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)

SUMMARY OF DOCTOR DISSERTATION IN INFORMATION SYSTEM

Hanoi, 2026

The dissertation was completed at: University of Engineering and Technology, Vietnam National University, Hanoi

Supervisors:	Assoc.Prof. Ha Quang Thuy	University of Engineering and Technology, Vietnam National University, Hanoi
Co-supervisors:	Dr. Le Duc Trong	University of Engineering and Technology, Vietnam National University, Hanoi
Reviewer 1:	Assoc.Prof. Bui Thu Lam	Academy of Cryptography Techniques
Reviewer 2:	Assoc.Prof. Le Hong Phuong	University of Science, Vietnam National University, Hanoi
Reviewer 3:	Assoc.Prof. Nguyen Duc Dung	Institute of Information Technology, Vietnam Academy of Science and Technology

The dissertation will be defended before the Vietnam National University-level PhD Thesis Evaluation Committee at the University of Engineering and Technology, Vietnam National University, Hanoi (VNU-UET) at ……..…., …… 2026

The dissertation can be found at:

- National Library of Vietnam

- Information Center - Library, Vietnam National University, Hanoi

Abstract

A central challenge in multimodal emotion recognition in conversation (MERC) is to design fusion models that can 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 degraded by missing modalities and imbalanced modality contributions, which substantially undermine the robustness and generalization of existing MERC systems.

This dissertation tackles these challenges by developing advanced deep multimodal fusion models specifically tailored for emotion recognition in conversation under incomplete and imbalanced modalities. The first part focuses on structured multimodal conversational modeling under full-modality conditions, 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, thereby improving emotion recognition in complex multi-speaker dialogues.

The second part of the dissertation addresses robust and balanced multimodal fusion under low-quality data conditions, where modality completeness and balance cannot be assumed. To cope with the missing modality problem, we propose Mi-CGA, a two-stage graph-based framework that first builds incomplete multimodal representations and then employs a Cross-modal Graph Attention Network (CGA-Net) with modality feature estimation, graph attention, and cross-modal attention to reconstruct

missing cues and propagate complementary information across modalities. To tackle modality imbalance, we develop two complementary strategies for MERC. Ada2I introduces Adaptive Feature Weighting (AFW) and Adaptive Modality Weighting (AMW) to dynamically re-balance feature- and modality-level contributions during fusion, while Self-Paced Curriculum Learning (SPCL) serves as a plug-and-play training scheme that gradually schedules samples based on dual-level difficulty measures, stabilizing multimodal learning under heterogeneous modality qualities.

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

The rapid growth of online communication platforms has transformed how people interact, share information, and express emotions, making digital conversations a central medium of everyday affective experience. In these settings, human emotions are conveyed through the coordinated use of language, voice, and facial expressions, which motivates Multimodal Emotion Recognition in Conversation (MERC) as a key task in affective computing. MERC aims to automatically infer the emotion of each utterance in a dialogue by jointly modeling multimodal signals together with conversational context and speaker interactions.

Effective multimodal fusion lies at the core of MERC, since textual, acoustic, and visual signals are structurally heterogeneous yet complementary in how they encode emotional cues. Existing deep models have advanced fusion for conversational emotion recognition, but many are designed under idealized conditions where all modalities are fully observed and equally reliable, and where conversational structure is only partially modeled. In real-world scenarios, however, multimodal conversational data are often low-quality: some modalities may be noisy, occluded, or temporally misaligned, and others may be partially or entirely missing or systematically weaker than the rest. These imperfections cause fusion models to overfit dominant modalities, rely on spurious correlations, and degrade sharply when modality availability or quality changes.

These observations motivate the central focus of this dissertation on advanced deep multimodal fusion for MERC under incomplete and imbalanced modalities. On the one hand, MERC requires structured conversational fusion that can integrate heterogeneous modalities while explicitly modeling temporal dependencies and speaker relations in dialogue. On the other hand, practical MERC systems must remain robust when modalities are missing and stable when modality contributions are imbalanced, avoiding dominance of a single channel and preserving useful signals from weaker modalities.

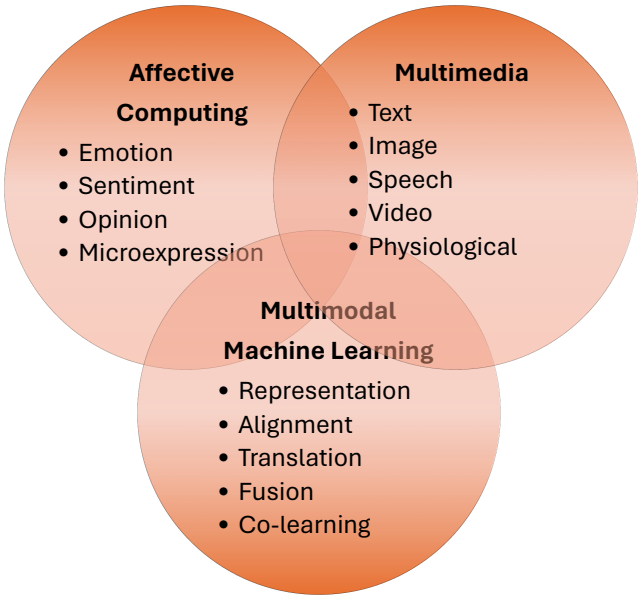

Figure 1: Overview of framework for Multimodal Affective Analytics

Within this context, the dissertation pursues two primary research directions: (1) designing multimodal fusion architectures that jointly model utterance-level features, cross-modal interactions, and conversational context for MERC; and (2) developing robust and balanced fusion strategies that handle incomplete and imbalanced modalities without sacrificing performance or generalization. Together, these directions aim to advance the reliability of multimodal fusion for emotion recognition in realistic conversational environments and to support the development of emotionally intelligent systems in applications such as human–computer interaction, social media analysis, and mental health support.

Research Scope, Objectives, and Questions

Existing approaches in Multimodal Emotion Recognition in Conversation (MERC) often struggle to (1) effectively align and fuse heterogeneous textual, acoustic, and visual features, (2) capture temporal and speaker-dependent conversational context, and (3) remain robust under low-quality multimodal conditions such as incomplete and imbalanced modalities. To address these gaps, this dissertation narrows its focus to advanced deep multimodal fusion for MERC under realistic data imperfections.

