VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

NGUYEN THI CAM VAN

ADVANCED DEEP MULTIMODAL FUSION MODELS FOR EMOTION RECOGNITION IN CONVERSATION UNDER INCOMPLETE AND IMBALANCED MODALITIES

(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)

DOCTOR OF PHILOSOPHY IN INFORMATION SYSTEM DISSERTATION

Hanoi, 2026

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

NGUYEN THI CAM VAN

DVANCED DEEP MULTIMODAL FUSION MODELS FOR EMOTION RECOGNITION IN CONVERSATION UNDER INCOMPLETE AND IMBALANCED MODALITIES

(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)

Major: Information System

Code: 9480104

DOCTOR OF PHILOSOPHY IN INFORMATION SYSTEM DISSERTATION

SUPERVISORS:

1. Assoc. Prof. Ha Quang Thuy
2. Dr. Le Duc Trong

Hanoi, 2026

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Ha Quang Thuy, for his continuous guidance and support throughout my Ph.D. journey and related research. He was the first person to guide me on my academic path, from my undergraduate thesis to my master's thesis and now to my doctoral dissertation. His mentorship has laid a strong foundation for my research career.

I am deeply thankful to Dr. Le Duc Trong, who has been not only a supervisor but also a senior brother figure to me. He accompanied me through the earliest stages of my doctoral research and encouraged me when I was still uncertain. From my first submissions, he believed in my ability and reminded me to be confident in my work. His guidance and encouragement have been invaluable throughout this journey.

My heartfelt thanks go to all members of the Data Science and Knowledge Technology Laboratory (DS&KTLab) and the Faculty of Information Technology at the VNU-University of Engineering and Technology. Their advice, collaboration, and encouragement have been a constant source of inspiration. I would also like to extend my deep appreciation to Assoc. Prof. Phan Xuan Hieu, Assoc. Prof. Tran Trong Hieu, Dr. Tran Mai Vu, Dr. Le Hoang Quynh, Dr. Vuong Thi Hai Yen, and other members of the laboratory for their insightful comments and kind support. I am equally grateful for the friendship I have received from everyone in the lab, including Quynh Trang, Cat Can, and Vuong Hong. We have laughed, argued, and supported one another, and these shared moments have made my doctoral years truly meaningful and memorable.

I also wish to thank my students, especially Tuan Mai, The Son, Phuong Anh, and my other students. Although I am a Ph.D. student myself, they have given me the opportunity to experience another meaningful role. Our discussions, meetings, and their positive energy have motivated me and enriched my academic life.

Finally, I would like to express my deepest gratitude to my parents, my sister, and my little brother. Their unwavering love, support, and encouragement have accompanied

me through every important stage of my life and given me the strength to pursue my Ph.D. journey. To my late father, I wish you could be here with me at this moment. It is deeply sad that you are no longer by my side, but I believe you have always been watching over me and that you would be proud of me.

I am sincerely grateful to my parents-in-law, who have supported me in countless ways. Their understanding, kindness, and willingness to share responsibilities have allowed me to devote more time and energy to my research.

Most importantly, I am profoundly thankful for my own little family. My husband, Dang Kieu, has been my steadfast source of love, patience, and encouragement. My two daughters, Mina and Zoe, have brought me immense joy and motivation, and their smiles have given me strength to keep moving forward. I cannot fully express my gratitude and love for them in words.

To myself, I am proud of you. Proud of the way you kept going, stayed strong, and did not give up, even when the journey was difficult. After all the hard days, doubts, and quiet struggles, you can finally say: you did it.

Declaration

I hereby declare that this Doctoral Dissertation was carried out by me for the degree of Doctor of Philosophy under the guidance and supervision of my supervisors.

This dissertation is my own work and includes nothing, which is the outcome of work done in collaboration except as specified in the text.

It is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university; and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

Hanoi, January 2026
Author

Nguyen Thi Cam Van

Table of Contents

ABBREVIATIONS

Acc.	Accuracy
Ada2I	Adaptive Imbalance-aware Learning
AFW	Adaptive Feature Weighting
AMW	Adaptive Modality Weighting
CM	Cross-Modal
CMA	Cross-Modal Attention
DAG	Directed Acyclic Graph
ER	Emotion Recognition
ERC	Emotion Recognition in Conversation
F1	F1-Score
FE	Feature Estimation
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
IMR	Incomplete Multimodal Representation
LSTM	Long Short-Term Memory Network
MAC	Multimodal Affective Computing
MAE	Mean Absolute Error

MER	Mutimodal Emotion Recognition
MERC	Multimodal Emotion Recognition in Conversation
MHA	Multi-Head Attention
MLP	Multilayer Perceptron
MML	Multimodal Machine Learning
MSA	Multimodal Sentiment Analysis
MSE	Mean Squared Error
PCM	Pairwise Cross-Modal Interaction
RNN	Recurrent Neural Network
RT-GCN	Relational Temporal Graph Convolutional Network
SPCL	Self-Paced Curriculum Learning
UA	Unweighted Accuracy
WA	Weighted Accuracy

Table of Notations

Symbol	Description
$\mathcal{D}$	Multimodal emotion recognition dataset
C	A conversation (dialogue) in the dataset
N	Number of utterances in a conversation
u_i	The i-th utterance in a conversation
S	A speaker in a conversation
p_i	Speaker associated with utterance u_i
y_i	Emotion label of utterance u_i
$\mathcal{M}$	Set of modalities (e.g., text, audio, visual)
m	Modality index, $m \in \mathcal{M}$
$\mathbf{x}_i^m$	Input feature of utterance u_i in modality m
$\mathbf{h}_i^m$	Latent representation of utterance u_i in modality m
$\mathbf{h}_i$	Fused multimodal representation of utterance u_i
$\mathbf{H}^m$	Sequence of utterance representations in modality m
$f_m(\cdot)$	Modality-specific encoding function
$f_{\text{fusion}}(\cdot)$	Multimodal fusion function
$\mathcal{G}$	Dialogue graph
$\mathcal{V}$	Set of nodes (utterances) in the dialogue graph
$\mathcal{E}$	Set of edges in the dialogue graph
e_{ij}	Directed edge from utterance u_i to u_j
$\mathcal{N}(i)$	Neighbor set of utterance u_i in the graph
$\mathbf{H}^{(l)}$	Node representations at the l-th graph layer
$\phi(\cdot)$	Message passing or aggregation function
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, Key, and Value matrices in attention mechanisms
α_{ij}	Attention weight from node i to node j

List of Figures

List of Tables

Abstract

A central challenge in multimodal emotion recognition in conversation (MERC) is to design fusion models that capture complex interactions among textual, acoustic, and visual modalities while respecting conversational dynamics and speaker-specific context. Conventional approaches often rely on early or late fusion strategies that treat modalities uniformly and overlook temporal structure, speaker relations, and higher-level contextual dependencies in dialogues. In realistic settings, conversational multimodal data are further affected by missing modalities and imbalanced modality contributions, which undermine the robustness and generalization of existing MERC systems.

This dissertation tackles these challenges by developing deep multimodal fusion models tailored for emotion recognition in conversation under full, incomplete, and imbalanced modality conditions. The first part focuses on structured multimodal conversational modeling under full-modality settings, aiming to learn expressive, context-aware representations for MERC. We propose CORECT, a relational–temporal graph-based framework that integrates a Relational Temporal Graph Convolutional Network (RT-GCN) with a Pairwise Cross-modal Feature Interaction module (P-CM) to jointly model utterance-level temporal dependencies, cross-modal interactions, and speaker-aware conversational relations. We further introduce MultiDAG+CL, which combines Directed Acyclic Graph–based contextual reasoning with curriculum learning to progressively handle emotional shifts and sample difficulty in multi-speaker dialogues.

The second part addresses robust and balanced multimodal fusion when modality completeness and balance cannot be assumed. To cope with missing modalities, we propose Mi-CGA, a graph-based framework that estimates missing modality features and propagates complementary cross-modal information to maintain reliable multimodal representations under incomplete conditions. To mitigate modality imbalance, we develop two complementary strategies: Ada2I introduces Adaptive Feature Weighting (AFW) and Adaptive Modality Weighting (AMW) to re-balance feature- and modality-level contributions during fusion, while Self-Paced Curriculum Learning (SPCL) pro-

vides a plug-and-play curriculum training scheme that stabilizes multimodal learning under heterogeneous modality conditions.

Extensive experiments on widely used MERC benchmarks, including IEMOCAP, CMU-MOSI, and CMU-MOSEI, together with additional evaluations under incomplete and imbalanced modality settings, demonstrate that the proposed models consistently outperform strong baselines in terms of fusion quality, contextual reasoning, robustness to missing modalities, and balanced learning dynamics. Overall, the dissertation contributes a unified multimodal fusion perspective for MERC: (1) structured conversational fusion with temporal and relational modeling (CORECT, MultiDAG+CL); (2) fusion that remains robust under missing modalities (Mi-CGA); and (3) fusion-aware training strategies that mitigate modality dominance (Ada2I, SPCL), thereby enabling more reliable emotion recognition in realistic multimodal conversations.

Preamble

Motivation and Research Challenges

The rapid growth of online communication platforms has fundamentally transformed how people interact, exchange information, and express emotions, making digital environments a primary locus of affective experience in contemporary life. With over five billion Internet users worldwide[1], emotions are now routinely expressed and perceived through social networks, video-sharing platforms, and conversational applications, where they shape user engagement, social interaction, and interpersonal understanding. In this context, automatic *emotion recognition* has become a core capability for intelligent systems in human–computer interaction and social computing, enabling machines to sense, interpret, and respond to users' affective states in a more human-centered manner. Recent discussions also highlight that emotional intelligence and the ability to understand others' feelings are increasingly important in many aspects of modern life, from interpersonal communication to decision making[2], [3].

In natural human communication, emotional states are rarely conveyed through a single channel. Instead, affect is expressed through the coordinated interaction of linguistic content, vocal characteristics, facial expressions, and visual context, which together give rise to rich and nuanced emotion displays. This inherently **multimodal** nature of emotional expression has motivated extensive research in Affective Computing, an interdisciplinary field that seeks to endow machines with the ability to recognize, interpret, and regulate human affective states by integrating insights from computer science, psychology, and cognitive science [12, 74, 89]. Building upon this foundation, **Multimodal Affective Computing (MAC)** has emerged as a prominent research direction that studies *emotion understanding* through the joint modeling of heterogeneous modalities such as text, audio, and visual signals [20, 22, 69, 140].

[1] https://www.statista.com/statistics/617136/digital-population-worldwide/
[2] https://www.mckinsey.com/about-us/new-at-mckinsey-blog/adam-grant-on-modern-leadership
[3] https://dilanconsulting.com/why-leaders-must-master-emotional-intelligence-in-2025/

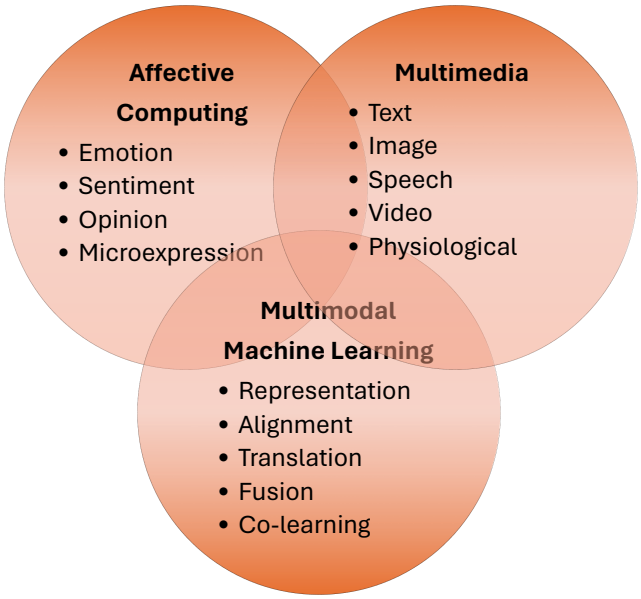


Figure 1: Overview of the framework for Multimodal Affective Analysis. The figure is integrated and adapted from [20, 94].

Human affect itself represents a complex psychological and physiological phenomenon that manifests through multiple expressive channels and gives rise to discrete and continuous emotional experiences [41]. While early affective analysis studies primarily focused on individual modalities such as text, speech, or facial images, advances in sensing technologies and deep learning have enabled increasingly effective **multimodal approaches** that more faithfully capture how emotions are communicated in real interactions [5]. This progress has led to the broader framework of **Multimodal Affective Analysis (MAA)**, which investigates how affective states, particularly emotions, can be modeled by integrating information from multiple heterogeneous sources [20, 94]. As illustrated in Figure 1, MAA lies at the intersection of Affective Computing, Multimedia, and Multimodal Machine Learning, and provides a unifying perspective for studying multimodal emotion analysis.

Within this broad framework, **this dissertation focuses on Multimodal Emotion Recognition in Conversation (MERC)** as a concrete and practically significant task. MERC aims to automatically identify the emotional state associated with each utterance in a dialogue by jointly modeling multimodal signals together with conversational context and speaker interactions. Compared with static emotion recognition settings, MERC places a stronger emphasis on the **dynamics of emotions across dialogue turns**, where temporal dependencies, speaker-specific behavior, and interpersonal relations jointly shape how emotions are expressed, perceived, and evolve over time [23]. Consequently, MERC constitutes a challenging instantiation of multimodal

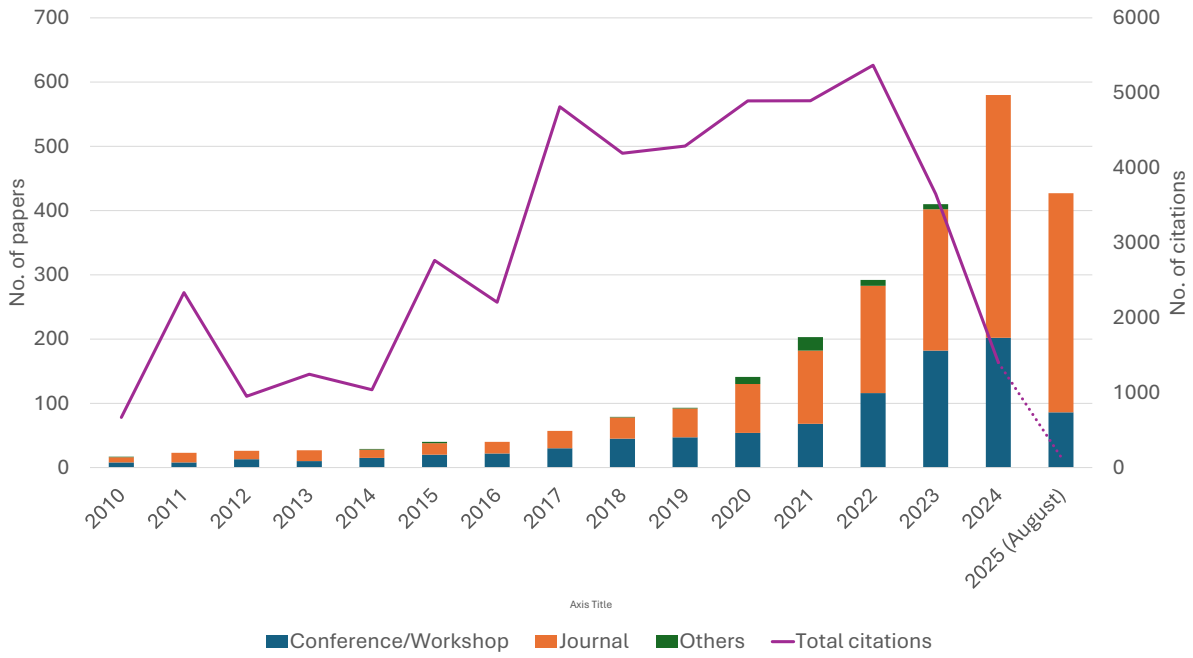


Figure 2: The rising wave of research in MERC. (Data source: openalex.org)

affective analysis and serves as the central task guiding the methodological design and experimental investigation throughout this dissertation.

The increasing importance of emotion-centric applications has coincided with a rapid expansion of research activities in multimodal affective computing and MERC. As shown in Figure 2, the number of publications and citations in this area has grown substantially in recent years, reflecting both methodological advances in multimodal learning and the demand for robust emotion-aware systems in real-world scenarios. This trend further underscores the need for **principled multimodal emotion modeling** that can cope with the complexity of conversational data.

Effective **multimodal fusion** is therefore a core component of MERC, as it determines how heterogeneous modality-specific representations are integrated to capture complementary emotional cues while preserving conversational context. Advances in multimodal representation learning and fusion strategies have significantly contributed to progress in this area [5, 48, 80]. However, many existing MERC models are developed under **idealized assumptions** in which all modalities are fully observed, equally reliable, and contribute uniformly to learning, which only partially reflects the conditions under which emotions are actually expressed and sensed in everyday interactions.

In the Vietnamese context, research on emotion recognition has also progressed across multiple modalities and application scenarios. Existing works cover speech emo-

tion recognition for Vietnamese using traditional models such as GMM/HMM and, more recently, deep learning pipelines developed in VLSP ASR/SER challenges, where models are trained on short utterances with a small number of emotion categories. Other studies explore emotion-aware dialogue systems or multimodal settings combining text and images, as well as systems that fuse facial expressions with EEG signals for emotion recognition in controlled laboratory environments.

In real-world conversational scenarios, multimodal data supporting emotion recognition are often **low-quality and imperfect**. Different modalities may vary substantially in reliability due to noise, occlusion, sensor failures, or asynchronous data acquisition, and emotional cues themselves can be subtle, context-dependent, and unevenly distributed across modalities. Empirical studies have shown that multimodal learning models can be misled by spurious correlations and biased modality contributions when trained under such conditions [139], leading to degraded emotion recognition performance and poor generalization. These issues are particularly pronounced in MERC, where correctly interpreting an utterance's emotion often requires integrating **weak or partially missing cues** from several modalities together with the surrounding conversational context.

Despite substantial advances enabled by deep learning, MERC therefore remains challenging under realistic conditions. To clearly position this dissertation within the MERC literature and to highlight unresolved challenges in **multimodal fusion for conversational emotion recognition**, Table 2 summarizes major categories of MERC approaches, their strengths, limitations, and open research gaps. In summary, existing MERC approaches continue to face **three fundamental challenges**: *(1) designing multimodal fusion mechanisms that effectively integrate complementary cues while preserving modality-specific characteristics; (2) modeling temporal and speaker-dependent conversational context in a structured and scalable manner; and (3) ensuring robustness under low-quality multimodal data, most notably in the presence of incomplete modalities and imbalanced modality contributions.* These challenges are closely intertwined and collectively limit the robustness and generalization of current multimodal emotion recognition systems in conversational settings.

More specifically, the limitations of existing studies can be grouped into **three problem axes** that motivate this dissertation. *First*, many multimodal ERC models improve performance by combining textual, acoustic, and visual features, but their fusion mechanisms often treat modality representations as uniformly reliable or mainly opti-

Method Category	Representative Approaches	Key Strengths	Main Limitations	Open Challenges in MERC
Unimodal ERC Models	Text-, audio-, or vision-based ERC models	Simple and efficient; effective when one modality dominates	Ignore complementary emotional cues from other modalities; highly sensitive to noise	Inability to model cross-modal interactions required for reliable MERC
Early Multimodal Fusion for ERC	Feature concatenation, early/late fusion	Straightforward multimodal integration; low implementation cost	Treat all modalities equally; neglect modality-specific reliability and interaction structure	Need principled fusion mechanisms tailored to conversational emotion recognition
Attention-based Multimodal ERC	Cross-attention and co-attention based ERC models	Explicitly model cross-modal interactions; improved alignment	Typically assume complete and clean modalities; vulnerable to noisy or dominant modalities	Robust multimodal fusion under incomplete and imbalanced conversational data
Context-aware ERC Models	RNN- or Transformer-based dialogue emotion models	Capture temporal dependencies across dialogue turns	Limited modeling of speaker interactions and structured conversational relations	Structured modeling of temporal and speaker-dependent context in MERC
Graph-based MERC Models	Graph neural networks for conversational emotion recognition	Flexible relational modeling; capture inter-utterance and inter-speaker dependencies	Graph construction and training stability remain challenging; often assume full modalities	Unified modeling of multimodal fusion and conversational structure
MERC with Incomplete Modalities	Modality dropout, reconstruction-based ERC methods	Improve robustness when some modalities are missing	Rely on strong assumptions or modality-specific heuristics	Reliable information propagation and fusion under severe modality incompleteness
MERC with Modality Imbalance	Reweighting, gradient balancing, curriculum-based ERC methods	Mitigate modality dominance; improve training stability	Often model-dependent; limited generalization across architectures	Model-agnostic and adaptive balancing strategies for MERC

Table 2: Comparison of major categories of MERC methods, highlighting their strengths, limitations, and open research challenges.

mize task accuracy without explicitly preserving modality-specific characteristics and cross-modal complementarity [24, 30, 31, 33, 60, 131, 132]. As a result, such models may fail to fully exploit weak but useful affective cues from non-dominant modalities. *Second*, although context-aware and graph-based ERC models have improved the modeling of dialogue history and speaker relations [35, 40, 91, 123, 126, 136], they are still often designed under full-modality assumptions and do not sufficiently consider how conversational structure interacts with multimodal fusion. *Third*, robust multimodal learning under low-quality data remains underexplored in MERC: methods for incomplete modalities frequently rely on reconstruction or imputation assumptions that may introduce noisy affective cues, while modality-balancing methods are often model-dependent or focus on a single level of imbalance [49, 71, 122, 135, 142, 150]. These observations indicate the need for MERC models that jointly address **structured multimodal fusion**, **robustness to incomplete modalities**, and **balanced learning under modality dominance**.

Motivated by these emotion-centered challenges, this dissertation aims to advance multimodal emotion recognition in conversation through a systematic investigation of **multimodal fusion and contextual modeling** under realistic data conditions. By addressing both representational design and learning robustness, this work seeks to improve the reliability of MERC systems in practical conversational scenarios and to contribute to the broader development of emotion-aware intelligent systems.

In addition to the international literature summarized above, several research efforts in Vietnam have also contributed to emotion recognition across different modalities and application scenarios. Early studies on Vietnamese speech emotion recognition (SER) [44] explored traditional statistical models such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), using prosodic and spectral features to distinguish a small set of basic emotions from read or semi-spontaneous speech. More recently, the VLSP 2023 [4] and VLSP 2025 [5] shared tasks on ASR/SER have fostered the development of deep learning-based pipelines for Vietnamese speech and speech emotion recognition, providing benchmark datasets and baselines for neutral versus negative affective states in short utterances.

Beyond speech-only SER, some works investigate multimodal or cross-modal settings in Vietnamese. For instance, Pham et al. [101] propose a lightweight confidence-based fusion framework that combines a HuBERT-based voice classifier and a PhoBERT-v2 text classifier for Vietnamese emotion recognition from voice and text, achieving robust performance on datasets such as UIT-VSMEC [67] and Bud500 [1]. Other studies focus on text-based emotion or sentiment classification for Vietnamese social media, such as the UIT-VSMEC [67] corpus, which supports neural models for recognizing multiple emotion categories in user-generated content. In another direction, multimodal settings combining facial expressions with EEG signals have been explored in controlled laboratory environments, where emotion is inferred from synchronized visual and brain-activity signals rather than from conversational context [66]. Overall, these efforts illustrate a growing interest in emotion recognition in Vietnam across speech, text, and physiological signals, but they predominantly operate in single-utterance, single-modality, or non-conversational scenarios, and thus only partially overlap with the multimodal, conversation-centric setting considered in this work.

[4] https://vlsp.org.vn/vlsp2023/eval/asr
[5] https://vlsp.org.vn/vlsp2025/eval/asr-ser

Research Scope, Objectives, and Questions

Guided by the challenges identified in the preceding subsection, this dissertation focuses on advancing MERC through principled multimodal fusion and contextual modeling under realistic data conditions. The scope is defined to ensure conceptual clarity and methodological coherence, while avoiding affective analysis tasks that are not central to conversational emotion recognition. It is specified along two complementary dimensions: **task scope** and **methodological scope**.

- **Task scope.** This dissertation investigates MERC, where emotional states are inferred from conversational data by jointly modeling heterogeneous multimodal signals, including language, acoustic cues, and visual expressions, together with dialogue context and speaker interactions. The focus is on conversational settings in which emotions evolve dynamically across dialogue turns and are shaped by both temporal dependencies and interpersonal relations among speakers.
- **Methodological scope.** This dissertation develops deep learning models for multimodal fusion in MERC, with emphasis on (i) learning expressive multimodal representations that capture complementary cross-modal interactions, (ii) modeling conversational context in a structured and scalable manner, and (iii) ensuring robustness and stability under low-quality multimodal data, including incomplete modalities and imbalanced modality contributions.

Within this scope, the dissertation is organized around two core research objectives that instantiate three problem axes introduced in the Motivation: (i) **structured multimodal fusion and conversational contextual modeling**, (ii) **robust fusion under incomplete modalities**, and (iii) **balanced learning under modality dominance**. **Objective O1** addresses the first axis by studying multimodal fusion and conversational modeling under full-modality conditions, leading to two research questions (RQ1–RQ2) on utterance-level fusion and dialogue-level context modeling. **Objective O2** covers the other two axes by focusing on robust fusion under incomplete modalities (RQ3) and balanced multimodal learning under modality dominance (RQ4).

Objective O1: Multimodal fusion and contextual modeling for MERC. The first objective is to design multimodal fusion architectures tailored to the characteristics of conversational emotion recognition. It focuses on learning multimodal representations that integrate heterogeneous modality-specific cues while preserving their individual characteristics, and on modeling temporal dependencies and speaker interactions

that govern emotional dynamics in conversations. This objective leads to the following research questions:

- **RQ1.** How can multimodal representations be effectively learned and fused at the utterance level by capturing both intra-modality characteristics and inter-modality complementary interactions for emotion recognition in conversation?
- **RQ2.** How can temporal dependencies and speaker interactions in conversations be modeled in a structured and progressive manner to enable effective contextual reasoning for multimodal emotion recognition in conversation?

These questions are addressed in Chapter 2.

Objective O2: Robust and balanced multimodal fusion for MERC. The second objective is to improve the robustness and reliability of MERC models in realistic scenarios where multimodal data are imperfect. It addresses challenges arising from incomplete modalities and imbalanced modality contributions, which frequently occur in real-world conversational data and can substantially degrade multimodal learning performance. Accordingly, the following research questions are formulated:

- **RQ3.** In MERC scenarios with incomplete modalities, how can missing multimodal features be effectively compensated and complementary information be propagated across modalities to maintain robust multimodal representations?

  This research question is addressed in Chapter 3.
- **RQ4.** How can training and optimization strategies be designed to mitigate imbalance across feature-, modality-, emotion-, and dialogue-level signals, thereby improving training stability and generalization performance in multimodal emotion recognition in conversation?

  This research question is addressed in Chapter 4.

Together, these objectives and research questions define a coherent and progressive research agenda for Multimodal Emotion Recognition in Conversation. The dissertation moves from multimodal fusion and contextual modeling under ideal full-modality conditions to robust and balanced learning under realistic data imperfections, providing the basis for the methodological contributions and experimental studies presented in the subsequent chapters.

Significance and Contribution

This dissertation contributes to MERC by developing a coherent set of modeling and learning capabilities that directly correspond to the three problem axes identified in the Motivation. Rather than treating individual models as isolated solutions, the contributions are organized around three complementary capability axes: **structured multimodal conversational modeling**, **robust representation learning under modality incompleteness**, and **balanced learning under modality dominance**. This organization clarifies how the proposed methods collectively address central limitations of prior MERC studies.

- **Structured multimodal conversational modeling.** This dissertation establishes a principled framework for jointly modeling multimodal fusion, temporal dynamics, and speaker-dependent conversational structure in MERC. Through relational–temporal graph modeling, modality-aware interaction mechanisms, and structured conversational reasoning, it provides a consistent approach to learning expressive and context-aware multimodal representations for conversational emotion recognition. These capabilities are realized by the proposed CORECT [**VanNTC 1**] framework and the MultiDAG+CL model [**VanNTC 2**]. CORECT is evaluated on IEMOCAP and CMU-MOSEI, while MultiDAG+CL is evaluated on IEMOCAP and MELD.

- **Robust multimodal representation under modality incompleteness.** Building on structured multimodal conversational modeling, this dissertation addresses the challenge of missing modalities by developing mechanisms that estimate and propagate complementary information across modalities. The proposed framework enhances the robustness of MERC systems by maintaining reliable multimodal representations even when one or more modalities are partially or entirely unavailable. This capability is realized by the Mi-CGA framework [**VanNTC 3**], which is evaluated under incomplete-modality settings on IEMOCAP, CMU-MOSI, and CMU-MOSEI.

- **Balanced learning dynamics under modality dominance.** Beyond representation design, this dissertation investigates how learning dynamics influence multimodal fusion in MERC. By introducing adaptive and curriculum-based balancing strategies, it mitigates modality dominance during training and improves the stability and generalization of MERC models across diverse data conditions. These

strategies are realized by Ada2I [**VanNTC 4**] and Self-Paced Curriculum Learning (SPCL) [**VanNTC 5**], which are evaluated on IEMOCAP, MELD, and CMU-MOSEI.

Together, these contributions define a unified research framework that advances MERC from structured multimodal conversational modeling to robust and balanced learning under realistic data imperfections.

Dissertation Structure

Figure 3 illustrates the overall organization of this dissertation, which consists of a **Preamble**, four main **Chapters**, and a **Conclusion**. The dissertation is centered on a unified research theme - MERC- and follows a progressive research agenda that advances from structured multimodal conversational modeling under full-modality conditions to robust and balanced learning under realistic data imperfections.

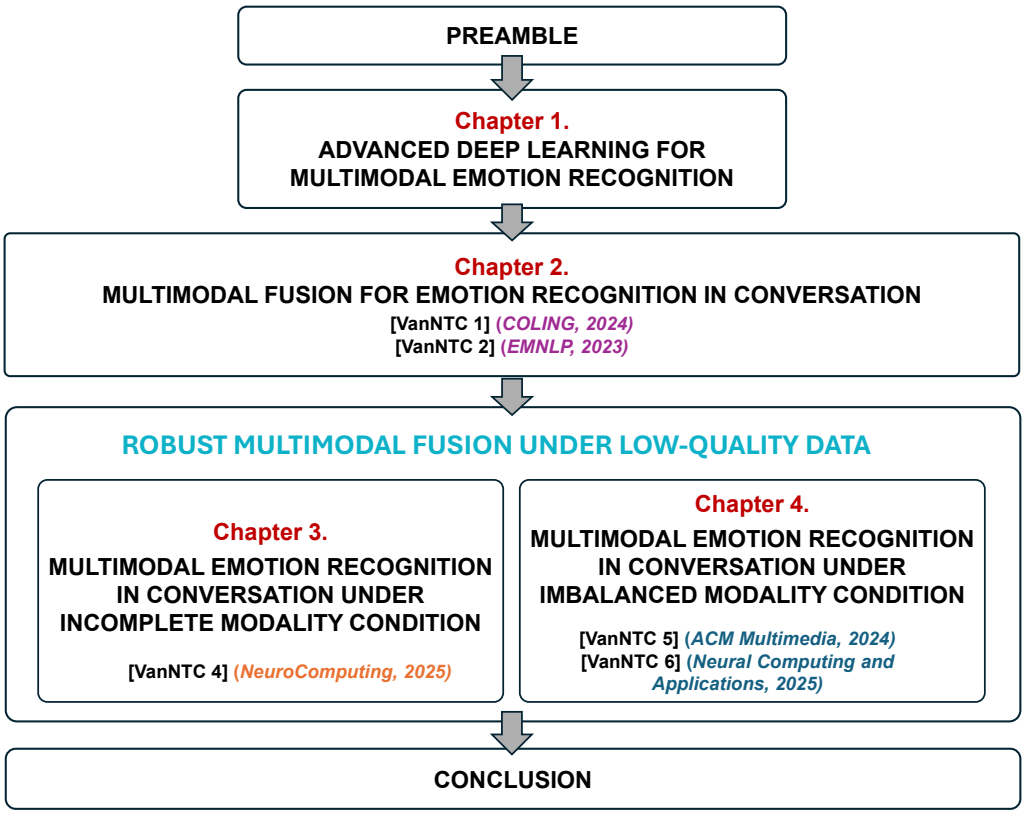


Figure 3: The overall organization of the dissertation. *Related publications are listed alongside their corresponding chapters.*

The dissertation is organized into two complementary parts. The first part establishes the methodological foundation for MERC by focusing on multimodal fusion and

conversational modeling under full-modality settings (Objective O1). Building on this foundation, the second part extends the investigation to robustness and learning stability under low-quality multimodal data conditions, where assumptions such as modality completeness and balanced modality contributions no longer hold (Objective O2). Together, these parts operationalize the three problem axes introduced in the Preamble: structured multimodal fusion and contextual modeling, robust fusion under incomplete modalities, and balanced learning under modality dominance.

The **Preamble** introduces the research motivation, scope, objectives, and key contributions of the dissertation. It positions MERC within the broader landscape of multimodal affective analysis, identifies fundamental challenges in multimodal fusion and conversational modeling, and outlines the overall research roadmap.

Chapter 1 provides the conceptual and methodological foundations required for this dissertation, including multimodal machine learning fundamentals, problem formulations for MERC, commonly used datasets, and evaluation protocols, as well as key challenges related to multimodal fusion, incomplete modalities, and modality imbalance.

Chapter 2 investigates structured multimodal fusion and conversational modeling for MERC under full-modality conditions. This chapter focuses on learning expressive multimodal representations while explicitly modeling temporal dependencies and speaker interactions, addressing the core challenges of multimodal representation learning and contextual reasoning.

Chapter 3 extends the study to MERC under incomplete modality conditions. It proposes robust fusion strategies that compensate for missing cues through cross-modal information propagation, aiming to maintain reliable multimodal representations and stable emotion recognition performance.

Chapter 4 focuses on learning stability under modality imbalance. This chapter investigates adaptive and curriculum-based strategies that regulate modality contributions during training, improving optimization stability and generalization in realistic conversational settings.

Finally, the **Conclusion and Future Work** summarizes the main findings and contributions of the dissertation, discusses its limitations, and outlines directions for future research, particularly toward more robust and scalable multimodal conversational intelligence systems.

Chapter 1

Advanced Deep Learning for Multimodal Emotion Recognition

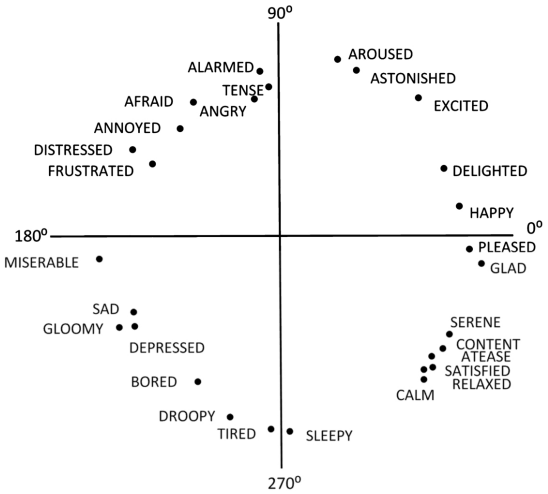
In this chapter, we will provide a brief introduction to several background topics, definitions of the key concepts and notation that will be extensively used throughout this dissertation. Additional background is introduced where necessary in later chapters.

1.1 Emotion and Multimodal Emotion Recognition

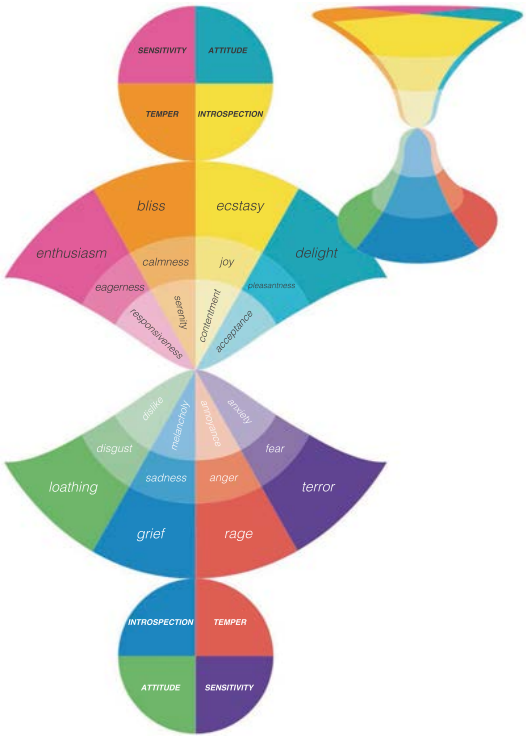
Emotions are discrete affective states that arise in response to specific events or stimuli and are characterized by coordinated changes in subjective experience, expressive behavior, and physiological response [86]. Unlike broader affective phenomena such as moods (which are diffuse and long-lasting) or attitudes (which reflect evaluative dispositions), emotions are *episodic*, *intense*, and *object-directed* [16]. Within affective computing, emotions are often distinguished from related constructs such as affect and sentiment: affect typically denotes pre-conscious feeling states, while sentiment reflects relatively stable evaluative attitudes toward entities or topics [63]. Emotions are therefore a primary focus of computational affective analysis.

Psychological research has proposed several influential emotion representation theories. Discrete emotion theories posit a small set of universally recognizable categories such as happiness, anger, or sadness [17]. Dimensional theories characterize emotions along continuous axes, most notably arousal and valence, as exemplified by Russell's Circumplex Model of Affect (Figure 1.1a) [84]. Dyadic or hybrid models, such as the Hourglass of Emotions (Figure 1.1b), integrate categorical structure with graded inten-

sity [99]. These schemes provide the conceptual foundation for defining learning objectives in affective analysis.



(a) The Circumplex Model of Affect exemplifies dimensional emotion representations and encompasses 28 affect words [84]. Arousal is depicted along the vertical axis, valence along the horizontal axis, with the neutral point at the center.



(b) The Hourglass of Emotions model represents a dyadic and categorisational emotion structure, commonly adopted in text sentiment analysis [99].

Figure 1.1: Illustration of (a) dimensional and (b) dyadic emotion representation models

Affective computing, introduced by Picard [74], aims to enable computational systems to recognize, interpret, and respond to human affective states. Within this landscape, **Multimodal Affective Analysis** focuses on understanding affective states by jointly leveraging textual, acoustic, and visual modalities, which naturally align with how emotions are expressed in human communication [111].

