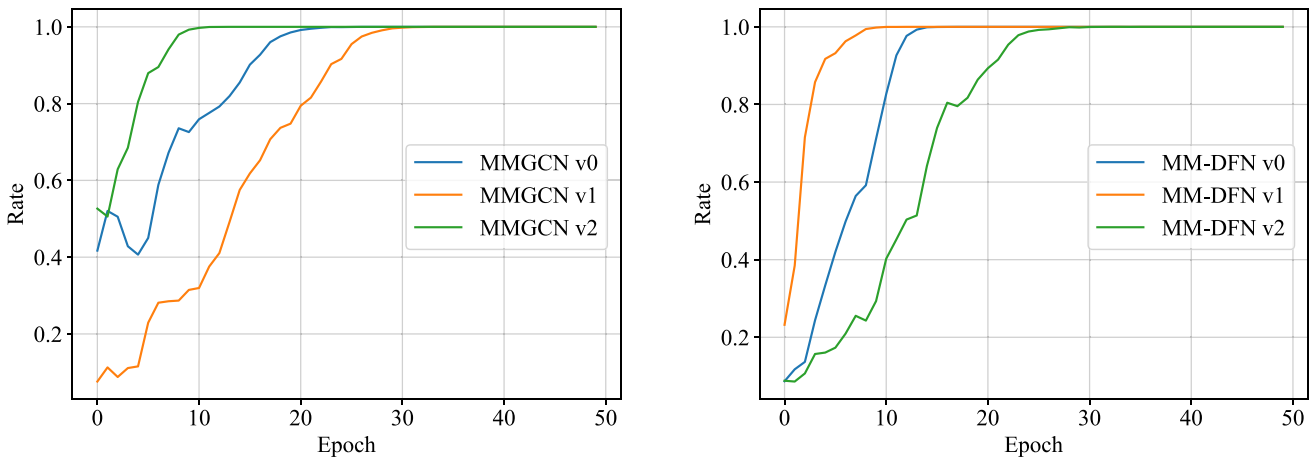


Fig. 3 The curricula expanding rate of the four baselines integrated on IEMOCAP

Table 6 Performance of MMGCN and MM-DFN on IEMOCAP under different hyper-parameter settings for our SPCL module, with v0 representing the best configuration

Model	Version	ε	α	w-F1 (%)	Acc (%)
MMGCN	v0	0.8	1.1	**67.84**	**67.84**
	v1	0.4	1.1	67.19	64.02
	v2	0.8	1.2	65.84	65.56
MM-DFN	v0	0.4	1.2	**67.92**	**68.21**
	v1	0.8	1.2	67.45	67.80
	v2	0.6	1.1	66.83	66.51



(a) MMGCN (b) MM-DFN

Fig. 4 Curricula expanding rate of MMGCN and MM-DFN under *SPCL hyper-parameters setting* specified in Table 6

introduction of complex samples enhances the learning progression of the model. Furthermore, this expanding rate highlights the critical role of early training phases in shaping overall model performance.

Since hyperparameter tuning is crucial, we further investigate this by conducting an ablation study on MMGCN and MMDFN using the IEMOCAP dataset. We experiment with different hyperparameter settings and analyze their impact on the curriculum expanding rate, as illustrated in the corresponding Table 6 and Fig. 4. Our findings indicate that ε and α are proportional to the expanding rate, meaning that the rate can be sped up by increasing these values or slowed down by decreasing them.

From our experiments, we observe that each model is optimized for a specific expanding rate. The v0 setting yields the best performance, whereas both speeding up and slowing down (v1+v2) the expanding rate result in

performance degradation. Our intuitive explanation for this phenomenon is that if the expanding rate is too fast, weak modalities with slower learning rates will fail to fully exploit easy samples, leading to an unreliable starting point and degrading the training process later on. Conversely, if the expanding rate is too slow, the model tends to overfit on easy samples and struggles to learn from hard examples due to mismatched data distribution, ultimately resulting in poor generalization.

5.2.3 Analysis of regularization strategy

We conducted a comprehensive comparison between our proposed hard regularizer and two alternative soft regularization strategies, namely the Linear and Logistic regularizers. The implementations of these soft regularizers follow the closed-form formulations described in [58].