The scope of this dissertation is defined along two complementary dimensions: task scope and methodological scope. In terms of **task scope**, the dissertation concentrates on MERC, where the goal is to infer utterance-level emotions in multi-speaker

dialogues by jointly modeling multimodal signals and conversational structure. In terms of **methodological scope,** the work develops deep learning models for multimodal fusion that (i) learn expressive multimodal representations tailored to conversational data, and (ii) ensure robustness and balance when modalities are missing or unequally informative.

Within these scopes, the dissertation is organized around two core research objectives (O) and their associated research questions (RQ):

- **O1**: Multimodal fusion and contextual modeling for MERC: This objective aims to design multimodal fusion architectures specifically adapted to conversational emotion recognition, focusing on how to integrate heterogeneous modality-specific cues while modeling dialogue dynamics.
  - **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 research questions is solved in **Chapter** 2.

- **O2**: Robust and balanced multimodal fusion for MERC: This objective focuses on improving the robustness and reliability of MERC models when multimodal data are imperfect, particularly under incomplete and imbalanced modality conditions.
  - **RQ3**: In MERC scenarios with incomplete modalities, how can missing multimodal features be effectively compensated and complementary information propagated across modalities to maintain robust multimodal representations? *This research question is solved 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 solved in* **Chapter** 4.

Significance and Contribution

Our main contributions stem from a unified research focus on advanced deep multimodal fusion for MERC, with particular emphasis on structured conversational modeling and robustness under incomplete and imbalanced modalities. Specifically:

1. **Structured multimodal conversational fusion**: We propose **CORECT** [VanNTC 1] and **MultiDAG+CL** [VanNTC 2] as fusion-centric MERC models that jointly consider multimodal features and dialogue structure. CORECT introduces relational–temporal graph modeling with auxiliary cross-modal interaction to capture utterance-level temporal dependencies and conversation-level cross-modal relations, while MultiDAG+CL leverages Directed Acyclic Graph–based contextual reasoning with curriculum learning to better handle emotional shifts and varying sample difficulty in multi-speaker dialogues.

2. **Robust fusion under incomplete modalities**: We develop **Mi-CGA** [VanNTC 3], a graph-based framework designed for MERC with missing modalities. By combining an Incomplete Multimodal Representation (IMR) stage with a Cross-modal Graph Attention Network (CGA-Net), Mi-CGA reconstructs missing modality features and propagates complementary information across modalities, enabling stable fusion when one or more modalities are partially or entirely unavailable.

3. **Balanced fusion under modality dominance**: We address modality imbalance in MERC through two complementary fusion-aware learning strategies. **Ada2I** [VanNTC 4] introduces Adaptive Feature Weighting (AFW) and Adaptive Modality Weighting (AMW) guided by a discrepancy ratio to dynamically re-balance feature- and modality-level contributions during fusion, while **Self-Paced Curriculum Learning (SPCL)** [VanNTC 5] acts as a lightweight plug-and-play module that progressively schedules training samples based on dual-level difficulty measures, stabilizing multimodal fusion across backbones.

Collectively, these contributions provide a coherent multimodal fusion perspective for MERC, spanning structured conversational fusion, robustness to missing modalities, and balanced learning under modality dominance in realistic conversational settings.

Chapter 1

Advanced Deep Learning for Multimodal Emotion Recognition

Emotion and Multimodal Emotion Recognition. Emotions are discrete affective states triggered by specific events and expressed through coordinated changes in experience, behavior, and physiology. In computational settings, emotion recognition aims to assign such states to data segments based on observable signals. This dissertation focuses on Multimodal Emotion Recognition in Conversation (MERC), where the goal is to predict an emotion label for each utterance in a multi-speaker dialogue given textual, acoustic, and visual inputs. A conversation is modeled as a sequence $C = \{u_1, \ldots, u_N\}$ of utterances with associated speakers and modality-specific features, and MERC seeks a function that maps these multimodal conversational inputs to utterance-level emotion labels while exploiting both local content and dialogue context.

Multimodal Machine Learning. Multimodal machine learning studies how to represent, align, and combine heterogeneous data sources such as language, audio, and vision within a unified framework. Central challenges include learning joint or coordinated representations, aligning asynchronous signals in time, and enabling translation or reasoning across modalities. For affective computing, and MERC in particular, these challenges are amplified by the strong dependence of emotional meaning on context, speaker identity, and interaction history. As a result, effective multimodal methods must simultaneously handle modality-specific structures and higher-level conversational dynamics.

Multimodal Data Fusion. Multimodal data fusion concerns how information from different modalities is integrated to support downstream tasks. Classical strategies range

from early fusion, where raw or low-level features are concatenated, to intermediate fusion, where shared latent spaces and cross-modal interactions are learned, and late fusion, where modality-specific predictions are combined. In practice, fusion must balance preserving discriminative properties of each modality with exploiting their complementary strengths. In conversational settings, this is further complicated by low-quality data: some modalities can be noisy, partially missing, or systematically weaker than others. Consequently, modern fusion mechanisms for MERC increasingly incorporate temporal modeling, attention-based cross-modal interaction, and robustness to incomplete or imbalanced modality conditions.

Multimodal Emotion Recognition in Conversation. Multimodal Emotion Recognition in Conversation (MERC) formalizes emotion prediction at the utterance level while explicitly accounting for dialogue structure. A conversation is typically modeled as a sequence $C = \{u_1, \ldots, u_N\}$, where each utterance u_i is associated with a speaker p_i, a set of modality-specific features $\{x_i^m\}_{m \in M}$ (e.g., text, audio, visual), and an emotion label $y_i \in Y$. The MERC task is to learn a function:

$$f : (C, \{x_i^m\}_{i=1..N,\, m \in M}) \rightarrow \{y_1, y_2, \ldots, y_N\},$$

that maps multimodal conversational inputs to utterance-level emotion labels while exploiting both local content and dialogue-level context.