Multimodal Emotion Recognition (MER) integrates signals from text, audio, and vision to infer emotional states. Compared to unimodal approaches, multimodal systems offer three key advantages [76]: (i) **complementarity**: different modalities capture distinct aspects of emotion; (ii) **redundancy**: when one modality is noisy or missing, others can compensate; and (iii) **disambiguation**: multimodal cues jointly resolve ambiguities inherent in single modalities (e.g., the utterance "I'm fine" can express happiness, frustration, or sarcasm depending on vocal tone and facial expression).

Effective MER requires addressing challenges in *multimodal representation learn-*

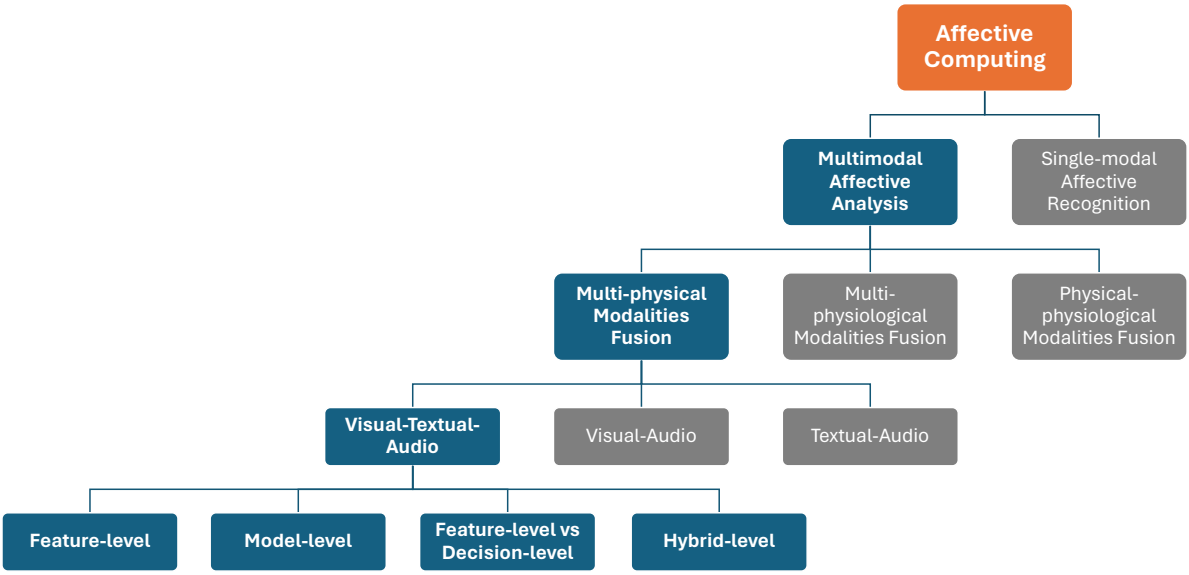


Figure 1.2: Overview of Affective Computing [111]. The highlighted area indicates the focus of this dissertation: multimodal affective analysis with textual, visual, and auditory modalities.

ing and *fusion*, as well as modeling temporal and speaker-dependent dynamics in conversational settings. Furthermore, real-world MER must handle **low-quality data**, including **incomplete modalities** and **imbalanced modality contributions**, which directly motivate the methods proposed in this dissertation.

1.2 Multimodal Machine Learning

Effective multimodal emotion recognition requires learning *expressive representations* that capture both intra-modal characteristics and inter-modal interactions. This problem is central to **Multimodal Machine Learning (MML)**, which studies computational systems that process and relate information from multiple modalities [5].

Definition 1.1. *A modality refers to a way in which a natural phenomenon is perceived or expressed.*

Definition 1.2. *Multimodal refers to situations where multiple modalities are involved.*

Definition 1.3. *Multimodal Deep Learning is a branch of machine learning that aims to learn representations and build models capable of processing and relating information from multiple modalities.*

Given a dataset $\mathcal{D} = \{(x_i^1, x_i^2, \ldots, x_i^M, y_i)\}_{i=1}^N$ with N samples and M modalities,

a multimodal model predicts $\hat{y}_i$ by encoding each modality independently and integrating the results through a fusion function:

$$\hat{y}_i = f_{\text{fusion}}\left(f_1(x_i^1),\, f_2(x_i^2),\, \ldots,\, f_M(x_i^M)\right), \tag{1.1}$$

where $f_m(\cdot)$ is a modality-specific encoder and $f_{\text{fusion}}(\cdot)$ integrates information across modalities. The training objective minimizes empirical risk:

$$\min_{\theta_{\text{fusion}},\theta_1,\ldots,\theta_M} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\hat{y}_i, y_i). \tag{1.2}$$

As outlined in Baltrušaitis et al. [5], MML presents five core technical challenges:

- **Representation**: summarizing heterogeneous data by exploiting complementarity and redundancy across modalities.
- **Translation**: mapping data from one modality to another, where multiple valid translations may exist.
- **Alignment**: identifying direct relations between elements of different modalities.
- **Fusion**: joining information from multiple modalities for prediction. *This is the central challenge addressed in this dissertation.*
- **Co-learning**: transferring knowledge between modalities, particularly when one modality has scarce data.

MML enables a broad spectrum of applications, from audio-visual speech recognition to image captioning, as summarized in Table 1.1. This dissertation focuses on **Multimodal Emotion Recognition in Conversation**, where multimodal **fusion** and **co-learning** play central roles, as highlighted in the table.

1.3 Multimodal Data Fusion

1.3.1 Multimodal Fusion Mechanisms

Multimodal fusion determines how heterogeneous information from multiple modalities is integrated to form task-relevant representations [5]. Beyond the classical distinction of early, late, and hybrid fusion strategies [80], recent advances have shifted fo-

Table 1.1: A summary of applications enabled by multimodal machine learning. For each application area we identify the core technical challenges that need to be addressed in order to tackle it [5].

	Challenges				
Applications	**Representation**	**Translation**	**Alignment**	**Fusion**	**Co-learning**
Speech recognition and synthesis					
Audio-visual speech recognition	✓		✓	✓	✓
(Visual) speech synthesis	✓	✓			
Event detection					
Action classification	✓			✓	✓
Multimedia event detection	✓			✓	✓
Emotion and affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media description					
Image description	✓	✓	✓		✓
Video description	✓	✓	✓	✓	✓
Visual question-answering	✓		✓	✓	✓
Media summarization	✓	✓		✓	
Multimedia retrieval					
Cross modal retrieval	✓	✓	✓		✓
Cross modal hashing	✓				✓

cus toward learnable and parameterized fusion mechanisms [68, 141] that model cross-modal interactions in an end-to-end manner. Contemporary methods are broadly categorized by their underlying interaction mechanisms [29]; in particular, *attention-based* and *graph neural network-based* paradigms have emerged as dominant frameworks for MERC.

1.3.1.1 Classical Multimodal Fusion Strategies

Classical fusion approaches are categorized by the stage at which modalities are integrated: *early (feature-level)*, *intermediate (joint)*, *late (decision-level)*, and *mixed (hybrid)* fusion.

Early Fusion (Feature-level Fusion). Early fusion, also referred to as data-level or feature-level fusion, integrates information from multiple modalities at the input stage or at very shallow representation levels. Formally, given modality-specific inputs $\{\mathbf{x}_i^m\}_{m\in\mathcal{M}}$ for an utterance u_i, early fusion constructs a joint representation by direct aggregation:

$$\mathbf{x}_i^{\text{early}} = \text{Concat}\Big(\{\mathbf{x}_i^m\}_{m\in\mathcal{M}}\Big), \qquad \mathbf{h}_i = f\Big(\mathbf{x}_i^{\text{early}}\Big), \tag{1.3}$$

where a single predictive model $f(\cdot)$ is trained on the concatenated multimodal input.

This strategy preserves low-level information from each modality and allows cross-modal interactions to emerge implicitly within a unified learning framework. However, early fusion assumes that heterogeneous modality features can be meaningfully aligned in a shared input space. In practice, low-level features from different modalities often differ substantially in scale, structure, and semantic richness, making early fusion highly sensitive to modality heterogeneity and representation imbalance. Moreover, since modality-specific characteristics are not explicitly modeled, early fusion offers limited flexibility when incorporating new modalities and may struggle to capture meaningful cross-modal semantics without carefully designed modality-dependent preprocessing or normalization (Figure 1.3).

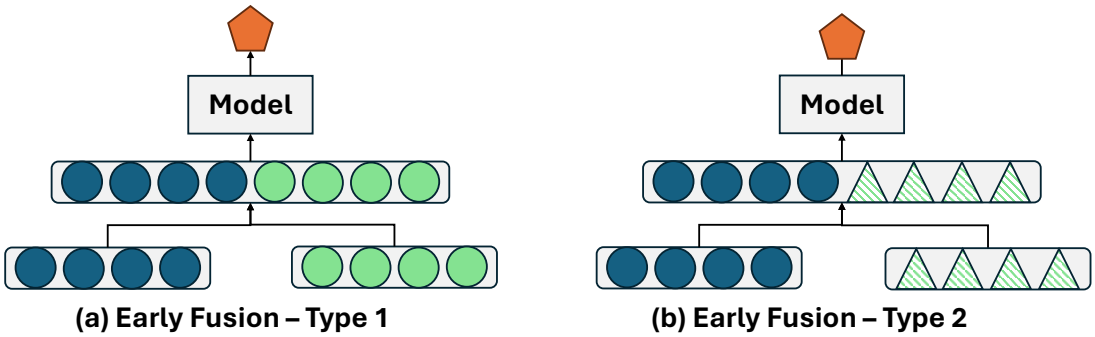


Figure 1.3: Classical multimodal fusion strategies: Early fusion

Intermediate (Joint) Fusion. Processing each modality through a dedicated encoder to learn modality-specific latent representations, which are then integrated and jointly optimized for the downstream prediction task. Formally, for an utterance u_i with multimodal inputs $\{\mathbf{x}_i^m\}_{m\in\mathcal{M}}$, modality-specific representations are obtained as:

$$\mathbf{h}_i^m = f_m(\mathbf{x}_i^m), \qquad m \in \mathcal{M}. \tag{1.4}$$

These representations are subsequently combined through a multimodal fusion function:

$$\mathbf{h}_i = f_{\text{fusion}}\Big(\{\mathbf{h}_i^m\}_{m\in\mathcal{M}}\Big), \tag{1.5}$$

and used for emotion prediction.

By decoupling unimodal representation learning from multimodal integration, intermediate fusion enables gradients to propagate back through modality-specific encoders during end-to-end training. This design allows the model to capture cross-modal

interactions at higher levels of abstraction, while preserving modality-dependent characteristics. As a result, intermediate fusion serves as the dominant formulation underlying most modern multimodal fusion architectures, including attention-based and graph-based methods.

Despite its expressive power, intermediate fusion typically introduces increased model complexity and a larger number of learnable parameters. Consequently, it often requires more training data and careful optimization, making it more challenging to train effectively in data-scarce or low-quality multimodal settings (Figure 1.4).

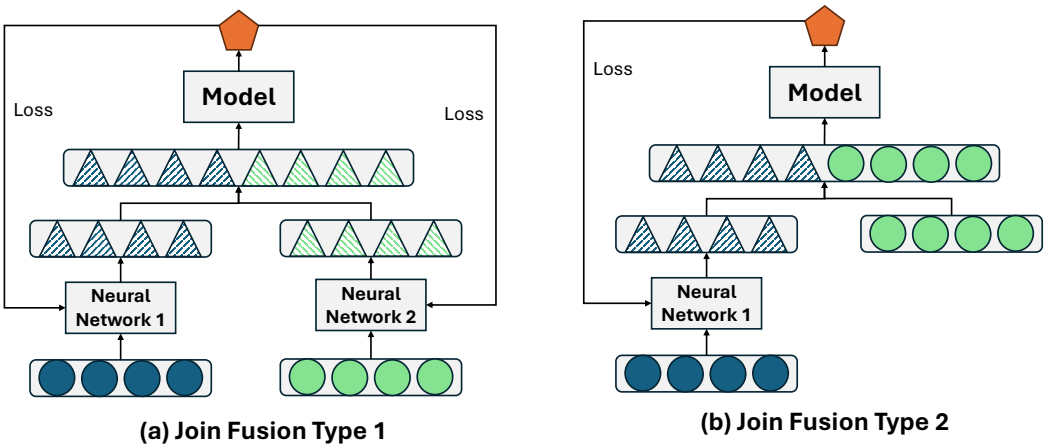


Figure 1.4: Classical multimodal fusion strategies: Join fusion

Late Fusion (Decision-level Fusion). Late fusion integrates modalities at the decision level by combining predictions from independently trained modality-specific models. In this paradigm, each modality is processed in isolation, and multimodal integration is deferred until after unimodal inference. Formally, for each modality $m \in \mathcal{M}$, a modality-specific predictor produces an intermediate decision:

$$\hat{y}_i^m = g_m(f_m(\mathbf{x}_i^m)) \,, \tag{1.6}$$

where $f_m(\cdot)$ denotes the modality encoder and $g_m(\cdot)$ is a modality-specific classifier. The final prediction is then obtained by aggregating unimodal decisions:

$$\hat{y}_i = f_{\text{fusion}}\Big(\{\hat{y}_i^m\}_{m\in\mathcal{M}}\Big), \tag{1.7}$$

where $f_{\text{fusion}}(\cdot)$ typically corresponds to weighted averaging, voting, or a meta-classifier.

By decoupling unimodal processing, late fusion offers high modularity and natural robustness to missing modalities, as each modality can be trained and evaluated inde-

pendently. However, because inter-modal interactions are not explicitly modeled at the representation level, late fusion is fundamentally limited in its capacity to capture cross-modal dependencies. This often leads to information loss and suboptimal performance in tasks such as multimodal emotion recognition, where complementary cues across modalities play a crucial role (Figure 1.5).

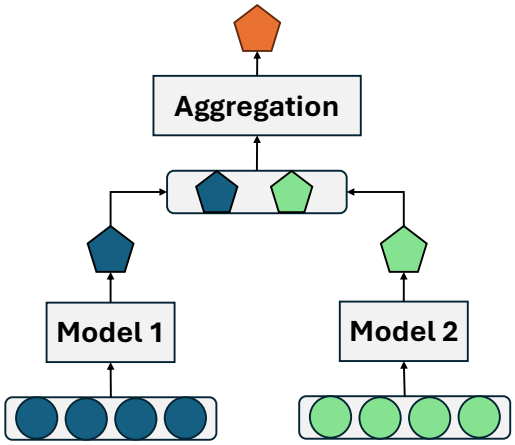


Figure 1.5: Classical multimodal fusion strategies: Late fusion

Mixed (Hybrid) Fusion. This mechanism combines multiple fusion paradigms within a single framework, allowing different modalities to be integrated at different stages of the learning pipeline. Rather than committing to a single fusion point, mixed fusion instantiates the fusion function $f_{\text{fusion}}(\cdot)$ at multiple levels, enabling modality-specific processing depths while still supporting cross-modal interaction.

Formally, mixed fusion can be viewed as a composition of intermediate representations and fusion operations:

$$\mathbf{h}_i = f_{\text{fusion}}^{(L)}\Big(\mathbf{h}_i^{(L-1)},\, f_{\text{fusion}}^{(L-1)}\big(\{\mathbf{h}_i^m\}_{m\in\mathcal{M}}\big)\Big), \tag{1.8}$$

where fusion is performed at multiple layers or stages with potentially different parameterizations.

By integrating modalities progressively, mixed fusion mitigates the representation imbalance often encountered in early fusion while enabling inter-modality interactions that are not captured by purely late fusion. This strategy provides greater flexibility by tailoring fusion depth and representation learning to the characteristics of each modality. However, mixed fusion introduces additional architectural complexity, as determining appropriate fusion stages, interaction mechanisms, and training strategies typically requires careful design choices and domain knowledge.

1.3.1.2 Attention-based Fusion

Attention mechanisms address a key limitation of classical fusion: the inability to adaptively regulate modality contributions under varying contextual conditions [3, 105]. In MERC, the informativeness of each modality varies across utterances, speakers, and dialogue turns, making data-driven relevance modeling essential.

Scaled Dot-product Attention. Given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, scaled dot-product attention is defined as:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1.9}$$

where d_k denotes the key dimension.

As defined in Section 1.2, a multimodal deep learning model consists of modality-specific encoders $\{f_m(\cdot)\}_{m=1}^{M}$ and a multimodal fusion function $f_{\text{fusion}}(\cdot)$. Let $\mathbf{X}^m$ denote the (utterance-level or sequence-level) representation of modality $m \in \mathcal{M}$ produced by its encoder, i.e., $\mathbf{X}^m = f_m(\mathbf{x}^m)$, with $\mathbf{X}^m \in \mathbb{R}^{N \times d_m}$ for a conversation of N utterances. Attention-based fusion instantiates f_{fusion} through learnable attention operations applied at different scopes. Below we describe three representative forms: intra-modal attention, cross-modal attention (directional and bidirectional), and stacked cross-modal interaction (Transformer-style).

Intra-modal Attention (within-modality refinement). Intra-modal attention models dependencies *within* each modality to emphasize salient elements before multimodal integration. Given a structured representation $\mathbf{X}^m$ (e.g., a temporal sequence), intra-modal attention produces a refined modality embedding:

$$\tilde{\mathbf{X}}^m = \text{Attn}_{\text{intra}}(\mathbf{X}^m), \tag{1.10}$$

where $\text{Attn}_{\text{intra}}(\cdot)$ can be implemented by applying attention over temporal/spatial tokens. The refined representations are then combined by a fusion function:

$$\mathbf{H} = f_{\text{fusion}}\left(\{\tilde{\mathbf{X}}^m\}_{m \in \mathcal{M}}\right). \tag{1.11}$$

While intra-modal attention improves unimodal representation quality, it does not explicitly model cross-modal complementarity.

Cross-modal Attention (directional information flow). Cross-modal attention explicitly models interactions *across* modalities by conditioning one modality on another. Given two modalities δ and γ, with representations $\mathbf{X}^{\delta} \in \mathbb{R}^{N \times d_\delta}$ and $\mathbf{X}^{\gamma} \in \mathbb{R}^{N \times d_\gamma}$, directional cross-modal attention from δ to γ (denoted "$\delta \to \gamma$") is computed by:

$$\mathbf{Q}^{\gamma} = \mathbf{X}^{\gamma}\mathbf{W}_{Q^{\gamma}}, \tag{1.12}$$

$$\mathbf{K}^{\delta} = \mathbf{X}^{\delta}\mathbf{W}_{K^{\delta}}, \tag{1.13}$$

$$\mathbf{V}^{\delta} = \mathbf{X}^{\delta}\mathbf{W}_{V^{\delta}}, \tag{1.14}$$

where $\mathbf{W}_{Q^{\gamma}} \in \mathbb{R}^{d_\gamma \times d_Q}$, $\mathbf{W}_{K^{\delta}} \in \mathbb{R}^{d_\delta \times d_K}$, and $\mathbf{W}_{V^{\delta}} \in \mathbb{R}^{d_\delta \times d_V}$ are learnable projection matrices. The cross-modal attention output is:

$$\mathbf{H}^{\gamma} = \text{softmax}\left(\frac{\mathbf{Q}^{\gamma}(\mathbf{K}^{\delta})^{\top}}{\sqrt{d_k}}\right)\mathbf{V}^{\delta}, \qquad \mathbf{H}^{\gamma} \in \mathbb{R}^{N \times d_V}. \tag{1.15}$$

This mechanism allows modality γ to selectively attend to information from modality δ based on relevance scores.

Bidirectional Pairwise Cross-modal Interaction. Since multimodal emotional cues are often complementary and mutually expressed, cross-modal attention is commonly applied in both directions:

$$\mathbf{H}^{\delta} = \mathbf{CMA}_{\gamma \to \delta}(\mathbf{X}^{\gamma}, \mathbf{X}^{\delta}), \qquad \mathbf{H}^{\gamma} = \mathbf{CMA}_{\delta \to \gamma}(\mathbf{X}^{\delta}, \mathbf{X}^{\gamma}). \tag{1.16}$$

The bidirectional interaction representation for the pair (δ, γ) can be formed by concatenation:

$$\mathbf{X}^{\delta \rightleftarrows \gamma} = [\mathbf{H}^{\delta}, \mathbf{H}^{\gamma}] \in \mathbb{R}^{N \times 2d_V}. \tag{1.17}$$

Extending to multiple modalities $\mathcal{M}$, bidirectional cross-modal interactions can be computed for all modality pairs and aggregated:

$$\mathbf{X}_{\text{Cross}} = \left[\mathbf{X}_{\text{Cross}}^{m_1 \rightleftarrows m_2}, \mathbf{X}_{\text{Cross}}^{m_2 \rightleftarrows m_3}, \dots\right], \qquad m_i \in \mathcal{M}, \tag{1.18}$$

where $[\cdot]$ denotes concatenation (or another aggregation operator depending on the architecture).

Stacked Cross-modal Interaction (Transformer-style). To capture higher-order and long-range cross-modal dependencies, cross-modal attention blocks are often stacked with residual connections, layer normalization, and position-wise feed-forward networks [2, 105]. Let $\mathbf{Z}^{[0]}_{\delta \rightleftarrows \gamma}$ denote an initial representation for a modality pair (e.g., $\mathbf{Z}^{[0]}_{\delta \rightleftarrows \gamma} = \mathbf{X}^{\gamma}$ when enriching γ using δ). At layer i, a Transformer-style update can be expressed as:

$$\overline{\mathbf{Z}}^{[i]}_{\delta \rightleftarrows \gamma} = \mathbf{CMA}^{[i]}_{\delta \rightarrow \gamma}\Big(\mathrm{LN}(\mathbf{Z}^{[i-1]}_{\delta \rightleftarrows \gamma}), \mathrm{LN}(\mathbf{X}^{\delta})\Big) + \mathbf{Z}^{[i-1]}_{\delta \rightleftarrows \gamma}, \tag{1.19}$$

$$\mathbf{Z}^{[i]}_{\delta \rightleftarrows \gamma} = \mathrm{FFN}\Big(\mathrm{LN}(\overline{\mathbf{Z}}^{[i]}_{\delta \rightleftarrows \gamma})\Big) + \overline{\mathbf{Z}}^{[i]}_{\delta \rightleftarrows \gamma}, \tag{1.20}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization [2] and $\mathrm{FFN}(\cdot)$ is a position-wise feed-forward network. Stacking D layers enables iterative refinement of cross-modal representations and facilitates deeper cross-modal reasoning.

Overall, attention-based fusion realizes $f_{\text{fusion}}(\cdot)$ through learnable relevance modeling at different scopes: (i) within-modality refinement, (ii) directional and bidirectional cross-modal interaction, and (iii) stacked Transformer-style cross-modal reasoning. In MERC, these mechanisms are particularly effective for capturing fine-grained cross-modal alignment and contextual relevance across dialogue turns. However, conversational emotions also involve structured dependencies (e.g., speaker interactions and long-range relational influence) that are not explicitly enforced by attention alone. This motivates fusion paradigms that incorporate explicit relational inductive biases, such as graph-based multimodal fusion methods discussed in the following subsection.

1.3.1.3 Graph Neural Network-based Fusion

Multimodal conversational data exhibit inherently relational structure: utterances, speakers, and modalities form interconnected entities whose dependencies are not adequately captured by flat fusion mechanisms. Graph Neural Networks (GNNs) [26, 85] provide a principled framework for modeling such structured dependencies through message passing over graph-structured representations.

General Formulation. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph constructed from multimodal data, where each node $v_i \in \mathcal{V}$ is associated with an initial feature representation $\mathbf{h}_i^{(0)}$ derived from one or more modalities. A generic GNN layer updates node representations

through message passing:

$$\mathbf{h}_i^{(l+1)} = \phi\Big(\mathbf{h}_i^{(l)}, \{\mathbf{h}_j^{(l)} \mid j \in \mathcal{N}(i)\}\Big), \tag{1.21}$$

where $\mathcal{N}(i)$ denotes the neighbor set of node i, $\phi(\cdot)$ is an aggregation function (e.g., mean, sum, attention-weighted aggregation), and l indexes the graph layers. By stacking multiple layers, information can be progressively propagated across multi-hop neighborhoods, enabling the modeling of higher-order dependencies. The final node or graph-level representations can then be used as fused multimodal representations:

$$\mathbf{h} = f_{\text{fusion}}\big(\{\mathbf{h}_i^{(L)}\}_{i \in \mathcal{V}}\big). \tag{1.22}$$

From an inductive bias perspective, GNN-based fusion explicitly assumes that multimodal interactions are structured and locally compositional. That is, complex multimodal dependencies can be decomposed into a sequence of local interactions and progressively integrated through message passing. Unlike attention-based fusion, which relies on similarity-based soft alignment in a shared latent space, GNNs encode interaction constraints directly through graph topology. This explicit relational bias enables GNNs to capture structured dependencies, long-range interactions, and hierarchical relationships, but also tightly couples representation learning to the quality of graph construction.

Instantiation with GCNs. GCNs [43] instantiate the generic GNN formulation by performing neighborhood aggregation through normalized graph convolution. Let $\mathbf{A}$ denote the adjacency matrix of graph $\mathcal{G}$ and $\mathbf{D}$ its degree matrix. A typical GCN layer updates node representations as:

$$\mathbf{H}^{(l+1)} = \sigma\Big(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\Big), \tag{1.23}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ adds self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, $\mathbf{W}^{(l)}$ is a learnable weight matrix, and $\sigma(\cdot)$ is a nonlinear activation function.

From a fusion perspective, GCN-based fusion assumes that all neighboring nodes contribute equally after normalization, making it well-suited for capturing smooth and globally consistent relational patterns. In MERC, this property is particularly effective for modeling emotion propagation across dialogue turns or within speaker-specific subgraphs. However, the uniform aggregation scheme limits the ability of GCNs to dif-

ferentiate the relative importance of heterogeneous neighbors, which may be suboptimal when multimodal or speaker interactions exhibit varying relevance.

Instantiation with GATs. GATs [8, 106] extend GCNs by introducing attention mechanisms to dynamically weight neighboring nodes during message passing. Instead of relying on fixed normalization, GATs learn attention coefficients that reflect the importance of each neighbor.

Given node representations $\mathbf{h}_i^{(l)}$ and $\mathbf{h}_j^{(l)}$, the attention coefficient from node j to node i is computed as:

$$e_{ij}^{(l)} = \text{LeakyReLU}\Big(\mathbf{a}^\top \big[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \parallel \mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\big]\Big), \tag{1.24}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik}^{(l)})}, \tag{1.25}$$

where $\mathbf{a}$ and $\mathbf{W}^{(l)}$ are learnable parameters and $\parallel$ denotes concatenation. The node update is then given by:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right). \tag{1.26}$$

By learning edge-specific importance weights, GAT-based fusion enables adaptive and heterogeneous message passing. This property is especially advantageous in multimodal conversational settings, where different speakers, modalities, or temporal relations may contribute unequally to emotion inference. As a result, GATs provide a more expressive relational fusion mechanism than GCNs, at the cost of increased computational complexity and potential sensitivity to noisy edges.

GNN-based Representation Learning for Individual Modalities. One common strategy applies GNNs independently to modality-specific graph-structured data in order to enhance representations within each modality. Let $\mathbf{h}_i^{m,(0)}$ denote the initial node representation of modality m. A modality-specific GNN computes:

$$\mathbf{h}_i^{m,(L)} = \text{GNN}_m\Big(\mathcal{G}^m, \mathbf{h}_i^{m,(0)}\Big), \tag{1.27}$$

where $\mathcal{G}^m$ represents the graph constructed for modality m. The resulting modality-enhanced representations are then integrated using a fusion function:

$$\mathbf{h} = f_{\text{fusion}}\Big(\{\mathbf{h}_i^{m,(L)}\}_{m\in\mathcal{M}}\Big). \tag{1.28}$$

This strategy preserves modality-specific relational structures and allows flexible downstream fusion. However, cross-modal interactions are only implicitly captured during the fusion stage, limiting the expressiveness of inter-modality dependency modeling.

GNN-based Representation Learning on Fused Multimodal Graphs. An alternative strategy constructs a unified multimodal graph prior to representation learning, where nodes and edges encode both intra-modality and inter-modality relationships. Given a multimodal graph $\mathcal{G}$ with initial node features $\mathbf{h}_i^{(0)}$, a GNN directly learns fused representations:

$$\mathbf{h}_i^{(L)} = \text{GNN}\Big(\mathcal{G}, \mathbf{h}_i^{(0)}\Big), \tag{1.29}$$

followed by a readout or aggregation operation:

$$\mathbf{h} = f_{\text{fusion}}\Big(\{\mathbf{h}_i^{(L)}\}_{i\in\mathcal{V}}\Big). \tag{1.30}$$

By enforcing interactions through graph topology, this approach explicitly models cross-modal dependencies and structured interactions. Nevertheless, its effectiveness is highly sensitive to graph construction choices, such as node definitions, edge types, and connectivity patterns, which may be task-dependent and difficult to generalize.

Implications for Multimodal Emotion Recognition in Conversation. In MERC, graph structures naturally arise from conversational organization, where utterances, speakers, and modalities form interconnected entities. Emotional states often propagate across dialogue turns and speaker interactions, leading to structured temporal and relational dependencies that are difficult to capture using flat fusion mechanisms alone. GNN-based fusion provides a principled framework to explicitly model such dependencies through relational edges and message passing, enabling structured contextual reasoning over multimodal conversational data. At the same time, challenges related to graph design, scalability, and robustness motivate careful architectural choices when applying GNN-based fusion in real-world MERC systems.

Overall, GNN-based multimodal fusion offers a powerful and conceptually grounded approach for integrating relational and structured multimodal data. By introducing an explicit relational inductive bias, GNNs complement attention-based fusion mechanisms and provide a natural foundation for modeling conversational emotion dynamics in this dissertation.

1.3.2 Multimodal Fusion on Low-quality Data

Multimodal fusion, which integrates information from multiple modalities, has become a cornerstone of MERC, as emotions in conversations are rarely expressed through language alone. Vocal prosody, facial expressions, and visual context often provide complementary affective cues that are critical for accurate emotion inference. However, recent studies have revealed a fundamental challenge: *traditional multimodal fusion models are vulnerable to spurious correlations and modality bias, particularly when multimodal data quality is low*.

In conversational scenarios, the quality of different modalities can vary significantly across utterances and speakers due to environmental noise, occlusions, sensor failures, or recording conditions. For MERC, such variability is especially problematic, as emotional cues are subtle, context-dependent, and unevenly distributed across modalities. Empirical and theoretical studies have demonstrated that conventional multimodal fusion strategies may fail under low-quality multimodal data conditions, including modality imbalance [37, 71, 109, 119], noisy signals [121], and corrupted inputs [38].

A recent survey by Zhang et al. [139] systematically identifies four primary technical challenges associated with multimodal fusion under low-quality data settings. As illustrated in Figure 1.6, these challenges are especially relevant to MERC due to the conversational nature of emotional expression:

- **Noisy multimodal data**: In MERC, background noise, overlapping speech, and visual distractions often introduce noise into acoustic and visual modalities. Effectively leveraging cross-modal correlations to suppress such noise is difficult due to modality heterogeneity.
- **Incomplete multimodal data**: Conversational datasets frequently exhibit missing modalities, e.g., absent facial cues in audio-only segments or missing audio in text

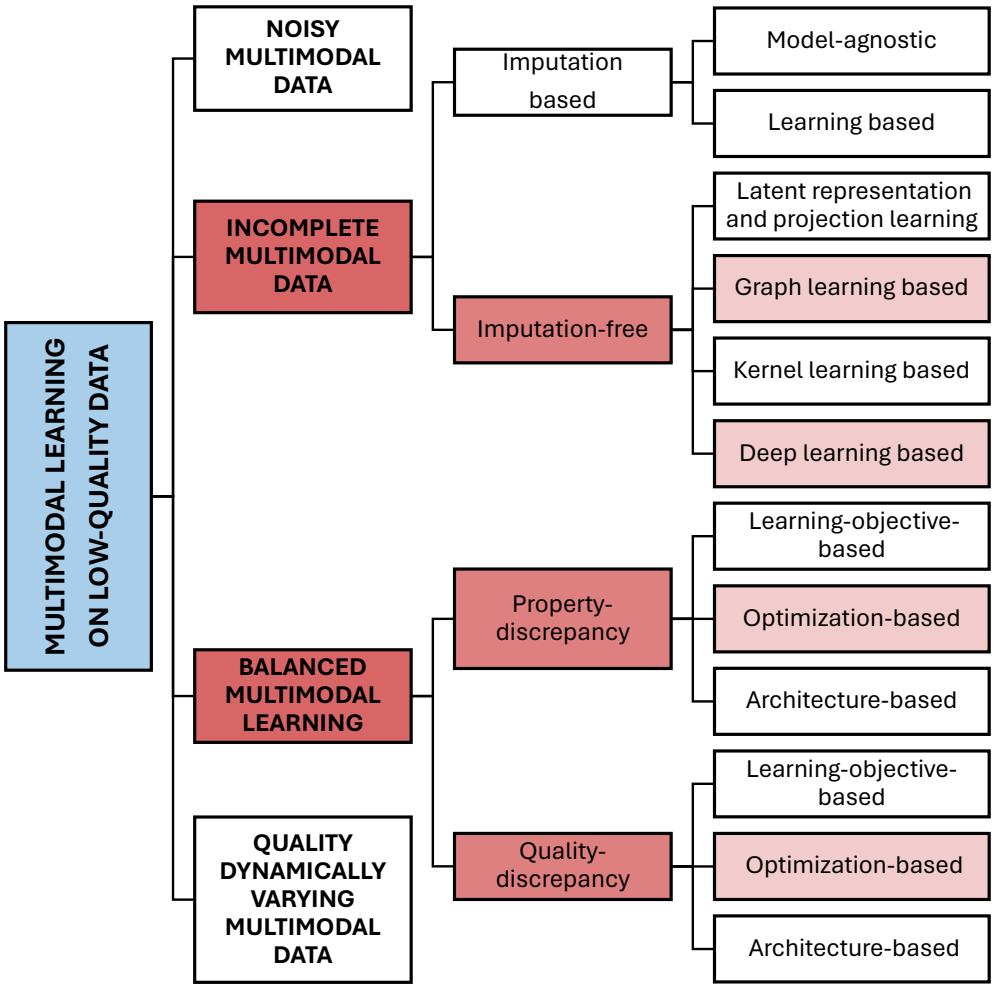


Figure 1.6: An overview of strategies addressing key challenges in multimodal learning under low-quality data. The dissertation focuses on two challenges that are particularly critical for MERC, highlighted in the red box.

transcripts. Robust MERC models must therefore operate under partial modality observability.

- **Imbalanced multimodal data**: Certain modalities, such as text, may dominate emotion prediction, causing models to ignore weaker but complementary acoustic or visual cues. This imbalance is particularly harmful in MERC, where emotional signals may shift across modalities over dialogue turns.

- **Dynamically varying modality quality**: In conversations, modality quality can fluctuate across utterances due to speaker movement, lighting changes, or recording artifacts, requiring MERC models to dynamically adapt their fusion strategies.

In this dissertation, we focus on two challenges that are most critical for conversational emotion understanding: **(1) incomplete multimodal data** and **(2) balanced**

modality learning. Both issues frequently arise in MERC benchmarks and directly affect the robustness and generalization of emotion recognition models in real-world conversational settings.

1.3.2.1 Incomplete Multimodal Learning

In MERC, missing modalities are a common occurrence rather than an exception. Facial expressions may be unavailable due to occlusion or camera absence, acoustic signals may be corrupted by background noise, and textual transcripts may omit paralinguistic cues. Most conventional multimodal emotion recognition models assume complete modality availability, which limits their applicability to real-world conversational scenarios [58, 98, 142].

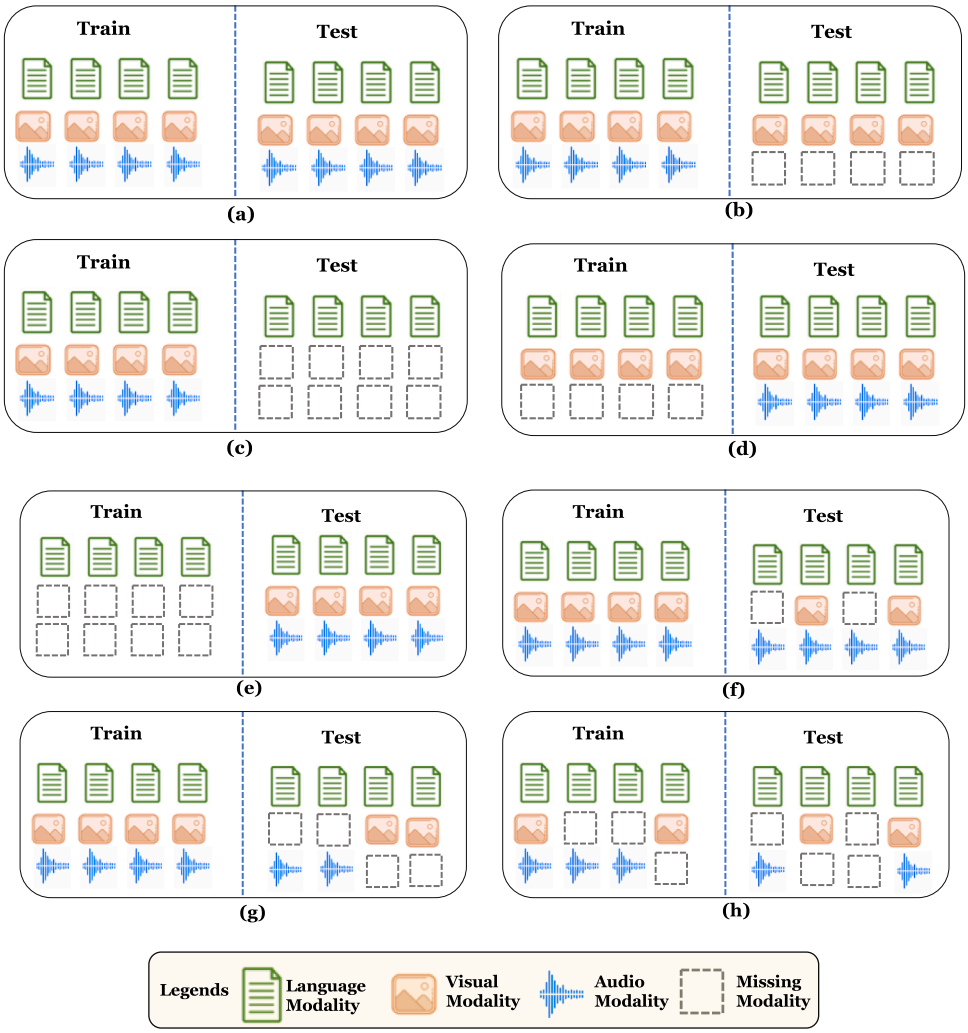


Figure 1.7: Modality availability conditions at training and testing time for language, visual, and audio modalities.