As shown in Table 7 and illustrated in Fig. 5, the hard regularizer consistently achieves superior or at least comparable performance across all evaluated backbones. For example, in the MMGCN model on the IEMOCAP dataset, the hard regularizer attains a weighted F1-score of 67.84%, outperforming both the Linear (65.20%) and Logistic (65.98%) regularizers. A similar trend is observed on the MELD dataset, where the hard regularizer achieves a weighted F1-score of 59.35% in MMGCN, surpassing the Linear (58.43%) and Logistic (59.02%) alternatives.

Beyond accuracy, the hard regularizer also leads to more stable performance across modalities, contributing to reducing the discrepancy caused by modality imbalance. These findings confirm that a hard regularization strategy is more effective for our dual objectives: improving overall model performance and managing modality imbalance in multimodal emotion recognition.

5.2.4 Analysis of Pacing Strategy

We further study alternative strategies for updating the difficulty threshold λ by adopting the following methods: cosine pacing, moving average(MA) pacing, and competence-based(CB) pacing [65]. The exponential pacing used in SPCL is described in Eq. 8, whereas the formulations of newly adopted strategies are described in Table 8.

As shown in Fig. 6, linear pacing strategies, i.e., exponential and cosine pacing, yield smoother updates of λ . In contrast, the two non-linear strategies, where λ is adaptively updated with regards to sample difficulty ρ_{ij} , exhibit larger fluctuations, particularly under competence-based pacing. Consequently, linear pacing results in a smoother curriculum expansion, indicating a more stable introduction of new samples. Table 9 further shows that exponential and cosine pacing achieve better overall performance. These results highlight the importance of selecting an appropriate pacing strategy to ensure stable curriculum progression.

Table 7 Performance comparison of four backbone models on IEMOCAP and MELD datasets using different types of regularizers for the learning scheduler

Regularizer	MMGCN		DialogueGCN		BiDDIN		MM-DFN	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
IEMOCAP								
Hard Regularizer	**67.84**	**68.02**	66.46	66.61	**59.98**	**60.07**	**67.92**	**68.21**
Soft Linear	65.20	65.13	64.87	64.70	56.77	57.18	65.94	65.80
Soft Logistic	65.98	66.17	**67.14**	**67.41**	58.16	59.52	64.79	64.88
MELD								
Hard Regularizer	**59.35**	**61.72**	55.37	60.38	**57.76**	**60.50**	**59.14**	**62.07**
Soft Linear	58.43	60.73	54.55	60.12	57.64	59.62	57.64	59.62
Soft Logistic	59.02	61.17	**56.07**	**60.84**	56.94	59.72	58.27	61.26