Task formulation typically assumes that each utterance is annotated with an emotion label from a fixed set, and models must exploit both its multimodal content and dependencies across neighboring turns and speakers. Early modeling paradigms rely on recurrent or Transformer-based architectures to capture context, whereas more recent approaches employ graph-based structures to represent utterance-utterance and speaker-speaker relations, often combined with attention to highlight salient cues. Across these paradigms, a central theme is how to couple expressive multimodal fusion with structured conversational modeling, and how to maintain performance when modalities are noisy, missing, or imbalanced.

Dataset and Evaluation Metrics This dissertation evaluates MERC models on multimodal datasets, primarily IEMOCAP, CMU-MOSI and CMU-MOSEI. We report utterance-level accuracy and weighted F1-scores to account for label imbalance.

Chapter 2

Multimodal Fusion for Emotion Recognition in Conversation

This chapter focuses on advancing multimodal fusion for MERC, where models must effectively reason over textual, acoustic, and visual cues within structured dialogues. Emotions in conversation evolve across turns and depend on contextual flow and speaker interactions, making MERC more challenging than isolated utterance classification. Existing recurrent, graph-based, and fusion-based approaches capture parts of this structure but often struggle with long-range dependencies, modality imbalance, and the preservation of modality-specific information.

To address these limitations under full-modality conditions, we propose **CORECT** [**VanNTC 1**], a relational-temporal graph-based framework that jointly models conversational context and cross-modal interactions. CORECT integrates a Relational Temporal Graph Convolutional Network (RT-GCN) to capture utterance-level temporal and speaker-dependent relations, together with a Pairwise Cross-modal Feature Interaction module (P-CM) that explicitly models interactions between modalities at the conversation level. This design aims to preserve modality-specific cues while enhancing cross-modal fusion in MERC.

Building on this, we further introduce **MultiDAG+CL** [**VanNTC 2**], which extends Directed Acyclic Graph (DAG)–based reasoning to multimodal inputs and incorporates a curriculum learning strategy. MultiDAG models information flow along directed conversational structures to better capture long-range dependencies and emotional shifts, while curriculum learning progressively exposes the model to more difficult conversations.

2.1 Multimodal Fusion with Relation and Temporal Conversational Modeling

CORECT integrates a Relational Temporal Graph Convolution Network (RT-GCN) with a Pairwise Cross-modal Feature Interaction (P-CM) module for multimodal ERC.

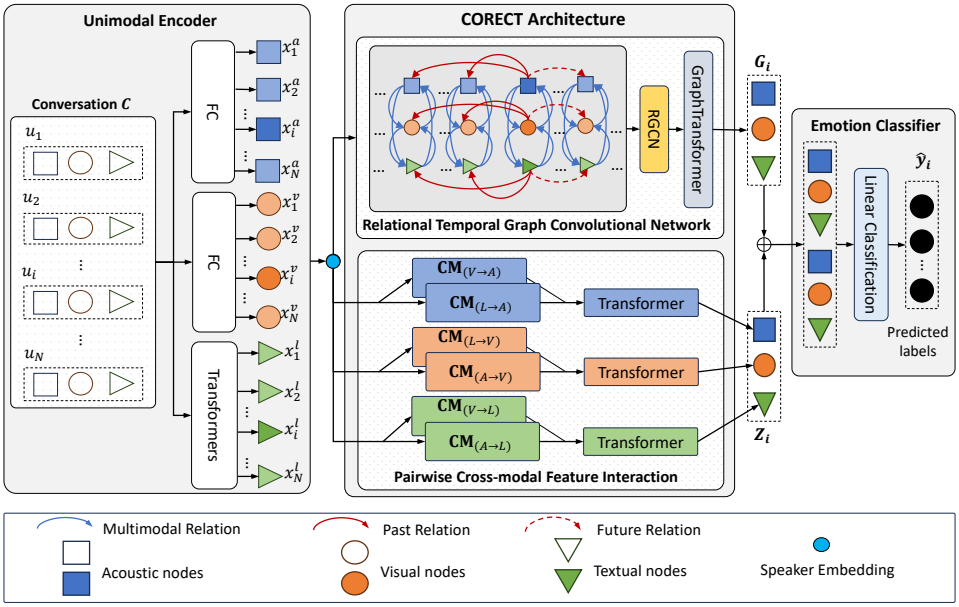

Figure 2.2: Framework illustration of CORECT

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

The CORECT framework processes a dialogue by first extracting **utterance-level unimodal features**. For each utterance u_i, we obtain modality-specific embeddings: $x_i^a,\ x_i^v,\ x_i^l$ for acoustic, visual, and lexical modalities respectively (via FC layers for a, v and a Transformer encoder for l). To incorporate speaker information, each modality embedding is enhanced using the speaker vector s_i, producing $X_i^\tau = x_i^\tau + \eta s_i$,

To capture local conversational structure, CORECT constructs a multimodal graph $\mathcal{G}(\mathcal{V}, \mathcal{R}, \mathcal{E})$, where each utterance yields three nodes $u_i^a,\ u_i^v,\ u_i^l$, and edges encode (1) *multimodal relations* $\mathcal{R}_{multi}$ linking different modalities of the same utterance, and (2) *temporal relations* $\mathcal{R}_{temp}$ linking past and future utterances within a window. A Relational Temporal GCN (RT-GCN) produces contextualized representations: $g_i^\tau,\quad \tau \in$

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** ($\uparrow$ 2.89)	**70.02** ($\uparrow$ 2.75)

$\{a, v, l\}$. A Graph Transformer layer further refines these representations, giving o_i^τ. To capture global cross-modal dependencies, CORECT applies a pairwise cross-modal transformer. Given two modalities α and β, the module computes attention-based interactions such as $\text{CM}^{\beta\to\alpha}$, and stacks D layers to obtain the final cross-modal features $Z_{\alpha\rightleftarrows\beta}^{[D]}$. This process is repeated for all modality pairs (a, v, l). The final representation for each utterance is formed by fusing: $H_i = \text{Fusion}(o_i^\tau,\ Z_{\alpha\rightleftarrows\beta}^{[D]})$. A classifier predicts the emotion label: $\hat{y}_i = \arg\max(\text{softmax}(W H_i + b))$.