As illustrated in Figure 1.7, missing modality patterns vary significantly between

training and inference in MERC datasets. This has motivated the development of **incomplete multimodal learning** techniques, which aim to leverage available modalities while mitigating the absence of others [70, 112, 143].

Existing methods broadly fall into two categories. **Imputation-based approaches** attempt to reconstruct missing modalities at the feature or latent level using statistical or learning-based techniques, including autoencoders [53], diffusion models [113], and adversarial learning [135]. While effective in some cases, such methods rely on strong assumptions about the underlying data distribution and risk introducing noisy or misleading affective signals, which can be particularly detrimental for emotion recognition tasks where emotional cues are subtle and context-dependent.

In contrast, **imputation-free approaches** avoid explicit reconstruction and instead learn robust representations directly from available modalities. These include latent alignment [116, 148], graph-based modeling [115, 118], kernel learning [55], and deep multimodal encoders [117, 143]. Such methods are well-suited to MERC, where emotional cues may be sporadic and unevenly distributed across modalities, but they still face difficulties in aligning heterogeneous modalities and exploiting conversational structure under diverse missingness patterns.

Despite notable progress, challenges remain in aligning heterogeneous modalities, handling modality imbalance, and ensuring robustness across varying missingness patterns in conversational emotion recognition [58, 82].

1.3.2.2 Balanced Modality Learning

In MERC, different modalities contribute unequally to emotion prediction. Text often conveys explicit semantic information, while acoustic and visual modalities capture subtler affective cues such as tone, facial expressions, and gestures. Differences in signal quality and convergence speed cause deep multimodal models to over-rely on dominant modalities, leading to **modality imbalance** [71, 109, 119]. Analyses further show that imbalance can manifest at multiple levels, including feature activations, gradient magnitudes, and logit contributions, making models systematically favor text-dominant solutions even when non-text modalities carry complementary emotional cues [119]. This issue is particularly pronounced in conversational settings, where the relevance of each modality can vary across dialogue turns.

Balanced modality learning methods aim to mitigate this bias by encouraging fair

utilization of all modalities. Existing approaches can be broadly categorized based on whether they address **learning-property discrepancies** or **quality discrepancies** across modalities.

Property-discrepancy methods focus on differences in learning dynamics. These include loss reweighting [109], adaptive learning rate scheduling [97], gradient modulation [71, 96], and architectural interventions such as modality dropout [120]. Such techniques are particularly beneficial in MERC, where acoustic and visual encoders often converge more slowly than text. However, many of these methods are tightly coupled to specific model architectures or operate at a single optimization level (e.g., loss or gradient), which may limit their generality across different MERC backbones.

Quality-discrepancy methods address differences in semantic informativeness. Contrastive learning [125], confidence-aware regularization [57], and knowledge distillation [15, 54] encourage the model to extract complementary affective signals from weaker modalities. Data augmentation strategies that mask or perturb dominant modalities [119, 149] further promote balanced multimodal reasoning. Nevertheless, these methods often assume the existence of a relatively reliable reference modality and may not fully account for dynamically varying modality relevance in long, multi-speaker conversations, as highlighted in recent benchmarks on multimodal imbalance learning [**?**].

In summary, balanced modality learning is essential for robust MERC, as conversational emotions are expressed through heterogeneous and dynamically varying multimodal cues. Addressing modality imbalance lays the foundation for emotion recognition models that generalize effectively across speakers, contexts, and real-world conditions.

1.4 Multimodal Emotion Recognition in Conversation

1.4.1 Task Formulation

Early computational approaches to emotion recognition predominantly relied on unimodal signals, including text, speech, or facial expressions. When applied to conversational data, this setting naturally leads to *Emotion Recognition in Conversation (ERC)*, where the goal is to predict an emotion label for each utterance within a dialogue based on its content and surrounding context [79]. Unlike isolated emotion classification, ERC explicitly accounts for the sequential nature of conversations, where emotional states evolve across dialogue turns and are influenced by speaker interactions.

Definition 1.4 (Emotion Recognition in Conversation (ERC)). *Given a conversation* $C = \{u_1, \ldots, u_N\}$ *consisting of a sequence of utterances, where each utterance* u_i *is associated with a speaker* p_i*, the goal of ERC is to predict an emotion label* y_i *for each utterance by modeling its linguistic content together with contextual information from the dialogue.*

A defining characteristic of ERC is that conversational emotions exhibit structured dependencies. Emotional states often demonstrate temporal continuity, speaker-dependent dynamics, and interaction-driven transitions, making them difficult to infer from isolated utterances alone. These properties fundamentally distinguish conversational emotion recognition from standard sequence classification tasks and motivate context-aware and relational modeling strategies.

Figure 1.8: An example of multimodal conversation from the MELD dataset [78].

However, real-world conversational emotions are rarely conveyed through a single modality. Linguistic content conveys semantic intent, vocal prosody reflects affective tone, and facial expressions and visual behaviors provide complementary non-verbal cues. As illustrated in Figure 1.2, multimodal affective analysis explicitly integrates these heterogeneous information sources. Figure 1.8 further provides a concrete example of such multimodal conversational data, where each utterance is temporally aligned with textual, acoustic, and visual signals. These observations motivate the transition from unimodal ERC to *Multimodal Emotion Recognition in Conversation (MERC)*, which generalizes ERC by jointly modeling multiple aligned modalities.

Definition 1.5 (Multimodal Emotion Recognition in Conversation (MERC)). *Given a*

conversation $C = \{u_1, \ldots, u_N\}$*, where each utterance* u_i *is associated with a speaker* p_i *and a set of multimodal observations* $\{\mathbf{x}_i^m\}_{m \in \mathcal{M}}$*, the goal of MERC is to predict an emotion label* y_i *for each utterance by jointly modeling multimodal evidence together with conversational context and speaker interactions.*

Formally, for each modality $m \in \mathcal{M}$ (e.g., text, audio, and vision), modality-specific features $\mathbf{x}_i^m$ are encoded into latent representations

$$\mathbf{h}_i^m = f_m(\mathbf{x}_i^m), \tag{1.31}$$

which are subsequently integrated through a multimodal fusion function

$$\mathbf{h}_i = f_{\text{fusion}}\Big(\{\mathbf{h}_i^m\}_{m \in \mathcal{M}}\Big). \tag{1.32}$$

Based on the fused representation $\mathbf{h}_i$, an utterance-level classifier predicts the corresponding emotion label. When the modality set $\mathcal{M}$ reduces to a single modality, this formulation degenerates to the unimodal ERC setting.

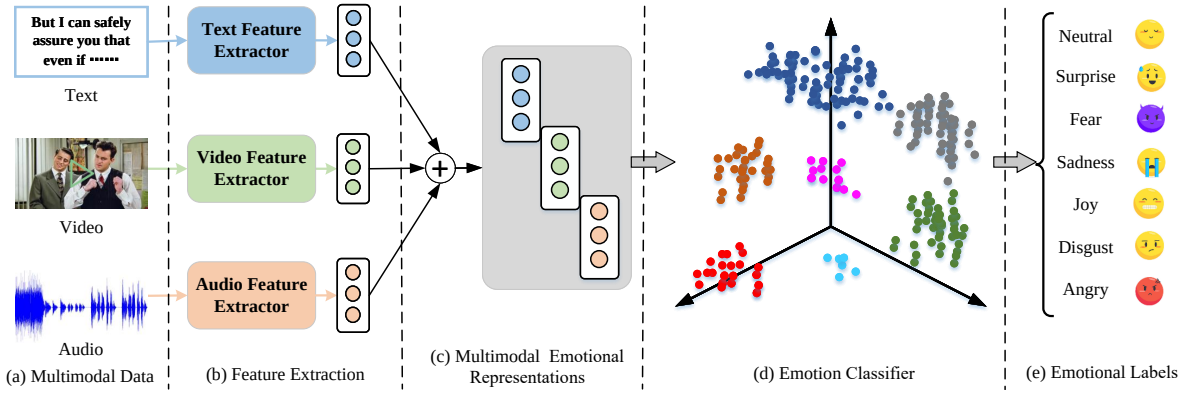


Figure 1.9: Typical steps for Multimodal Emotion Recognition in Conversation [93].

Viewed from this perspective, MERC represents a focused yet challenging instantiation of multimodal affective analysis. It inherits the multimodal nature of affective computing, the psychological grounding of emotion recognition, and the contextual complexity of conversational interaction. Under this unified formulation, MERC places three fundamental requirements on learning models: (i) effective multimodal fusion to capture complementary emotional cues while preserving modality-specific characteristics, (ii) explicit modeling of conversational context, including temporal dependencies and speaker interactions, and (iii) robustness to realistic multimodal data conditions, such as missing or imbalanced modalities. These requirements motivate the advanced

multimodal fusion and contextual modeling approaches investigated in the subsequent chapters of this dissertation.

From a modeling perspective, most MERC approaches can be abstracted into a common processing paradigm. As illustrated in Figure 1.9, the overall process typically consists of three conceptual stages: (i) unimodal feature extraction, (ii) multimodal representation learning and fusion, and (iii) utterance-level emotion classification. This abstraction does not prescribe a specific architecture, but rather highlights the key components that any MERC system must address.

1.4.2 Modeling Paradigms

Having formalized MERC as a task that integrates multimodal evidence with conversational context, we now review the major modeling paradigms adopted in existing MERC systems. From a modeling perspective, MERC can be decomposed into three tightly coupled components: (i) unimodal representation learning, (ii) multimodal representation learning and fusion, and (iii) contextual and conversational modeling. These components correspond to different stages of the MERC pipeline (Figure 1.9) and jointly determine a model's capacity to capture emotionally salient cues, cross-modal interactions, and conversational dynamics. Rather than focusing on implementation details, this subsection organizes prior work according to how these components are modeled, following common taxonomies in recent surveys such as Geetha et al. [23].

1.4.2.1 Unimodal Representation Learning

Unimodal representation learning constitutes the foundation of MERC systems, as all multimodal fusion and contextual reasoning mechanisms operate on modality-specific representations. For each modality, the objective is to transform raw signals into latent representations that encode emotionally relevant information while preserving modality-specific characteristics.

Textual Feature Extraction. With the advancement of deep learning techniques, particularly neural networks, word embedding technology has become pivotal in text feature extraction. Word embedding methods, such as NNLM [7], HLBL [11], and Word2Vec [62], utilize neural networks to capture the semantic relationships between words, representing them as dense vectors in a lower-dimensional space. Recent advancements have

seen the adoption of GloVe [72] and large pre-trained models like BERT, which capture rich contextual information using attention mechanisms.

Visual Feature Extraction. This step is crucial for analyzing facial expressions and body posture from video, as these elements are key indicators of the speaker's sentiment. Public libraries such as OKAO, CERT [52], OpenFace [6], and Facet provide facial landmarks, facial action units (AUs), head pose, gaze, and other features that serve as inputs for visual encoders.

Acoustic Feature Extraction. Open-source toolkits such as OpenEAR [18], openSMILE [19], LibROSA [61], and COVAREP [13] are widely used to extract acoustic features. These toolkits compute prosodic, spectral, and cepstral descriptors (e.g., MFCCs, pitch, energy) and their statistics, providing rich representations of vocal prosody and timbre.

Building upon these extracted features, unimodal encoders such as recurrent neural networks, convolutional neural networks, or transformer-based architectures are employed to learn higher-level representations. These encoders aim to abstract away noise and redundancy while emphasizing emotionally salient patterns, forming the basis for subsequent multimodal fusion and conversational modeling in MERC.

1.4.2.2 Multimodal Representation Learning and Fusion

Multimodal representation learning and fusion lie at the core of MERC, as emotions in conversation are rarely conveyed through a single modality. Linguistic content, vocal prosody, and visual expressions jointly contribute complementary and sometimes redundant emotional cues. The objective of multimodal fusion is therefore to integrate unimodal representations into a unified representation that captures cross-modal interactions while preserving modality-specific information.

Early multimodal approaches adopt simple fusion strategies, such as feature concatenation or weighted averaging, which provide a straightforward means of combining modalities but often fail to capture complex cross-modal dependencies. More expressive fusion mechanisms have since been proposed. Tensor Fusion Networks (TFN) [130] introduce tensor-based outer-product fusion to model high-order cross-modal interactions, while Memory Fusion Networks (MFN) [131] incorporate multimodal memory modules to capture temporal cross-modal dependencies. Subsequent extensions, such as

MARN [132] and RMFN [50], introduce modality-specific attention and hierarchical fusion structures to progressively integrate multimodal information.

From a learning perspective, multimodal fusion can be performed at different stages of the pipeline, giving rise to early, late, and hybrid fusion paradigms as discussed in Section 1.2. Contemporary MERC models increasingly favor intermediate or hybrid fusion schemes, where modality-specific encoders and fusion modules are jointly optimized in an end-to-end manner. While these approaches substantially improve representational expressiveness, they also introduce challenges related to modality heterogeneity, spurious correlations, and imbalance in modality contributions, which are particularly pronounced in conversational settings with low-quality multimodal data.

1.4.2.3 Contextual and Conversational Modeling

Beyond utterance-level multimodal fusion, MERC fundamentally differs from isolated emotion recognition due to its conversational nature. Emotional states in dialogue evolve over time, are influenced by speaker interactions, and often depend on long-range contextual information. Consequently, effective MERC systems must incorporate explicit mechanisms for contextual and conversational modeling.

Sequential context modeling represents an early paradigm in this direction. Approaches such as those proposed by Zhou et al. [146] apply recurrent neural networks over utterance sequences to capture emotional flow across conversations. DialogueRNN [60] further advances this paradigm by maintaining global dialogue context while tracking speaker-specific states. ICON [30] introduces interactive memory networks to model emotional influence between speakers, while DialogueXL [90] and HiTrans [47] extend transformer architectures to better capture long-range conversational dependencies.

Speaker-aware modeling is particularly important in multi-party conversations, where emotional dynamics are strongly speaker-dependent. Methods such as DialogueRNN [60], COSMIC [25], and DAG-ERC [91] explicitly distinguish between self-speaker history and interlocutor context, enabling fine-grained modeling of interpersonal emotional influence.

More recent work adopts graph-based conversational modeling, representing conversations as graphs in which utterances and speakers are treated as nodes connected by edges encoding temporal, structural, or interaction-based relations. DialogueGCN [24] applies graph convolution to propagate contextual information, while DAG-ERC [91]

introduces directed acyclic graphs to model causal emotional influence. MMGCN [35] and MM-DFN [34] further integrate multimodal representations into graph structures, enabling richer reasoning over multimodal conversational context.

To summarize, these modeling paradigms provide the building blocks for MERC systems: unimodal encoders for each modality, multimodal fusion mechanisms to integrate cross-modal information, and contextual/conversational models to capture dialogue-level dependencies. They form a crucial foundation for the advanced graph-based and robustness-oriented approaches developed in the subsequent chapters of this dissertation, which specifically target the three core challenges identified earlier: structured multimodal fusion and contextual modeling, robustness under incomplete modalities, and balanced learning under modality dominance.

1.5 Dataset and Evaluation Metrics

1.5.1 Dataset

Many other datasets are used, depending on the purpose of the proposed model verification. The detailed information of these datasets are described in the following sections. Table 1.2 presents the publicly available benchmark datasets for mutlimodal emotion recognition utilized in this dissertation. The table includes information regarding the release time, modality, and open-source URL for each dataset.

Table 1.2: Multimodal datasets

Name	Year	Modalities	Sent.?	Emo.?	Samples	Speakers	Source	Language	Topics Covered	URL
IEMOCAP [9]	2008	T+A+V	×	✓	10,000	10	Lab data	English	General	https://sail.usc.edu/iemocap/
MELD [78]	2019	T+A+V	×	✓	13,708	407	TV-series Friends	English	General	http://affective-meld.github.io/
CMU-MOSI [129]	2016	T+A+V	✓	×	2,199	98	YouTube	English	General	http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/
CMU-MOSEI [133]	2018	T+A+V	✓	✓	23 453	1,000	YouTube	English	Reviews, Debate, Consulting	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/

Additionally, Table 1.3 outlines the distribution of the dataset across different emotional labels, revealing a long-tail distribution.

Table 1.3: Distribution of conversational emotion recognition datasets on different emotion labels

Labels	IEMOCAP	MELD	CMU-MOSEI
Neutral	1,708	6,436	-
Happiness/Joy	648	2,308	11,982
Surprise/Powerful	-	1,636	2,242
Sadness	1,084	1,002	5,816
Anger/Mad	1,103	1,607	4,831
Disgust	-	361	3,994
Fear/Scared	-	358	1,837
Frustrated	1,849	-	-
Excited	1,041	-	-
Other	-	-	-

Multimodal Corpus of Sentiment Intensity (CMU-MOSI). CMU-MOSI [129] comprises 2199 opinion video clips, with each clip annotated for sentiment on a scale ranging from -3 to 3. This dataset is characterized by comprehensive annotations, including labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-millisecond annotated audio features. Thanks to its fine-grained annotations and heterogeneous recording conditions, CMU-MOSI serves as a representative benchmark for evaluating **robust multimodal fusion** under realistic noise and modality degradation, especially when some modalities are only weakly informative.

Interactive Emotional Dyadic Motion Capture (IEMOCAP). IEMOCAP [9] encompasses approximately 12 hours of acted, multimodal, and multispeaker data. This dataset includes audio-visual components like video, speech, and motion capture of facial expressions, along with corresponding text transcriptions. The database is structured around dyadic sessions where actors engage in improvisations or scripted scenarios carefully chosen to evoke emotional expressions. The clear speaker turns, rich temporal structure, and fully synchronized modalities make IEMOCAP a key testbed for **structured multimodal fusion and conversational context modeling** under full-modality conditions, as well as for controlled studies on **incomplete** and **imbalanced** modality settings in later chapters.

Multimodal EmotionLines Dataset (MELD). MELD [78] is an augmented version of the EmotionLines [32] dataset. While it retains the same dialogue instances present in EmotionLines, MELD goes beyond by incorporating audio and visual modalities in addition to text. Derived from the *Friends* TV series, MELD comprises over 1400 dia-

logues and 13000 utterances. The dialogues involve multiple speakers, and each utterance within a dialogue is labeled with one of seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Compared to IEMOCAP, MELD features more speakers per dialogue and more natural, overlapping interactions, which accentuate challenges in modeling **speaker-dependent context and emotion dynamics** for multimodal conversational emotion recognition.

Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). CMU-MOSEI [133] is notable for being one of the largest collections for multimodal sentiment analysis and emotion recognition available. It consists of over 23,500 sentence-utterance videos presented by more than 1000 YouTube speakers. The dataset is designed to be gender-balanced, with sentence utterances randomly chosen from a variety of topics and monologue videos. The videos are meticulously transcribed and punctuated for comprehensive analysis. Its scale, speaker diversity, and real-world artifacts (e.g., occlusions, background noise, variable expressiveness) make CMU-MOSEI particularly suitable for assessing the **scalability and generalization** of the proposed methods, especially with respect to **robustness to missing modalities** and **balanced learning under modality dominance** in large-scale settings.

Overall, these datasets are chosen because they collectively cover (i) multi-speaker conversational settings with rich temporal and speaker interactions (IEMOCAP, MELD) and (ii) large-scale, noisy multimodal opinion data with heterogeneous modality quality (CMU-MOSI, CMU-MOSEI). This combination enables a systematic evaluation of the proposed models along the three main problem axes of this dissertation: structured multimodal fusion and contextual modeling, robustness under incomplete modalities, and balanced learning under modality dominance.

Limitations. Despite their widespread use, existing datasets also have notable limitations: (1) IEMOCAP, though widely used, is recorded in controlled lab settings with scripted/improvised dialogues, limiting ecological validity. (2) MELD and CMU-MOSEI exhibit class imbalance, with neutral or joy dominating, making minority emotions (e.g., disgust, fear) harder to recognize. (3) Most datasets assume complete modalities at test time, which does not reflect real-world conditions (e.g., poor lighting, audio dropout). These limitations further motivate the robustness-focused contributions of this dissertation, particularly the methods for handling **incomplete modalities** and **modality imbalance**.

1.5.2 Evaluation metrics

In the field of MERC, four commonly used metrics are employed to evaluate model performance: Accuracy (ACC), Weighted Average Accuracy (WA), F1 Score (F1), and Weighted Average F1 Score (WF1). ACC and WF1 are the primary metrics for overall evaluation. However, due to the imbalanced distribution of emotion labels, models often exhibit varying performance across different categories, performing well on some while underperforming on others. To assess the impact of data imbalance on model performance, researchers also report ACC and F1 scores for each individual emotion category. These four evaluation metrics are defined as follows:

We assume that N represents the number of emotion labels in the dialogue emotion dataset, and E_j denotes the total number of samples of emotion labels in the j-th category, $j \in [1, N]$.

1. **Accuracy (Acc)** signifies the model's emotion recognition precision, and its formula is defined as follows:

$$\mathbf{Acc}_j = \frac{\sum_{n=1}^{\vartheta_2} E_j^i}{\sum_{m=1}^{\vartheta_1} S_i^m} \tag{1.33}$$

   where ϑ_1 is the number of labels for a certain category of emotion, ϑ_2 is the number predicted by the model for that category of emotion. E_j^i denotes that the i-th sample in the j-th category is predicted correctly, with $E_j^i \in [0, 1]$. S_j^m represents the m-th sample of the j-th emotion. A larger value of Acc_j implies a better recognition effect of the model on the j-th type of emotion.

2. **F1** value corresponds to the F1-score for each emotion, and its formula is delineated as follows:

$$\mathbf{F1}_j = \frac{2 \times Recall\left(E_{TP}^j, E_{FP}^j\right) \times Precision\left(E_{TP}^j, E_{FN}^j\right)}{Recall\left(E_{TP}^j, E_{FP}^j\right) + Precision\left(E_{TP}^j, E_{FN}^j\right)} \tag{1.34}$$

   where

$$\begin{aligned} Precision\left(E_{TP}^j, E_{FN}^j\right) &= \frac{\mid E_{TP}^j \mid}{\mid E_{TP}^j \cup E_{FN}^j \mid} \\ Recall\left(E_{TP}^j, E_{FP}^j\right) &= \frac{\mid E_{TP}^j \mid}{\mid E_{TP}^j \cup E_{FP}^j \mid} \end{aligned} \tag{1.35}$$

   Where E_{TP}^j represents the number of samples that the model correctly predicts

for the j-th category of emotion, E^j_{FP} denotes the number of samples that the model incorrectly predicts for the j-th category of emotion, and E^j_{FN} signifies the number of emotions from other categories that the model incorrectly predicts as the j-th category of emotion. $Precision(E^j_{TP}, E^j_{FN})$ denotes the precision of the model on the j-th category of emotion, while $Recall(E^j_{TP}, E^j_{FP})$ signifies the recall of the model on the j-th emotion. The F1 value combines the effects of both precision and recall metrics. Typically, a higher F1 value indicates better model prediction.

3. **Weighted accuracy (WA)** is the weighted average of the classification accuracy across all emotion categories. The weight assigned to each sample decreases as the number of samples for the j-th emotion increases. The formula is defined as follows:

$$\mathbf{WA} = \frac{\sum_{m=1}^{\vartheta_1} S_j * \mathbf{Acc}_j}{\sum_{j=1}^{N} \sum_{m=1}^{\vartheta_1} S_j} \tag{1.36}$$

   WA represents the classification accuracy of the model across all emotions combined. A higher WA indicates superior performance of the model on average across all classes.

4. **Weighted F1 (WF1)** is the weighted F1 value across all emotion categories. The weight assigned to each sample decreases as the number of samples for the j-th emotion increases. The formula is defined as follows:

$$\mathbf{WF1} = \frac{\sum_{m=1}^{\vartheta_1} S_j * \mathbf{F1}_j}{\sum_{j=1}^{N} \sum_{m=1}^{\vartheta_1} S_j} \tag{1.37}$$

   WF1 represents the F1 value where the model integrates all emotions. It serves as another effective index for evaluating the model's effectiveness. Typically, a higher WF1 indicates better average performance of the model across all classes.

1.6 Chapter Summary

This chapter established the conceptual and methodological foundations for this dissertation. We formalized MERC as a task requiring joint modeling of multimodal signals, conversational context, and robustness under low-quality data. We reviewed classical and modern fusion paradigms, highlighting the shift from fixed strategies (early/late fusion) to learnable mechanisms (attention, GNNs). We further identified two critical gaps: (1) incomplete modalities, where missing cues degrade fusion, and (2) modality imbalance, where dominant modalities bias learning. Building on this foundation, Chapter 2 addresses structured multimodal fusion and conversational modeling under full-modality settings, proposing graph-based architectures (CORECT, MultiDAG+CL) that explicitly model temporal dependencies and cross-modal interactions. Chapters 3–4 then extend these models to handle incomplete and imbalanced data, respectively, completing the dissertation's progressive research agenda.

Chapter 2

Multimodal Fusion for Emotion Recognition in Conversation

2.1 Introduction

Multimodal Emotion Recognition in Conversation (MERC) aims to identify the emotional state of each utterance in a dialogue by jointly reasoning over conversational context and heterogeneous multimodal signals, including language, acoustic cues, and visual expressions. Unlike emotion recognition from isolated utterances, MERC is inherently contextual and dynamic: emotional meanings emerge from interactions between speakers, long-range temporal dependencies across dialogue turns, and complementary yet uneven multimodal cues [79].

This chapter addresses **Objective O1** of the dissertation, which focuses on developing effective multimodal fusion and contextual modeling approaches for MERC under full-modality settings. Building upon the multimodal learning foundations introduced in Chapter 1, we narrow our scope to conversational scenarios, where multimodal fusion must explicitly account for dialogue structure, temporal dynamics, and speaker interactions. In this context, multimodal learning is not merely a problem of feature aggregation, but a structured reasoning process that requires preserving modality-specific information while modeling cross-modal interactions within evolving conversations.

MERC presents several fundamental challenges. First, conversational emotions are highly context-dependent, often requiring access to long-range dialogue history and speaker-specific information for correct interpretation [24, 60]. Second, emotional states in conversation evolve over time and may change abruptly, making temporal model-

ing essential [91]. Third, even under full-modality settings, modality contributions are frequently uneven and label distributions can be skewed, which complicates effective multimodal fusion and can bias representation learning [31]. These challenges jointly motivate the need for principled multimodal fusion mechanisms together with explicit conversational modeling.

Existing MERC approaches can be broadly categorized into recurrent-based, graph-based, and fusion-centric methods. Recurrent models such as DialogueRNN [60] and its variants track speaker and dialogue states with gated recurrent units, and thus capture local contextual dependencies across utterances. However, their reliance on sequential recurrence makes it difficult to model long-range, multi-speaker interactions and complex relational structures in extended conversations. Fusion-centric approaches, including early/late fusion and attention-based multimodal models originally designed for non-conversational settings, integrate multimodal features directly but often risk collapsing modality-specific affective cues into a single embedding, limiting their effectiveness in complex, multi-party conversational scenarios.

Graph-based methods provide a more structured view of dialogue dynamics by explicitly encoding conversational relations and temporal dependencies. DialogueGCN [24] and related models construct graphs over utterances and speakers, enabling relational reasoning across dialogue turns and participants. Nevertheless, many of these methods either (i) operate on unimodal or loosely fused representations, or (ii) adopt coarse multimodal fusion strategies that do not explicitly model fine-grained cross-modal interactions. As a result, they only partially exploit complementary cues from heterogeneous modalities and do not clearly disentangle the roles of conversational structure and multimodal fusion.

From a methodological perspective, the proposed models in this chapter are situated within the graph-based paradigm for MERC, while explicitly addressing key limitations of existing graph-based and fusion-centric approaches. Traditional graph-based MERC models primarily focus on encoding conversational structure and temporal dependencies, but often entangle conversational modeling and multimodal fusion in a single stage. This design makes it difficult to preserve unimodal representations, to reason about pairwise modality interactions, or to control how contextual information propagates across the dialogue. In contrast, this chapter advances graph-based MERC along two complementary axes: (i) by disentangling relational–temporal conversational structure modeling from modality-aware interaction learning, and (ii) by explicitly investigat-

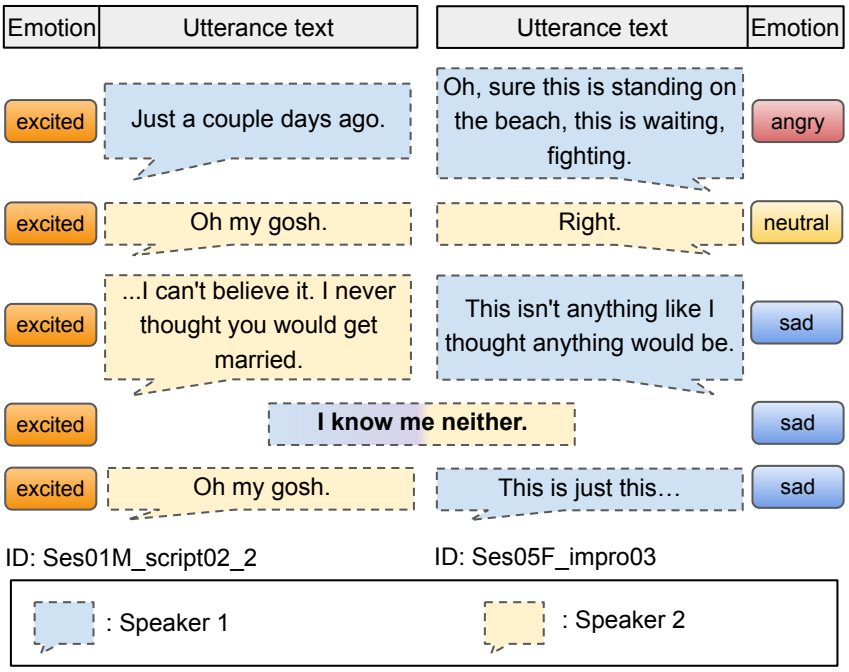


Figure 2.1: Examples of temporal effects in conversation. The emotional meaning of an utterance may change depending on surrounding context.

ing how conversational dependencies should be learned under varying levels of temporal and contextual complexity.

Recent studies have attempted to incorporate multimodal information into graph-based conversational models. MMGCN [35] augments dialogue graphs with multimodal node representations and modality-aware graph convolution, thereby improving contextual modeling compared with purely textual graph-based models. COGMEN [40] further integrates gated mechanisms and multimodal attention into graph-based architectures, achieving strong performance on benchmarks such as IEMOCAP and MELD. These approaches demonstrate that combining graph-based conversational structure with multimodal signals is effective for MERC. However, they still exhibit two important limitations with respect to the objectives of this chapter. First, they typically fuse modalities within a single graph propagation process, which couples multimodal fusion with conversational message passing and makes it difficult to explicitly capture pairwise modality interactions across the entire dialogue. Second, they rely on fixed graph structures and uniform training strategies, which can be insufficient for capturing diverse conversational dynamics and varying temporal patterns, particularly when emotional transitions and speaker interactions become complex.

As illustrated in Figure 2.1, the emotional interpretation of an utterance can vary substantially depending on surrounding context and speaker interactions, highlighting

the necessity of temporally aware and modality-sensitive MERC models. Addressing these requirements calls for architectures that (i) represent conversational context in a structured and interpretable manner, (ii) perform explicit, modality-aware fusion that preserves unimodal characteristics, and (iii) learn conversational dependencies in a way that adapts to the difficulty and complexity of dialogue patterns.

To address **RQ1**, we first propose **CORECT**, a graph-based framework for multimodal representation learning and context-aware fusion in MERC. CORECT explicitly separates conversational structure modeling from multimodal interaction learning. It introduces a relational–temporal graph convolutional network that encodes speaker-dependent and long-range conversational dependencies over a structured multimodal dialogue graph, while maintaining modality-specific utterance representations. On top of these contextualized representations, CORECT employs a stacked, pairwise cross-modal interaction module that models fine-grained interactions between text, audio, and visual modalities across the entire dialogue. This design allows CORECT to preserve unimodal information, capture complementary cross-modal cues, and provide a robust representational foundation for multimodal fusion in conversational emotion recognition.

Building upon this representational foundation, we then address **RQ2** by investigating how conversational dependencies should be learned under varying levels of temporal and contextual complexity. We introduce **MultiDAG+CL**, which formulates MERC as a structured reasoning problem over directed acyclic graphs and integrates curriculum learning into the training process. Instead of increasing model expressiveness solely through more complex graph architectures, MultiDAG+CL focuses on the learning dynamics: it constructs directed acyclic graphs that encode directed emotional and contextual influences among utterances, and employs a curriculum scheduler that progressively organizes training samples from easy to hard conversational patterns. By aligning graph-based contextual reasoning with a principled, difficulty-aware training schedule, MultiDAG+CL enables more stable optimization and more effective modeling of emotional transitions and speaker dynamics in complex dialogues.

Formally, MERC is solved by learning a model that maps a multimodal conversation to utterance-level emotion labels.

Input. A labeled multimodal dataset $\mathcal{D} = \{(C^{(k)}, \mathbf{y}^{(k)})\}_{k=1}^{K}$, where each conversation $C^{(k)} = [u_1^{(k)}, \ldots, u_{N_k}^{(k)}]$ is a sequence of N_k utterances. Each utterance $u_i^{(k)}$ is associated with a speaker identity $p_i^{(k)}$, modality-specific representations $\{(u_i^{(k)})^a, (u_i^{(k)})^v, (u_i^{(k)})^t\}$

for audio, visual, and text, and an emotion label $y_i^{(k)}$. The sequence of emotion labels for $C^{(k)}$ is $\mathbf{y}^{(k)} = [y_1^{(k)}, \ldots, y_{N_k}^{(k)}]$.

Output. A learned multimodal emotion recognition model that predicts utterance-level emotion labels for new conversations.

In the following sections, we instantiate our multimodal emotion recognition model with two fusion-centric architectures, CORECT and MultiDAG+CL, which differ in how they model conversational context and cross-modal interactions.

2.2 Multimodal Fusion with Relation and Temporal Conversational Modeling

2.2.1 Overview

This subsection presents **CORECT**, a graph-based framework for multimodal emotion recognition in conversation [**VanNTC 1**]. CORECT is designed to address **Research Question RQ1** by focusing on multimodal representation learning and context-aware fusion under conversational settings.

As illustrated in Figure 2.2, CORECT jointly models conversational structure and multimodal interactions at the utterance level, while keeping these two aspects conceptually separated. *Local conversational context* is captured by modeling relational and temporal dependencies among utterances and speakers, whereas *global multimodal context* is captured by explicitly learning cross-modal interactions across the entire dialogue.

Concretely, CORECT consists of two main components. The *Relational–Temporal Graph Convolutional Network (RT-GCN)* encodes local contextual information for each utterance on a multimodal conversational graph, leveraging both inter-speaker relations and temporal dependencies across dialogue turns. On top of these contextualized representations, the *Pairwise Cross-modal Feature Interaction (P-CM)* module learns global dialogue-level representations by modeling pairwise interactions between modalities (text, audio, visual) in a modality-aware manner. This design decouples structural conversational modeling from cross-modal interaction learning and provides a principled basis for multimodal fusion in conversational emotion recognition.

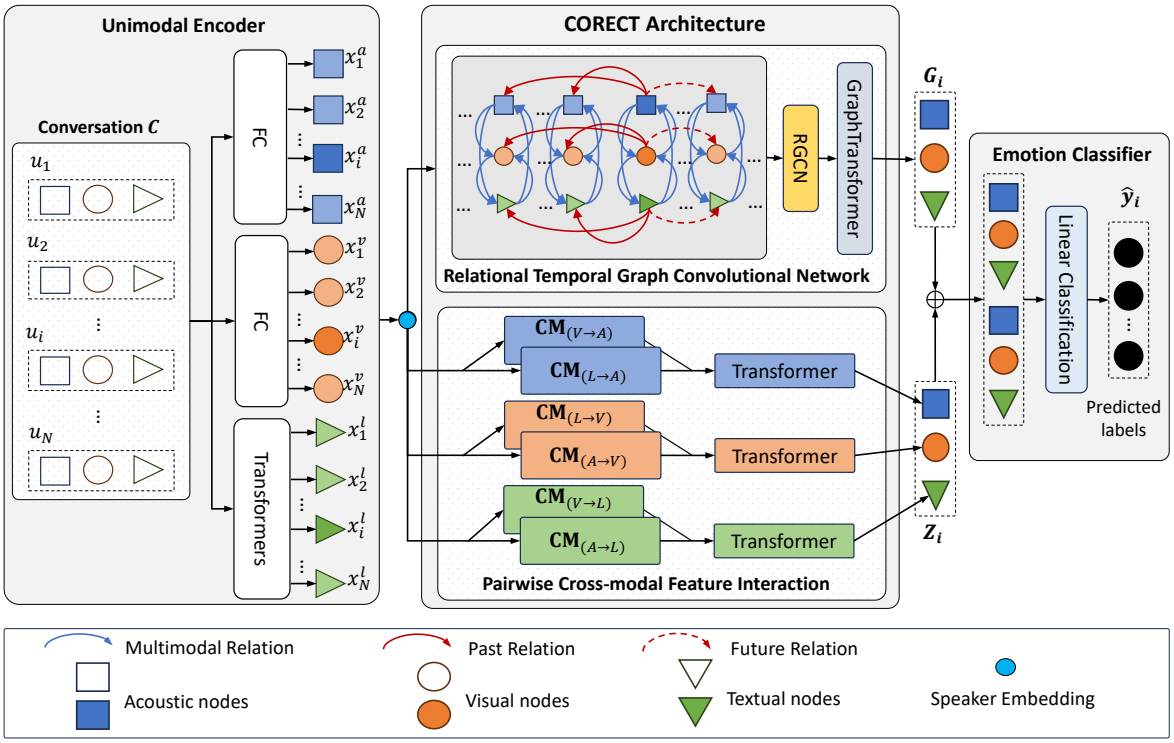


Figure 2.2: Framework illustration of CORECT

2.2.2 Utterance-level Feature Extraction

This subsection describes the extraction of utterance-level multimodal features that serve as the input to the CORECT framework. Given a multi-speaker conversation consisting of N utterances, the goal of this step is to obtain modality-specific representations for each utterance while preserving information that is essential for subsequent conversational modeling and multimodal fusion.