The best performance for each dataset and backbone is highlighted in **bold**

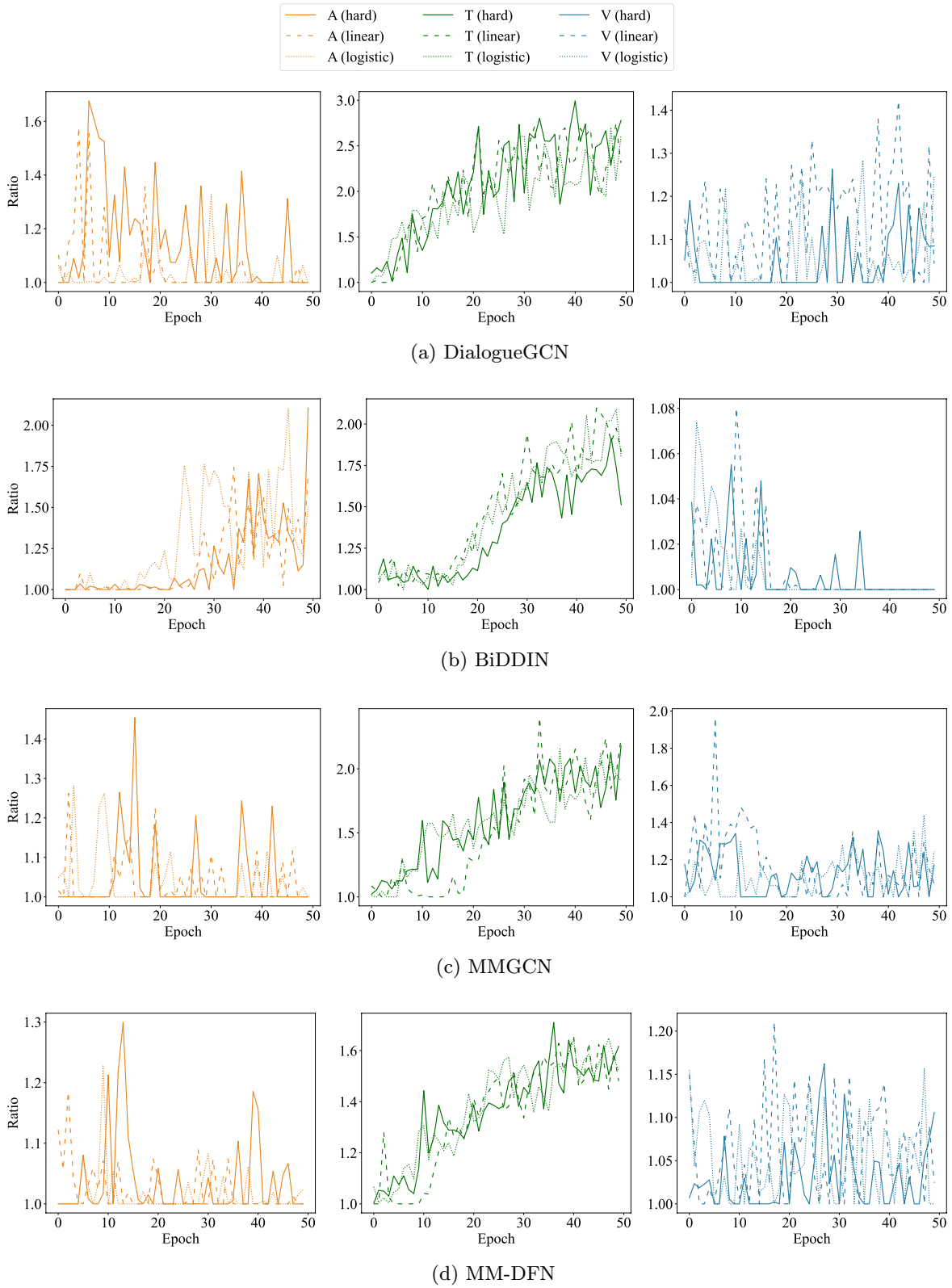


Fig. 5 Modality ratio of the four backbones during training on the IEMOCAP dataset using different types of regularizer for the Learning Scheduler

 https://doi.org/10.1007/s00521-026-12160-6

Table 8 Formulations of experimented pacing strategies. T and t denote total training epoch and current training epoch, respectively

Strategy	Formulation
Cosine	$\lambda^{(t)} = \lambda_{\min} + \frac{\lambda_{\max} - \lambda_{\min}}{2} \cdot (1 - \cos(\frac{\pi \cdot t}{T}))$
MA	$\lambda^{(t)} = \begin{cases} \alpha\lambda^{(t-1)} + (1-\alpha) \cdot \sum_i^{\lvert\mathcal{D}\rvert} \sum_j^{N_i} \rho_{ij} & \text{if } t < t_0 \\ \max \rho_{ij} & \text{if } t \geq t_0 \end{cases}$
CB	$c_t = \min\left(1, \sqrt{t\frac{1-c_0^2}{T} + c_0^2}\right)$
	$\lambda^{(t)} = \text{Quantile}(\rho_{ij}, c_t)$