Results. On IEMOCAP, CORECT consistently outperforms prior baselines, achieving notable gains in across 6-way and 4-way settings. On CMU-MOSEI, CORECT likewise demonstrates strong performance across sentiment (2-class and 7-class) and emotion classification tasks, matching or surpassing state-of-the-art models. These results collectively confirm the robustness and effectiveness of CORECT for multimodal conversation understanding.

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	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE	82.40	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN	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**

2.2 Multimodal Fusion with Directed Acyclic Graph Modeling and Curriculum Learning

Given a conversation $C = \{u_1, \ldots, u_N\}$ where each utterance u_i is represented by multimodal features $u_i = \{u_i^a, u_i^v, u_i^l\}$, the goal of MultiDAG+CL is to perform multi-

modal ERC by combining **(1) MultiDAG**, a directed acyclic graph model for contextual reasoning, and **(2) Curriculum Learning (CL)**, a difficulty-aware training strategy.

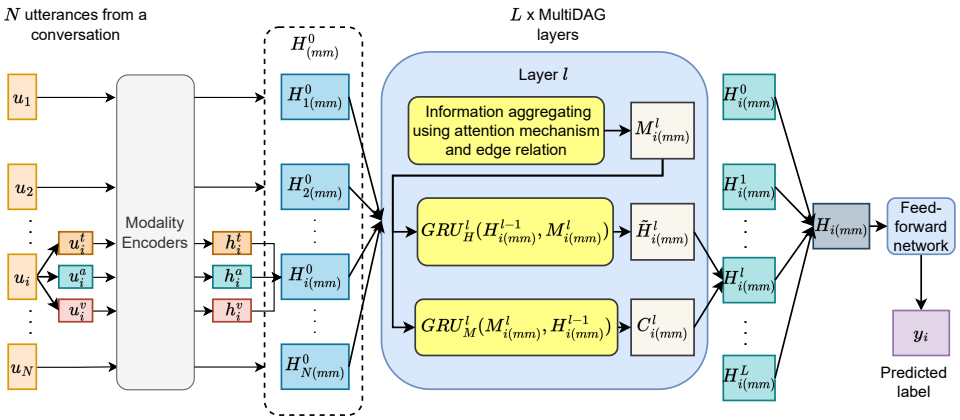

Figure 2.7: Overall structure of MultiDAG.

Each modality is encoded using modality-specific encoders: $h_i^a = Enc_A(u_i^a)$, $h_i^v = Enc_V(u_i^v)$, $h_i^l = Enc_L(u_i^l)$, and concatenated into the multimodal utterance representation $H_{i(mm)}^0 = h_i^a \oplus h_i^v \oplus h_i^l$.

The conversation is modeled as a DAG in which each utterance receives information from its predecessors only. A DAG-GNN iteratively updates the hidden states layer by layer through attention over preceding nodes and GRU-based integration, yielding contextualized representations $H_{i(mm)} = \|_{l=0}^{L}(\tilde{H}_{i(mm)}^l + C_{i(mm)}^l)$. This captures both local and long-range dependencies along the directed conversational flow. The final representation $H_{i(mm)}$ is passed to a Feed-Forward Network for emotion prediction.

Difficulty Measure. A Difficulty Measure Function (DMF) computes the difficulty of each conversation based on the frequency of *emotion shifts*: $\mathcal{DIF}(c_i) = \frac{N_{\text{shift}}(c_i)+N_{\text{sp}}(c_i)}{N_u(c_i)+N_{\text{sp}}(c_i)}$. Higher emotional variability implies a more challenging conversation.

Training Scheduler. Conversations are sorted by difficulty and split into ordered bins $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$. Training begins with the easiest bin and progressively incorporates harder bins, forming a curriculum: $\mathcal{D}^{train} \leftarrow \mathcal{D}_1 \rightarrow \mathcal{D}_1 \cup \mathcal{D}_2 \rightarrow \cdots \rightarrow \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_k$. After all bins are included, standard training continues. We evaluate our approach on the following two ERC datasets: IEMOCAP and MELD .

Results. Table 2.8 presents a comprehensive comparison against state-of-the-art multimodal ERC models. Our proposed *MultiDAG+CL* achieves the best overall performance on both IEMOCAP and MELD, surpassing previous SOTAs by 1.05% and 0.34%, respectively. The model yields consistent gains across most emotion categories, particularly for *Sad*, *Neutral*, and *Angry*.

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	-	-
MultiDAG (Ours)	49.65	79.83	66.40	67.59	71.78	67.90	68.30	68.45	64.29	63.87
MultiDAG+HCL (Ours)	45.26	**81.40**	**69.53**	**70.33**	71.61	66.94	**69.11**	**69.08**	**64.41**	**64.00**

2.3 Chapter Summary

In this chapter, we addressed the dissertation's first research focus: multimodal fusion and contextual modeling for Emotion Recognition in Conversation. Section 2.1 introduced **CORECT**, which models relational and temporal conversational structure through RT-GCN and the P-CM module to enhance cross-modal fusion under full-modality conditions. Section 2.2 presented **MultiDAG+CL**, which combines DAG-based context modeling with curriculum learning to better handle emotional shifts and varying dialogue difficulty.

Experiments on IEMOCAP and CMU-MOSEI (for CORECT) and on IEMOCAP and MELD (for MultiDAG+CL) demonstrate consistent gains over strong baselines, directly supporting **Objective O1** by providing structured multimodal conversational fusion for MERC. Collectively, these methods establish the foundational fusion and context-modeling capabilities of the dissertation and pave the way for the next chapter, which investigates robust multimodal fusion under real-world low-quality data conditions, including incomplete and imbalanced modalities.

Chapter 3

Multimodal Emotion Recognition in Conversation under Incomplete Modality Condition

Introduction. Human communication naturally integrates language, voice, facial expressions, and other behavioral cues, but real-world affective systems rarely observe all modalities reliably. In conversational settings, multimodal streams are often partially missing, corrupted, or asynchronously recorded, making incomplete multimodal data a pervasive challenge for MERC. Nevertheless, many existing models still assume full modality availability or treat missingness with simple heuristics.