2.2.2.1 Unimodal Encoder

Given an utterance u_i, each data modality provides a distinct view of its affective content. To preserve modality-specific characteristics, we employ dedicated unimodal encoders to extract utterance-level features for each modality. Specifically, the acoustic, visual, and lexical modalities are represented as $\mathbf{x}_i^a \in \mathbb{R}^{d_a}$, $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$, and $\mathbf{x}_i^l \in \mathbb{R}^{d_l}$, respectively, where d_a, d_v, and d_l denote the corresponding feature dimensions.

For the textual modality, we adopt a Transformer-based encoder [105] to capture contextualized semantic representations. The lexical feature $\mathbf{x}_i^l$ is obtained from the

textual input u_i^l as follows:

$$\mathbf{x}_i^l = \textbf{Transformer}(u_i^l; \mathbf{W}_{\text{trans}}^l), \tag{2.1}$$

where $\mathbf{W}_{\text{trans}}^l$ denotes the learnable parameters of the Transformer encoder.

For the acoustic and visual modalities, we employ modality-specific fully connected networks to project the raw modality features into latent feature spaces:

$$\mathbf{x}_i^\tau = \textbf{FC}(u_i^\tau; \mathbf{W}_{\text{fc}}^\tau), \quad \tau \in \{a, v\}, \tag{2.2}$$

where $\textbf{FC}(\cdot)$ denotes a fully connected network, $\mathbf{W}_{\text{fc}}^\tau \in \mathbb{R}^{d_\tau \times d_{\text{in}}^\tau}$ are learnable parameters, and d_{in}^τ is the input dimensionality of modality τ.

2.2.2.2 Speaker Embedding

Speaker identity plays an important role in conversational emotion recognition, as emotional expressions are often conditioned on the speaker. Inspired by MMGCN [35], we incorporate speaker information by learning speaker embeddings and integrating them into utterance-level representations.

Let p_i denote the speaker associated with utterance u_i. We define a learnable embedding function that maps speaker identities to latent representations:

$$\mathbf{s}_{p_i} = \textbf{Embedding}(p_i) \in \mathbb{R}^{d_s}. \tag{2.3}$$

To inject speaker information into modality-specific utterance features, the speaker embedding is projected into each modality space and combined with the corresponding unimodal representation:

$$\mathbf{x}_i^\tau \leftarrow \mathbf{x}_i^\tau + \eta \mathbf{W}_s^\tau \mathbf{s}_{p_i}, \quad \tau \in \{a, v, l\}, \tag{2.4}$$

where $\mathbf{W}_s^\tau \in \mathbb{R}^{d_\tau \times d_s}$ are modality-specific projection matrices, and $\eta \in [0, 1]$ controls the contribution of speaker information.

After this step, we obtain speaker-aware unimodal representations for all utterances, which are subsequently used for relational–temporal graph construction and cross-modal interaction modeling in CORECT.

2.2.3 Relational Temporal Graph Convolutional Network (RT-GCN)

RT-GCN is designed to capture *local conversational context* for each utterance by explicitly modeling structured interactions among utterances and modalities in a multimodal dialogue graph. By incorporating both relational and temporal dependencies, RT-GCN enables fine-grained contextual reasoning over multimodal conversational data.

2.2.3.1 Multimodal Graph Construction

We construct a multimodal dialogue graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ from each conversation, where $\mathcal{V}$ denotes the set of nodes, $\mathcal{E}$ denotes the set of edges, and $\mathcal{R}$ denotes the set of relation types associated with edges. Each utterance is represented by modality-specific nodes, resulting in $|\mathcal{V}| = 3N$ nodes for a conversation with N utterances. Figure 2.3 illustrates an example of the constructed multimodal graph.

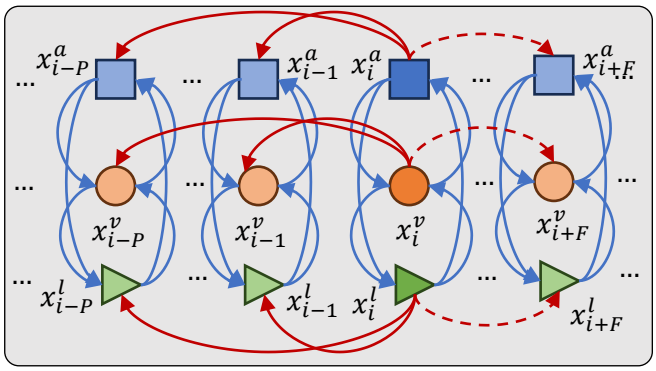


Figure 2.3: An example of multimodal graph construction. Audio, visual, and textual nodes are represented by squares, circles, and triangles, respectively. The temporal window is set to $[\mathcal{P}, \mathcal{F}] = [2, 1]$ for the query utterance u_i. Solid blue arrows indicate cross-modal connections, while solid and dashed red arrows denote past and future temporal relations.

Nodes. For each utterance u_i, we create three modality-specific nodes u_i^a, u_i^v, and u_i^l, corresponding to the acoustic, visual, and lexical modalities, respectively. Each node is initialized with the speaker-aware unimodal feature vectors $\mathbf{x}_i^a$, $\mathbf{x}_i^v$, and $\mathbf{x}_i^l$ obtained from the utterance-level feature extraction stage.

Edges. An edge $(u_i^\tau, u_j^\tau, r_{ij}) \in \mathcal{E}$, where $\tau \in \{a, v, l\}$, represents an interaction between nodes u_i^τ and u_j^τ with relation type $r_{ij} \in \mathcal{R}$. We define two groups of relations:

multimodal relations $\mathcal{R}_{multi}$ and *temporal relations* $\mathcal{R}_{temp}$. Together, these relations allow the graph to model both intra-utterance multimodal interactions and inter-utterance temporal dependencies.

Multimodal Relations. Emotions in conversation are rarely conveyed through a single modality. To capture interactions across modalities within the same utterance, we define multimodal relations that connect modality-specific nodes of an utterance. In addition, self-loop relations are included to preserve modality-specific information. Specifically, we define the following nine multimodal relation types:

$$\mathcal{R}_{multi} = \begin{cases} \{(u_i^a, u_i^v), (u_i^v, u_i^a), (u_i^a, u_i^a)\} \\ \{(u_i^v, u_i^l), (u_i^l, u_i^v), (u_i^v, u_i^v)\} \\ \{(u_i^l, u_i^a), (u_i^a, u_i^l), (u_i^l, u_i^l)\} \end{cases} \tag{2.5}$$

Temporal Relations. To model the evolution of emotions over time, it is essential to distinguish interactions occurring in different temporal orders [77]. We define a sliding temporal window $[\mathcal{P}, \mathcal{F}]$ to control the number of past and future utterances connected to a given node u_i^τ. This window enables the graph to capture local temporal context while avoiding overly dense connections. Accordingly, we define six temporal relation types as:

$$\mathcal{R}_{temp} = \begin{cases} \{(u_j^\tau \overset{\text{past}}{\rightarrow} u_i^\tau) \mid i - \mathcal{P} < j < i\} \\ \{(u_i^\tau \overset{\text{future}}{\leftarrow} u_j^\tau) \mid i < j < i + \mathcal{F}\} \end{cases} \tag{2.6}$$

where $\tau \in \{a, v, l\}$ and $i, j \in \{1, \ldots, N\}$. The operators $\overset{\text{past}}{\rightarrow}$ and $\overset{\text{future}}{\leftarrow}$ denote directed edges from past and future utterances, respectively.

2.2.3.2 Graph Learning

To leverage the heterogeneous interactions encoded in the multimodal dialogue graph, we adopt Relational Graph Convolutional Networks (R-GCN) [87]. R-GCN allows relation-specific message passing, enabling the model to learn distinct transformation patterns for different types of multimodal and temporal relations.

For each relation type $r \in \mathcal{R}$, node representations are updated through relation-specific transformations. The intermediate representation of node u_i^τ after R-GCN ag-

gregation is computed as:

$$\mathbf{g}_i^\tau = \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \mathbf{W}_r \mathbf{x}_j^\tau + \mathbf{W}_0 \mathbf{x}_i^\tau, \tag{2.7}$$

where $\mathcal{N}_r(i)$ denotes the set of neighbors of node u_i^τ under relation r, $\mathbf{W}_r$ and $\mathbf{W}_0$ are learnable weight matrices, and $\mathbf{x}_j^\tau$ is the input feature of neighbor node u_j^τ.

To further enhance contextual representation learning, we employ a Graph Transformer layer [128] on top of the R-GCN outputs. The self-attention mechanism enables each node to selectively attend to its neighbors and capture both local and global interaction patterns. Given the R-GCN output $\mathbf{g}_i^\tau$, the Graph Transformer updates the representation as:

$$\mathbf{o}_i^\tau = \Big\|_{c=1}^{C} \left[\mathbf{W}_1 \mathbf{g}_i^\tau + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}^\tau \mathbf{W}_2 \mathbf{g}_j^\tau \right], \tag{2.8}$$

where $\|\cdot\|$ denotes concatenation over C attention heads, $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable projection matrices, and $\mathcal{N}(i)$ denotes the set of neighboring nodes of u_i^τ.

The attention coefficient $\alpha_{i,j}^\tau$ is computed as:

$$\alpha_{i,j}^\tau = \text{softmax}\left(\frac{(\mathbf{W}_3 \mathbf{g}_i^\tau)^\top (\mathbf{W}_4 \mathbf{g}_j^\tau)}{\sqrt{d}} \right), \tag{2.9}$$

where $\mathbf{W}_3$ and $\mathbf{W}_4$ are learnable parameters and d is the feature dimensionality used for scaling.

After graph learning, we obtain modality-specific graph-enhanced representations for the entire conversation:

$$\mathbf{G}^\tau = \{\mathbf{o}_1^\tau, \mathbf{o}_2^\tau, \ldots, \mathbf{o}_N^\tau\}, \tag{2.10}$$

where $\mathbf{G}^\tau \in \mathbb{R}^{N\times d_{h_2}}$ represents the sequence of contextualized utterance representations for modality $\tau \in \{a, v, l\}$.

2.2.4 Pairwise Cross-modal Feature Interaction

Multimodal conversational signals are heterogeneous and often temporally misaligned, which poses challenges for effective fusion. To explicitly model cross-modal dependencies and dialogue-level interactions, we introduce the *Pairwise Cross-modal*

Feature Interaction (P-CM) module, inspired by cross-modal attention mechanisms [103]. P-CM instantiates directional cross-modal attention in a stacked pairwise interaction block tailored for context-aware fusion in MERC. Figure 2.4 illustrates the overall architecture of the P-CM module.

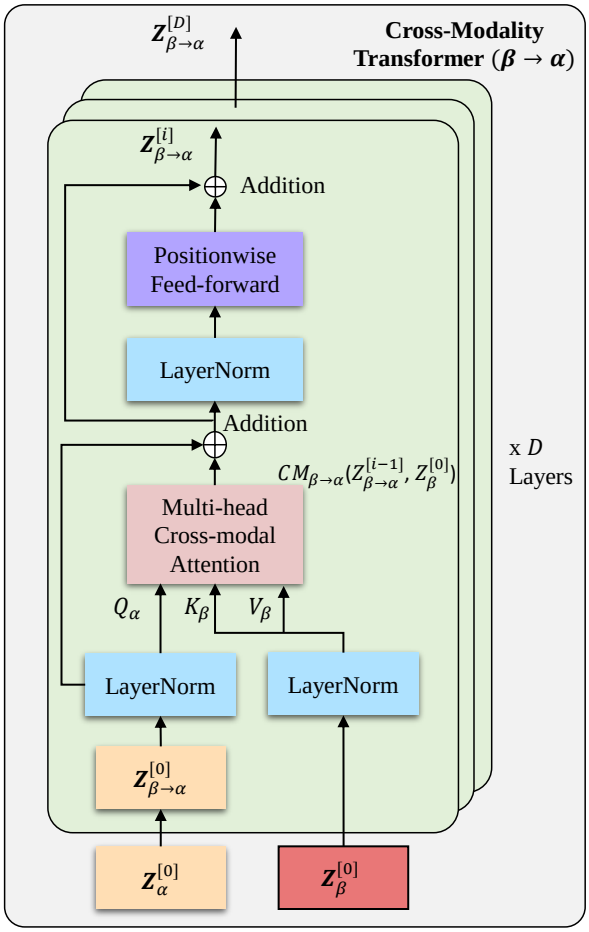


Figure 2.4: Illustration of the P-CM module.

Cross-modal Attention. Given two modalities, e.g., acoustic modality a and lexical modality l, let $\mathbf{X}^a \in \mathbb{R}^{N\times d_a}$ and $\mathbf{X}^l \in \mathbb{R}^{N\times d_l}$ denote their utterance-level representations after unimodal encoding and contextualization by RT-GCN. P-CM enriches one modality by attending to another modality through cross-modal attention, enabling it to incorporate complementary cues while preserving its own representation space.

Following the Transformer attention mechanism [105], we define:

$$\mathbf{Q}^a = \mathbf{X}^a\mathbf{W}_{Q^a}, \tag{2.11}$$

$$\mathbf{K}^l = \mathbf{X}^l\mathbf{W}_{K^l}, \tag{2.12}$$

$$\mathbf{V}^l = \mathbf{X}^l\mathbf{W}_{V^l}, \tag{2.13}$$

where $\mathbf{W}_{Q^a} \in \mathbb{R}^{d_a \times d_k}$, $\mathbf{W}_{K^l} \in \mathbb{R}^{d_l \times d_k}$, and $\mathbf{W}_{V^l} \in \mathbb{R}^{d_l \times d_v}$ are learnable parameters.

The cross-modal attention from modality l to modality a is computed as:

$$\mathbf{CM}^{l \to a} = \text{softmax}\left(\frac{\mathbf{Q}^a (\mathbf{K}^l)^\top}{\sqrt{d_k}}\right) \mathbf{V}^l. \tag{2.14}$$

This operation allows acoustic representations to selectively focus on lexical cues that are most informative for emotion recognition.

Stacked Pairwise Cross-modal Interaction. To capture higher-order and long-range cross-modal dependencies, P-CM stacks D layers of pairwise cross-modal interaction for each modality pair. Let $\mathbf{Z}^{[i]}_{a \rightleftarrows l} \in \mathbb{R}^{N \times d_v}$ denote the bidirectional acoustic–lexical cross-modal representation at the i-th layer, initialized with $\mathbf{Z}^{[0]}_{a \rightleftarrows l} = \mathbf{X}^a$. The update at layer i is defined as:

$$\overline{\mathbf{Z}}^{[i]}_{a \rightleftarrows l} = \mathbf{CM}^{l \to a}_{[i]}\Big(\text{LN}(\mathbf{Z}^{[i-1]}_{a \rightleftarrows l}), \text{LN}(\mathbf{X}^l)\Big) + \mathbf{Z}^{[i-1]}_{a \rightleftarrows l}, \tag{2.15}$$

$$\mathbf{Z}^{[i]}_{a \rightleftarrows l} = \text{FFN}\Big(\text{LN}(\overline{\mathbf{Z}}^{[i]}_{a \rightleftarrows l})\Big) + \overline{\mathbf{Z}}^{[i]}_{a \rightleftarrows l}, \tag{2.16}$$

where $\text{LN}(\cdot)$ denotes layer normalization [2] and $\text{FFN}(\cdot)$ is a position-wise feed-forward network.

The same procedure is applied symmetrically to obtain $\mathbf{Z}^{[D]}_{a \rightleftarrows v}$ and $\mathbf{Z}^{[D]}_{v \rightleftarrows l}$ for the remaining modality pairs. Finally, the resulting pairwise cross-modal representations are aggregated to form global, modality-aware dialogue representations used for emotion classification. By operating at the pairwise level and stacking multiple interaction layers, P-CM can model fine-grained, complementary relations between modalities without collapsing their individual characteristics.

2.2.5 Multimodal Emotion Classification

The final emotion prediction is performed by integrating the local conversational context captured by the RT-GCN module and the global pairwise cross-modal representations learned by the P-CM module. These two complementary representation are fused to construct utterance-level multimodal representations for emotion classification.

Let $\mathbf{G} = \{o^\tau_1, o^\tau_2, \ldots, o^\tau_N\}$, $\tau \in \{a, v, l\}$, denote the collection of graph-enhanced utterance representations produced by RT-GCN, and let $\mathbf{Z} = \{\mathbf{Z}^{[D]}_{a \rightleftarrows v}, \mathbf{Z}^{[D]}_{a \rightleftarrows l}, \mathbf{Z}^{[D]}_{v \rightleftarrows l}\}$ denote

the set of pairwise cross-modal representations obtained from the final P-CM layer. The fusion operation is defined as:

$$\text{Fusion}([\mathbf{G}, \mathbf{Z}]) = \left[\{o_1^\tau, o_2^\tau, \ldots, o_N^\tau\},\ \{\mathbf{Z}_{a\rightleftarrows v}^{[D]}, \mathbf{Z}_{a\rightleftarrows l}^{[D]}, \mathbf{Z}_{v\rightleftarrows l}^{[D]}\}\right]. \tag{2.17}$$

For each utterance u_i, the corresponding multimodal representation $\mathbf{h}_i$ is constructed by concatenating the i-th elements from the fused representations:

$$\mathbf{h}_i = \left[o_i^a \parallel o_i^v \parallel o_i^l \parallel \mathbf{Z}_{a\rightleftarrows v,i}^{[D]} \parallel \mathbf{Z}_{a\rightleftarrows l,i}^{[D]} \parallel \mathbf{Z}_{v\rightleftarrows l,i}^{[D]}\right], \tag{2.18}$$

where $\mathbf{Z}_{\alpha\rightleftarrows\beta,i}^{[D]}$ denotes the pairwise cross-modal representation corresponding to utterance u_i.

The fused representation $\mathbf{h}_i$ is then fed into a feed-forward classification network to predict the emotion label:

$$\mathbf{v}_i = \text{ReLU}(\mathbf{\Phi}_0\mathbf{h}_i + \mathbf{b}_0), \tag{2.19}$$

$$\mathbf{p}_i = \text{softmax}(\mathbf{\Phi}_1\mathbf{v}_i + \mathbf{b}_1), \tag{2.20}$$

$$\hat{y}_i = \arg\max(\mathbf{p}_i), \tag{2.21}$$

where $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$ are learnable weight matrices, $\mathbf{b}_0$ and $\mathbf{b}_1$ are bias terms.

Algorithm 1 presents the detailed training and inference procedures of CORECT.

2.2.6 Implementation

Dataset. We investigate two public real-life datasets for the multimodal ERC task including IEMOCAP [9] and CMU-MOSEI [133].

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 [19]; visual features, with a size of 512, are extracted using OpenFace [6]; textual features, with a size of 768, are derived using sBERT [81].

For MOSEI, the audio features are extracted using librosa [61] with 80 filter banks,

Algorithm 1 CORECT: Training and Inference Procedure

Require: Pre-extracted unimodal features for all utterances; past window size $\mathcal{P}$; future window size $\mathcal{F}$; number of epochs T

Ensure: Predicted emotion labels $\{\hat{y}_i\}$

for epoch $t = 1$ to T **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Construct the multimodal graph $\mathcal{G}$ using temporal windows $[\mathcal{P}, \mathcal{F}]$ and relation types $R_{\text{multi}}, R_{\text{temp}}$
    Update node representations via RT-GCN
    Compute pairwise cross-modal interaction via P-CM
    Fuse representations and predict $\hat{y}_i = \text{softmax}(\text{MLP}(h_i^{\text{fused}}))$
    Compute cross-entropy loss and update parameters
  end for
end for
for each test conversation C **do** ▷ Inference
  Apply RT-GCN and P-CM with fixed $[\mathcal{P}, \mathcal{F}]$ to obtain h_i^{fused}
  Output predicted labels $\{\hat{y}_i\}_{i=1}^{N}$
end for

resulting in a feature vector size of 80.The visual features, with a size of 35, are obtained from [133]. The textual features, with a size of 768, are obtained using sBERT [81].

Evaluation Metrics. We use *weighted F1-score* (w-F1) and *Accuracy* (Acc.) as evaluation metrics.

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. Referring 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.

[1] https://pytorch.org/

Table 2.1: The results on IEMOCAP (6-way) multimodal (A+V+T) setting.

Methods	IEMOCAP (6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)
bc-LSTM	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN	47.10	**80.88**	58.71	66.08	70.97	61.21	65.54	65.04
MMGCN	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
DialogueCRN	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN	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)

*Results in **bold** denote the best performance, underlined values indicate the second-best.*
↑ marks improvements over the previous SOTA.

2.2.7 Results

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 2.1), 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 2.5).

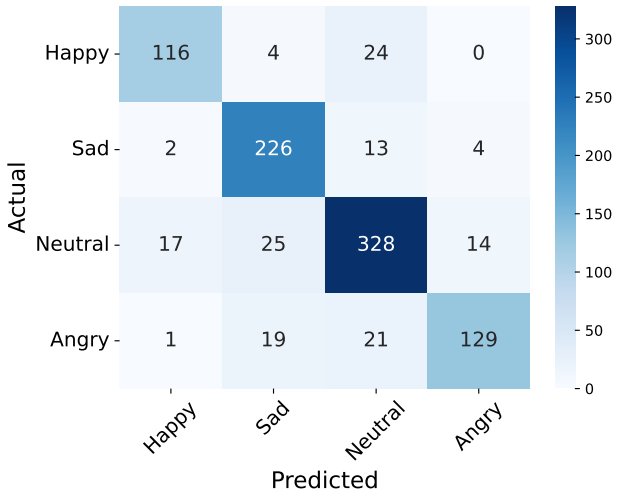
Nevertheless, the accuracy and weighted F1 score of CORECT are 2.89% and 2.75% higher than all baseline models on average. Likewise, we observe the similar phenomena on the IEMOCAP (4-way) dataset with a 2.49% improvement over the previous state-of-the-art models as Table 2.2. These results affirm the efficiency of CORECT for the multimodal ERC task.

Table 2.2: The results on the IEMOCAP (4-way) dataset in the multimodal setting.

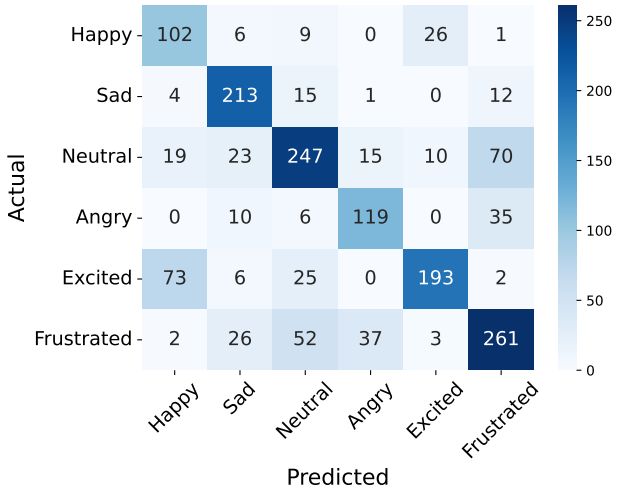
Modality Settings	IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)
bc-LSTM [77]	75.20	75.13
CHFusion [59]	76.59	76.80
COGMEN [40]	82.29	82.15
CORECT (Ours)	**84.73** (↑ 2.44)	**84.64** (↑ 2.49)

The ↑ indicates the improvement over the previous SOTA model.

CMU-MOSEI: Table 2.3 presents a comparison of the CORECT model on the CMU-MOSEI dataset with current SOTA models in two settings: Sentiment Classifica-



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



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

Figure 2.5: 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*.

tion (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. We provide additional experiments on the CMU-MOSEI dataset for all possible combinations of modalities in Table 2.4.

Table 2.3: Results on CMU-MOSEI dataset compared with previous works.

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 [92]	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE [14]	82.40	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN [40]	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**

Bold *results indicate the best performance, underlined results denote the second-best.*

Table 2.4: Ablation study on CMU-MOSEI dataset.

Datasets	Modality Settings	Sentiment Class Accuracy (%)		Emotion Class weighted F1-score (%)					
		2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilogue-Net [92]	A+T+V	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE [14]	A+T	82.4	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN [40]	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**

Bold *results indicate the best performance, underlined results denote the second-best.*

Effect of Main Components. The impact of main components in our CORECT model is presented via Table 2.5. 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*.

Table 2.5: The performance of CORECT in different strategies under the fully multimodal (A+V+T) setting.

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**

Bold *results represent the best performance, underlined results denote the second-best,* ↓ *indicates the performance drop when ablating a module relative to CORECT model.*

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.

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 2.6: Ablation study on IEMOCAP dataset.

Table 2.6 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 [40, 46] 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 2.6 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 2.7 presents the performance of the CORECT model in different modality combinations on both the IEMOCAP and CMU-MOSEI datasets.

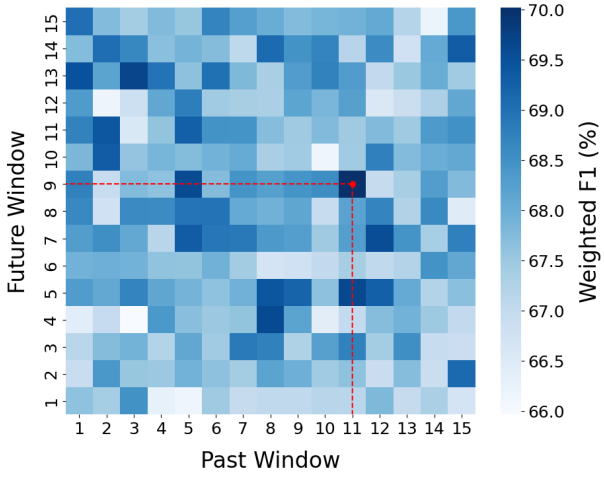


Figure 2.6: The effects of $\mathcal{P}$ and $\mathcal{F}$ nodes in the past and future of CORECT model on the IEMOCAP (6-way) The red-dash line implies our best setting for $\mathcal{P}$ and $\mathcal{F}$.

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 2.7: The performance of CORECT under various modality settings.

For IEMOCAP (Table 2.1 and Table 2.2), the textual modality performs the best among the uni-modal 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 2.3), where fusing all modalities together leads to a better result compared to using individual or paired modalities.

2.2.8 Discussion

While CORECT demonstrates strong performance for multimodal emotion recognition in conversation, several limitations remain that motivate further investigation. First, its performance is sensitive to several hyper-parameters, such as the number of attention heads in the P-CM module and the sizes of the past and future temporal win-

dows in RT-GCN. Given the limited computational budget, exhaustive hyper-parameter search is impractical, which may lead to suboptimal configurations.

More importantly, although CORECT models relational and temporal dependencies through a bidirectional graph structure, it still treats conversational context in a relatively symmetric manner. In real-world dialogues, emotional dynamics are often asymmetric over time, where past utterances usually exert a stronger and more structured influence on the current emotional state than future utterances. Such directional dependencies are difficult to capture with fixed temporal windows and densely connected graph structures.

These limitations suggest that, beyond improving fusion and representation learning, effective multimodal emotion recognition in conversation also requires directional modeling of emotional transitions and learning strategies that can adapt to varying sample difficulty. Motivated by these challenges, we next investigate a complementary framework that formulates conversational context as a directed acyclic graph and explicitly models emotion propagation over time. This leads to the proposed **MultiDAG+CL** [**VanNTC 2**] framework, which emphasizes temporal directionality and curriculum-based learning to better capture the evolving nature of emotions in conversational settings.

2.3 Multimodal Fusion with Directed Acyclic Graph Modeling and Curriculum Learning

2.3.1 Overview

Early recurrent-based methods, such as DialogueRNN [60] and DialogueCRN [33], capture speaker states and short-term dependencies but rely on sequential propagation, which limits their ability to reason over complex temporal interactions. More recent graph-based approaches, including DialogueGCN [24] and DAG-ERC [91], introduce relational structures to model conversational dependencies. However, these methods primarily focus on graph expressiveness and often treat all training samples uniformly, making learning dynamics sensitive to noisy or complex dialogue patterns.

While CORECT [**VanNTC 1**] already incorporates relational and temporal graph modeling to encode long-range conversational context, its primary focus lies in mul-

timodal representation learning and context-aware fusion. In contrast, MultiDAG+CL is designed to further investigate how temporal and speaker-dependent conversational dependencies should be *learned* under varying levels of dialogue complexity. Rather than introducing additional fusion mechanisms, MultiDAG+CL emphasizes structured temporal reasoning and learning dynamics, thereby directly targeting **RQ2**.

MultiDAG+CL consists of two core components: *MultiDAG* and *Curriculum Learning (CL)*. The *MultiDAG* component models conversational context as a *directed acyclic graph*, where utterances are organized according to their temporal order and speaker interactions. By enforcing directionality and acyclicity, MultiDAG enables explicit modeling of asymmetric emotion propagation across dialogue turns, allowing past utterances to exert structured influence on subsequent emotional states. This formulation extends prior DAG-based ERC models (e.g., DAG-ERC [91]) by providing a more explicit and flexible representation of temporal dependencies in multimodal conversations.

Building upon the MultiDAG structure, the *CL* component introduces a curriculum-based training strategy that addresses the instability commonly observed in training graph-based conversational models. Rather than treating all utterances and dialogues equally, CL progressively organizes training samples from emotionally stable and contextually simple cases to more challenging ones involving rapid emotion shifts or long-range dependencies. This design allows MultiDAG+CL to focus not only on *what* conversational structure to model, but also on *how* such structure should be learned, leading to improved stability and generalization in realistic MERC settings.

2.3.2 MERC with Directed Acyclic Graph

Modality Encoder. We first employ modality-specific encoders to obtain utterance-level multimodal representations as the input to the MultiDAG framework. For the textual modality, a bidirectional LSTM captures sequential contextual information, while fully connected networks are adopted for the acoustic and visual modalities. Formally, given an utterance u_i, the modality-specific encodings are computed as:

$$\mathbf{h}_i^a = \text{Enc}_A(u_i^a), \quad \mathbf{h}_i^v = \text{Enc}_V(u_i^v), \quad \mathbf{h}_i^l = \text{Enc}_L(u_i^l), \tag{2.22}$$

where Enc_A, Enc_V, and Enc_L denote the encoders for acoustic, visual, and textual modalities, respectively. These encoders produce context-aware unimodal representations $\mathbf{h}_i^a$, $\mathbf{h}_i^v$, and $\mathbf{h}_i^l$.

The multimodal representation of utterance u_i, denoted as $u_{i(mm)}$, is obtained by concatenating the available modality features:

$$\mathbf{H}^0_{i(mm)} = \mathbf{h}^a_i \oplus \mathbf{h}^v_i \oplus \mathbf{h}^l_i. \tag{2.23}$$

MultiDAG Construction. In conversational emotion recognition, the emotional state of an utterance is primarily influenced by preceding utterances. We therefore model each conversation using a directed acyclic graph (DAG), where information flows strictly from past utterances to future ones. This structure allows each utterance to aggregate information not only from immediate neighbors but also from more distant historical context while avoiding cyclic dependencies.

Based on the multimodal input representations, we construct the MultiDAG using a Directed Acyclic Graph Gated Neural Network (DAG-GNN) [126], following a similar formulation to DAG-ERC [91], but extending it to multimodal node representations and richer speaker-aware relations. The overall architecture of MultiDAG is illustrated in Figure 2.7.

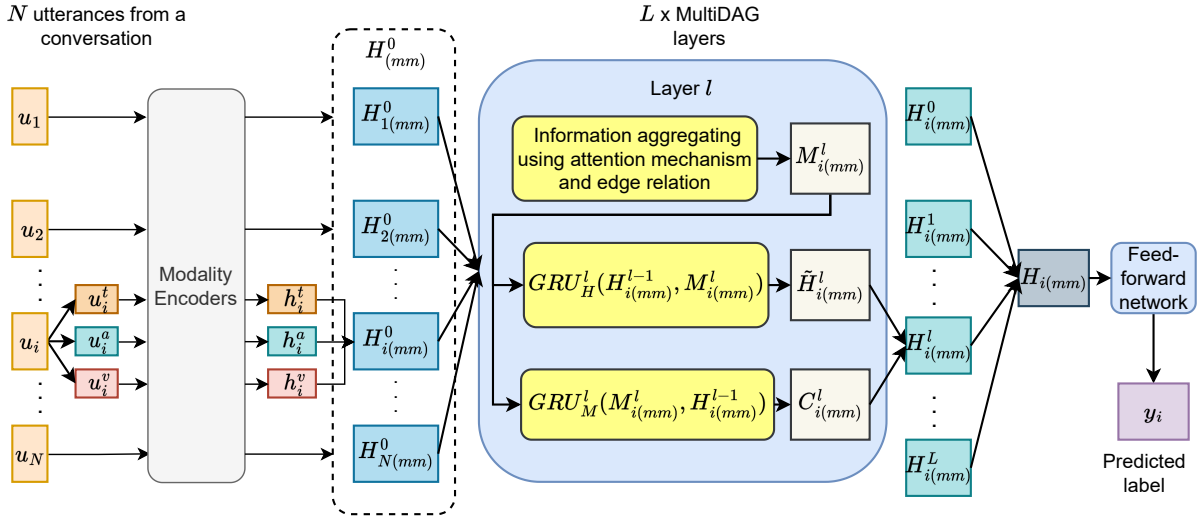


Figure 2.7: Overall structure of the MultiDAG component.

At each layer l, utterance representations are updated sequentially from the first utterance to the last one. For a target utterance $u_{i(mm)}$, attention weights over its preceding nodes are computed as:

$$a^l_{ij(mm)} = \text{softmax}_{j \in \mathcal{N}_{i(mm)}} \left(\mathbf{W}^l_\alpha [\mathbf{H}^{l-1}_{j(mm)} \parallel \mathbf{H}^{l-1}_{i(mm)}] \right), \tag{2.24}$$

where $\mathcal{N}_{i(mm)}$ denotes the set of preceding utterances of $u_{i(mm)}$, $\mathbf{W}^l_\alpha$ is a learnable

weight matrix, and $\|$ denotes vector concatenation. This step assigns different importance to historical utterances depending on their relevance to the current target.

The attention weights are used to aggregate information from neighboring nodes while considering speaker-related edge types:

$$\mathbf{M}^l_{i(mm)} = \sum_{j \in \mathcal{N}_{i(mm)}} a^l_{ij(mm)} \mathbf{W}^l_{r_{ij}} \mathbf{H}^{l-1}_{j(mm)}, \tag{2.25}$$

where $\mathbf{W}^l_{r_{ij}}$ is a relation-specific transformation matrix associated with the edge type r_{ij}, which distinguishes speaker relations between utterances. This design allows MultiDAG to encode both temporal proximity and speaker-dependent influence.

The aggregated message $\mathbf{M}^l_{i(mm)}$ is then integrated with the previous hidden state $\mathbf{H}^{l-1}_{i(mm)}$ through a GRU-based node information unit:

$$\tilde{\mathbf{H}}^l_{i(mm)} = \text{GRU}^l_H(\mathbf{H}^{l-1}_{i(mm)}, \mathbf{M}^l_{i(mm)}). \tag{2.26}$$

To further model the flow of contextual information across layers, a second GRU is used as a context information unit, in which the roles of the input and hidden state are exchanged:

$$\mathbf{C}^l_{i(mm)} = \text{GRU}^l_M(\mathbf{M}^l_{i(mm)}, \mathbf{H}^{l-1}_{i(mm)}). \tag{2.27}$$

Together, these two GRU units jointly update node states by combining node-specific history and aggregated contextual messages.

The final representation of utterance $u_{i(mm)}$ is obtained by concatenating representations from all layers:

$$\mathbf{H}_{i(mm)} = \Big\|_{l=0}^{L} \big(\tilde{\mathbf{H}}^l_{i(mm)} + \mathbf{C}^l_{i(mm)}\big), \tag{2.28}$$

which is subsequently fed into a feed-forward network for emotion classification. The model is trained using the standard cross-entropy loss.

2.3.3 Curriculum Learning

To further enhance the learning of temporal emotion dynamics, we integrate curriculum learning into the MultiDAG framework. Specifically, we design a *Difficulty Measure Function (DMF)* to quantify the difficulty of each conversation and a *training*

Algorithm 2 Curriculum Learning with Difficulty Measure Function (DMF)

Input: training set $\mathcal{D} = \{c_1, c_2, \ldots, c_m\}$ with emotion labels, model M, number of curriculum stages k, training epochs T
Initialize: $S \leftarrow \emptyset$
for each conversation $c \in \mathcal{D}$ **do**
    Group utterances in c by speaker identity
    Compute number of emotion shifts $N_{shift}(c)$
    Compute number of speakers $N_{sp}(c)$ and utterances $N_u(c)$
    Compute difficulty score $\mathcal{DIF}(c)$ using Eq. (16)
end for
Sort conversations by difficulty: $\mathcal{D}' \leftarrow$ sort($\mathcal{D}$, $\mathcal{DIF}$)
Split $\mathcal{D}'$ into k ordered buckets: $\mathcal{D}^1, \ldots, \mathcal{D}^k$
Initialize training set: $\mathcal{D}^{train} \leftarrow \emptyset$
for $t = 1$ to T **do**
    if $t \leq k$ **then**
        $\mathcal{D}^{train} \leftarrow \mathcal{D}^{train} \cup \mathcal{D}^t$
    end if
    Train model M on $\mathcal{D}^{train}$
end for
Return: trained model M^*

scheduler that organizes the training process according to a predefined learning curriculum. This allows the model to progressively learn from simple to complex conversational patterns, thereby improving training stability and generalization.

Difficulty Measure Function (DMF). The key challenge in curriculum learning is defining what constitutes an easier or more difficult training sample. In MERC, conversations with frequent emotional transitions and multiple speakers are generally more challenging to model. Inspired by prior work [124], we define a conversation-level difficulty measure based on the frequency of emotional shifts.