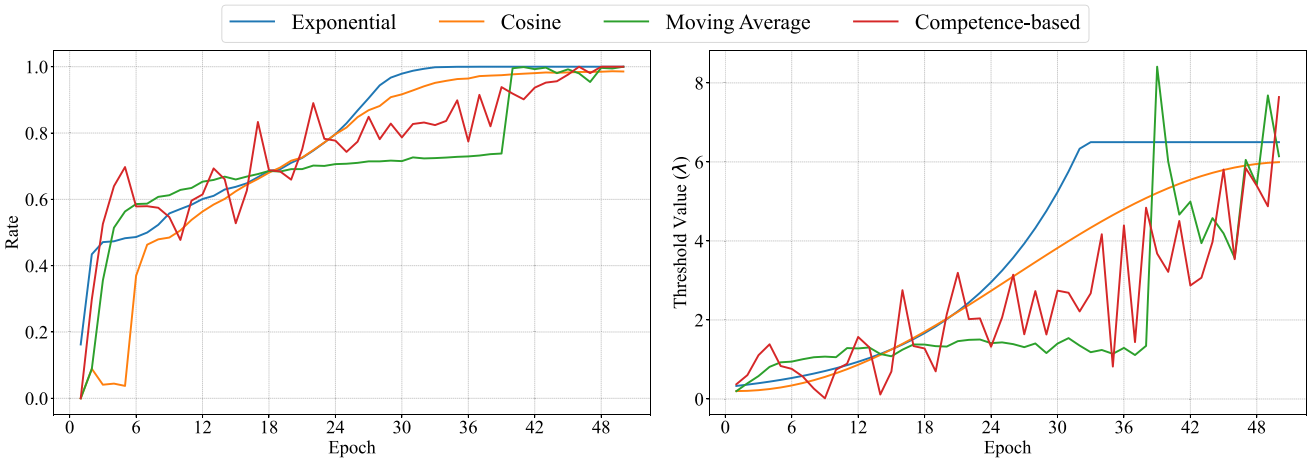


Fig. 6 Curricula expanding rate and respective threshold value of MMGCN on MELD under different pacing strategies

Table 9 Performance comparison of MMGCN and MM-DFN on IEMOCAP and MELD dataset using different pacing strategies. The best and second-best performances for each dataset and backbone are highlighted in **bold** and underline

Strategy	*IEMOCAP*				*MELD*			
	MMGCN		MM-DFN		MMGCN		MM-DFN	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
Exponential	**67.84**	**68.02**	**67.92**	**68.21**	**59.35**	61.72	**59.14**	**62.07**
Cosine	65.47	65.00	67.63	67.53	58.10	61.23	57.06	61.11
MA	62.63	62.91	67.55	67.34	57.78	60.77	56.46	57.78
CB	63.80	63.52	66.68	66.42	58.71	62.34	58.44	60.61

5.2.5 Modality ratio

Our study aims to achieve two key objectives: *(1) enhancing the performance of tri-modal models relative to their bi-modal and uni-modal counterparts and (2) mitigating modality imbalance during training*. To further investigate the latter, we analyze the modality ratio, which quantifies each modality's contribution relative to the weakest modality throughout training.

As depicted in Fig. 7, the integration of our SPCL module effectively reinforces the weaker modalities across all baseline models. Specifically, we observe an increase in the audio modality ratio by 0.2 to 0.5 and an increase of 0.15 in the visual modality ratio. Concurrently, our approach reduces the dominance of the strongest modality (i.e. text). This effect is particularly notable in MMGCN, where the text modality ratio decreases from 3 to 2, indicating a more balanced learning process. These findings confirm that our method successfully addresses modality imbalance by narrowing the gap between strong and weak modalities, ensuring a more equitable contribution from all modalities.