Current solutions for missing modalities fall into a few main categories but remain limited for conversational affective modeling. Typical methods assume that at most one modality is missing, discard incomplete samples, or overlook conversational structure, and only a few jointly consider feature reconstruction, cross-modal reasoning, and context modeling. As a result, rich relational dependencies in dialogue are not fully exploited when modalities are absent.

To overcome these challenges, we introduce **Mi-CGA** (Figure 3.2), a unified graph-driven framework for robust multimodal ERC under incomplete modalities. Mi-CGA operates in two stages: an Incomplete Multimodal Representation (IMR) stage that simulates missing patterns and learns modality-robust representations for arbitrary modality subsets, followed by a Cross-modal Graph Attention Network (CGA-Net) that performs context-sensitive multimodal reasoning over conversations. This study advances **Objective O2** by improving robustness to low-quality multimodal data with missing modalities, directly contributing to **RQ3** through graph- and attention-based feature reconstruction and more stable multimodal learning under variable-modality conditions.

Mi-CGA: Cross-Modal Graph Attention Network for Robust Emotion Recognition in the Presence of In- complete Modalities. Let us consider a conversation C of a sequence of N utterances $\{U_1, U_2, \ldots, U_N\}$. Each utterance U_i in C associates with three data modalities including audio (a), visual (v), and text (t), i.e., $U_i = \{u_i^{(a)}, u_i^{(v)}, u_i^{(t)}\}$. Suppose that there exists an overall random missing rate of $\rho \in [0, 1]$ on multimodal information, let us denote $x_i^{(a)miss} \in \mathbb{R}^{d_a}$, $x_i^{(v)miss} \in \mathbb{R}^{d_v}$, and $x_i^{(t)miss} \in \mathbb{R}^{d_t}$ are incomplete representation of $u_i^{(a)}, u_i^{(v)}$ and $u_i^{(t)}$ with latent dimensions d_a, d_t, d_v. Considering the multimodal emotion recognition task in the presence of incomplete modalities, Mi-CGA takes the tuple $(x_i^{(a)miss}, x_i^{(v)miss}, x_i^{(t)miss})$ as input and seeks to predict the corresponding emotion label $\hat{y}_i$ of U_i from a predefined emotion label set $E = \{y_1, y_2, \ldots, y_{|E|}\}$.

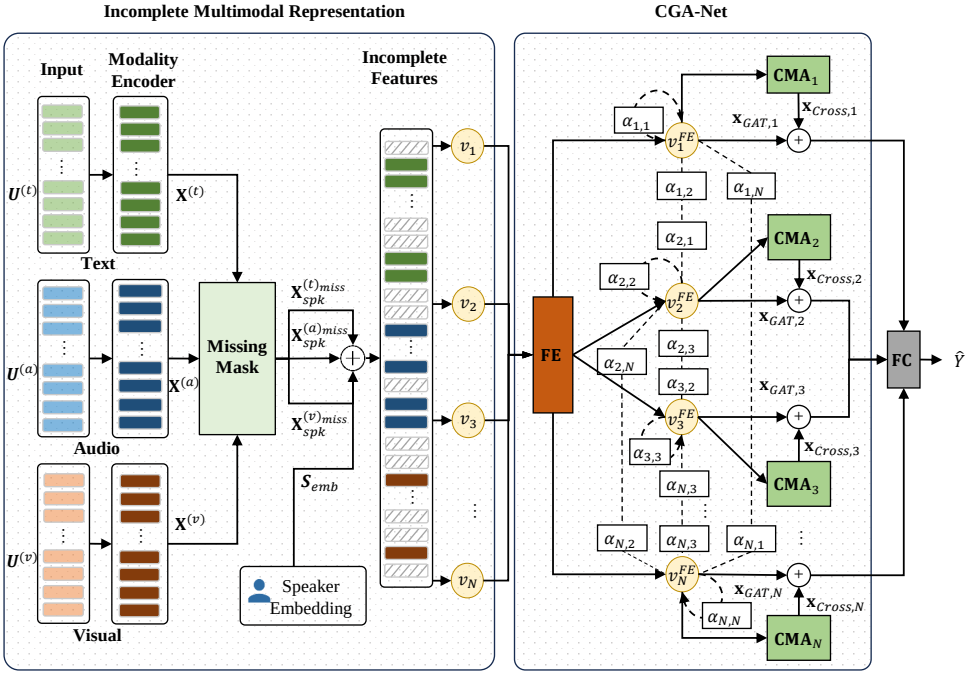

Figure 3.2: Overall Architecture of Mi-CGA model.

Mi-CGA is designed to address multimodal emotion recognition in conversations under incomplete-modality conditions. Given a conversation with N utterances $\{U_1, \ldots, U_N\}$, where each utterance U_i contains audio, visual, and textual features $U_i = \{u_i^{(a)}, u_i^{(v)}, u_i^{(t)}\}$, a global missing-rate ρ is applied to simulate incomplete multimodal inputs the model receives at training and inference time. The resulting masked tuple $(x_i^{(a)miss}, x_i^{(v)miss}, x_i^{(t)miss})$ is used to predict the emotion label $\hat{y}_i$ from a predefined label set E.

The model comprises two major stages. The first stage, *Incomplete Multimodal Representation (IMR)*, produces robust utterance-level features under arbitrary missing-modality patterns. Modality-specific encoders generate contextual embeddings $x_i^{(a)}$, $x_i^{(v)}$, and $x_i^{(t)}$, where BiLSTM is used for text and fully connected networks for audio

and visual signals. A missing mask $\mathcal{M}$ ensures that missing patterns follow the global rate ρ, producing masked features $\mathbf{X}^{miss} = \mathbf{X} \odot \mathcal{M}$. Speaker embeddings S_{emb} are then concatenated with $\mathbf{X}^{miss}$ to form enhanced representations $\mathbf{X}^{miss}_{spk} = \eta S_{emb} \oplus \mathbf{X}^{miss}$.