An emotional shift occurs when two consecutive utterances by the same speaker express different emotions. Formally, for two utterances u_i and u_k in a conversation c, an emotional shift is detected if $e(u_i) \neq e(u_k)$, $p(u_i) = p(u_k)$, and there exists no utterance u_j such that $i < j < k$ and $p(u_j) = p(u_i)$, where $e(u)$ denotes the emotion label of utterance u.

The difficulty of conversation c_i is then defined as:

$$\mathcal{DIF}(c_i) = \frac{N_{shift}(c_i) + N_{sp}(c_i)}{N_u(c_i) + N_{sp}(c_i)}, \tag{2.29}$$

where $N_{shift}(c_i)$ denotes the number of emotional shifts, $N_u(c_i)$ is the total number of utterances, and $N_{sp}(c_i)$ is the number of speakers in conversation c_i. The speaker term serves as a smoothing factor to avoid bias toward short conversations.

Training Scheduler. Based on the computed difficulty scores, the training scheduler organizes the dataset into k ordered subsets $\{\mathcal{D}^1, \ldots, \mathcal{D}^k\}$, where conversations with similar difficulty levels are grouped together. Training begins with the easiest subset $\mathcal{D}^1$ and progressively incorporates more difficult subsets as training proceeds. Once all subsets have been introduced, additional epochs are performed on the full training set. This progressive scheduling strategy enables the model to first learn stable conversational patterns and then adapt to more complex emotional dynamics, which empirically leads to more stable optimization and better generalization in MERC.

2.3.4 Implementation

Dataset. We evaluate our approach on the following two ERC datasets: IEMOCAP [9] and MELD [78].

Implementation Details. MultiDAG+CL is implemented using Pytorch [2], and run experiments on Google Colab Pro. 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.

2.3.5 Results

Comparision with Baselines. We conducted a comprehensive comparison of our proposed approach with SOTA multimodal ERC methods, and the results are summarized in Table 2.8. 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

[2]https://pytorch.org/

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. 2.8). Such phenomenon is related to imbalanced class distribution.

Table 2.8: Performance of approaches on IEMOCAP and MELD datasets.

Model	**IEMOCAP**								**MELD**	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
bc-LSTM	33.82	78.76	56.75	64.35	60.25	60.75	60.51	60.42	59.62	57.29
MFN	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	60.80	57.80
ICON	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	58.20	56.30
DialogueRNN	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	60.31	57.66
DialogueGCN	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	58.62	56.36
DAG-ERC	47.59	79.83	69.36	66.67	66.79	**68.66**	67.53	68.03	61.04	63.66
MMGCN	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	60.42	58.31
CTNet	51.3	79.9	65.8	67.2	**78.7**	58.8	68.0	67.5	62.0	60.5
DAG-ERC+HCL	-	-	-	-	-	-	68.73	-	63.89	
COGMEN	-	-	-	-	-	-	68.2	67.6	-	-
CORECT	**59.30**	80.53	66.94	69.59	72.69	68.50	**69.93**	**70.02**	-	-
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**

Bold *denotes the best performance, underlining marks the second-best.*
"-" indicates missing values from the original papers.

As shown in Table 2.8 the gains are relatively modest compared to the performance gap between MultiDAG-based models and CORECT, suggesting that learning-dynamic refinement alone is insufficient to close the gap without stronger multimodal fusion mechanisms. These results further confirm that CORECT provides a more effective representational foundation, while MultiDAG+CL plays a complementary role by improving the learning behavior of temporal structures.

Effect of Modality. Table 2.9 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 combination of visual and acoustic modalities yields the lowest result. Similar observations are made in the MELD dataset.

Effect of Curriculum Learning. The MultiDAG+CL model demonstrates notable performance improvement by incorporating curriculum learning for both the IEMOCAP

Table 2.9: Results of MultiDAG and MultiDAG+CL under different modality settings.

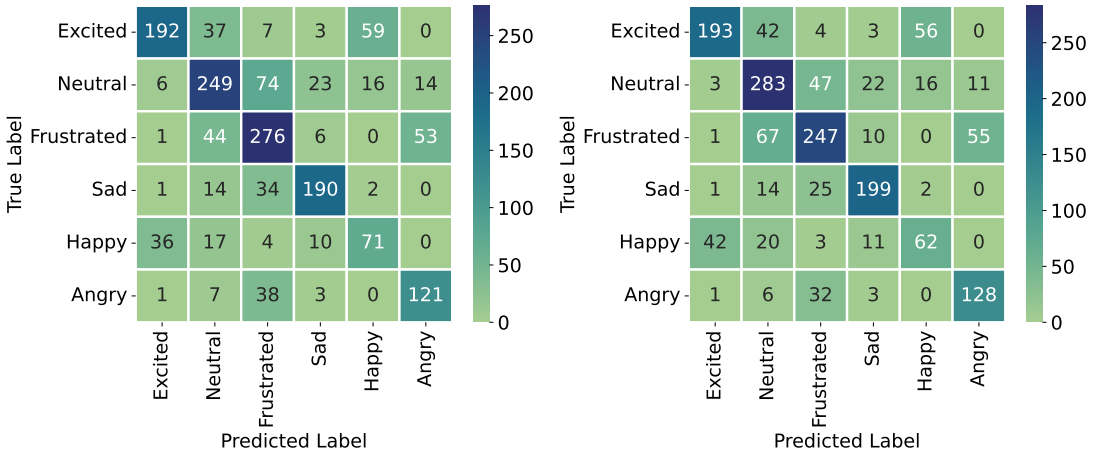
Modality	MultiDAG		MultiDAG+CL	
	IEMOCAP	MELD	IEMOCAP	MELD
T	68.17	63.66	67.12	63.47
A	49.37	40.27	50.58	40.17
V	33.79	31.27	36.69	31.27
T + A	68.42	63.61	68.45	63.56
T + V	67.56	63.69	67.40	63.62
A + V	52.40	40.54	51.86	39.99
T + V + A	**68.45**	**63.87**	69.08	**64.00**

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 2.10, 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.

Table 2.10: Results of MultiDAG+CL for different number of buckets in CL training scheduler.

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

Performance for Emotion-shift. From the confusion matrices of the MultiDAG and MultiDAG+CL models (Figure 2.8), 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 2.8: The confusion matrices on the IEMOCAP.

2.3.6 Discussion

MultiDAG+CL represents a complementary modeling direction for MERC that extends beyond representation-centric fusion design. While CORECT already incorporates relational–temporal graph modeling to encode long-range conversational context and multimodal interactions, MultiDAG+CL shifts the focus toward *how* temporal dependencies should be learned effectively under varying levels of conversational complexity.

Specifically, MultiDAG+CL emphasizes two orthogonal yet critical aspects of conversational emotion recognition: **directed acyclic conversational structure** and **curriculum-based optimization**. The directed acyclic graph formulation enforces a unidirectional information flow from past to future utterances, enabling structured modeling of hierarchical and asymmetric emotional dependencies across dialogue turns. In parallel, curriculum learning explicitly addresses optimization dynamics by progressively organizing training samples from emotionally stable conversations to those exhibiting frequent emotional shifts. This learning strategy allows MultiDAG+CL to alleviate optimization instability and class imbalance that commonly arise in complex conversational settings, thereby complementing CORECT from an optimization-centric perspective rather than a purely representational one.

As shown in Table 2.11, transformer-based multimodal baselines such as COGMEN [40] benefit from highly optimized dense computation but incur substantially higher GPU memory consumption due to activation storage and attention operations. In contrast, graph-based approaches, including CORECT and MultiDAG+CL, operate

Table 2.11: Model complexity and efficiency comparison on the IEMOCAP dataset.

Metric	**COGMEN**	**CORECT**	**MultiDAG+CL**
Parameters	55.93M	17.72M	19.71M
Train time per epoch (s)	8.40	6.93	10.20
Inference time (s)	0.62	0.205	0.41
GPU memory	957.48MB	952.43MB	1350.2MB

on sparse relational structures, resulting in significantly lower memory footprints while maintaining competitive training and inference efficiency. Notably, MultiDAG+CL introduces only a moderate increase in computational cost over CORECT, reflecting the additional overhead of directed acyclic reasoning, while remaining substantially more memory-efficient than the closest transformer-based baseline.

Nevertheless, MultiDAG+CL also has limitations. First, the construction of directed acyclic graphs relies on predefined temporal ordering and speaker interaction assumptions, which may restrict flexibility in highly spontaneous or overlapping conversational scenarios. Second, the effectiveness of curriculum learning depends on the quality of the difficulty estimation, and suboptimal curricula may reduce training efficiency or slow convergence. Third, the current curriculum is defined at the conversation level and does not explicitly adapt to modality-specific noise or label uncertainty, leaving room for more fine-grained or data-driven pacing strategies. These limitations suggest that MultiDAG+CL is most effective when employed as a structured optimization strategy on top of well-defined conversational graphs, rather than as a fully adaptive temporal modeling solution.

2.4 Chapter Summary

In this chapter, we addressed the first major research focus of this dissertation: **Multimodal Emotion Recognition in Conversation**. Specifically, we investigated how multimodal representations and conversational context can be jointly modeled to capture the dynamic and structured nature of emotions in dialogue.

In Section 2.2, we introduced **CORECT**, a relational-temporal graph-based framework that explicitly models both temporal dependencies and cross-modal interactions. Through its Relational Temporal Graph Convolution Network (RT-GCN) and Pairwise Cross-modal Feature Interaction (P-CM) modules, CORECT enables fine-grained context-aware multimodal fusion while preserving modality-specific characteristics. This design effectively addresses the challenge of integrating heterogeneous multimodal cues within conversational context.

In Section 2.3, we presented **MultiDAG+CL**, which approaches MERC from a complementary perspective. By integrating Directed Acyclic Graph (DAG) modeling with Curriculum Learning, MultiDAG+CL explicitly captures directional temporal dependencies and progressively guides the learning process from simpler to more complex conversational patterns. This framework is particularly effective in handling emotional shifts and alleviating data imbalance during training.

Extensive experiments conducted on benchmark datasets such as IEMOCAP, CMU-MOSI, CMU-MOSEI and MELD demonstrate that both CORECT and MultiDAG+CL consistently outperform strong baselines, validating their effectiveness in modeling multimodal conversational emotions under complex dynamics.

Together, CORECT and MultiDAG+CL contribute to **Objective O1** by advancing multimodal fusion and contextual modeling for MERC, while also partially supporting **Objective O2** through improved robustness to challenging learning conditions, such as emotional variability and imbalance. CORECT focuses on representation learning with relational fusion, whereas MultiDAG+CL emphasizes structured temporal modeling and curriculum-driven learning dynamics. Despite their distinct methodological designs, both frameworks aim to enhance multimodal emotion recognition in complex conversational settings. This complementarity naturally points to a promising future direction: integrating structure-aware graph modeling with curriculum-based training, for example by progressively exposing conversational subgraphs according to temporal depth or emotional complexity to enable more adaptive and fine-grained learning.

Chapter 3

Multimodal Emotion Recognition in Conversation under Incomplete Modality Condition

3.1 Introduction

This chapter focuses on multimodal emotion recognition under incomplete modality conditions, a critical challenge in real-world conversational affective analysis. The core methodological contribution presented in this chapter is based on our previously proposed framework, **Mi-CGA**, which was originally introduced as *"Mi-CGA: Cross-modal Graph Attention Network for Robust Emotion Recognition in the Presence of Incomplete Modalities" - Neurocomputing* (SCIE Q1 Journal) in 2025 [**VanNTC 3**]. This chapter directly addresses **Research Question RQ3**, which investigates how robust multimodal representations can be learned under incomplete-modality conditions.

Unlike the full-modality assumption adopted in the previous chapter, incomplete-modality settings reflect practical scenarios in which multimodal signals are only partially available during training or inference. Such conditions frequently arise in conversational data due to sensor failures, occlusions, transmission noise, privacy constraints, or domain shifts, making incomplete multimodal data the norm rather than the exception. Importantly, *missing modality* represents a special case of this broader setting, typically referring to the complete absence of an entire modality, whereas *incomplete modality* provides a more general and fine-grained formulation that also encompasses partially observed or corrupted features, heterogeneous modality availability across di-

alogue turns, and varying degrees of information loss within a modality. Figure 3.1 illustrates representative examples of uncertain and incomplete modality observations in multimodal emotion recognition tasks.

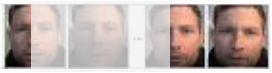
Modality	Demonstration	Possible Reasons
Text	…They act like they are too cool to talk to me…	• Unfamiliar terms • Automated speech recognition fault
Audio		• Background noise • Sensor failure
Video		• Face undetected • Fast motions

Figure 3.1: Illustration of uncertain missing modalities in Multimodal Emotion Recognition Task.

Despite increasing attention to multimodal learning, many existing approaches struggle to generalize under incomplete-modality conditions. A major limitation is that most models assume only a single modality is missing at a time, ignoring more realistic scenarios in which multiple modalities may be partially or entirely unavailable simultaneously. In addition, incomplete samples are often discarded or handled using shallow imputation strategies, which underutilize the remaining modalities and fail to preserve meaningful cross-modal interactions [98, 142].

Moreover, several prior work relies on rigid fusion schemes or processes each modality in isolation. Such designs limit feature-level cross-modal reasoning, particularly when modality observations are incomplete or unreliable. As a result, these models are ill-suited to capture the complex interdependencies among modalities and conversational context that are essential for robust emotion recognition.

As summarized in Table 3.1, existing methods for handling incomplete or missing modalities can be broadly categorized into three groups. The first group focuses on **data augmentation**, where artificial modality dropout is introduced during training to improve robustness to missing inputs. Representative examples include random modality ablation strategies [70] and iterative dropout mechanisms such as PANet and M2R2 [108]. While effective in improving robustness to a limited extent, these approaches do not explicitly model missing information nor adapt to varying incomplete-modality patterns at inference time.

The second category consists of **generative methods** that aim to reconstruct miss-

Table 3.1: A chronological summary of related works on missing modalities.

Model	Approach	Main Technique	Studied Problem	Modality Missing	Datasets	Advantages	Disadvantages
CRA [102]	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 [135]	Generative	GANs	Multi-view Learning	Arbitrary view-missing	Handwritten, 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.
MCTN [73]	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 [98]	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 [10]	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 [127]	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 [142]	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 [58]	Generative	Bayesisan, 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 [134]	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 [108]	Data Augumentation	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 [28]	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 [56]	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 [49]	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 [112]	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 [113]	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.

ing modalities based on available ones. Early approaches such as CRA [102] and CPM-Net [135] exploit cross-modal dependencies or adversarial learning for reconstruction. More recent methods, including DiCMoR [112] and IMDer [113], employ probabilistic modeling and diffusion-based sampling to recover missing features. However, these approaches typically rely on strong assumptions about modality alignment and distributional similarity, which are difficult to satisfy in conversational settings characterized by temporal dynamics and heterogeneous modality quality.

The third line of work explores **joint representation learning**, where models aim to learn shared latent spaces that are resilient to incomplete modality observations. Representative examples include MMIN [142], which integrates reconstruction and representation learning objectives, TFR-Nets [127], which employ Transformer-based cross-modal reconstruction, and GCNet [49], which introduces graph-based reasoning for conversational data with incomplete modalities. Despite their promise, many of these models treat missingness as a secondary issue rather than explicitly incorporating incomplete-modality awareness into the core model architecture.

Despite these advances, several challenges remain unresolved. First, most existing methods do not jointly consider local modality reconstruction and global conversational context modeling under incomplete-modality conditions. Second, cross-modal reasoning at the feature level is often shallow or heuristically designed. Third, the potential of graph neural networks for modeling multimodal conversational structures in the presence of incomplete modalities has not been fully explored. To address these challenges, we introduce **Mi-CGA**, a graph-based framework that jointly integrates local modality reconstruction and global conversational context modeling for robust multimodal emotion recognition under incomplete modality conditions.

In this chapter, we formalize the task of multimodal emotion recognition in conversation under incomplete-modality conditions as follows.

Input. A labeled multimodal dataset $\mathcal{D} = \{(C^{(k)}, \mathbf{y}^{(k)})\}_{k=1}^{K}$, where each conversation $C^{(k)} = [u_1^{(k)}, \ldots, u_{N_k}^{(k)}]$ is a sequence of utterances. Each utterance $u_i^{(k)}$ is associated with speaker identities and full-modality representations $\{(u_i^{(k)})^a, (u_i^{(k)})^v, (u_i^{(k)})^t\}$, together with emotion labels $\mathbf{y}^{(k)} = [y_1^{(k)}, \ldots, y_{N_k}^{(k)}]$. During training, incomplete-modality conditions are simulated by stochastically masking one or more modalities according to predefined missing patterns.

Output. A learned model that performs robust multimodal fusion under arbitrary missing-modality patterns.

3.2 Mi-CGA: Cross-Modal Graph Attention Network for Robust Emotion Recognition in the Presence of Incomplete Modalities

3.2.1 Overview

Figure 3.2 illustrates the overall architecture of the proposed Mi-CGA framework for multimodal emotion recognition in the presence of incomplete modalities. The model is organized into two main stages that operate sequentially from raw multimodal inputs to utterance-level emotion predictions.

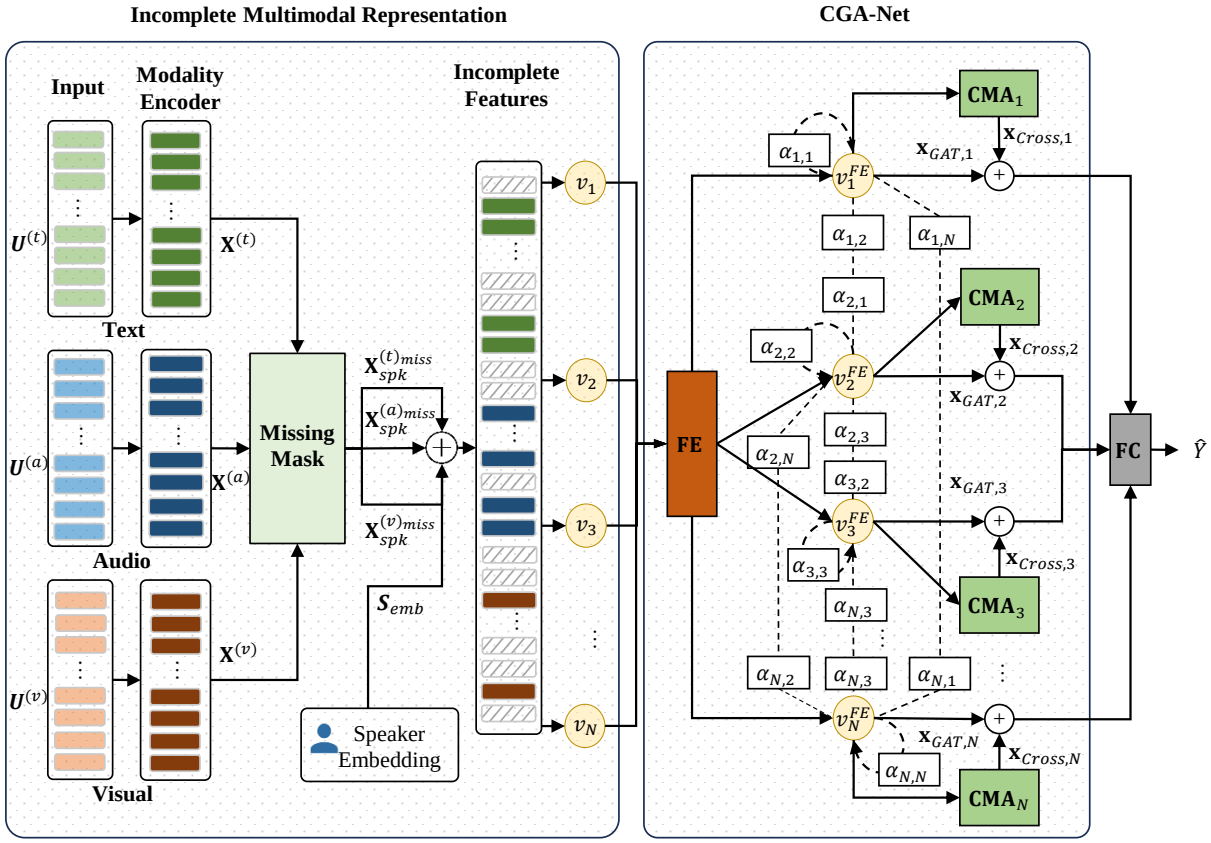


Figure 3.2: Overall Architecture of Mi-CGA model.

The first stage, termed **Incomplete Multimodal Representation** (IMR), aims to handle diverse missing scenarios by encoding incomplete multimodal signals into a unified, speaker-aware representation. Given raw textual, acoustic, and visual features, IMR first employs modality-specific encoders to derive utterance-level embeddings for each modality, then applies a stochastic missing-mask generation procedure to simulate or reflect incomplete-modality patterns, and finally augments the resulting representations

with speaker embeddings. This process yields an incomplete yet structured representation $X_s^{miss}pk$ that captures both modality availability and speaker-dependent conversational context, and serves as the input to the subsequent graph-based reasoning stage.

The second stage, referred to as the **Cross-modal Graph Attention Network** (**CGA-Net**), performs context-aware multimodal reasoning over $X_s^{miss}pk$ through three integrated components. First, the *Modality Feature Estimation* (FE) module leverages a graph convolutional encoder–decoder architecture to reconstruct missing or corrupted modality features from neighboring utterances in the dialogue graph, and refines them via a smoothing mechanism that balances reconstructed and original information. Second, a *Multi-head Graph Attention Network* (MulGAT) models utterance-level and conversation-level dependencies by attending over the fully connected dialogue graph, enabling topology-aware aggregation of contextual cues across time and speakers. Third, a *Cross-modal Attention Network* (CMA) performs directional cross-modal attention between modality-specific sequences to dynamically propagate complementary information across modalities under incomplete-modality conditions.

By jointly modeling feature reconstruction, graph-based contextual dependencies, and cross-modal interactions, Mi-CGA produces a final utterance representation that is both robust to incomplete modalities and sensitive to conversational structure. This representation is then fed into an utterance-level classifier for emotion prediction. In the following sections, each module of IMR and CGA-Net is described in detail.

3.2.2 Incomplete Multimodal Representation (IMR)

Unimodal Encoder: The Unimodal Encoder generates utterance-level embeddings tailored to specific modalities from raw modality features. For text, we employ a bi-directional Long Short-Term Memory (BiLSTM) network to capture sequential contextual information. For visual and acoustic modalities, following [35], we use a Fully Connected Network to independently encode contextual features for each modality.

The multimodal context-aware feature encoding for each utterance is formulated as follows:

$$\mathbf{x}_i^{(a)} = W_e^a \, u_i^{(a)} + \mathbf{b}^{(a)} \tag{3.1}$$

$$\mathbf{x}_i^{(v)} = W_e^v \, u_i^{(v)} + \mathbf{b}^{(v)} \tag{3.2}$$

$$\mathbf{x}_i^{(t)} = \left[\overrightarrow{\text{LSTM}}(u_i^{(t)}, \mathbf{x}_{i-1}^{(t)}), \overleftarrow{\text{LSTM}}(u_i^{(t)}, \mathbf{x}_{i+1}^{(t)})\right] \tag{3.3}$$

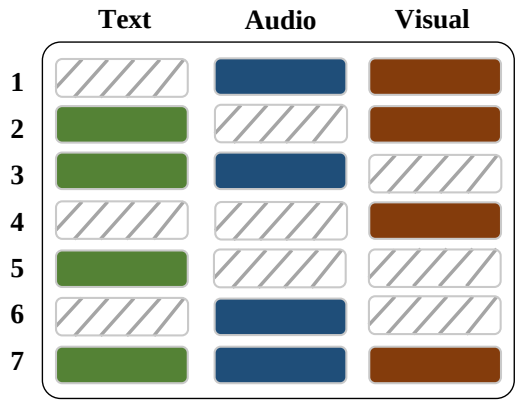


Figure 3.3: Seven missing patterns for $M = 3$. Each row illustrates a missing pattern, in which a rectangle with diagonal lines implies the missing modality.

where $\mathbf{x}_i^{(a)} \in \mathbb{R}^{d_a}$, $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$, and $\mathbf{x}_i^{(t)} \in \mathbb{R}^{d_t}$ are context-aware representations for audio, visual, and text modalities, respectively; d_a, d_v, d_t denote the corresponding latent dimensions; and W_e^a, W_e^v are learnable parameters.

For a given conversation C, three modality-specific sequences are constructed as:

$$\mathbf{X}^{(a)} = [\mathbf{x}_1^{(a)}, \mathbf{x}_2^{(a)}, \ldots, \mathbf{x}_N^{(a)}] \tag{3.4}$$

$$\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_N^{(v)}] \tag{3.5}$$

$$\mathbf{X}^{(t)} = [\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}, \ldots, \mathbf{x}_N^{(t)}] \tag{3.6}$$

Missing Mask Generation: Similar to GCNet [49], we simulate the presence of incomplete modalities across the entire conversation with an overall missing rate $\rho \in [0, 1]$. To ensure that each training sample retains at least one available modality, the missing rate is bounded within $[0, \frac{M-1}{M}]$, where M denotes the number of modalities. Figure 3.3 illustrates the possible missing patterns for a trimodal setting ($M = 3$).

Given a predefined missing rate ρ, a binary mask matrix $\mathbf{\Omega}$ is generated for the entire conversation C to indicate the availability of multimodal features. Specifically, for each utterance with M modalities, the modalities to be masked are randomly sampled according to ρ. The same missing rate is consistently applied across the training, validation, and testing phases to ensure fair evaluation.

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 under incomplete-modality conditions, where $d = d_a + d_t + d_v$. The

incomplete representation is obtained as:

$$\mathbf{X}^{miss} = \mathbf{X} \odot \mathbf{\Omega}, \tag{3.7}$$

where $\mathbf{X} = [\mathbf{X}^{(a)}, \mathbf{X}^{(t)}, \mathbf{X}^{(v)}] \in \mathbb{R}^{N \times d}$ denotes the complete multimodal feature matrix of the input conversation, $\mathbf{\Omega} \in \{0, 1\}^{N \times d}$ is the binary availability mask, and $\odot$ represents element-wise multiplication. An entry $\mathbf{\Omega}_{ij} = 0$ indicates that the corresponding feature dimension is unobserved due to incomplete modality, while $\mathbf{\Omega}_{ij} = 1$ indicates an available observation.

Enhancing Conversation-level Representation with Speaker Embedding: Recent studies [35, 42] validate the significance of speaker information in improving utterance representations. Inspired by this observation, we employ a speaker embedding procedure, denoted as **S-Emb**, to generate latent representations based on speaker identities. Given a conversation C and its corresponding speaker set S, the speaker embedding matrix $\mathbf{S}_{emb} \in \mathbb{R}^{N \times |S|}$ is obtained as:

$$\mathbf{S}_{emb} = \textbf{S-Emb}(S) \tag{3.8}$$

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

$$\mathbf{X}^{miss}_{spk} = \eta\, \mathbf{S}_{emb} \oplus \mathbf{X}^{miss}, \tag{3.9}$$

where $\eta \in [0, 1]$ is a scalar coefficient controlling the contribution of speaker information, and $\oplus$ denotes the concatenation operation.

3.2.3 Cross-modal Graph Attention Network (CGA-Net)

We construct a dialogue graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent multimodal conversational data, where each utterance is treated as a node $v_i \in \mathcal{V}$. Each node is associated with multimodal features $\mathbf{x}_i = [\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(v)}, \mathbf{x}_i^{(t)}]$. Edges are constructed between all utterances within the same conversation, resulting in a fully connected graph represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Building upon the speaker-aware incomplete representation $\mathbf{X}^{\text{miss}}_{\text{spk}}$ produced by IMR, CGA-Net performs graph-based and cross-modal reasoning through three key modules: (1) the *Modality Feature Estimation* module, (2) the *Multi-*

head Graph Attention Network, and (3) the *Cross-modal Attention Network*. In what follows, we detail each of these components.

Modality Feature Estimation (FE). Interconnected nodes corresponding to neighboring utterances and sharing mutual multimodal information often exhibit underlying similarities, which can be exploited to reconstruct missing features. Motivated by this observation, we propose a multimodal Feature Estimation (FE) module consisting of a GNN-based encoder f_ϕ and an MLP-based decoder g_θ, as illustrated in Figure 3.4. The encoder f_ϕ generates node embeddings by leveraging the conversational graph structure, while the decoder g_θ aims to approximate missing modality features. In this way, FE provides a principled alternative to naive imputation strategies such as zero or mean filling.

Formally, the coarse reconstructed representation $\mathbf{X}^{\text{coarse}} \in \mathbb{R}^{N \times \bar{d}}$ of a conversation C is computed as:

$$\mathbf{X}^{\text{coarse}} = g_\theta\Big(f_\phi(\mathbf{X}^{\text{miss}}_{\text{spk}}, \tilde{\mathbf{A}})\Big), \tag{3.10}$$

where ϕ and θ denote the learnable parameters of the encoder and decoder, respectively.

Specifically, we instantiate the encoder f_ϕ using a Graph Convolutional Network (GCN)[1] [43], where $\tilde{\mathbf{A}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}$ denotes the symmetrically normalized adjacency matrix, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, and $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix. For the decoder g_θ, we adopt a linear transformation defined as

$$g_\theta(\mathbf{z}) = \mathbf{W}_\theta \mathbf{z} + \mathbf{b}_\theta, \tag{3.11}$$

where $\mathbf{W}_\theta$ and $\mathbf{b}_\theta$ are learnable parameters. Through Equation (3.10), the missing positions indicated by the corresponding mask are filled with the reconstructed values at the same locations.

Although the imputed features reconstructed from neighboring utterances may contain useful information, the reconstructed values and the originally observed values in $\mathbf{X}^{\text{coarse}}$ can differ significantly, which may negatively affect the final classification performance. To mitigate this issue, we apply a smoothing step by introducing a normalization layer that blends the reconstructed features with the original incomplete rep-

[1] The proposed framework is flexible and can be extended to other graph neural network variants.

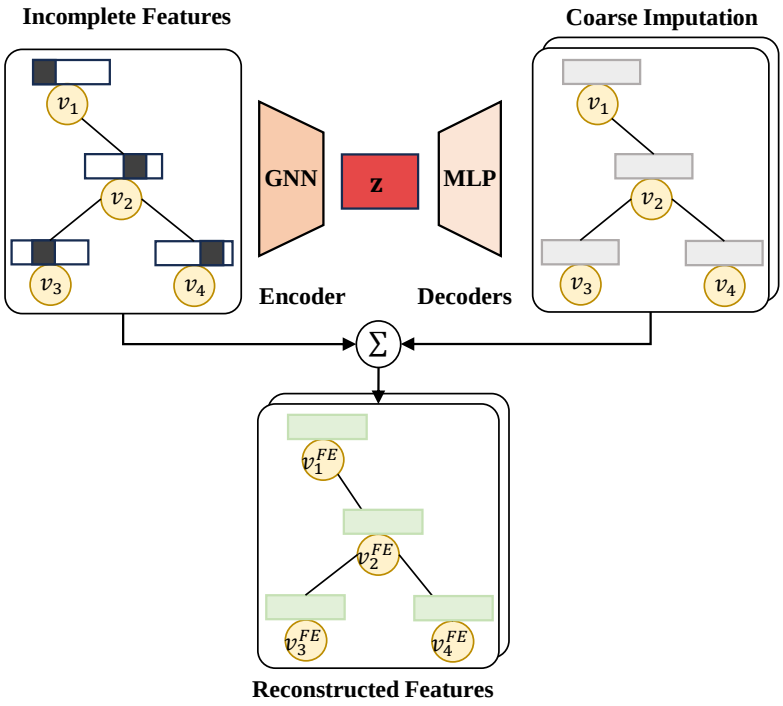


Figure 3.4: Multimodal Feature Estimation Module (FE)

resentations, yielding the final imputed conversation-level representation $\mathbf{X}^{\text{FE}} \in \mathbb{R}^{N \times \bar{d}}$:

$$\mathbf{X}^{\text{FE}} = \text{Norm}(\mathbf{X}^{\text{coarse}}, \mathbf{X}^{\text{miss}}_{\text{spk}}) \tag{3.12}$$

$$= (1 - \lambda)\, \mathbf{X}^{\text{coarse}} + \lambda\, \mathbf{X}^{\text{miss}}_{\text{spk}}, \tag{3.13}$$

where $\lambda \in [0, 1]$ is a hyperparameter controlling the contribution of the original incomplete representation during feature refinement. This smoothing mechanism allows FE to strike a balance between leveraging information from neighboring nodes and preserving useful cues from the original node.

To further encourage accurate reconstruction, we incorporate a reconstruction objective for the FE module into the overall training loss, which is detailed in Section 3.2.5.

Multi-head Graph Attention Network (MulGAT). Graph neural networks (GNNs) are a class of deep learning methods that leverage topological information on graphs to enhance representation learning over graph entities such as nodes, edges, and subgraphs for downstream tasks, e.g., node classification [147]. GNNs typically consist of two fundamental operations: information aggregation from neighboring nodes and state updating. Among various GNN architectures, the Graph Attention Network (GAT) [106] introduces an attention mechanism to adaptively weight neighborhood information dur-

ing aggregation.

Inspired by GAT, we introduce a single-head graph attention sublayer, denoted as **S-GAT**, to refine utterance-level representations within a conversation. Formally, given the feature-estimated representation $\mathbf{X}^{\text{FE}}$, the output of **S-GAT** is computed as:

$$\hat{\mathbf{X}}^{\text{FE}} = \textbf{S-GAT}(\mathbf{X}^{\text{FE}}, \Theta), \tag{3.14}$$

where $\hat{\mathbf{X}}^{\text{FE}} \in \mathbb{R}^{N \times \bar{d}}$ denotes the enhanced representation and Θ represents the learnable parameters of the **S-GAT** module.

For each node $v_i \in \mathcal{V}$, let $\mathbf{x}_i \in \mathbb{R}^{\bar{d}}$ denote its input feature vector, and let $\mathcal{N}(i) = \{v_j \in \mathcal{V} \mid (v_j, v_i) \in \mathcal{E}\}$ denote the set of its neighboring nodes. The **S-GAT** layer operates through the following two steps.

Aggregation. The aggregated representation $\mathbf{x}_{agg,i} \in \mathbb{R}^{\bar{d}}$ of node v_i is obtained by attending over its neighbors:

$$\mathbf{x}_{agg,i} = \sum_{v_j \in \mathcal{N}(i)} \alpha_{ij}\, \mathbf{W}_{agg} \mathbf{x}_j, \tag{3.15}$$

where $\mathbf{W}_{agg} \in \mathbb{R}^{\bar{d} \times \bar{d}}$ is a learnable weight matrix and α_{ij} denotes the attention coefficient measuring the importance of node v_j to node v_i. Following GATv2 [8], the attention coefficients are computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{v_j \in \mathcal{N}(i)} \exp(e_{ij})}, \tag{3.16}$$

where the unnormalized attention score e_{ij} is defined as:

$$e_{ij} = \boldsymbol{\phi}_{att}^{\top} \sigma(\boldsymbol{\Theta}_{att}[\mathbf{x}_i \oplus \mathbf{x}_j])\,, \tag{3.17}$$

with $\boldsymbol{\Theta}_{att} \in \mathbb{R}^{d_{att} \times 2\bar{d}}$ and $\boldsymbol{\phi}_{att} \in \mathbb{R}^{d_{att}}$ being learnable parameters, $\sigma(\cdot)$ a nonlinear activation function (e.g., LeakyReLU), and $\oplus$ denoting concatenation.

State updating. The enhanced representation $\mathbf{x}_{\text{com},i} \in \mathbb{R}^{\bar{d}}$ of $\mathbf{x}_i$ is computed by aggregating the transformed neighbor representations:

$$\mathbf{x}_{\text{com},i} = \sigma\left(\sum_{v_j \in \mathcal{N}(i)} \alpha_{ij}\, \mathbf{W}_{\text{com}} \mathbf{x}_{\text{agg},j} \right), \tag{3.18}$$

where α_{ij} is the attention coefficient defined in Equation (3.16), $\mathbf{W}_{\text{com}}$ is a learnable weight matrix, and σ is a non-linear activation function. As a result, $\hat{\mathbf{X}}^{\text{FE}}$ is built by concatenating all $\mathbf{x}_{\text{com},i}$, $i \in [1, N]$, in the order of utterances within the conversation C.

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

$$\mathbf{X}_{\text{GAT}} = \textbf{MulGAT}(\mathbf{X}^{\text{FE}}, \Psi) = \Psi\left[\hat{\mathbf{X}}_1^{\text{FE}}, \ldots, \hat{\mathbf{X}}_H^{\text{FE}}\right], \tag{3.19}$$

where H is the number of attention heads, $\hat{\mathbf{X}}_i^{\text{FE}}$ is the output from the i-th **S-GAT**, $[\cdot]$ denotes concatenation, and Ψ is a trainable projection. In this way, MulGAT refines the FE outputs by aggregating long-range conversational context over the dialogue graph.

Cross-modal Attention Network (CMA). Inspired by the decoder architecture of the Transformer in neural machine translation [105], enabling latent adaptation across modalities provides an effective mechanism for cross-modal information fusion. CMA follows the general directional cross-modal attention formulation (Section 1.3.1.2) and is used here to propagate complementary information under incomplete-modality settings.

Given two modalities δ and γ, we extract the corresponding modality-specific feature sequences $\mathbf{X}^{\delta} \in \mathbb{R}^{N \times d_{\delta}}$ and $\mathbf{X}^{\gamma} \in \mathbb{R}^{N \times d_{\gamma}}$. To capture interactions between each modality pair, we employ a cross-modal attention mechanism (CMA), as illustrated in Figure 3.5, to model the directional information flow from modality δ to modality γ, denoted as "$\delta \rightarrow \gamma$".