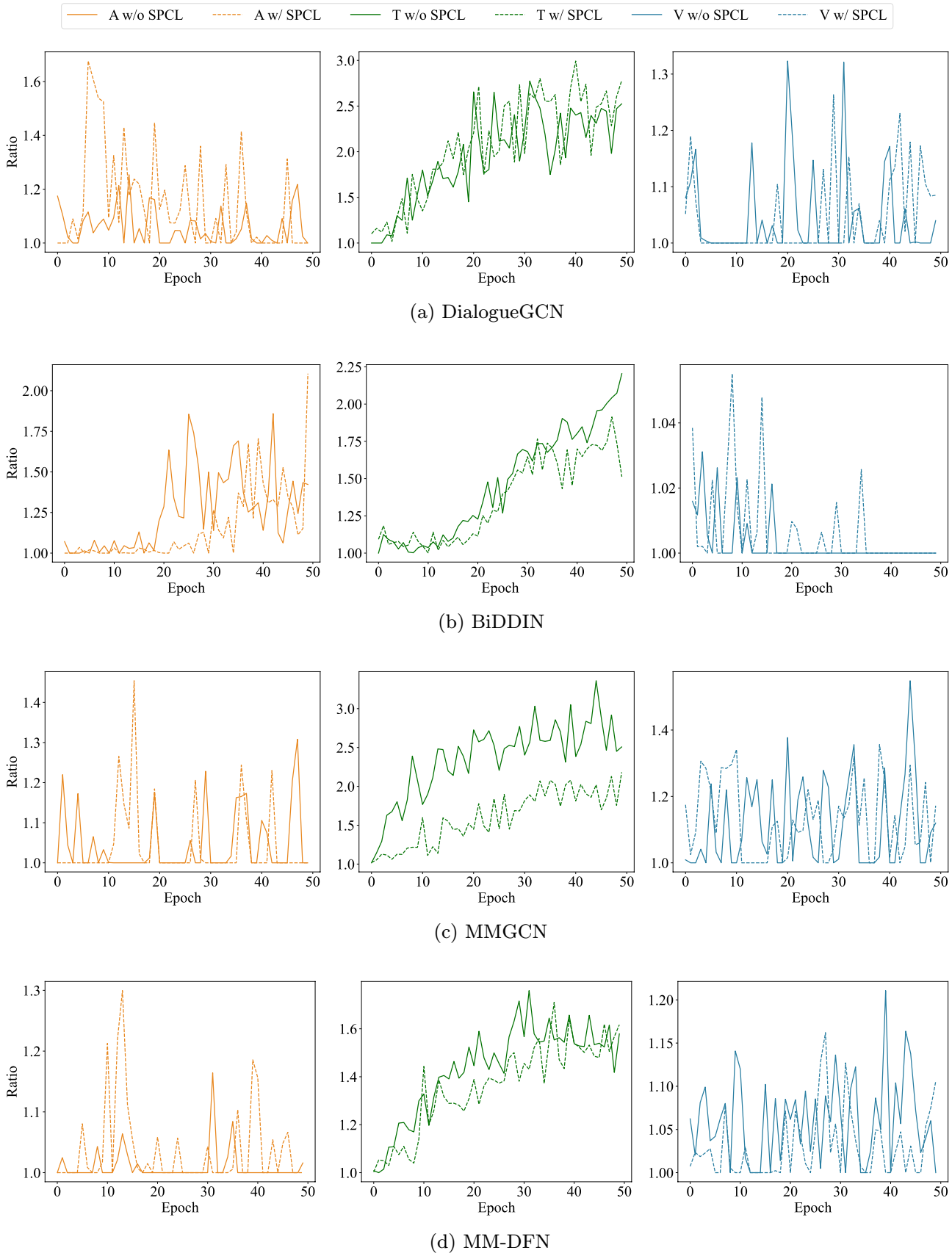


Fig. 7 Modality ratio of the four backbones during training on the IEMOCAP dataset

5.2.6 Limitations

Although SPCL introduces negligible architectural overhead, several practical considerations remain. First, its effectiveness is sensitive to backbone architectures and dataset characteristics, often requiring extensive hyperparameter tuning. As discussed in Sect. 5.2.2, brute-force strategies such as grid search are impractical for large-scale models, underscoring the need for a more general and adaptive tuning strategy. Second, the Difficulty Measurer operates at the batch level rather than over the entire dataset, which may lead to inaccurate difficulty estimation when easy or hard samples are unevenly distributed. Increasing batch size can mitigate this, though it also raises computational demands, highlighting a trade-off between estimation reliability and resource efficiency.

From a scalability perspective, SPCL was designed as a plug-in module with minimal computational overhead. At each training step, the Difficulty Measurer reuses model outputs and losses to compute utterance- and conversation-level scores, while the Learning Scheduler adjusts sample weights without redundant computation. With efficient masking and matrix operations implemented in PyTorch and NumPy, the additional training time remains moderate (e.g., an average increase of around 10 s per epoch for MMGCN on MELD compared to IEMOCAP).

Data-wise, SPCL's curriculum progression may not always align with datasets containing brief or fragmented dialogues. Moreover, both benchmark datasets (IEMOCAP and MELD) consist of scripted dialogues, which may not fully capture the dynamics of spontaneous emotional interactions. Future work will extend SPCL to more naturalistic datasets (e.g., K-EmoCon [66]) to evaluate its robustness and generalizability in real-world scenarios.

Finally, since SPCL operates solely during the training phase and leaves the baseline model architecture unchanged, it can be seamlessly integrated into various multimodal frameworks and does not affect model deployment, further demonstrating its extensibility and practicality for large-scale conversational emotion recognition.