The second stage, *Cross-modal Graph Attention Network (CGA-Net)*, integrates reconstructed features, conversational structure, and cross-modal interactions. First, the *Feature Estimation (FE)* module reconstructs missing modalities using a GCN-based encoder f_ϕ and an MLP decoder g_θ, producing a coarse reconstruction $\mathbf{X}^{coarse} = g_\theta(f_\phi(\mathbf{X}^{miss}_{spk}, \tilde{\mathbf{A}}))$. A smoothing step combines reconstructed and observed values to obtain final imputed features $\mathbf{X}^{FE} = (1-\lambda)\mathbf{X}^{coarse} + \lambda \mathbf{X}^{miss}_{spk}$. Next, a multi-head graph attention network (*MulGAT*) models intra-conversation dependencies. For each utterance node, attention coefficients α_{ij} are computed over neighboring utterances, and multi-head aggregation yields a topology-aware representation $\mathbf{X}_{GAT} = \text{MulGAT}(\mathbf{X}^{FE})$. To capture fine-grained cross-modal interactions, the *Cross-modal Attention (CMA)* module applies bidirectional scaled dot-product attention between every modality pair, generating representations such as $\mathbf{X}^{a \rightleftarrows t}$, $\mathbf{X}^{t \rightleftarrows v}$, and $\mathbf{X}^{v \rightleftarrows a}$, which are concatenated into a global cross-modal feature vector $\mathbf{X}_{Cross}$.

The final utterance representation is formed by concatenating the graph-based and cross-modal features as $\mathbf{X}_{Final} = [\mathbf{X}_{GAT}, \mathbf{X}_{Cross}]$, and a feed-forward classifier with softmax is used to produce emotion predictions $\hat{y}_i$. The model is optimized using a dual-loss objective $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rct}$, where $\mathcal{L}_{cls}$ is cross-entropy over emotion labels, and $\mathcal{L}_{rct}$ is a reconstruction loss computed either via mean squared error or KL-divergence at masked positions. This formulation jointly enhances classification robustness and missing-modality recovery, enabling Mi-CGA to perform reliably under diverse incomplete-modality scenarios.

Results. Table 3.2 shows that Mi-CGA consistently surpasses all SOTA baselines across multiple datasets under missing-modality conditions. On IEMOCAP (4-way), Mi-CGA achieves a +6.30% improvement in average w-F1 over GCNet, and on IEMOCAP (6-way), it sets a new SOTA with 62.43% accuracy (+6.25%). Similar gains on CMU-MOSI and CMU-MOSEI further confirm the robustness of Mi-CGA for incomplete multimodal ERC. Across missing-rate settings from complete data to severe incompleteness, Mi-CGA consistently outperforms prior methods. Moreover, Mi-CGA suffers substantially smaller performance degradation as ρ increases. These trends hold across all datasets, confirming the robustness of Mi-CGA under both complete and incomplete multimodal conditions.

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

3.1 Chapter summary

In this chapter, we addressed the challenge of learning from incomplete modalities in low-quality multimodal data by introducing Mi-CGA, a dedicated framework for Multimodal Emotion Recognition in Conversations (MER). By jointly leveraging intra-modal and cross-modal structure through graph-based reasoning, Mi-CGA directly advances **Objective O2** and addresses **RQ3** by providing a principled solution for reconstructing missing modality features and enabling stable affective understanding from incomplete multimodal data.

Chapter 4

Multimodal Emotion Recognition in Conversation under Imbalanced Modality Condition

The second focus of this dissertation on learning from low-quality multimodal data is **balanced multimodal learning**, which aims to ensure that all modalities contribute meaningfully during training rather than allowing one dominant modality to suppress others. This challenge—widely referred to as *modality imbalance*—is prevalent in multimodal ERC and related tasks, where textual signals often dominate optimization while acoustic and visual modalities remain under-optimized. To address this challenge, this dissertation introduces two complementary approaches: (1) **Ada2I**, an end-to-end model that re-balances multimodal learning at both the feature and modality levels; and (2) **SPCL**, a lightweight plug-and-play module that progressively mitigates modality imbalance and can be seamlessly integrated into various multimodal architectures.

Together, these contributions directly support **Objective O2**, which focuses on developing robust multimodal learning strategies for low-quality or imperfect data. Specifically, they tackle two key aspects: (i) by enhancing or recovering under-optimized modality features, they address **RQ3** on improving multimodal representations under degraded or incomplete signals; and (ii) by introducing adaptive, curriculum-based balancing mechanisms across feature, modality, emotion, and dialogue levels, they improve model stability and generalization as required by **RQ4**.

These works have been validated through publications in *ACM Multimedia 2024 (CORE A*)* for Ada2I [**VanNTC 4**], and an extended study on modality balancing using self-paced curriculum learning (SPCL), currently under major revision at *Neural Computing and Applications (Scopus Q1)* [**VanNTC 5**].

4.1 Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition

Ada2I is an end-to-end framework designed to mitigate *modality imbalance* in multimodal ERC by ensuring that all modalities contribute proportionally during representation learning. The model consists of three main stages: (1) *Modality Encoding*, (2) *Adaptive Feature Weighting (AFW)*, and (3) *Adaptive Modality Weighting (AMW)*, followed by a training strategy that dynamically balances gradients across modalities.