Specifically, the cross-modal attention output $\mathbf{H}^{\gamma}$ from modality δ to modality γ is computed as:

$$\mathbf{H}^{\gamma} = \textbf{CMA}_{\delta \rightarrow \gamma}(\mathbf{X}^{\delta}, \mathbf{X}^{\gamma}) \tag{3.20}$$

$$= \text{Softmax}\left(\frac{Q^{\gamma}(K^{\delta})^{\top}}{\sqrt{d_k}}\right) V^{\delta}, \tag{3.21}$$

where scaled dot-product attention [105] is adopted with Queries $Q^{\gamma} = \mathbf{X}^{\gamma}\mathbf{W}_{Q^{\gamma}}$, Keys $K^{\delta} = \mathbf{X}^{\delta}\mathbf{W}_{K^{\delta}}$, and Values $V^{\delta} = \mathbf{X}^{\delta}\mathbf{W}_{V^{\delta}}$. Here, $\mathbf{W}_{Q^{\gamma}} \in \mathbb{R}^{d_{\gamma} \times d_Q}$, $\mathbf{W}_{K^{\delta}} \in \mathbb{R}^{d_{\delta} \times d_K}$, and $\mathbf{W}_{V^{\delta}} \in \mathbb{R}^{d_{\delta} \times d_V}$ are learnable projection matrices, where d_Q, d_K, and d_V denote the dimensions of the query, key, and value spaces, respectively. The scaling factor $\sqrt{d_k}$ is

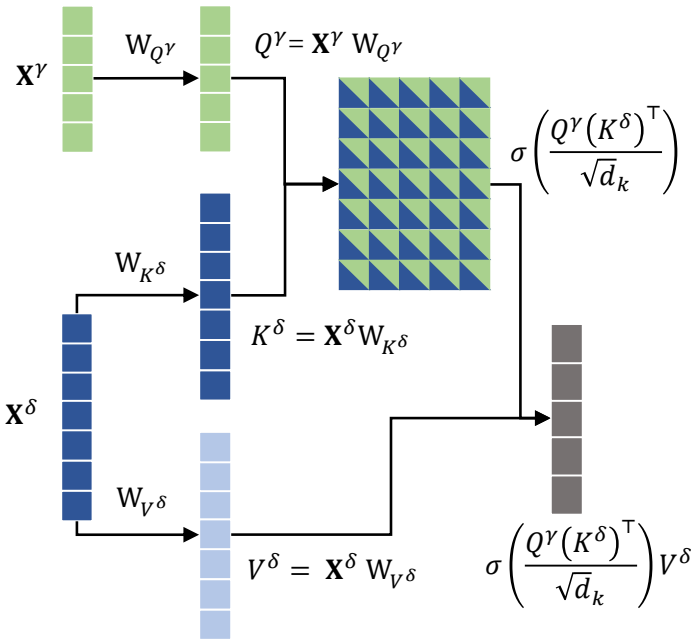


Figure 3.5: Crossmodal attention between sequence $\mathbf{X}^{\delta}$ and $\mathbf{X}^{\gamma}$

used to stabilize training, and the resulting representation satisfies $\mathbf{H}^{\gamma} \in \mathbb{R}^{N \times d_V}$.

Likewise, the reverse cross-modal attention from modality γ to modality δ is computed as:

$$\mathbf{H}^{\delta} = \mathbf{CMA}_{\gamma \to \delta}(\mathbf{X}^{\gamma}, \mathbf{X}^{\delta}) \in \mathbb{R}^{N \times d_V}. \tag{3.22}$$

These bidirectional cross-modal attention representations are then concatenated to form the cross-modal interaction representation between modalities δ and γ:

$$\mathbf{X}^{\delta \rightleftarrows \gamma} = [\mathbf{H}^{\gamma}, \mathbf{H}^{\delta}] \in \mathbb{R}^{N \times 2d_V}. \tag{3.23}$$

Revisiting the incomplete-modality setting, we compute cross-modal attention representations for all modality pairs in a conversation C as:

$$\begin{aligned}
\mathbf{X}_{\text{Cross}}^{a \rightleftarrows t} &= \mathbf{CMA}_{a \rightleftarrows t}(\mathbf{X}^{(a)_{\text{FE}}}, \mathbf{X}^{(t)_{\text{FE}}}), \\
\mathbf{X}_{\text{Cross}}^{t \rightleftarrows v} &= \mathbf{CMA}_{t \rightleftarrows v}(\mathbf{X}^{(t)_{\text{FE}}}, \mathbf{X}^{(v)_{\text{FE}}}), \\
\mathbf{X}_{\text{Cross}}^{v \rightleftarrows a} &= \mathbf{CMA}_{v \rightleftarrows a}(\mathbf{X}^{(v)_{\text{FE}}}, \mathbf{X}^{(a)_{\text{FE}}}),
\end{aligned} \tag{3.24}$$

where $\mathbf{X}^{(a)_{\text{FE}}} \in \mathbb{R}^{N \times d_a}$, $\mathbf{X}^{(t)_{\text{FE}}} \in \mathbb{R}^{N \times d_t}$, and $\mathbf{X}^{(v)_{\text{FE}}} \in \mathbb{R}^{N \times d_v}$ denote the modality-specific representations inferred from the FE module described in Section **??**. The aggregated cross-modal attention representation of conversation C is finally constructed

as:

$$\mathbf{X}_{\text{Cross}} = [\mathbf{X}_{\text{Cross}}^{a \rightleftarrows t}, \mathbf{X}_{\text{Cross}}^{t \rightleftarrows v}, \mathbf{X}_{\text{Cross}}^{v \rightleftarrows a}], \tag{3.25}$$

where $[\cdot]$ denotes the concatenation operation. Together with $\mathbf{X}_{\text{GAT}}$, this cross-modal representation will be used to form the final utterance-level features for emotion classification.

3.2.4 Emotion Classification

To jointly exploit topology-aware contextual representations and cross-modal interaction cues, we concatenate the outputs of the graph attention and cross-modal attention modules to form the final representation:

$$\mathbf{X}_{\text{Final}} = [\mathbf{X}_{GAT}, \mathbf{X}_{Cross}], \tag{3.26}$$

where $\mathbf{X}_{GAT}$ and $\mathbf{X}_{Cross}$ denote the enhanced conversation-level representations produced by the multi-head graph attention module (Equation 3.19) and the cross-modal attention module (Equation 3.25), respectively.

For each utterance $u_i \in C$, the corresponding feature vector $\hat{\mathbf{x}}_i \in \mathbf{X}_{\text{Final}}$ is fed into a feed-forward classification network to predict its emotion label:

$$\begin{aligned} \mathbf{l}_i &= \text{ReLU}(\mathbf{W}_l \hat{\mathbf{x}}_i + \mathbf{b}_l), \\ \mathbf{s}_i &= \text{Softmax}(\mathbf{W}_s \mathbf{l}_i + \mathbf{b}_s), \\ \hat{y}_i &= \underset{k}{\text{argmax}}\, \mathbf{s}_i[k], \end{aligned} \tag{3.27}$$

where $\mathbf{W}_l \in \mathbb{R}^{\bar{d} \times \bar{d}}$ and $\mathbf{b}_l \in \mathbb{R}^{\bar{d}}$ are the parameters of the hidden layer, and $\mathbf{W}_s \in \mathbb{R}^{|E| \times \bar{d}}$ and $\mathbf{b}_s \in \mathbb{R}^{|E|}$ are the parameters of the output layer.

3.2.5 Model Training

To jointly optimize emotion classification and feature reconstruction, we adopt a dual-loss training objective that combines a classification loss and a reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rct}. \tag{3.28}$$

Here, $\mathcal{L}_{cls}$ drives the model to correctly predict utterance-level emotion labels, while $\mathcal{L}_{rct}$ regularizes the FE module to produce meaningful reconstructions of incomplete features.

The classification loss $\mathcal{L}_{cls}$ is defined as the cross-entropy loss between the predicted emotion distribution and the ground-truth label. Let $\mathbf{s}_i \in \mathbb{R}^{|E|}$ denote the softmax output for utterance u_i, and let $\mathbf{y}_i \in \{0,1\}^{|E|}$ be its one-hot ground-truth label. The classification loss is computed as:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i^{\top} \log \mathbf{s}_i. \tag{3.29}$$

The reconstruction loss $\mathcal{L}_{rct}$ is introduced to encourage accurate feature estimation in the FE module. We investigate two alternative reconstruction objectives: Mean Squared Error (MSE) and Kullback–Leibler (KL) divergence.

The MSE-based reconstruction loss is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{|\mathbf{X}|}\left\|\mathbf{X} - \mathbf{X}^{\text{FE}}\right\|_F^2, \tag{3.30}$$

where $\mathbf{X}$ and $\mathbf{X}^{\text{FE}}$ denote the original and reconstructed conversation-level feature matrices, respectively, $|\mathbf{X}|$ denotes the total number of feature elements, and $\|\cdot\|_F$ is the Frobenius norm.

Alternatively, we adopt a KL-divergence-based reconstruction loss inspired by the sparse autoencoder framework [64]. Instead of reconstructing the entire feature values, this objective encourages the reconstructed features to match the activation distribution of the original representations. Let $\hat{\mathbf{p}} = \text{Sigmoid}(\mathbf{X}^{\text{FE}})$ denote the estimated activation probabilities, and let p be a predefined sparsity parameter. The KL-divergence loss is defined as:

$$\mathcal{L}_{KL} = \sum_{j}\left[p\log\frac{p}{\hat{p}_j} + (1-p)\log\frac{1-p}{1-\hat{p}_j}\right], \tag{3.31}$$

where $\hat{p}_j$ denotes the average activation of the j-th feature dimension. By minimizing $\mathcal{L}_{KL}$, the model is encouraged to recover meaningful activation patterns while avoiding overfitting to exact feature values.

In our experiments, we evaluate three training variants: (i) $\mathcal{L}_1 = \mathcal{L}_{cls} + \mathcal{L}_{MSE}$, (ii) $\mathcal{L}_2 = \mathcal{L}_{cls} + \mathcal{L}_{KL}$, and (iii) $\mathcal{L}_3 = \mathcal{L}_{cls}$, which excludes the reconstruction loss. As reported in Figure 3.7, $\mathcal{L}_1$ generally yields the best trade-off between recognition

accuracy and robustness under different missing-rate settings; therefore, we adopt $\mathcal{L}_1$ as the default training objective for Mi-CGA unless otherwise stated.

3.3 Experiments and Results

3.3.1 Implementation

Dataset. Mi-CGA is evaluated on IEMOCAP, CMU-MOSI and CMU-MOSEI datasets.

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 [135], AE [4], CRA [102], MMIN [142], GCNet [49], DiCMoR [112], and IMDer [113].

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.

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 8GB of memory. For the structure of Mi-CGA, we stack 2 layers of GAT along with 4-head attentions. The coefficient λ and p is set to default as 0.5 and 0.2 respectively.

3.3.2 Results

Comparison with baseline models. Table 3.2 shows the comparision 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 propose model, Mi-CGA, consistently outperforms all other competing approaches across multiple datasets. Specifically, on the IEMOCAP (4-way) dataset, our method exhibits a substantial performance gain of 6.30% over the leading model, GCNet on average w-F1 score. Moreover,

on the IEMOCAP (6-way), Mi-CGA establishes a new SOTA record with an impressive accuracy of 62.43%, signifying a noteworthy improvement of 6.25% compared to the prior best-performing model (GCNet). We consistently observe similar performance improvements on both the CMU-MOSI and CMU-MOSEI datasets, further validating the robustness and effectiveness of our approach in the realm of incomplete multimodal learning for Multimodal ERC.

Table 3.2: Comparision with existing works for various missing rates.

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

Bold *marks the best result, underline the second-best.*
Δ *shows Mi-CGA's improvements over the second-best model.*

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 3.2 clearly demonstrate that Mi-CGA exhibits significantly smaller performance degradation compared to the baseline models as the missing rate increases. In IEMOCAP (4-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.

Importance of the Modalities. 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 3.2.2 (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 3.3 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 3.3 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

Table 3.3: Results of modality ablation experiments on IEMOCAP dataset.

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**

"A", "V", and "T" denote audio, visual, and textual modalities, respectively.
***Bold** indicates the best result, underline the second-best.*
"-" means no results were reported at that missing rate.

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.

Effects of MulGAT and CMA in CGA-Net. 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) **CGA$_{\text{-G}}$**, which omits the Multi-head Graph Attention Network from CGA-Net; (2) **CGA$_{\text{-C}}$**, which excludes the Cross-modal Attention Network from CGA-Net.

The module ablation experiment results, as presented in Table 3.4, offer insights into the performance of the different model variations. For the IEMOCAP (4-way)

Table 3.4: Effectiveness of MulGAT and CMA.

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

CGA$_{\text{-G}}$ *and* ***CGA$_{\text{-C}}$*** *denote CGA-Net without the MulGAT and CMA modules, respectively. The best results are shown in* ***bold****.*

dataset, we observe that **CGA$_{\text{-G}}$** experiences an average reduction of 0.91% across all missing rates, while **CGA$_{\text{-C}}$** 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, **CGA$_{\text{-G}}$** 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, **CGA$_{\text{-G}}$** 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.

Advantage of the Multimodal FE Module. 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 [83]. 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**). Figure 3.6 shows the performance comparison between our FE and these imputation strategies.

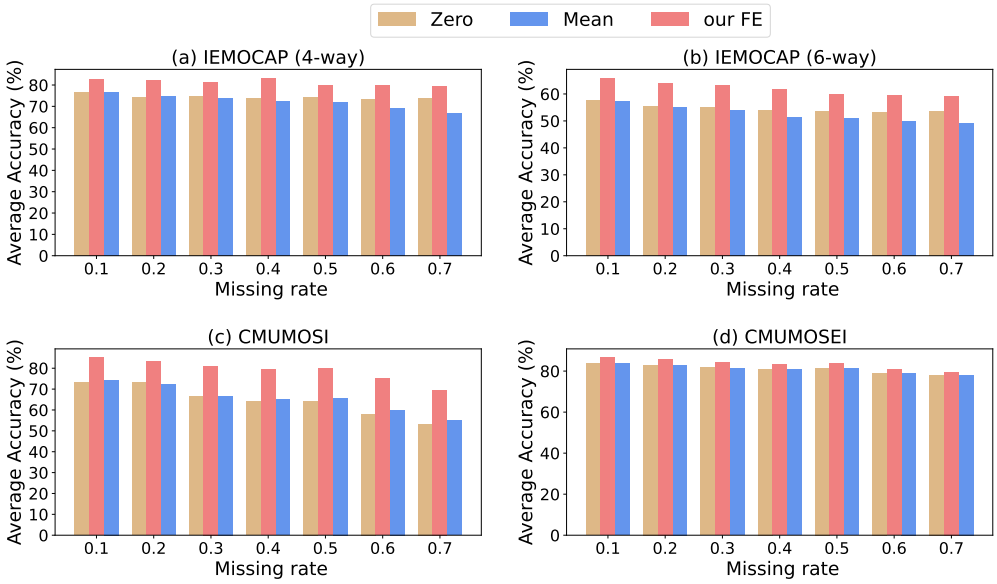


Figure 3.6: Illustration of the robustness in performance of our proposed feature estimation against basis approaches in the different rate of missing in modalities.

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.

Effects of smooth factor λ in FE. 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 3.5 shows the results on the IEMOCAP dataset for different λ settings.

Table 3.5: An investigation of the impact of the smoothing factor λ.

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	<u>79.73</u>
$\lambda = 0.2$	82.12	<u>81.98</u>	80.44	82.25	79.77	75.92	78.51
$\lambda = 0.5$	**82.83**	**82.27**	<u>81.50</u>	<u>83.17</u>	<u>80.08</u>	**77.86**	**80.79**
$\lambda = 0.9$	<u>82.58</u>	81.87	80.73	81.38	78.79	<u>76.75</u>	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	<u>60.40</u>	<u>58.44</u>
$\lambda = 0.1$	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**	<u>63.08</u>	**61.72**	**59.96**	**61.32**	**59.18**
$\lambda = 0.9$	<u>65.76</u>	<u>63.60</u>	**63.93**	<u>61.10</u>	<u>58.03</u>	58.92	56.09

*The best results are **bolded**,*
the second-highest result is denoted by the <u>underline</u>.

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.

Effect of Different Losses. 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 Error (MSE) loss and the Kullback–Leibler (KL) divergence loss, as described in Section 3.2.5. Additionally, we assess our model's performance against GCNet [49], 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. Figure 3.7 shows the performance comparison among different loss settings. KL-

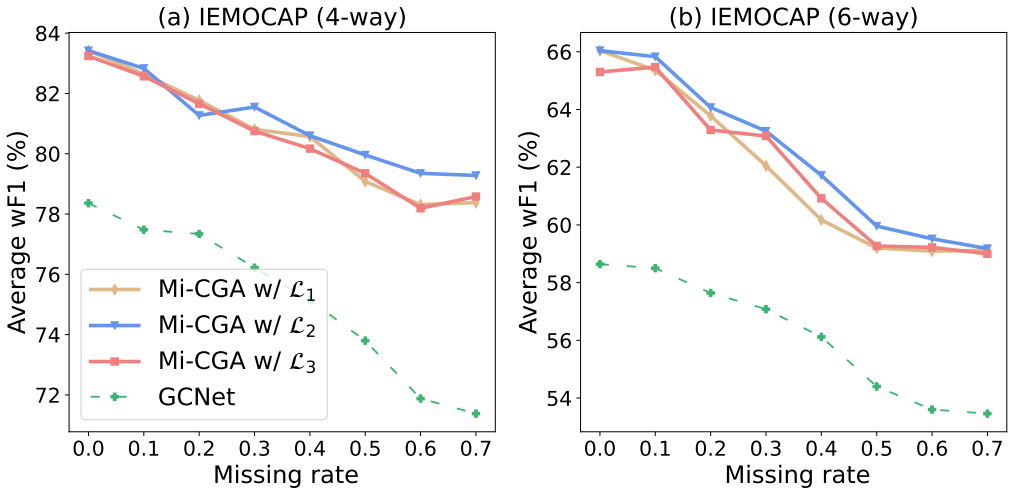


Figure 3.7: Illustration of our Mi-CGA performance with different types of objective function on IEMOCAP datasets.

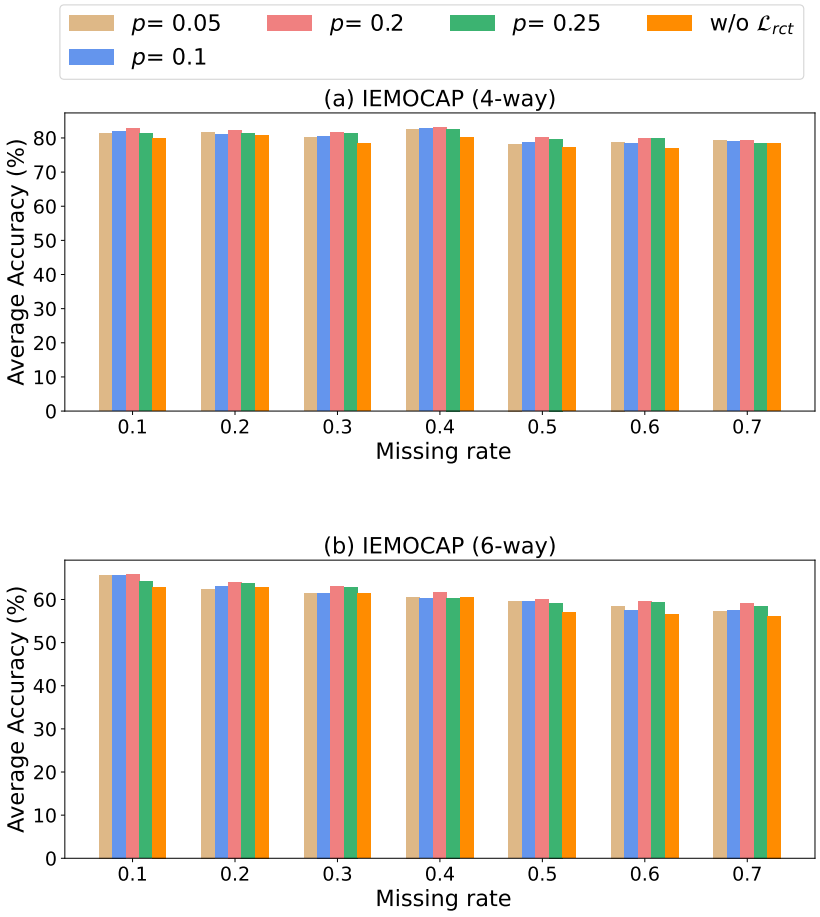


Figure 3.8: A comparison of the impact of the p value in $\mathcal{L}_{rct}$. The default setting in Mi-CGA is $p = 0.2$. The results for different p values show minimal variation, but they consistently outperform the scenario without $\mathcal{L}_{rct}$.

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 Figure 3.8 show that incorporating $\mathcal{L}_{KL}$ as the reconstruction loss in Eq. (3.28) 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.

3.3.3 Discussion

The experimental results confirm that explicitly modeling cross-modal relationships is an effective strategy for improving robustness under incomplete modality settings. By enabling structured information propagation across available modalities, Mi-CGA allows complementary emotional cues to compensate for missing signals at the representation level. This design is particularly well suited to conversational emotion recognition, where emotional information is often unevenly distributed and not simultaneously observable across modalities.

However, the results also highlight the inherent limitations of representation-level compensation. The effectiveness of Mi-CGA strongly depends on the quality of the remaining modalities: when the available signals contain sufficiently informative emotional cues, cross-modal propagation can recover a substantial portion of the missing information. In contrast, when multiple modalities are missing or the remaining cues are weak, the performance gains become smaller. This indicates that cross-modal representation learning alone cannot fully substitute missing information in challenging multimodal scenarios.

From an efficiency perspective, Table 3.6 shows that the robustness gains of Mi-CGA are achieved without prohibitive computational cost. Mi-CGA is substantially more parameter-efficient, using approximately $100\times$ fewer parameters than GCNet and about $16\times$ fewer parameters than MMIN. It also requires around $2.2\times$ less GPU memory than GCNet and $1.5\times$ less memory than MMIN, while maintaining comparable training

Metric	GCNet	MMIN	Mi-CGA
#Parameters	34M	5.44M	0.34M
Train time / epoch	0.8877 s	2.4819 s	0.8467 s
Inference time	0.1190 s	0.9702 s	0.1977 s
GPU memory usage	1305.85 MB	888.06 MB	583.27 MB

Table 3.6: Computational cost comparison on the IEMOCAP dataset.

time and significantly faster inference than MMIN. These results suggest that Mi-CGA improves robustness primarily through structured cross-modal reasoning rather than increased model capacity.

An important implication of these findings is that modality imbalance remains a critical issue even in incomplete modality settings. When some modalities are missing, the learning process tends to over-rely on the remaining dominant modalities, which can further amplify imbalance during optimization. While Mi-CGA addresses the availability of multimodal information, it does not explicitly regulate modality contributions during training. This observation motivates the focus of the following chapter, which studies modality imbalance from an optimization perspective.

3.4 Chapter Summary

In this chapter, we addressed the challenge of learning from incomplete modalities in low-quality multimodal data by introducing **Mi-CGA**, a framework specifically designed for MERC. At the core of Mi-CGA is the **Cross-modal Graph Attention Network (CGA-Net)**, which integrates three key components: (i) *Modality Feature Estimation (FE)* for reconstructing missing features and mitigating information loss, (ii) *Graph Attention Network (GAT)* for modeling contextual relationships between utterances, and (iii) *Cross-modal Attention (CMA)* for capturing semantic correlations between modalities.

Extensive experiments on benchmark datasets, including IEMOCAP, CMU-MOSI, and CMU-MOSEI, demonstrate that Mi-CGA effectively improves robustness in scenarios with incomplete modalities. By jointly leveraging intra-modal and cross-modal relationships through graph-based modeling, Mi-CGA significantly enhances the model's ability to learn meaningful representations despite severe data incompleteness, thereby contributing to the broader objective of developing robust multimodal learning strategies for low-quality data.

Limitations and Future Directions. While Mi-CGA demonstrates strong performance, several limitations remain. First, the *Missing Mask* is randomly generated, meaning the model cannot explicitly control which modality or specific positions are missing. Although this randomness simulates real-world data loss, it can also lead to extreme cases where an entire modality is nearly masked out, severely reducing the available information for the CGA-Net module. Second, each modality differs in feature length, and this heterogeneity can make the feature estimation task challenging, especially when large portions of weaker modalities are masked. Third, the performance of Mi-CGA is sensitive to hyperparameters such as the number of GAT layers and the smoothing factor in the Feature Estimation Module. Additionally, the framework imposes a non-trivial computational cost, which may limit its suitability for real-time applications.

In future work, we plan to explore (1) adaptive strategies for generating Missing Masks that better balance information loss across modalities, (2) automated hyperparameter optimization to improve model robustness and reduce manual tuning, and (3) lightweight graph-based architectures or pruning techniques to reduce computational overhead while maintaining performance.

Chapter 4

Multimodal Emotion Recognition in Conversation under Imbalanced Modality Condition

4.1 Introduction

Following the investigation of incomplete-modality learning in the previous chapter, this chapter focuses on another fundamental challenge in learning from low-quality multimodal data, namely **modality imbalance**. While incomplete-modality settings arise from partial or missing inputs, modality imbalance occurs even when all modalities are available, but contribute unequally during training.

In this chapter, we study multimodal emotion recognition in conversation under imbalanced modality conditions. The task input and output remain identical to the standard MERC formulation, while the learning process is explicitly designed to regulate modality contributions during optimization.

Input. A labeled multimodal dataset $\mathcal{D} = \{(C^{(k)}, \mathbf{y}^{(k)})\}_{k=1}^{K}$, where each conversation $C^{(k)} = [u_1^{(k)}, \ldots, u_{N_k}^{(k)}]$ consists of a sequence of utterances. Each utterance $u_i^{(k)}$ is associated with a speaker and three modalities (text t, audio a, visual v), i.e., $u_i^{(k)} = \{(u_i^{(k)})^t, (u_i^{(k)})^a, (u_i^{(k)})^v\}$, and $\mathbf{y}^{(k)} = [y_1^{(k)}, \ldots, y_{N_k}^{(k)}]$ denotes the corresponding sequence of emotion labels.

Output. A learned multimodal emotion recognition model that predicts an emotion label for each utterance in every conversation under modality-imbalanced conditions, while achieving more balanced and stable multimodal fusion.

Under this formulation, modality imbalance does not alter the task definition, but manifests during training as uneven modality contributions and gradient dominance. Accordingly, this chapter focuses on optimization-level strategies that can be flexibly integrated as plug-in modules to existing MERC models, without modifying their input–output interfaces.

In multimodal emotion recognition, different modalities often exhibit distinct learning dynamics and discriminative capacities. In practice, dominant modalities tend to converge faster and exert a disproportionate influence on gradient updates, whereas weaker modalities such as audio and visual signals remain under-optimized. As illustrated in Figure 4.1, this imbalance leads to suboptimal multimodal fusion, where complementary information from non-dominant modalities is insufficiently exploited. Prior studies have characterized this phenomenon using terms such as the *greedy nature* of dominant modalities [119], *modality collapse* [39], and *modality imbalance* [21, 51]. Existing explanations attribute this issue to heterogeneous convergence rates [109], gradient suppression of slower modalities [71], and diminishing marginal utility from additional modalities [114].

To address modality imbalance, this chapter presents two complementary approaches that aim to achieve **balanced and stable multimodal learning**. The first approach, **Ada2I**, introduces an end-to-end framework that explicitly re-balances learning at both the feature and modality levels. The second approach, **SPCL**, proposes a lightweight and plug-and-play self-paced curriculum learning strategy that progressively mitigates modality imbalance and can be integrated into a wide range of multimodal architectures.

Together, Ada2I and SPCL contribute to **Objective O2** of this dissertation by improving the robustness of multimodal learning under low-quality conditions. Specifically, these methods address **RQ3** by enhancing the contribution of under-optimized modalities, and **RQ4** by promoting balanced and stable optimization across modalities throughout training. The studies presented in this chapter are based on works published at the ACM International Conference on Multimedia (ACM MM 2024) [**VanNTC 4**] and an extended study currently accepted at Neural Computing and Applications [**VanNTC 5**].

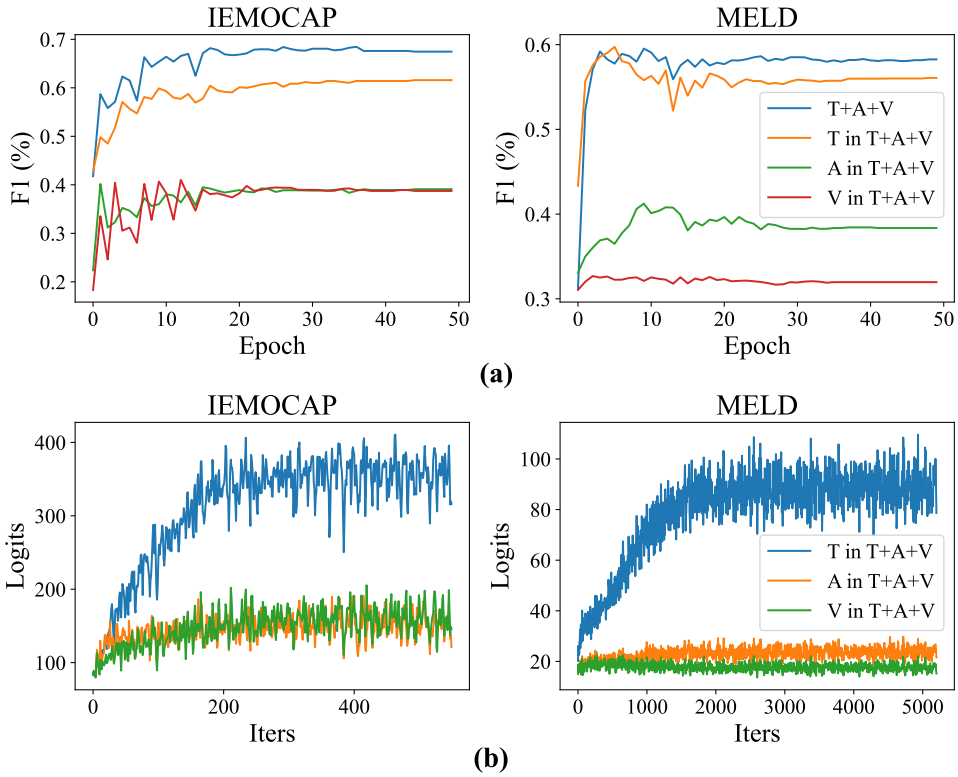


Figure 4.1: (a) Weighted F1 scores for the multimodal setting (T+A+V) compared with each unimodal encoder, and (b) batch-average unimodal-logit scores.

4.2 Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition

While recent approaches such as OGM-GE [71] and MMCosine [122] attempt to balance learning across modalities through gradient modulation or normalization, they often focus on pairwise modality interactions, leading to high computational cost and neglecting holistic multi-modality balancing. Moreover, these methods have primarily been validated on tasks like audio-visual learning [71, 122] or sentiment analysis [150], with limited attention to multimodal ERC on benchmark datasets such as IEMOCAP [9], MELD [78], and CMU-MOSEI [133].

Additionally, some recent studies [104, 114] report improved ERC performance with multimodal fusion but still reveal a persistent gap in the contribution of weaker modalities, highlighting the need for strategies that enhance all modalities simultaneously rather than optimizing only dominant ones. This gap is particularly critical for practical multimodal ERC applications, where unbalanced modality contributions can degrade overall model robustness and generalization.

To address these challenges, we propose **Ada2I**, an end-to-end framework that comprehensively tackles modality imbalance in multimodal ERC. Ada2I introduces two complementary modules: (i) **Adaptive Feature Weighting (AFW)** for feature-level balancing, and (ii) **Adaptive Modality Weighting (AMW)** for modality-level balancing. AFW applies tensor contraction to derive feature-aware attention weights, enhancing the representation of each modality, while AMW normalizes and re-weights modality-level representations to mitigate dominance. Additionally, we extend the disparity ratio from OGM-GE [71] to simultaneously handle all three modalities (text, audio, and visual), thereby reducing model complexity and improving learning efficiency for MERC.

Figure 3.2 summarizes the overall Ada2I architecture: modality-specific Transformer encoders first produce unimodal features, which are refined by AFW at the feature level and then re-balanced by AMW at the modality level before being fed into the final classifier. In this way, AFW focuses on correcting intra-modality feature imbalance, whereas AMW and the extended disparity-ratio based training strategy explicitly address inter-modality dominance.

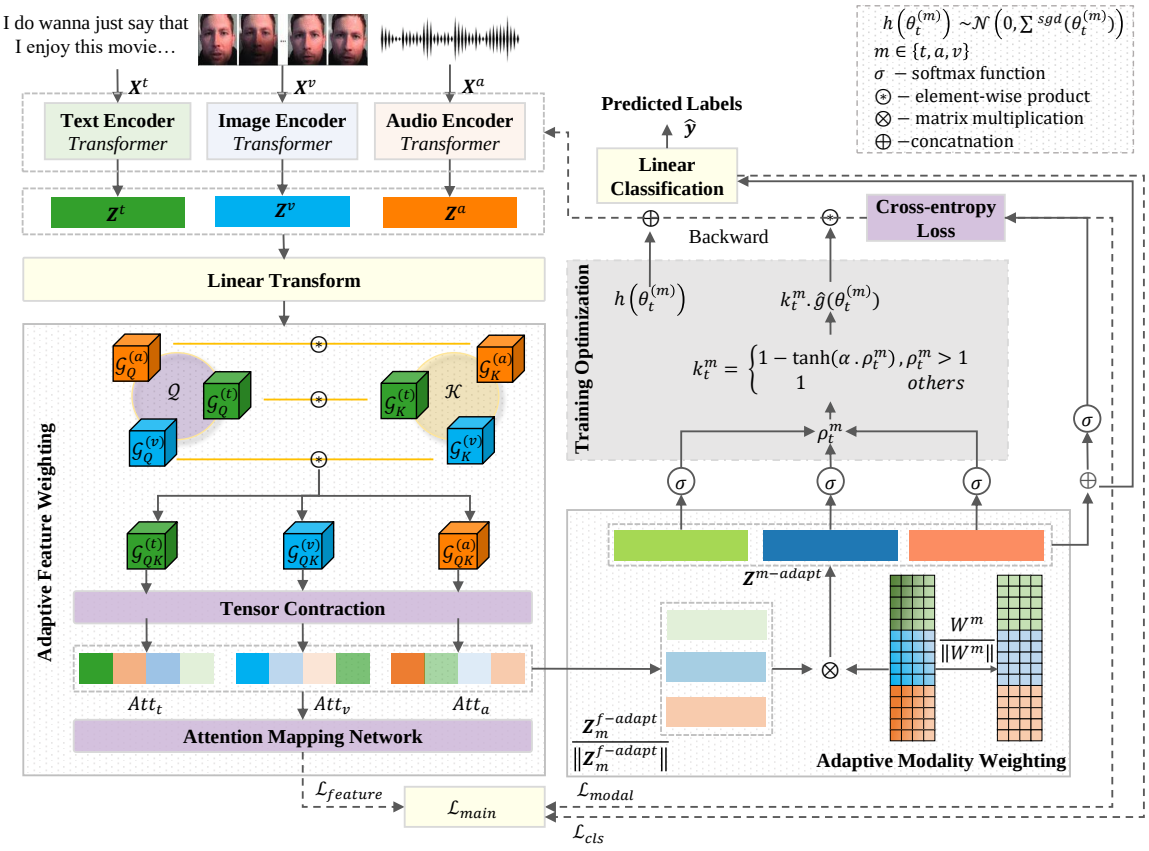


Figure 4.2: Illustration of Ada2I framework

4.2.1 Problem Definition

Consider a conversation C consisting of N utterances $\{u_1, u_2, \ldots, u_N\}$, where each utterance is associated with M modalities, including text (t), audio (a), and visual (v). Each utterance is represented as:

$$u_i = \{u_i^t, u_i^a, u_i^v\}, \quad i \in \{1, \ldots, N\}. \tag{4.1}$$

For each modality $m \in \{t, a, v\}$, a modality-specific encoder extracts a sequence of features $\mathbf{X}^m = [\mathbf{x}_1^m, \ldots, \mathbf{x}_N^m] \in \mathbb{R}^{N \times d_m}$, where d_m denotes the feature dimension of modality m.

4.2.2 Modality Encoder

Given a conversation C, a **Transformer** [105] 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:

$$\mathbf{Z}^m = \phi(\theta^{(m)}, \mathbf{X}^m), m \in \{t, a, v\} \tag{4.2}$$

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

4.2.3 Adaptive Feature Weighting (AFW)

Tensor-based Multimodal Interaction Representation

Motivated by the tensor-ring decomposition method introduced by [144], 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 [100], 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:

$$\begin{cases} \mathcal{Q} = \mathbf{Tr}\{\mathcal{G}_Q^{(t)}, \mathcal{G}_Q^{(a)}, \mathcal{G}_Q^{(v)}\} \in \mathbb{R}^{d_t \times d_a \times d_v} \\ \mathcal{K} = \mathbf{Tr}\{\mathcal{G}_K^{(t)}, \mathcal{G}_K^{(a)}, \mathcal{G}_K^{(v)}\} \in \mathbb{R}^{d_t \times d_a \times d_v} \end{cases} \tag{4.3}$$

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:

$$\Theta^m = \text{softmax}\left(\frac{1}{\sqrt{d_k}} \mathcal{G}_Q^m \odot \mathcal{G}_K^m\right) \tag{4.4}$$

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 4.3), as expressed below:

$$\begin{cases} \mathcal{G}_Q^m = reshape\Big((\mathbf{Z}^m W_{Q_m}^{(1)}) \otimes_1 (\mathbf{Z}^m W_{Q_m}^{(2)})\Big) \\ \mathcal{G}_K^m = reshape\Big((\mathbf{Z}^m W_{K_m}^{(1)}) \otimes_1 (\mathbf{Z}^m W_{K_m}^{(2)})\Big) \end{cases} \tag{4.5}$$

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.

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 [100], the *feature*-aware attention matrix $Att_m \in \mathbb{R}^{N \times d_m}$ for a given modality m is computed as follows:

$$Att_m = \text{Linear}\left(\Theta^m \times_3^1 \mathbf{A}^{(t)} \times_3^1 \mathbf{A}^{(a)} \times_3^1 \mathbf{A}^{(v)}\right) \tag{4.6}$$

where $\times_3^1$ is the $mode - (\frac{1}{3})$ tensor contraction.