6 Conclusion

In this work, we have introduced SPCL, a plug-and-play module designed to address modality imbalance in Multimodal Emotion Recognition in Conversation (MERC). Our approach leverages Self-Paced Curriculum Learning to dynamically mitigate modality discrepancies during training, thereby promoting more balanced multimodal representation learning. Specifically, SPCL comprises two key components: (1) a Difficulty Measurer, which quantifies sample complexity at both the utterance and conversation levels based on loss dynamics and modality alignment, and (2) a Learning Scheduler, which adaptively regulates the training curriculum, progressing from easier to more complex samples. Extensive experiments on IEMOCAP and MELD validate the effectiveness of SPCL, demonstrating consistent improvements in both w-F1 and accuracy across multiple baselines. Moreover, our approach effectively strengthens weaker modalities while mitigating the over-reliance on dominant ones, leading to a more balanced multimodal representation. Ablation studies further highlight the necessity of integrating both utterance- and conversation-level information and maintaining a progressive curriculum expansion rate for optimal convergence.

We further show that maintaining an appropriate curriculum expanding rate is essential. A rate that grows steadily over training yields better results, as it allows weaker modalities sufficient exposure to easier samples while preventing overfitting on simple instances. Additionally, our comparison between regularization strategies demonstrates that the hard regularizer consistently achieves superior or comparable performance to soft regularizers, offering more stable learning across modalities and better managing modality imbalance. Moreover, our analysis of modality ratios confirms that SPCL effectively reduces the dominance of strong modalities, such as text, and enhances contributions from weaker ones like audio and visual inputs, thereby addressing modality imbalance more comprehensively.

In future work, we aim to enhance our approach by adapting the curriculum scheduling strategy to better reflect dataset-specific traits, exploring alternative difficulty measures suited to brief dialogues, and integrating

adaptive pacing mechanisms to ensure greater compatibility with diverse conversational formats. Further extensions include expanding SPCL for Large Multimodal Model (LMMs), and to cross-lingual datasets. To support these developments, comprehensive studies on hyperparameter behaviors and evaluations on more diverse datasets will be conducted to address current limitations and ensure robust scalability. Additionally, we find extending SPCL to the task of sentiment analysis in conversations presents a feasible and meaningful direction, given its close similarity in data characteristics and backbone architectures to multimodal emotion recognition.

Data availability The experimental data in this study are from the IEMOCAP (https://sail.usc.edu/iemocap/iemocap.htm) and MELD (https://affective-meld.github.io/).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence this work.

Code availability The code will be made available upon publications.

References

1. Yuan Y, Li Z, Zhao B (2025) A survey of multimodal learning: methods, applications, and future. ACM Comput Surv 57(7):1–34
2. Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
3. Liang PP, Lyu Y, Fan X, Wu Z, Cheng Y, Wu J, Chen L, Wu P, Lee MA, Zhu Y et al (2021) Multibench: Multiscale benchmarks for multimodal representation learning. Adv Neural Inf Process Syst 2021(DB1):1
4. Liang PP, Zadeh A, Morency LP (2024) Foundations & trends in multimodal machine learning: principles, challenges, and open questions. ACM Comput Surv 56(10):1–42
5. Gladys AA, Vetriselvi V (2023) Survey on multimodal approaches to emotion recognition. Neurocomputing 556:126693
6. Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S (2020) COSMIC: COmmonSense knowledge for eMotion identification in conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2470–2481. Association for Computational Linguistics, Online https://doi.org/10.18653/v1/2020.findings-emnlp.224
7. Hu J, Liu Y, Zhao J, 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), pp 5666–5675
8. Nguyen CVT, Mai T, The S, Kieu D, Le DT (2023) Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 15154–15167. https://doi.org/10.18653/v1/2023.emnlp-main.937
9. Peng X, Wei Y, Deng A, Wang D, Hu D (2022) Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8238–8247
10. Wang W, Tran D, Feiszli M (2020) What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12695–12705
11. Shi Q, Ye M, Huang W, Du B, Zong X (2025) Gradient and structure consistency in multimodal emotion recognition. IEEE Trans Image Process. https://doi.org/10.1109/TIP.2025.3608664
12. Wang Y, Liu M, Li Z, Hu Y, Luo X, Nie L (2023) Unlocking the power of multimodal learning for emotion recognition in conversation. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 5947–5955
13. Zhang Q, Wei Y, Han Z, Fu H, Peng X, Deng C, Hu Q, Xu C, Wen J, Hu D et al (2024) Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv:2404.18947
14. Du C, Teng J, Li T, Liu Y, Yuan T, Wang Y, Yuan Y, Zhao H (2023) On uni-modal feature learning in supervised multi-modal learning. In: International Conference on Machine Learning, PMLR, pp 8632–8656
15. Wu N, Jastrzebski S, Cho K, Geras KJ (2022) Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In: International Conference on Machine Learning, PMLR, pp 24043–24055
16. Xu R, Feng R, Zhang SX, Hu D (2023) Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5