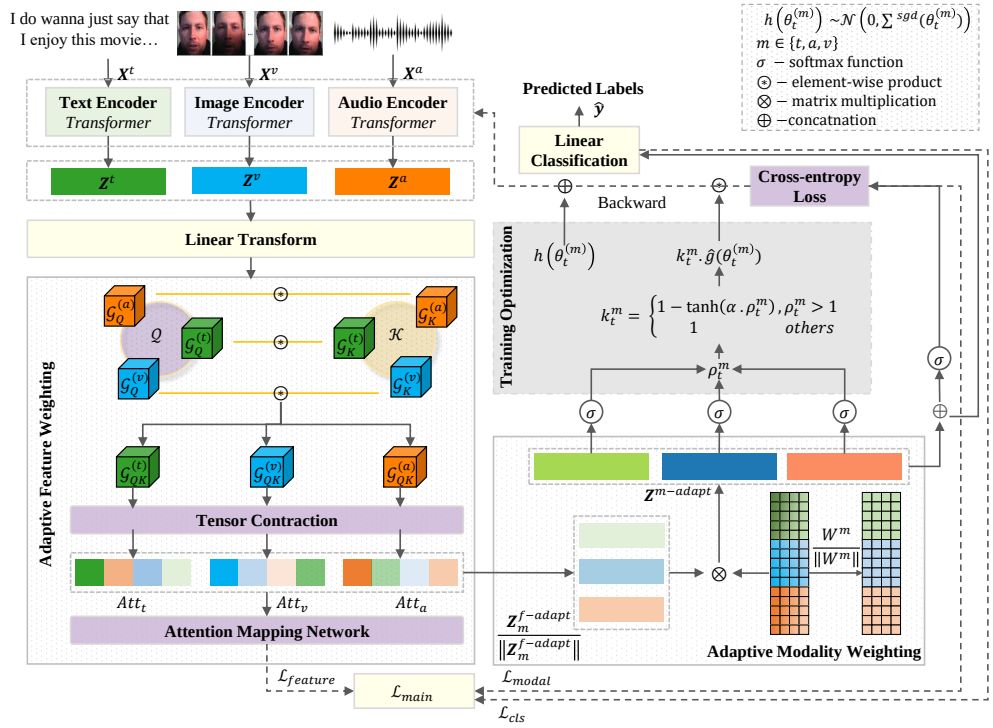

Figure 4.2: Illustration of Ada2I framework

Given a conversation with unimodal sequences $\mathbf{X}^m$ for $m \in \{t, a, v\}$, a Transformer encoder $\phi(\theta^{(m)})$ produces context-aware modality representations $\mathbf{Z}^m = \phi(\theta^{(m)}, \mathbf{X}^m)$. These serve as the initial unimodal features for subsequent balancing.

To correct imbalances arising at the *feature level*, Ada2I models multi-way cross-modal interactions using tensor-ring decomposition. Each modality yields tensor-ring query and key cores $\mathcal{G}_Q^m$ and $\mathcal{G}_K^m$, which are fused into higher-order tensors $\mathcal{Q}$ and $\mathcal{K}$. Attention coefficients Θ^m are computed via scaled element-wise interactions and summarized into pooling matrices $\mathbf{A}^{(m)}$. These yield feature-aware attention weights Att_m, and the balanced feature representation is computed as $\mathbf{Z}_m^{f\text{-}adapt} = Att_m \mathbf{Z}^m + \beta \mathbf{Z}^m$, where β regulates residual contribution from the original features.

To prevent dominant modalities (e.g., text) from overwhelming weaker ones, Ada2I applies modality-wise L2 normalization and cosine-based reweighting. For each modality m, normalized logits are computed using $\mathbf{Z}_m^{f\text{-}adapt}$ and output weights W^m, and the balanced multimodal representation is aggregated as $\mathbf{Z}^{m\text{-}adapt} = \sum_m \frac{W^m \mathbf{Z}_m^{f\text{-}adapt}}{\|W^m\| \, \|\mathbf{Z}_m^{f\text{-}adapt}\|} + b$. The resulting vector is fed into an MLP to predict the emotion label $\hat{y}$.

Ada2I uses a joint loss combining classification loss $\mathcal{L}_{cls}$, feature-level alignment loss $\mathcal{L}_{feature}$, and modality-level balancing loss $\mathcal{L}_{modal}$. To address optimization imbalance, Ada2I extends the discrepancy ratio to three modalities by computing s_t^m (from cosine similarities between logits and features) and evaluating the relative dominance $\rho_t^m = s_t^m / \min_j s_t^j$. A modulation coefficient $k_t^m = 1 - \tanh(\alpha \rho_t^m)$ suppresses gradients of dominant modalities during encoder updates, ensuring that weaker modalities receive sufficient optimization throughout training.

The balanced multimodal representation $\mathbf{Z}^{m\text{-}adapt}$ is passed through an MLP to obtain the predicted emotion label $\hat{y}$. Together, AFW, AMW, and gradient modulation enable Ada2I to maintain stable, balanced multimodal learning and improve emotion recognition performance across all modalities.

Results. As shown in Table 4.2, Ada2I consistently surpasses previous SOTA baselines across all modality combinations on both datasets. Notably, for the weakest modality pair (A+V) on MELD, Ada2I achieves substantial gains of 10.77% in w-F1 and 6.98% in Accuracy, significantly narrowing the gap to text-dominant settings.

Table 4.2: Comparison of results in the multimodal setting of Ada2I with the modality-balanced baseline model enhanced by FAGM (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	64.36	64.39	61.25	62.23	49.20	49.85
BiDDIN†	58.81	58.84	58.88	58.16	59.04	58.96	46.36	46.77
MM-DFN†	64.92	64.57	63.91	64.20	61.02	60.60	54.48	55.03
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**
Δ(%)	*↑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	44.39	48.62
MM-DFN†	55.75	60.80	57.10	60.00	57.73	60.65	42.05	48.66
MMGCN†	58.48	61.15	57.59	60.69	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**
Δ(%)	*↑1.90*	*↑1.88*	*↑2.49*	*↑1.95*	*↑0.89*	*↑1.30*	*↑10.77*	*↑6.98*

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

Several approaches have attempted to mitigate modality imbalance, ranging from improving unimodal feature quality to adding auxiliary objectives or balancing gradients across modalities. However, these methods often require heavy pre-training, task-specific losses, or complex gradient manipulations, limiting their generalizability and scalability. To overcome these limitations, we introduce **Self-Paced Curriculum Learning (SPCL)**, a simple yet adaptive training strategy that balances modality contributions by progressively guiding the model from easy to harder samples. SPCL consists of: (1) a **Difficulty Measurer** that evaluates sample complexity using recognition performance and modality discrepancy, and (2) a **Learning Scheduler** that adaptively selects samples to ensure weaker modalities receive adequate optimization. By integrating these components, SPCL offers a unified, architecture-agnostic solution for addressing modality imbalance, improving training stability and enhancing contributions from under-optimized modalities. Figure 4.7 presents the overall SPCL pipeline and its seamless integration with existing MER models.