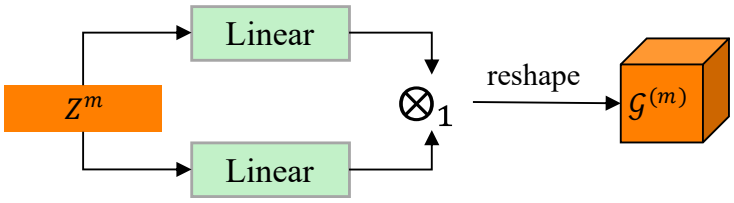


Figure 4.3: Linear Transform block to compute core tensor.

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:

$$\mathbf{Z}_m^{f-adapt} = Att_m \mathbf{Z}^m + \beta \mathbf{Z}^m \tag{4.7}$$

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

4.2.4 Adaptive Modality Weighting (AMW)

Our key focus is to achieve balanced contributions from each modality during the training. Similar to [122], 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, i.e., audio and visual, leading to divergent unimodal logit scores and distorting the joint fusion representation. Inspired by [107, 145], 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:

$$\mathbf{Z}^{m-adapt} = \sum_{m}^{\{t,a,v\}} \frac{W^m \mathbf{Z}_m^{f-adapt}}{\|W^m\| \|\mathbf{Z}_m^{f-adapt}\|} + b \tag{4.8}$$

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 $\hat{y}_i \in \mathbb{R}^{N\times|\mathcal{E}|}$.

$$\hat{y}_i = \text{MLP}(\mathbf{Z}^{m-adapt}) \tag{4.9}$$

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

4.2.5 Learning

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

$$\mathcal{L}_{cls} = -\frac{1}{B}\sum_{i}^{B} y_i log\hat{y}_i \tag{4.10}$$

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 (4.6), we employ Attention Mapping Network as follows:

$$\hat{Att}_m = \Phi_m(\mathbf{Z}_m, \psi^{(m)}), m \in \{t, a, v\} \tag{4.11}$$

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:

$$\mathcal{L}_{feature} = \frac{1}{B}\sum_{i}^{B}\left(\sum_{m}^{\{t,a,v\}} |Att_m^i - \hat{Att}_m^i|\right) \tag{4.12}$$

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

$$\mathcal{L}_{modal} = -\frac{1}{B}\sum_{i}^{B}\log\frac{e^{\mathbf{Z}_i^{m-adapt}}}{\sum_{j=1}^{|\mathcal{E}|} e^{\mathbf{Z}_j^{m-adapt}}} \tag{4.13}$$

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:

$$\mathcal{L}_{main} = \mathcal{L}_{modal} + \mathcal{L}_{feature} + \mathcal{L}_{cls} \tag{4.14}$$

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. [71] 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:

$$\theta_{t+1}^{(m)} = \theta_t^{(m)} - \eta.\hat{g}(\theta_t^{(m)}) \tag{4.15}$$

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

$$s_t^m = \sum_{j=1}^{L}\sum_{k=1}^{\mathcal{E}} \mathbb{I}_{k=y_j} \text{softmax}(cos\langle W_k^m, \mathbf{Z}_k^m\rangle + \frac{b_k}{M})jk \tag{4.16}$$

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:

$$\rho_t^m = \frac{s_t^m}{\min_{m\in\{t,a,v\}}(s_t^j)} \tag{4.17}$$

The learnable parameters are updated according to:

$$\theta_{t+1}^{(m)} = \theta_t^{(m)} - \eta.\hat{g}(\theta_t^{(m)}).k_t^m \tag{4.18}$$

where the modulation coefficient k_t^m is defined as:

$$k_t^m = \begin{cases} 1 - \tanh(\alpha \cdot \rho_t^m), & \text{if } \rho_t^m > 1, \\ 1, & \text{otherwise.} \end{cases} \tag{4.19}$$

Algorithm 3 Ada2I Training Procedure

Require: Training set $\mathcal{D} = \{(x_i^t, x_i^a, x_i^v), y_i\}_{i=1}^N$, $m \in \{t, a, v\}$
Ensure: Predicted emotion label $\hat{y}$

1: **for** each training epoch **do**
2: **for** minibatch $\mathcal{B} = \{(x_i^t, x_i^a, x_i^v), y_i\}_{i=1}^B$ sampled from $\mathcal{D}$ **do**
3: # *Refer to Subsection 4.2.2*
4: Encode unimodal features $\mathbf{X}^m$ to $\mathbf{Z}^m$ (Eq. 4.2)
5: # *Refer to Subsection 4.2.3*
6: Compute multimodal interaction representation (Eq. 4.3)
7: Calculate coefficient matrix Θ^m (Eq. 4.4)
8: Calculate modality-aware attention Att_m (Eq. 4.6)
9: Fuse features to obtain $\mathbf{Z}_m^{f\text{-}adapt}$ with β (Eq. 4.7)
10: # *Refer to Subsection 4.2.4*
11: Compute $\mathbf{Z}^{m\text{-}adapt}$ with modality-wise L2 normalization (Eq. 4.8)
12: Predict label $\hat{y}_i$ (Eq. 4.9)
13: # *Refer to Subsection 4.2.5*
14: Compute $\mathcal{L}_{cls}$ using cross-entropy (Eq. 4.10)
15: Compute $\mathcal{L}_{feature}$ using L_1 loss (Eq. 4.12)
16: Compute $\mathcal{L}_{modal}$ using cross-entropy (Eq. 4.13)
17: Aggregate to obtain $\mathcal{L}_{main}$ (Eq. 4.14)
18: Compute discrepancy ratio $\rho_t^m = \frac{s_t^m}{\min_{j\in\{t,a,v\}} s_t^j}$
19: Compute modulation coefficient k_t^m (Eq. 4.19)
20: Update parameters:
21: $\theta_{t+1}^{(i)} = \theta_t^{(i)} - \eta \cdot \hat{g}(\theta_t^{(i)}) \cdot k_t^m + \eta \cdot h(\theta_t^{(i)})$
22: **end for**
23: **end for**

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:

$$\theta_{t+1}^{(i)} = \theta_t^{(i)} - \eta \cdot \hat{g}(\theta_t^{(i)}) \cdot k_t^i + \eta \cdot h(\theta_t^{(i)}) \tag{4.20}$$

Training Optimization Strategy. The training process of Ada2I is illustrated in Algorithm 3.

In practice, each training step of Ada2I proceeds in five stages that match Algorithm 3. First, the modality encoders transform the raw input features $\mathbf{X}^m$ into unimodal representations $\mathbf{Z}^m$ (encoding stage). Second, the Adaptive Feature Weighting (AFW) module computes tensor-based multimodal interactions, derives modality-aware attention weights Att_m, and produces feature-balanced representations $\mathbf{Z}_m^{f\text{-}adapt}$ (lines 4–7).

Third, the Adaptive Modality Weighting (AMW) module applies modality-wise L2 normalization and re-weighting to obtain $\mathbf{Z}^{m\text{-}adapt}$ and the corresponding logits (prediction stage). Fourth, three losses $\mathcal{L}_{cls}$, $\mathcal{L}_{feature}$ and $\mathcal{L}_{modal}$ are computed and aggregated into the main objective $\mathcal{L}_{main}$ (lines 11–15). Finally, the training optimization strategy computes the discrepancy ratio ρ_t^m and modulation coefficient k_t^m for each modality, and updates the parameters with modulated gradients and Gaussian noise.

4.2.6 Implementation

Baseline. 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 [60], DialogueGCN [24], MMGCN [35], BiDDIN [137], and MM-DFN [34]. We report the best results obtained from [114], 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 [40], CORECT [65], GraphMFT [46], DF-ERC [45], and AdaIGN [104].

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 [92], TBJE [14], COGMEN [40], CORECT [65], OGM-GE [71], and I^{2}MCL [150]. Notably, OGM-GE and I^{2}MCL specifically address the issue of imbalanced modalities in MERC, whereas the others do not.

Evaluation Metrics. Similar to prior studies [35, 60, 114], we evaluate the effectiveness of emotion recognition using Accuracy (Acc) and Weighted F1 Score (W-F1) as our primary evalucation metrics.

Hyperparametes. 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 4.1. We employ PyTorch[3] for training our architecture on Google Colab Pro and Comet[4] for logging all experiments, leveraging its Bayesian optimizer for

[3]https://pytorch.org/
[4]https://comet.ml

Table 4.1: Hyper-parameter settings

Parameter/Module	IEMOCAP	MELD	CMU-MOSEI
Text Feature Extraction	sBERT[1]		
Audio Feature Extraction	Wave2vec-Large [88], OpenSmile [19]		
Visual Feature Extraction	MTCNN [138], MA-Net[2], DenseNet [36]		
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

hyperparameter tuning. Additional parameters can be found in Table 4.1.

4.2.7 Results

Performance Comparison on IEMOCAP and MELD dataset. As depicted in Table 4.2, our model Ada2I performs better than the previous SOTA baselines in the context of balanced modality consideration on all modality combinations on both IEMOCAP and MELD dataset.

Indeed, in the (A+V) 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 4.4b 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 sub-

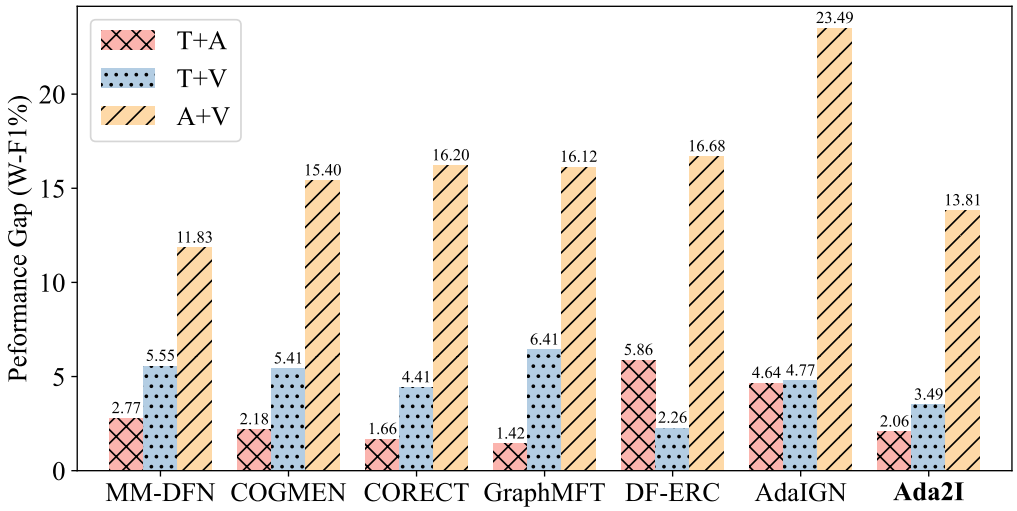
Table 4.2: Comparison of results in the multimodal setting of Ada2I with the modality-balanced baseline model enhanced by FAGM [114] (denoted by †).

IEMOCAP								
Methods	T+A+V		T+A		T+V		A+V	
	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
DialogueGCN†	62.76	63.22	<u>64.36</u>	<u>64.39</u>	<u>61.25</u>	<u>62.23</u>	49.20	49.85
BiDDIN†	58.81	58.84	58.88	58.16	59.04	58.96	46.36	46.77
MM-DFN†	<u>64.92</u>	<u>64.57</u>	63.91	64.20	61.02	60.60	<u>54.48</u>	<u>55.03</u>
MMGCN†	64.53	64.51	63.25	63.40	61.02	61.06	54.14	54.90
Ada2I (Ours)	**68.97**	**68.76**	**66.91**	**67.28**	**65.48**	**65.43**	**55.16**	**55.64**
$\Delta(\%)$	*↑4.05*	*↑4.19*	*↑2.55*	*↑2.89*	*↑4.23*	*↑3.20*	*↑0.68*	*↑0.61*
MELD								
Methods	T+A+V		T+A		T+V		A+V	
	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc
DialogueRNN†	56.42	58.05	56.46	58.01	55.67	57.39	40.46	45.39
DialogueGCN†	54.61	58.96	54.80	57.28	55.26	57.10	10.02	44.44
BiDDIN†	57.47	59.18	56.56	58.05	56.93	58.10	<u>44.39</u>	48.62
MM-DFN†	55.75	60.80	57.10	60.00	<u>57.73</u>	<u>60.65</u>	42.05	<u>48.66</u>
MMGCN†	<u>58.48</u>	<u>61.15</u>	<u>57.59</u>	<u>60.69</u>	57.14	59.46	43.49	48.43
Ada2I (Ours)	**60.38**	**63.03**	**60.08**	**62.64**	**58.62**	**61.95**	**55.16**	**55.64**
$\Delta(\%)$	*↑1.90*	*↑1.88*	*↑2.49*	*↑1.95*	*↑0.89*	*↑1.30*	*↑10.77*	*↑6.98*

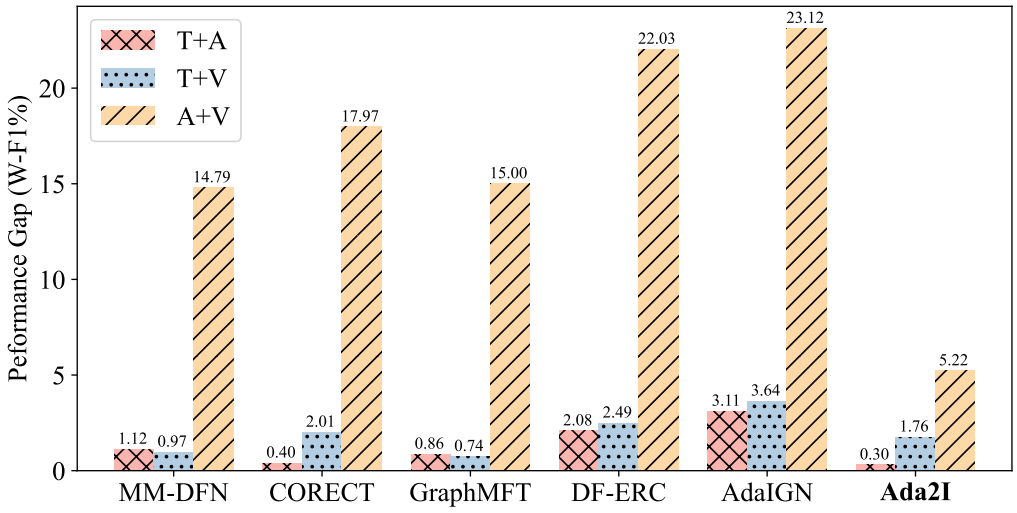
*The best performance is indicated in **bold**.*
The second-best performance is <u>underlined</u>.

stantially 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 4.4a.

Performance Comparison on CMU-MOSEI dataset Table 4.3 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 4.4: 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.

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 4.5 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.

Table 4.3: Results on the CMU-MOSEI dataset.

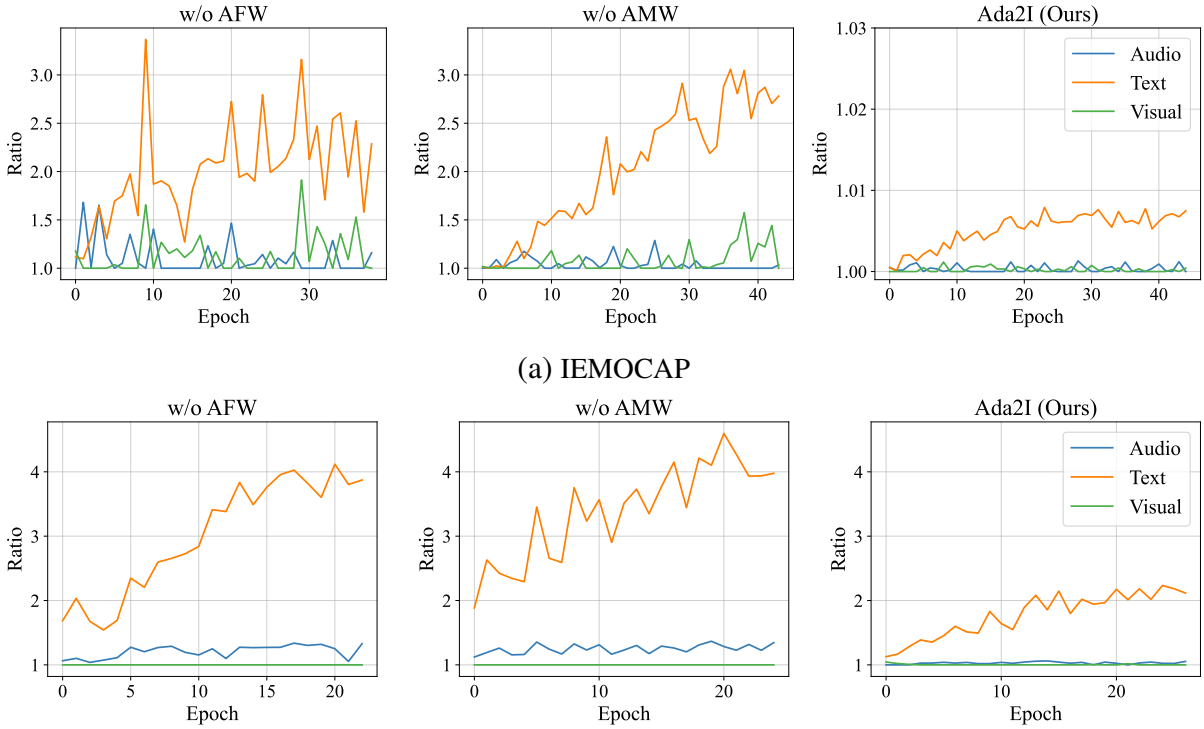
Methods	2-class				7-class			
	T+A+V	T+A	T+V	A+V	T+A+V	+TA	T+V	A+V
Multilouge-Net [92]	82.10	80.18	80.06	**75.16**	44.83	-	-	-
TBJE [14]	81.50	82.40	-	-	44.40	45.50	-	-
COGMEN† [40]	82.95	85.00	82.99	65.95	43.90	44.31	42.68	24.27
CORECT† [65]	83.98	84.28	82.83	68.89	46.31	44.89	43.76	24.55
I^2MCL [150]	81.05	-	-	-	-	-	-	-
OGM-GE† [71]	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**
$\Delta(\%)$	↑0.67	↑0.08	↑1.54	↓0.23	↑2.28	↑1.85	↑2.93	↑3.11

*The best performance is in **bold**. Cells with "-" indicate missing results, and † denotes results reproduced from the code provided in the original paper.*

Similarly, on the MELD dataset, our proposed model Ada2I has reduced this discrepancy 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.

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 4.6 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.

Effect of Module. Table 4.4 provides an ablation on the modules. AFW and AMW are two closely linked and crucial modules in 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.



(a) IEMOCAP

(b) MELD

Figure 4.5: 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.

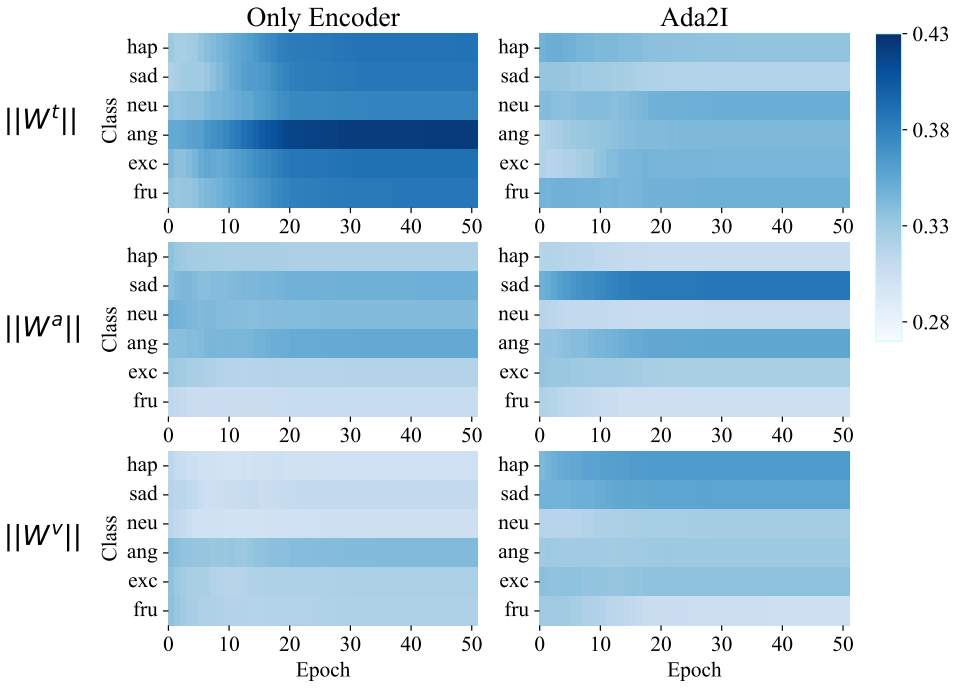


Figure 4.6: Modality-wise weights of each label normalized for the IEMOCAP dataset

Table 4.4: Ablation studies of Ada2I on AFW, AMW, and training strategy.

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**

$\downarrow$ denotes the reduction in performance of the variants compared to Ada2I.

4.2.8 Discussion

The proposed Ada2I framework addresses modality imbalance from a *representation and optimization* perspective by explicitly re-balancing learning at both the feature level and the modality level. By adaptively adjusting feature importance and modality contributions during training, Ada2I prevents dominant modalities from overwhelming the fusion process and encourages more equitable learning across modalities.

A key strength of Ada2I lies in its fine-grained control over learning dynamics. Rather than relying on heuristic fusion or post-hoc reweighting, Ada2I directly modulates how features and modalities participate in optimization, enabling weaker modalities to receive sufficient gradient signals. This is particularly beneficial for MERC, where subtle emotional cues may appear sparsely or inconsistently across modalities but remain critical for accurate recognition. As a result, Ada2I produces more balanced and robust multimodal representations that better capture complementary emotional information.

Nevertheless, Ada2I introduces additional architectural components and hyperparameters, which may increase training complexity and require careful tuning across datasets. More importantly, while Ada2I effectively mitigates modality imbalance at the representation and optimization levels, it does not explicitly regulate the *order* in which training samples are presented to the model. This observation naturally motivates a complementary perspective that focuses on *training dynamics and sample scheduling*, rather than architectural re-design. As will be discussed in the next section, this motivation leads to the proposed SPCL approach, which addresses modality imbalance from an optimization-level curriculum learning viewpoint. Overall, Ada2I demonstrates that explicitly re-balancing feature- and modality-level learning is an effective foundation for enhancing multimodal emotion recognition in conversational settings.

4.3 SPCL: Leveraging Self-Paced Curriculum Learning for Enhanced Modality Balance in Multimodal Conversational Emotion Recognition

Several approaches have been proposed to mitigate modality imbalance. Pre-trained unimodal networks [15, 119] improve modality-specific feature quality but require large-scale labeled data and significant computational resources. Auxiliary learning objectives [122, 150] use additional constraints such as contrastive or self-supervised learning to enhance modality-specific representations but often fail to generalize across tasks. Optimization-based methods [21, 71, 114] focus on balancing gradients between modalities, but they involve complex gradient manipulations that increase implementation difficulty and may not adapt well to diverse learning scenarios. *These limitations highlight the need for a simpler and more adaptive solution to mitigate modality imbalance.*

To address this challenge, we introduce a novel approach based on **Self-Paced Curriculum Learning (SPCL)** [**VanNTC 5**]. Unlike existing gradient-based or auxiliary-objective solutions, SPCL focuses on dynamically balancing the learning process by progressively guiding the model from easier to harder training samples while accounting for modality discrepancies. Our SPCL framework incorporates two key components: (1) a **Difficulty Measurer**, which estimates sample complexity using both utterance-level recognition performance and modality discrepancy scores, allowing the model to better prioritize informative but balanced samples, and (2) a **Learning Scheduler**, which adaptively refines sample selection according to the model's learning progress, ensuring that weaker modalities receive sufficient optimization throughout training.

By introducing these components, SPCL provides a unified and adaptive training strategy that directly addresses modality imbalance without requiring major architectural modifications or complex gradient operations. This design improves the stability of multimodal learning and enhances the contributions of weaker modalities, ultimately leading to more robust and balanced multimodal representation learning.

Figure 4.7 illustrates the overall pipeline of our approach, showcasing how the SPCL module seamlessly integrates with existing MER models. 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*.

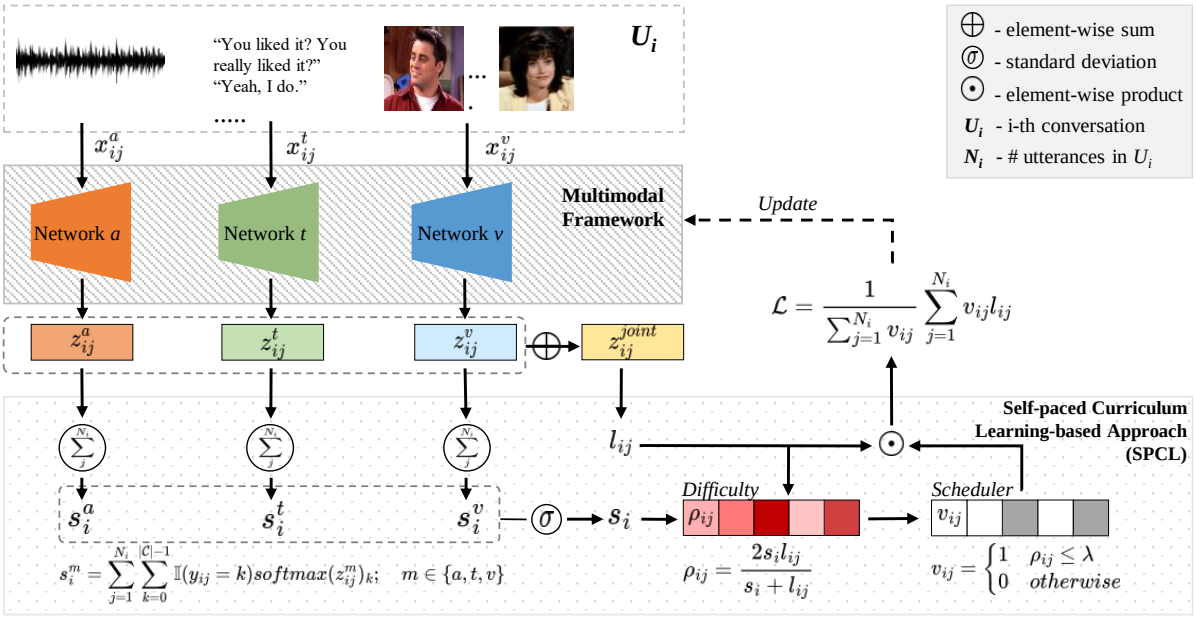


Figure 4.7: Our framework pipeline with integrated SPCL module.

4.3.1 Modality Prediction

Given a conversation C consisting of utterances $\{u_1, u_2, \ldots, u_N\}$, we employ a modality-specific emotion prediction network to estimate unimodal prediction logits for each utterance. For each modality $m \in \{a, t, v\}$, the unimodal logit of utterance u_i is computed as:

$$\mathbf{z}_i^m = \phi_m(\mathbf{x}_i^m; \boldsymbol{\theta}^m), \tag{4.21}$$

where $\phi_m(\cdot) : \mathbb{R}^{d_m} \to \mathbb{R}^{|\mathcal{C}|}$ denotes the unimodal prediction network for modality m with learnable parameters $\boldsymbol{\theta}^m$, $\mathbf{x}_i^m \in \mathbb{R}^{d_m}$ is the feature representation of utterance u_i in modality m, and $\mathbf{z}_i^m \in \mathbb{R}^{|\mathcal{C}|}$ is the corresponding unimodal logit vector.

To obtain the joint prediction for utterance u_i, we adopt a simple yet effective fusion strategy by summing the unimodal logits across modalities. The cross-modal logit is computed as:

$$\mathbf{z}_i^{\text{joint}} = \sum_{m \in \{a,t,v\}} \mathbf{z}_i^m, \tag{4.22}$$

where $\mathbf{z}_i^{\text{joint}} \in \mathbb{R}^{|\mathcal{C}|}$ denotes the cross-modal logit prediction for utterance u_i.

The emotion prediction loss for utterance u_i is then computed based on the cross-modal logit using the negative log-likelihood of the ground-truth label:

$$\ell_i = -\log\left(\text{softmax}(\mathbf{z}_i^{\text{joint}})_{y_i}\right), \tag{4.23}$$

where $y_i \in \mathcal{C}$ denotes the ground-truth emotion label of utterance u_i, and $\ell_i \in \mathbb{R}$ is the corresponding utterance-level loss.

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

To mitigate modality imbalance during training while improving performance on multimodal emotion recognition, we adopt a self-paced curriculum learning strategy. Following prior curriculum learning paradigms, SPCL organizes the training process by gradually introducing more challenging samples. The curriculum is constructed using two components: (1) a *Difficulty Measurer* that quantifies sample difficulty, and (2) a *Learning Scheduler* that controls the learning pace.

Difficulty Measurer

Traditional difficulty measurers typically operate at the utterance level, treating each utterance as an independent sample and computing a scalar difficulty score accordingly. In line with this paradigm, we adopt the utterance-level loss defined in Section 4.3.1 as an *utterance-level difficulty score*, under the assumption that higher loss indicates greater prediction uncertainty or misalignment among unimodal representations.

Specifically, for each utterance u_{ij}, we define a binary mask variable v_{ij} using a hard regularization function $g(\rho_{ij}, \lambda)$:

$$v_{ij} = g(\rho_{ij}, \lambda) = \begin{cases} 1, & \rho_{ij} \leq \lambda, \\ 0, & \text{otherwise}, \end{cases} \tag{4.24}$$

where ρ_{ij} denotes the estimated difficulty of utterance u_{ij} and $\lambda > 0$ is a threshold parameter that separates easy samples from hard ones.

For a given conversation C_i consisting of N_i utterances, we first compute a unimodal score for each modality $m \in \{a, t, v\}$ based on the unimodal logits:

$$s_i^m = \sum_{j=1}^{N_i} \sum_{k=1}^{|\mathcal{C}|} \mathbb{I}(y_{ij} = k) \, \text{softmax}(\mathbf{z}_{ij}^m)_k, \tag{4.25}$$

where s_i^m denotes the unimodal score of conversation C_i with respect to modality m, $\mathbf{z}_{ij}^m$ is the unimodal logit defined in Equation (4.21), and $\mathbb{I}(\cdot)$ is the indicator function.

Next, we quantify inter-modality discrepancy by computing the standard deviation of unimodal scores across modalities. This statistic reflects the degree of modality imbalance within a conversation: larger values indicate stronger disparities among modalities, whereas smaller values suggest more balanced contributions. We therefore define the conversation-level difficulty score as:

$$s_i = \sigma(s_i^a, s_i^t, s_i^v), \tag{4.26}$$

where $\sigma(\cdot)$ denotes the standard deviation operator.

Finally, we integrate utterance-level difficulty and conversation-level modality misalignment to obtain the overall difficulty of utterance u_{ij}. Specifically, we compute the harmonic mean of the utterance loss ℓ_{ij} and the conversation-level score s_i:

$$\rho_{ij} = \frac{2\, s_i\, \ell_{ij}}{s_i + \ell_{ij}}, \tag{4.27}$$

where $\rho_{ij} \in \mathbb{R}$ denotes the difficulty of utterance u_{ij}. The harmonic mean penalizes extreme values and prevents either component from dominating the difficulty estimation, ensuring that both recognition difficulty and modality imbalance are jointly considered.

Learning Scheduler

To organize training samples according to their estimated difficulty, we employ a learning scheduler based on a hard regularization strategy. This scheduler enforces a strictly progressive training process, where samples are either included or excluded depending on their difficulty level ρ_{ij}. In contrast to soft curriculum learning approaches that continuously adjust sample weights [27], the hard regularizer completely excludes difficult samples until the model becomes sufficiently capable of handling them. Such a strict progression has been shown to promote more stable optimization and prevent catastrophic forgetting [95].

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:

$$v_{ij} = g(\rho_{ij}, \lambda) = \begin{cases} 1 & \rho_{ij} \leq \lambda, \\ 0 & otherwise \end{cases} \tag{4.28}$$

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

splitting easy and hard samples.

The difficulty threshold λ is initialized with a small value and gradually increased over training epochs to progressively admit more challenging samples:

$$\lambda^{(t)} = \begin{cases} \varepsilon, & t = 0, \\ \alpha\, \lambda^{(t-1)}, & t > 0, \end{cases} \tag{4.29}$$

where $\lambda^{(t)}$ denotes the threshold at epoch t, ε is a small positive constant, and $\alpha > 1$ is an aging parameter that controls the learning pace. Both ε and α are selected empirically.

4.3.3 Multi-modal Learning with SPCL

In a standard multimodal emotion recognition setting, the training objective is defined as the average negative log-likelihood loss over all utterances in the dataset. Let ℓ_{ij} denote the utterance-level loss for the j-th utterance in conversation C_i. The conventional training objective is given by:

$$\mathcal{L} = \frac{1}{\sum_{i=1}^{|\mathcal{D}|} N_i} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{N_i} \ell_{ij}, \tag{4.30}$$

where $\sum_{i=1}^{|\mathcal{D}|} N_i$ denotes the total number of utterances in the dataset.

At each training step, the parameters of the multimodal framework, including the learnable parameters $\boldsymbol{\theta}^m$ of the unimodal prediction networks $\phi_m(\cdot)$, are updated via gradient-based optimization:

$$\boldsymbol{\theta}^{m\,(t+1)} \leftarrow \boldsymbol{\theta}^{m\,(t)} - \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}^{m\,(t)}}, \quad m \in \{a, t, v\}, \tag{4.31}$$

where η denotes the learning rate.

With the integration of the proposed SPCL module, this training procedure is refined by selectively excluding difficult samples from the loss computation. Specifically, each utterance-level loss ℓ_{ij} is modulated by a binary mask v_{ij}, which acts as a gating mechanism that controls whether a sample participates in training. The resulting SPCL-aware training objective is defined as:

$$\mathcal{L}_{\text{SPCL}} = \frac{1}{\sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{N_i} v_{ij}} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{N_i} v_{ij}\, \ell_{ij}, \tag{4.32}$$

Algorithm 4 Training a Multimodal ERC Framework with SPCL Integration

Input: Multimodal dataset $\mathcal{D} = \{C_1, C_2, \ldots, C_{|\mathcal{D}|}\}$
Initialize difficulty threshold $\lambda \leftarrow \varepsilon$, aging parameter α
for epoch $t = 1$ to T **do**
    for mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
        Compute unimodal logits $\mathbf{z}_{ij}^{m}$ using Eq. (4.21)
        Compute joint logits $\mathbf{z}_{ij}^{\text{joint}}$ using Eq. (4.22)
        Compute utterance-level loss ℓ_{ij} using Eq. (4.23)
        Compute unimodal conversation scores s_i^m using Eq. (4.25)
        Compute conversation-level score s_i using Eq. (4.26)
        Compute utterance difficulty ρ_{ij} using Eq. (4.27)
        Obtain sample mask v_{ij} using Eq. (4.28)
        Compute SPCL loss $\mathcal{L}_{\text{SPCL}}$ using Eq. (4.32)
        Update model parameters by minimizing $\mathcal{L}_{\text{SPCL}}$ (Eq. 4.33)
    end for
    Update difficulty threshold λ using Eq. (4.29)
end for

where $\sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{N_i} v_{ij}$ denotes the number of utterances selected by the curriculum at the current training stage.

Accordingly, the parameters of the unimodal networks are updated based on $\mathcal{L}_{\text{SPCL}}$:

$$\boldsymbol{\theta}^{m\,(t+1)} \leftarrow \boldsymbol{\theta}^{m\,(t)} - \eta \frac{\partial \mathcal{L}_{\text{SPCL}}}{\partial \boldsymbol{\theta}^{m\,(t)}}, \quad m \in \{a, t, v\}, \tag{4.33}$$

which ensures that the model initially learns from low-difficulty, well-aligned samples and gradually incorporates more challenging utterances as training progresses.

Overall, the whole training process is described in Algorithm 4.

4.3.4 Implementation

Datasets. We conduct experiments on two benchmark datasets for ERC task that support multi-modal, namely: IEMOCAP [9], MELD [78].

Baselines. To evaluate the robustness and stability of our proposed method, we incorporate it into 4 existing models for ERC task, namely: DialogueGCN [24], BiDDIN [137], MMGCN [35], MM-DFN [34]. In particular, these models are used as our emotion prediction network $\phi_m(\cdot)$ in Equation 4.21.

Reproducibility. SPCL is implemented using Pytorch [5], 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 $[1.05, 1.4]$. As a plug-in training strategy, SPCL does not introduce additional learnable parameters.

4.3.5 Results

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.

Analysis of Experimental Results on IEMOCAP. Table 4.5 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

[5] https://pytorch.org/

Table 4.5: Performance comparison of baseline models with our SPCL module and other plug-in methods on IEMOCAP.

Model	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [24]								
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**$^{\dagger}_{\pm 0.95}$	**65.32**$^{\dagger}_{\pm 0.99}$	**65.46**$^{\dagger}_{\pm 1.15}$	**64.47**$^{\dagger}_{\pm 0.21}$	**64.46**$^{\dagger}_{\pm 0.20}$	**57.89**$^{\dagger}_{\pm 1.00}$	**58.59**$^{\dagger}_{\pm 0.49}$
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 [137]								
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**$^{\dagger}_{\pm 0.56}$	**60.24**$^{\dagger}_{\pm 1.11}$	**60.43**$^{\dagger}_{\pm 0.99}$	**61.10**$^{\dagger}_{\pm 0.82}$	**61.91**$^{\dagger}_{\pm 0.58}$	46.34$_{\pm 0.43}$	**49.11**$^{\dagger}_{\pm 0.71}$
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 [35]								
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**$^{\dagger}_{\pm 0.64}$	**66.75**$^{\dagger}_{\pm 0.42}$	**66.51**$^{\dagger}_{\pm 0.42}$	**65.00**$^{\dagger}_{\pm 1.05}$	**65.09**$^{\dagger}_{\pm 1.11}$	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 [34]								
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**$^{\dagger}_{\pm 0.54}$	**66.03**$^{\dagger}_{\pm 0.86}$	**66.09**$^{\dagger}_{\pm 0.65}$	**64.31**$^{\dagger}_{\pm 0.66}$	**64.70**$^{\dagger}_{\pm 0.63}$	**53.38**$^{\dagger}_{\pm 0.67}$	**53.47**$^{\dagger}_{\pm 0.93}$
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*

Bold *and* underlined *denote the best and second-best results, respectively.*
Δ *indicates the performance gap to the previous SOTA.*
Δ_{Base} *measures the improvement of SPCL over the original baseline.*
† *denote value with statistically significant improvements* ($p < 0.05$) *based on paired t-tests.*

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 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.