17. Zhou Y, Wang X, Chen H, Duan X, Zhu W (2023) Intra-and inter-modal curriculum for multimodal learning. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 3724–3735
18. Fan Y, Xu W, Wang H, Wang J, Guo S (2023) Pmr: Prototypical modal rebalance for multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20029–20038
19. Nguyen DA, Kamboj A, Do MN (2025) Robult: Leveraging redundancy and modality-specific features for robust multimodal learning. In: IJCAI
20. Ismail AA, Hasan M, Ishtiaq F (2020) Improving multimodal accuracy through modality pre-training and attention. arXiv preprint arXiv:2011.06102
21. Huang C, Wei Y, Yang Z, Hu D (2025) Adaptive unimodal regulation for balanced multimodal information acquisition. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 25854–25863
22. Wu Q, Shao Y, Wang J, Sun X (2025) Learning optimal multimodal information bottleneck representations. In: Forty-second International Conference on Machine Learning
23. Li H, Li X, Hu P, Lei Y, Li C, Zhou Y (2023) Boosting multi-modal model performance with adaptive gradient modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 22214–22224
24. Liu F, Fu Z, Wang Y (2025) Reward-based gradient modulation for multimodal emotion recognition with lora. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/tcss.2025.3566373
25. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: Interactive emotional dyadic motion capture database. Lang Resour Eval 42:335–359
26. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, 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, pp 527–536
27. Ghosal D, Majumder N, Poria S, Chhaya N, 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), pp 154–164
28. Nguyen CVT, Nguyen CB, Le DT, Ha QT (2024) Curriculum learning meets directed acyclic graph for multimodal emotion recognition. In: Calzolari, N, Kan, M.-Y, Hoste, V, Lenci, A, Sakti, S, Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, pp 4259–4265
29. Hu D, Wei L, 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), pp 7042–7052. Association for Computational Linguistics, Online https://doi.org/10.18653/v1/2021.acl-long.547
30. Shen W, Chen J, Quan X, Xie Z (2021) Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35, pp 13789–13797
31. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP (2017) Context-dependent sentiment analysis in user-generated videos. In: Barzilay, R, Kan, M.-Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 873–883. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-1081
32. Hazarika D, Poria S, Zadeh A, Cambria E, Morency LP, Zimmermann R (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Walker, M, Ji, H, 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), pp 2122–2132. Association for Computational Linguistics, New Orleans, Louisiana. https://doi.org/10.18653/v1/N18-1193
33. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R (2018) ICON: Interactive conversational memory network for multimodal emotion detection. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 2594–2604. Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1280
34. Li B, Fei H, Liao L, Zhao Y, Teng C, Chua TS, Ji D, Li F (2023) Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 5923–5934
35. Shi T, Huang SL (2023) MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 14752–14766. Association for Computational Linguistics, Toronto, Canada. https://doi.org/10.18653/v1/2023.acl-long.824
36. Delbrouck JB, Tits N, Brousmiche M, Dupont S (2020) A transformer-based joint-encoding for emotion recognition and sentiment analysis. In: Zadeh A, Morency L-P, Liang PP, Poria S (eds) Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Association for Computational Linguistics, Seattle, USA, pp 1–7. https://doi.org/10.18653/v1/2020.challengehml-1.1

37. Tu G, Xie T, Liang B, Wang H, Xu R (2024) Adaptive graph learning for multimodal conversational emotion detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38, pp 19089–19097
38. Joshi A, Bhat A, Jain A, Singh A, Modi A (2022) Cogmen: Contextualized gnn based multimodal emotion recognition. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 4148–4164
39. Huang Y, Lin J, Zhou C, Yang H, Huang L (2022) Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In: International Conference on Machine Learning, PMLR, pp 9226–9259
40. Fan Y, Xu W, Wang H, Liu J, Guo S (2024) Detached and interactive multimodal learning. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 5470–5478
41. Wei Y, Li S, Feng R, Hu D (2025) Diagnosing and re-learning for balanced multimodal learning. In: European Conference on Computer Vision, Springer, pp 71–86
42. Guo Z, Jin T, Zhao Z (2024) Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1726–1736
43. Nguyen CVT, Le TS, Mai AT, Le, DT (2024) Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 9330–9339
44. Planamente M, Plizzari C, Alberti E, Caputo B (2022) Domain generalization through audio-visual relative norm alignment in first person action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1807–1818
45. Hua C, Xu Q, Bao S, Yang, , Huang Q (2024) Reconboost: Boosting can achieve modality reconcilement. In: The Forty-first International Conference on Machine Learning
46. Du C, Li T, Liu Y, Wen Z, Hua T, Wang Y, Zhao H (2021) Improving multi-modal learning with uni-modal teachers. arXiv preprint arXiv:2106.11059
47. Wei Y, Hu D, Du H, Wen JR (2024) On-the-fly modulation for balanced multimodal learning, IEEE Transactions on Pattern Analysis and Machine Intelligence
48. Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp 41–48
49. Soviany P, Ionescu RT, Rota P, Sebe N (2022) Curriculum learning: a survey. Int J Comput Vision 130(6):1526–1565
50. Wang X, Zhou Y, Chen H, Zhu W (2024) Curriculum learning for multimedia in the era of large language models. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 11296–11297
51. Tong A, Tang C, Wang W (2022) Semi-supervised action recognition from temporal augmentation using curriculum learning. IEEE Trans Circuits Syst Video Technol 33(3):1305–1319
52. Yu T, Wang J, Luo J, Wang J, Zhou G (2025) Tacl: a trusted action-enhanced curriculum learning approach to multimodal affective computing. Neurocomputing 620:129195
53. Narvekar S, Peng B, Leonetti M, Sinapov J, Taylor ME, Stone P (2020) Curriculum learning for reinforcement learning domains: A framework and survey. J Mach Learn Res 21(181):1–50
54. Qian C, Han K, Wang J, Yuan Z, Lyu C, Chen J, Liu Z (2025) Dyncim: Dynamic curriculum for imbalanced multimodal learning. arXiv preprint arXiv:2503.06456
55. Tullis JG, Benjamin AS (2011) On the effectiveness of self-paced learning. J Mem Lang 64(2):109–118
56. Han K, Lyu C, Ma L, Qian C, Ma S, Pang Z, Chen J, Liu Z (2025) Climd: A curriculum learning framework for imbalanced multimodal diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 65–74
57. Zhou Y, Liang X, Xu Y, Gao B (2025) Sample-level self-paced learning to tackle multimodal imbalance problem. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1–5
58. Wang X, Chen Y, Zhu W (2021) A survey on curriculum learning. IEEE Trans Pattern Anal Mach Intell 44(9):4555–4576
59. Hacohen G, Weinshall D (2019) On the power of curriculum learning in training deep networks. In: International Conference on Machine Learning, PMLR, pp 2535–2544
60. Zhang D, Zhang W, Li S, Zhu Q, Zhou G (2020) Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 503–511
61. Hu D, Hou X, Wei L, Jiang L, 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), IEEE, pp 7037–7041
62. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp 1459–1462
63. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, pp 59–66

64. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 3982–3992. Association for Computational Linguistics, Hong Kong, China . https://doi.org/10.18653/v1/D19-1410
65. Platanios EA, Stretcu O, Neubig G, Poczos B, Mitchell T (2019) Competence-based curriculum learning for neural machine translation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp 1162–1172
66. Park CY, Cha N, Kang S, Kim A, Khandoker AH, Hadjileontiadis L, Oh A, Jeong Y, Lee U (2020) K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci Data 7(1):293

Authors and Affiliations

Phuong-Anh Nguyen[1] · The-Son Le[1] · Duc-Trong Le[1] · Cam-Van Thi Nguyen[1]

✉ Cam-Van Thi Nguyen
vanntc@vnu.edu.vn

Phuong-Anh Nguyen
22028332@vnu.edu.vn

The-Son Le
21020089@vnu.edu.vn

Duc-Trong Le
trongld@vnu.edu.vn

[1] Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi 100000, Vietnam