Analysis of Experimental Results on MELD. Table 4.6 presents a comparative performance analysis of our proposed SPCL module against multiple baselines on the

Table 4.6: Performance comparison of baseline models with our SPCL module and other plug-in methods on MELD.

Model	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [**?**]								
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**$^{\dagger}_{\pm 1.21}$	58.04$_{\pm 0.56}$	60.84$_{\pm 0.69}$	**56.18**$_{\pm 1.38}$	**58.61**$^{\dagger}_{\pm 1.43}$	42.28$_{\pm 0.79}$	46.64$_{\pm 1.38}$
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 [137]								
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**$^{\dagger}_{\pm 0.30}$	**58.08**$^{\dagger}_{\pm 0.30}$	**61.22**$^{\dagger}_{\pm 0.26}$	**58.10**$^{\dagger}_{\pm 0.43}$	**61.00**$^{\dagger}_{\pm 0.74}$	42.30$_{\pm 0.23}$	48.15$_{\pm 0.43}$
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 [35]								
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**$^{\dagger}_{\pm 0.48}$	**58.93**$^{\dagger}_{\pm 0.29}$	**61.65**$^{\dagger}_{\pm 0.39}$	**58.14**$^{\dagger}_{\pm 1.17}$	**60.64**$^{\dagger}_{\pm 1.81}$	**43.79**$^{\dagger}_{\pm 0.31}$	49.10$_{\pm 0.28}$
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 [34]								
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**$^{\dagger}_{\pm 0.90}$	**59.11**$^{\dagger}_{\pm 0.32}$	**62.31**$^{\dagger}_{\pm 0.32}$	**58.91**$^{\dagger}_{\pm 0.17}$	**61.94**$^{\dagger}_{\pm 0.34}$	43.32$_{\pm 0.57}$	48.59$^{\dagger}_{\pm 0.55}$
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*

Bold *and* underlined *denote the best and second-best results, respectively.*
Δ *indicates the performance gap to the previous SOTA.*
Δ_{Base} *measures the improvement of SPCL over the original baseline.*
† *denote value with statistically significant improvements* ($p < 0.05$) *based on paired t-tests.*

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.

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 Equation 4.27 and assess the resulting performance, as summarized in Table 4.7.

Table 4.7: Ablation study on IEMOCAP for our proposed SPCL module.

Method	TAV		TA		TV		AV	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
DialogueGCN [24]	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 [137]	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 [35]	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 [34]	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

$\downarrow$ or $\uparrow$ denotes the performance change compared to SPCL when a sub-module is ablated.

The results clearly demonstrate that these components are significantly more effective when combined. Our full SPCL module consistently outperforms all other configurations across all scenarios. For instance, in the TAV setting with MMGCN, our model achieves 67.66% w-F1 score and 67.71% accuracy, compared to 65.78% w-F1 when removing the conversation-level score, and 64.60% w-F1 when removing the utterance-level score, indicating the substantial contribution of both components. Similarly, in the TA setting with DialogueGCN, omitting the utterance-level score results in a negligible decrease from 65.32% to 65.31% w-F1, whereas removing the conversation-level score

causes a more noticeable reduction to 64.87%.

The complementary nature of the utterance- and conversation-level difficulty measures is evident, as removing either component leads to performance drops of up to 3.06% in w-F1 score (in MMGCN, TAV setting) and 4.87% in accuracy (in MM-DFN, AV setting). Interestingly, in certain cases, omitting one component still yields improvements over the original baselines. For example, in the TV setting with BiDDIN, the model without the conversation-level score achieves 59.48% w-F1, exceeding BiDDIN's original performance of 58.57%.

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 Figure 4.8. This rate is directly influenced by the tuning of ε and α, which can be explained through the updating of λ in Equation. 4.29, and varies depending on the baseline architecture, as different models exhibit unique sensitivity to data distribution.

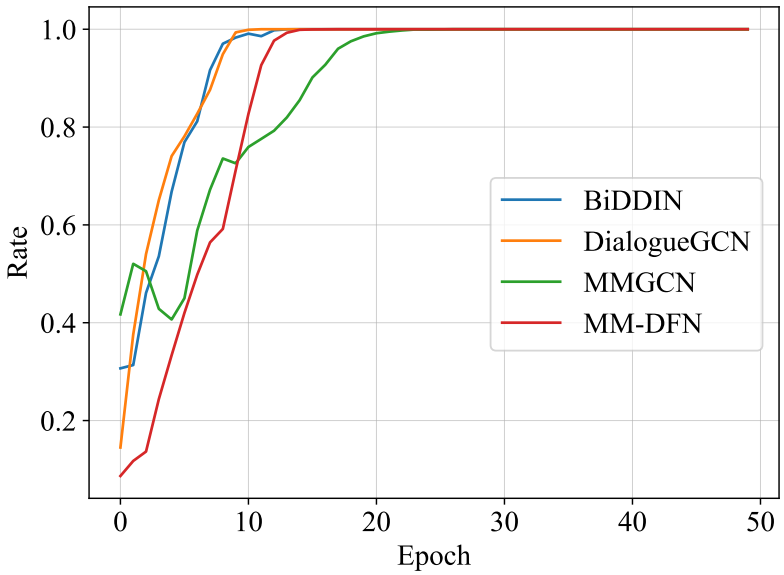


Figure 4.8: The curricula expanding rate of the four baselines integrated on IEMOCAP.

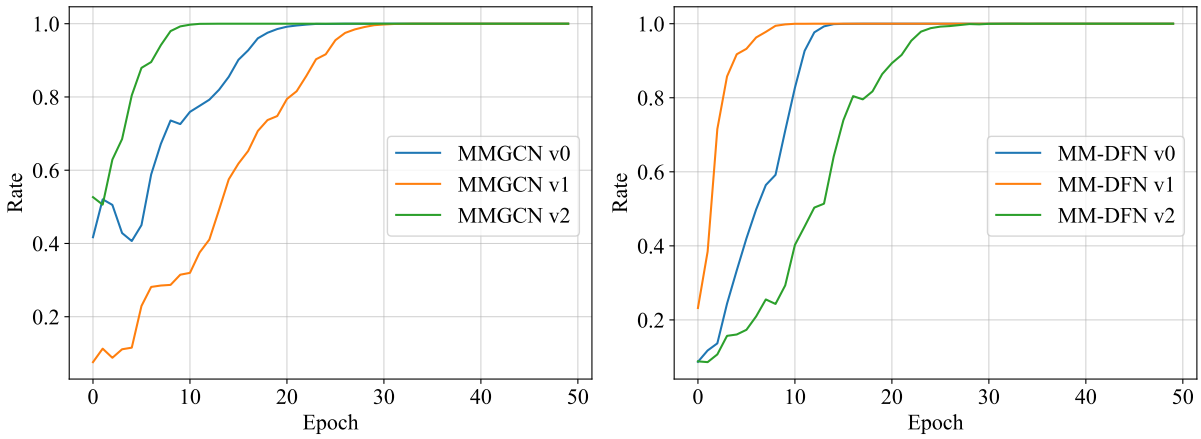
Our study on the curriculum expanding rate reveals that the best performance is achieved when the rate maintains a consistently increasing trend, as exemplified by DialogueGCN. This suggests that a gradual yet steady introduction of complex samples

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 4.8 and Figure 4.9. 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.

Table 4.8: Performance of MMGCN and MM-DFN on IEMOCAP under different hyper-parameter settings for our SPCL module.

Model	Version	ε	α	w-F1 (%)	Acc (%)
MMGCN	v0 (best)	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 (best)	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

Figure 4.9: Curricula expanding rate of MMGCN and MM-DFN under *SPCL hyper-parameters setting* specified in Table 4.8.

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

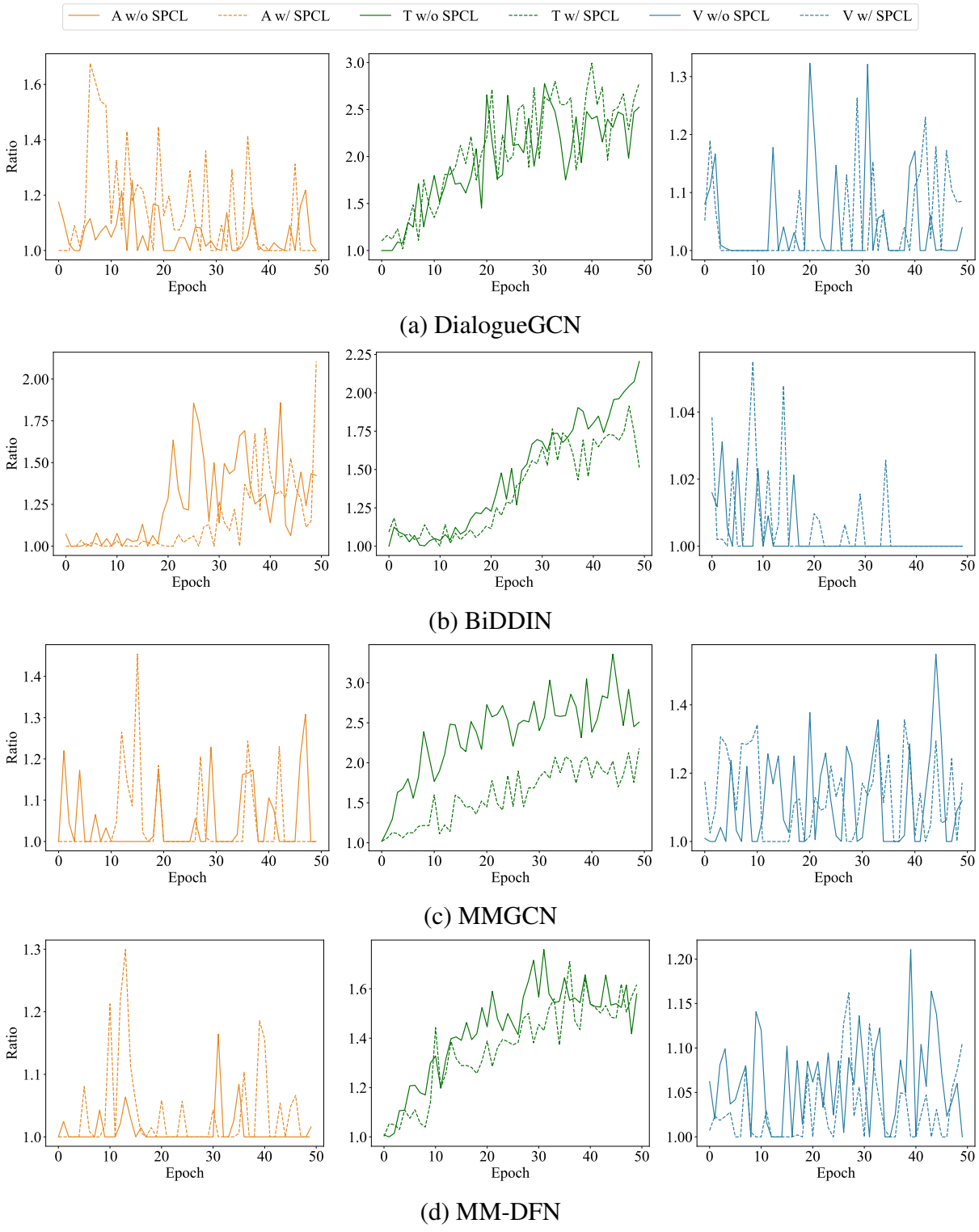


Figure 4.10: Modality ratio of the four backbones during training on IEMOCAP dataset.

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.

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 Figure 4.10, 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—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. This balance ultimately enhances the model's capacity to leverage multimodal information effectively, leading to improved robustness and generalizability in emotion recognition.

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 [75]. The exponential pacing used in SPCL is described in Eq. 4.29, whereas the formulations of newly adopted strategies are described in Table 4.9. As shown in Figure 4.11, linear pacing strategies, i.e.,

Table 4.9: 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	$c_t = \min\left(1, \sqrt{t\frac{1-c_0^2}{T} + c_0^2}\right)$ $\lambda^{(t)} = \text{Quantile}(\rho_{ij}, c_t)$

exponential and cosine pacing, yield smoother updates of λ. In contrast, the two nonlinear strategies, where λ is adaptively updated with regards to sample difficulty ρ_{ij}, exhibit larger fluctuations, particularly under competence-based pacing. Consequently,

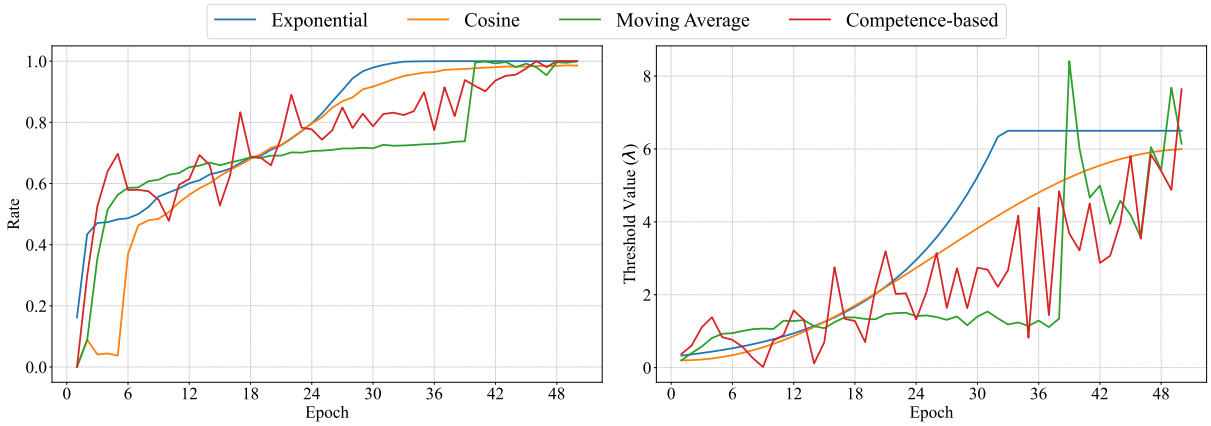


Figure 4.11: Curricula expanding rate and respective threshold value of MMGCN on MELD under different pacing strategies.

linear pacing results in a smoother curriculum expansion, indicating a more stable introduction of new samples. Table 4.10 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 4.10: Performance comparison of MMGCN and MM-DFN on IEMOCAP and MELD dataset using different pacing strategies.

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

*The best and second-best performances for each dataset and backbone are highlighted in **bold** and underline.*

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 [110].

As shown in Table 4.11 and illustrated in Figure 4.12, 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 MERC.

Table 4.11: Performance comparison of four backbone models on IEMOCAP and MELD datasets using different types of regularizers for the learning scheduler.

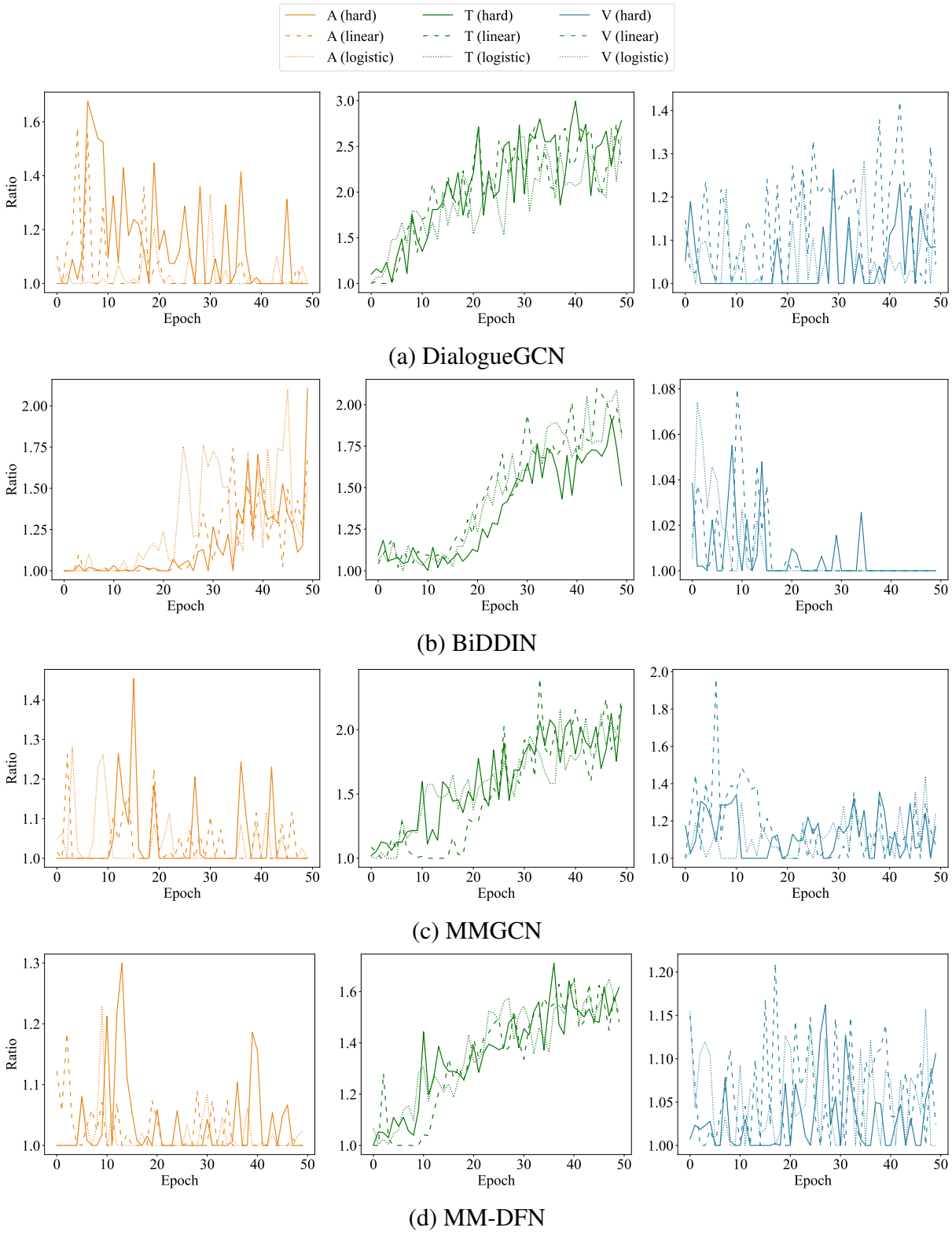
Regularizer	MMGCN		DialogueGCN		BiDDIN		MM-DFN	
	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc
IEMOCAP								
Hard Regularizer	**67.84**	**68.02**	66.46	66.61	**59.98**	**60.07**	**67.92**	**68.21**
Soft Linear	65.20	65.13	64.87	64.70	56.77	57.18	65.94	65.80
Soft Logistic	65.98	66.17	**67.14**	**67.41**	58.16	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**.*

4.3.6 Discussion

The proposed Self-Paced Curriculum Learning-based approach (SPCL) addresses modality imbalance from an *optimization* perspective, without modifying the underlying multimodal architecture. This design is particularly suitable for MERC, where the relative reliability of text, audio, and visual modalities varies substantially across utterances and conversations. By progressively introducing training samples from easier and more modality-aligned instances to harder and more imbalanced ones, SPCL stabilizes early-stage optimization and mitigates the tendency of dominant modalities, such as text, to drive the learning process.

A key strength of SPCL lies in its difficulty measurer, which jointly considers utterance-level recognition loss and conversation-level modality misalignment. The latter captures dialogue-level discrepancies among unimodal predictions, which naturally arise in MERC due to emotional dynamics and heterogeneous signal quality across speakers and turns. By combining these two factors through the harmonic mean, SPCL avoids biasing the curriculum toward either recognition difficulty or modality discrep-



(a) DialogueGCN

(b) BiDDIN

(c) MMGCN

(d) MM-DFN

Figure 4.12: Modality ratio of the four backbones during training on IEMOCAP dataset using different types of regularizer for the Learning Scheduler.

ancy alone, resulting in a more balanced and robust sample scheduling strategy.

From a practical perspective, SPCL retains a plug-and-play property: it does not introduce additional learnable parameters and incurs no extra inference cost. Empirically,

integrating SPCL into MMGCN increases the training time on MELD by approximately 10 seconds per epoch (batch size 32), indicating that the benefits of improved learning stability are achieved with only modest computational overhead.

Nevertheless, SPCL relies on a hard sample selection mechanism, which may delay the utilization of informative but difficult samples, particularly in the early training stages. Moreover, its dependence on unimodal confidence estimates can be sensitive when unimodal predictions are unreliable. These limitations point to promising future directions, such as soft weighting schemes or adaptive curriculum schedules.

Overall, SPCL demonstrates that explicitly regulating training dynamics through self-paced curriculum learning is an effective and model-agnostic strategy for alleviating modality imbalance in multimodal emotion recognition.

4.4 Chapter Summary

This chapter addresses **Objective O2** of the dissertation by systematically investigating balanced multimodal learning for MERC under modality imbalance. Rather than assuming uniformly informative modalities, the proposed approaches explicitly account for the uneven and dynamic nature of multimodal signals in conversational settings.

Specifically, Ada2I and SPCL address modality imbalance from complementary perspectives. Ada2I focuses on representation-level re-balancing by regulating feature- and modality-level learning dynamics, while SPCL operates at the optimization level by stabilizing training through difficulty-aware curriculum scheduling. As summarized in Table 4.12, the two approaches are effective under different imbalance scenarios and can be viewed as complementary rather than competing strategies.

Extensive experiments on benchmark MERC datasets demonstrate that both approaches consistently improve robustness and generalization under imbalanced multimodal conditions. Collectively, this chapter establishes a coherent MERC-oriented framework for understanding and addressing modality imbalance, laying the foundation for subsequent investigations into robustness and generalization in multimodal affective learning.

Learning Intervention Axis	**Ada2I**	**SPCL**
Feature-level representation calibration	✓	–
Modality-level contribution regulation	✓	–
Utterance- & conversation-level difficulty control	–	✓
Curriculum-based optimization scheduling	–	✓
Primary target of imbalance	Representation bias	Optimization instability
Model dependency	Architecture-aware	Architecture-agnostic
Training overhead	Moderate	Negligible

Table 4.12: Comparison of Ada2I and SPCL from the perspective of their primary learning intervention axes in MERC.

Conclusion and Future Work

This dissertation presents works into MERC, addressing both the modeling of multimodal conversational understanding and the robustness of learning under realistic low-quality data conditions. Rather than treating these challenges in isolation, the dissertation adopts a progressive research framework that connects multimodal fusion, conversational context modeling, and robust learning dynamics within a unified methodological scope.

Summary of Contributions. To achieve **Objective O1**, this dissertation investigates how multimodal representations and conversational context should be modeled for effective emotion recognition in dialogue. Specifically, CORECT addresses RQ1 by introducing a relational-temporal graph-based framework that enables context-aware multimodal fusion while preserving modality-specific representations. Building upon this foundation, MultiDAG+CL addresses RQ2 by further examining how temporal and speaker-dependent conversational dependencies should be learned under varying levels of dialogue complexity, through directed acyclic graph modeling and curriculum-based optimization. Together, these methods establish a principled approach to multimodal fusion and contextual reasoning in conversational emotion recognition.

To fulfill **Objective O2**, the dissertation further extends the investigation to MERC under practical low-quality data conditions. Mi-CGA addresses RQ3 by enabling robust multimodal fusion when one or more modalities are missing, through graph-based information propagation and cross-modal reasoning. In addition, Ada2I and SPCL address RQ4 by regulating modality imbalance from complementary perspectives: Ada2I focuses on representation-level re-balancing of modality contributions, while SPCL stabilizes learning dynamics through curriculum-based optimization. Collectively, these methods provide a coherent framework for improving the robustness and stability of MERC models in realistic conversational settings.

Limitations. Despite the effectiveness of the proposed approaches, several limitations remain. First, many methods rely on predefined conversational structures or modality availability assumptions, which may restrict flexibility in highly spontaneous, noisy, or open-domain conversational scenarios. Second, graph construction strategies and curriculum design depend on heuristic or dataset-specific criteria, and their effectiveness may vary across domains. Third, experimental evaluations are primarily conducted on benchmark datasets, and further validation in large-scale, real-world conversational systems is necessary to assess scalability and deployment feasibility.

Future Work. These limitations suggest several promising directions for future research. Potential extensions include adaptive graph construction and curriculum strategies that dynamically respond to conversational complexity, tighter integration of fusion and robustness mechanisms within unified learning frameworks, and exploration of MERC in broader conversational intelligence tasks such as empathetic dialogue systems and human-AI interaction. Investigating scalable and trustworthy MERC models for real-world deployment also remains an important avenue for future work.

Overall, this dissertation advances the understanding of MERC by providing a unified and progressive research framework that connects multimodal fusion, conversational modeling, and robustness under low-quality data conditions. The proposed methodologies contribute toward the development of reliable, context-aware, and practically applicable multimodal emotion recognition systems.

List of Publications

- [**VanNTC 1**] "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 (EMNLP 2023)*, pages 15154–15167, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.937 – **CORE Rank A* Conference**

- [**VanNTC 2**] "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, Italy. ELRA and ICCL. https://doi.org/10.63317/3eikm2yttbsc– **CORE Rank B Conference**

- [**VanNTC 3**] "Mi-CGA: Cross-modal Graph Attention Network for Robust Emotion Recognition in the Presence of Incomplete Modalities". *Neurocomputing*, 623: 129342. https://doi.org/10.1016/j.neucom.2025.129342 – **SCIE Q1 Journal, Impact Factor: 6.5**

- [**VanNTC 4**] "Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition". In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM 2024)*, pages 9330–9339. https://doi.org/10.1145/3664647.3681648 – **CORE Rank A* Conference**

- [**VanNTC 5**] "Leveraging Self-Paced Curriculum Learning for Enhanced Modality Balance in Multimodal Conversational Emotion Recognition". *Neural Computing and Applications*. 38, 459 (2026). https://doi.org/10.1007/s00521-026-12160-6 – **Scopus Q1 Journal**

References

[1] L. N. T. D. C. P. P. D. A. N. Anh Pham, Khanh Linh Tran, "Bud500: A comprehensive vietnamese asr dataset," 2024. [Online]. Available: https://github.com/quocanh34/Bud500

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[3] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[4] 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.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[6] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[7] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[8] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *International Conference on Learning Representations*, 2022.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[10] J. Chen and 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.

[11] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[12] R. J. Davidson, K. R. Sherer, and H. H. Goldsmith, *Handbook of affective sciences*. Oxford University Press, 2009.

[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[14] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Seattle, USA: Association for Computational Linguistics, Jul. 2020, pp. 1–7.

[15] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao, "On uni-modal feature learning in supervised multi-modal learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8632–8656.

[16] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[17] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–6.

[19] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[20] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Information Fusion*, p. 101847, 2023.

[21] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 029–20 038.

[22] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[23] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: advancements, challenges, and future directions," *Information Fusion*, vol. 105, p. 102218, 2024.

[24] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 154–164. [Online]. Available: https://aclanthology.org/D19-1015/

[25] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2470–2481.

[26] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2. IEEE, 2005, pp. 729–734.

[27] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *International conference on machine learning*. PMLR, 2019, pp. 2535–2544.

[28] W. Han, H. Chen, M.-Y. Kan, and 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*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10 498–10 511.

[29] S. Hangloo and B. Arora, "Multimodal fusion techniques: review, data representation, information fusion, and application areas," *Neurocomputing*, p. 130827, 2025.

[30] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2594–2604. [Online]. Available: https://aclanthology.org/D18-1280

[31] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and 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)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2122–2132. [Online]. Available: https://aclanthology.org/N18-1193/

[32] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[33] D. Hu, L. Wei, and X. Huai, "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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 7042–7052. [Online]. Available: https://aclanthology.org/2021.acl-long.547/

[34] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7037–7041.

[35] J. Hu, Y. Liu, J. Zhao, and 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)*, 2021, pp. 5666–5675.

[36] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[37] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably)," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9226–9259.

[38] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 406–29 419, 2021.

[39] A. Javaloy, M. Meghdadi, and I. Valera, "Mitigating modality collapse in multimodal VAEs via impartial optimization," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 9938–9964. [Online]. Available: https://proceedings.mlr.press/v162/javaloy22a.html

[40] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "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*, 2022, pp. 4148–4164.

[41] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[42] T. Kim and P. Vossen, "Emoberta: Speaker-aware emotion recognition in conversation with roberta," *CoRR*, vol. abs/2108.12009, 2021.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[44] . T. Lê Thuy, T. Van Loan, and N. H. Quang, "Gmm for emotion recognition of vietnamese," *Journal of Computer Science and Cybernetics*, vol. 33, no. 3, pp. 229–246, 2017.

[45] B. Li, H. Fei, L. Liao, Y. Zhao, C. Teng, T.-S. Chua, D. Ji, and F. Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5923–5934.

[46] J. Li, X. Wang, G. Lv, and Z. Zeng, "Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation," *Neurocomputing*, vol. 550, p. 126427, 2023.

[47] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, "Hitrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations," in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 4190–4200.

[48] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[49] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "Gcnet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on pattern analysis and machine intelligence*, 2023.

[50] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 150–161.

[51] X. Lin, S. Wang, R. Cai, Y. Liu, Y. Fu, Z. Yu, W. Tang, and A. Kot, "Suppress and rebalance: Towards generalized multi-modal face anti-spoofing," *arXiv preprint arXiv:2402.19298*, 2024.

[52] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 298–305.

[53] C. Liu, J. Wen, Z. Wu, X. Luo, C. Huang, and Y. Xu, "Information recovery-driven deep incomplete multiview clustering network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[54] S. Liu, L. Li, J. Song, Y. Yang, and X. Zeng, "Multimodal pre-training with self-distillation for product understanding in e-commerce," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 1039–1047.

[55] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, "Late fusion incomplete multi-view clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 10, pp. 2410–2423, 2018.

[56] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, p. 101973, 2023.

[57] H. Ma, Q. Zhang, C. Zhang, B. Wu, H. Fu, J. T. Zhou, and Q. Hu, "Calibrating multimodal learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 429–23 450.

[58] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[59] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[60] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[61] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Repre-*

sentations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[63] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[64] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[65] C.-V. T. Nguyen, A.-T. Mai, T.-S. Le, H.-D. Kieu, and 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*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 154–15 167.

[66] H. T. Nguyen *et al.*, "Combination of facial expressions and eeg for multimodal emotion recognition," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 41, no. 2, 2025.

[67] K. P.-Q. Nguyen and K. Van Nguyen, "Exploiting vietnamese social media characteristics for textual emotion recognition in vietnamese," in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 276–281.

[68] N. M. Nguyen, T. T. Nguyen, P.-N. Tran, C. P. Lim, N. T. Pham, and D. N. M. Dang, "Multimodal fusion in speech emotion recognition: A comprehensive review of methods and technologies," *Engineering Applications of Artificial Intelligence*, vol. 163, p. 112624, 2026.

[69] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, p. 126866, 2023.

[70] S. Parthasarathy and 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.

[71] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8238–8247.

[72] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[73] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and 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.

[74] R. W. Picard, *Affective computing*. MIT press, 2000.

[75] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. Mitchell, "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)*, 2019, pp. 1162–1172.

[76] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[77] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "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)*, 2017, pp. 873–883.

[78] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 527–536.

[79] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE access*, vol. 7, pp. 100 943–100 953, 2019.

[80] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[81] N. Reimers and I. Gurevych, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[82] Y. Ren, X. Chen, J. Xu, J. Pu, Y. Huang, X. Pu, C. Zhu, X. Zhu, Z. Hao, and L. He, "A novel federated multi-view clustering method for unaligned and incomplete data fusion," *Information Fusion*, vol. 108, p. 102357, 2024.

[83] E. Rossi, H. Kenlay, M. I. Gorinova, B. P. Chamberlain, X. Dong, and 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.

[84] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[85] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[86] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[87] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "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*. Springer, 2018, pp. 593–607.

[88] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[89] R. K. Shelly, "Emotions, sentiments, and performance expectations," in *Theory and research on human emotions*. Emerald Group Publishing Limited, 2004, pp. 141–165.

[90] W. Shen, J. Chen, X. Quan, and 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. 13 789–13 797.

[91] W. Shen, S. Wu, Y. Yang, and 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1551–1560. [Online]. Available: https://aclanthology.org/2021.acl-long.123

[92] A. Shenoy and A. Sardana, "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)*. Seattle, USA: Association for Computational Linguistics, Jul. 2020, pp. 19–28.

[93] Y. Shou, T. Meng, W. Ai, F. Fu, N. Yin, and K. Li, "A comprehensive survey on multi-modal conversational emotion recognition with deep learning," *ACM Transactions on Information Systems*, 2023.

[94] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, vol. 149, p. 102447, 2020.

[95] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022.

[96] T. Sun, Z. Qian, P. Li, and Q. Zhu, "Graph interactive network with adaptive gradient for multi-modal rumor detection," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 316–324.

[97] Y. Sun, S. Mai, and H. Hu, "Learning to balance the learning rates between various modalities via adaptive tracking factor," *IEEE Signal Processing Letters*, vol. 28, pp. 1650–1654, 2021.

[98] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities." in *IJCAI*, vol. 3534, 2019, p. 3540.

[99] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[100] J. Tang, K. Li, M. Hou, X. Jin, W. Kong, Y. Ding, and Q. Zhao, "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, 2022, pp. 3458–3465.

[101] N. T. H. Thi, M. D. Pham, S. T. Le, D. D. Pham, K.-T. Huynh, N. T. V. Tuyen, and T. D. Le, "Vietnamese emotion recognition from voice and text: A confidence-based approach," in *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2025, pp. 1–6.

[102] L. Tran, X. Liu, J. Zhou, and 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.

[103] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[104] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 089–19 097.

[105] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[106] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[107] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.

[108] N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, and J. Zhang, "M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation," *IEEE Transactions on Artificial Intelligence*, 2022.

[109] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.

[110] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[111] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[112] Y. Wang, Z. Cui, and Y. Li, "Distribution-consistent modal recovering for incomplete multimodal learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 025–22 034.

[113] Y. Wang, Y. Li, and Z. Cui, "Incomplete multimodality-diffused emotion recognition," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[114] Y. Wang, M. Liu, Z. Li, Y. Hu, X. Luo, and L. Nie, "Unlocking the power of multimodal learning for emotion recognition in conversation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5947–5955.

[115] J. Wen, Y. Xu, and H. Liu, "Incomplete multiview spectral clustering with adaptive graph learning," *IEEE transactions on cybernetics*, vol. 50, no. 4, pp. 1418–1429, 2018.

[116] J. Wen, Z. Zhang, Z. Zhang, L. Fei, and M. Wang, "Generalized incomplete multiview clustering with flexible locality structure diffusion," *IEEE transactions on cybernetics*, vol. 51, no. 1, pp. 101–114, 2020.

[117] J. Wen, Z. Zhang, Z. Zhang, Z. Wu, L. Fei, Y. Xu, and B. Zhang, "Dimc-net: Deep incomplete multi-view clustering network," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3753–3761.

[118] J. Wu, W. Zhuge, H. Tao, C. Hou, and Z. Zhang, "Incomplete multi-view clustering via structured graph learning," in *PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, Proceedings, Part I 15*. Springer, 2018, pp. 98–112.

[119] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24 043–24 055.

[120] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[121] B. Xu, S. Huang, M. Du, H. Wang, H. Song, C. Sha, and Y. Xiao, "Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 1855–1864.

[122] R. Xu, R. Feng, S.-X. Zhang, and D. Hu, "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, 2023, pp. 1–5.

[123] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "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*, 2021, pp. 1009–1021.

[124] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 595–11 603.

[125] L. Yang, Z. Wu, J. Hong, and J. Long, "Mcl: A contrastive learning method for multimodal data fusion in violence detection," *IEEE Signal Processing Letters*, 2022.

[126] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 7154–7163.

[127] Z. Yuan, W. Li, H. Xu, and 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.

[128] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," *Advances in neural information processing systems*, vol. 32, 2019.

[129] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, pp. 82–88, 11 2016.

[130] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[131] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[132] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and 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.

[133] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and 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.

[134] J. Zeng, T. Liu, and 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.

[135] C. Zhang, Z. Han, H. Fu, J. T. Zhou, Q. Hu *et al.*, "Cpm-nets: Cross partial multi-view networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[136] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *IJCAI*, 2019, pp. 5415–5421.

[137] D. Zhang, W. Zhang, S. Li, Q. Zhu, and G. Zhou, "Modeling both intra-and inter-modal influence for real-time emotion detection in conversations," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 503–511.

[138] K. Zhang, Z. Zhang, Z. Li, and Q. Yu, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[139] Q. Zhang, Y. Wei, Z. Han, H. Fu, X. Peng, C. Deng, Q. Hu, C. Xu, J. Wen, D. Hu *et al.*, "Multimodal fusion on low-quality data: A comprehensive survey," *arXiv preprint arXiv:2404.18947*, 2024.

[140] Y. Zhang, X. Yang, X. Xu, Z. Gao, Y. Huang, S. Mu, S. Feng, D. Wang, Y. Zhang, K. Song *et al.*, "Affective computing in the era of large language models: A survey from the nlp perspective," *CoRR*, 2024.

[141] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM computing surveys*, vol. 56, no. 9, pp. 1–36, 2024.

[142] J. Zhao, R. Li, and 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.

[143] L. Zhao, Z. Chen, Y. Yang, Z. J. Wang, and V. C. Leung, "Incomplete multi-view clustering via deep semantic mapping," *Neurocomputing*, vol. 275, pp. 1053–1062, 2018.

[144] Q. Zhao, G. Zhou, S. Xie, L. Zhang, and A. Cichocki, "Tensor ring decomposition," *arXiv preprint arXiv:1606.05535*, 2016.

[145] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5089–5097.

[146] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[147] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[148] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2411–2422, 2019.

[149] Y. Zhou, X. Liang, S. Zheng, H. Xuan, and T. Kumada, "Adaptive mask co-optimization for modal dependence in multimodal learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[150] Y. Zhou, X. Wang, H. Chen, X. Duan, and W. Zhu, "Intra-and inter-modal curriculum for multimodal learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3724–3735.