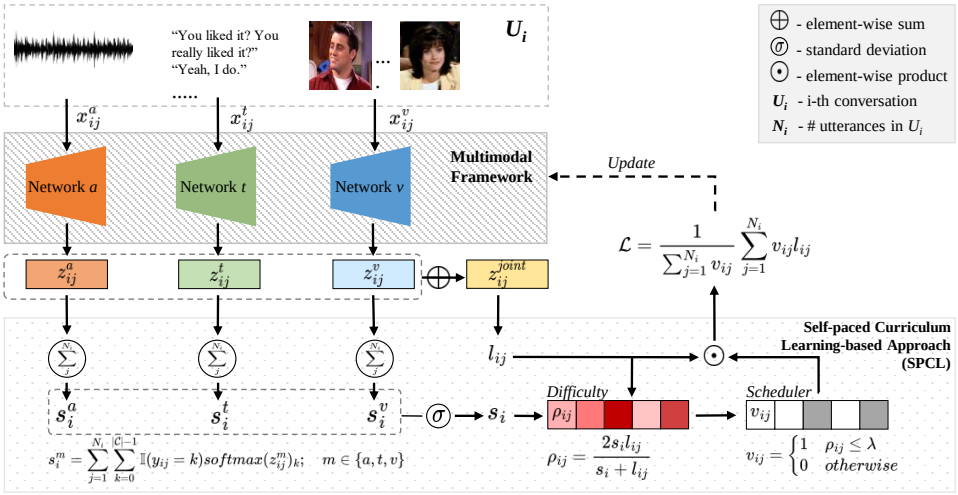

Figure 4.7: Our framework pipeline with integrated SPCL module.

Given a multimodal conversation dataset $\mathcal{D}$ and emotion category set $\mathcal{C}$, each utterance x_{ij} is represented through three modalities (audio, text, visual). Our method first derives uni-modal logits via modality-specific predictors, then fuses them into a cross-modal logit for emotion recognition. To further address modality imbalance during train-

ing, we incorporate a Self-Paced Curriculum Learning (SPCL) module that dynamically selects training samples based on their difficulty, encouraging balanced optimization across modalities.

For each utterance x_{ij}, uni-modal predictors generate logits $z_{ij}^m = \phi_m(x_{ij}^m; \theta^m)$ for $m \in \{a, t, v\}$. The final cross-modal logit is the sum of uni-modal logits, $z_{ij}^{joint} = \sum_m z_{ij}^m$, and the utterance-level loss is computed as $l_{ij} = -\log(\mathrm{softmax}(z_{ij}^{joint})_{y_{ij}})$.

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 [**?**]								
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 [**?**]								
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								
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								
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*

SPCL introduces adaptive sample selection through two components:

(1) *Difficulty Measurer*: Each utterance is assigned a difficulty score ρ_{ij}: an utterance-level term l_{ij} capturing classification difficulty and a conversation-level term s_i based on the standard deviation of uni-modal scores, estimating modality discrepancy. The final difficulty is the harmonic mean $\rho_{ij} = \frac{2 s_i l_{ij}}{s_i + l_{ij}}$, ensuring neither term dominates.

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								
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								
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								
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*

(2) *Learning Scheduler*: A hard regularizer selects samples with $\rho_{ij} \leq \lambda$ using a binary mask $v_{ij} \in \{0, 1\}$, enabling the model to focus on easier and more balanced samples in early epochs. The threshold λ is gradually increased following $\lambda^{(t)} = \alpha\lambda^{(t-1)}$, progressively incorporating harder samples as training advances.

SPCL modifies the standard MERC objective by weighting utterance losses with v_{ij}: $\mathcal{L}_{SPCL} = \frac{1}{\sum v_{ij}} \sum v_{ij}\, l_{ij}$. Model parameters are updated using the gradient of $\mathcal{L}_{SPCL}$, ensuring that optimization is guided by progressively more challenging but modality-balanced samples.

Results. Across both IEMOCAP and MELD, SPCL consistently boosts performance over baseline backbones and existing balancing techniques, particularly in the full-modality TAV setting. In IEMOCAP, SPCL delivers clear gains across all models—for instance, improving MM-DFN by +1.65% and MMGCN by +1.40% over their

baselines—achieving an average +0.85% over the second-best method and +2.25% over backbone models.

On MELD, SPCL also shows competitive improvements, especially within transformer-based architectures such as MM-DFN, where it surpasses OPM by +0.42%. However, in DialogueGCN, gains are smaller (e.g., +0.14% over OGM-GE), likely due to MELD's shorter and more fragmented conversational structure, which may not fully align with SPCL's progressive learning schedule. Nonetheless, SPCL remains robust overall, delivering consistent and meaningful improvements across both benchmarks.

4.3 Chapter Summary

In this chapter, we addressed the second focus of this dissertation: **balanced multimodal learning for low-quality multimodal data**. In Section 4.1, we introduced **Ada2I**, an end-to-end framework that explicitly re-balances learning at both the feature and modality levels to mitigate modality dominance. In Section 4.2, we presented **SPCL**, a lightweight plug-and-play module that progressively guides model training from easy to difficult samples while dynamically reducing modality imbalance, making it adaptable to various model architectures. We further conducted extensive experiments on benchmark datasets, including IEMOCAP, MELD, and CMU-MOSEI, demonstrating that both Ada2I and SPCL consistently enhance the performance of multimodal models under imbalanced learning scenarios.

Collectively, these contributions advance the goal of **Objective O2** of this dissertation by addressing two critical aspects of balanced multimodal learning. First, Ada2I improves the optimization of weaker modalities through feature-level and modality-level re-balancing, ensuring robust multimodal representations. Second, SPCL introduces an adaptive and architecture-agnostic training strategy that mitigates modality dominance and improves model stability. Together, these approaches provide effective and complementary solutions for overcoming modality imbalance, thereby improving the resilience and generalization of multimodal models in real-world low-quality data settings.

Conclusion

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. – **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*. https://doi.org/10.1007/s00521-026-12160-6 – **Scopus Q1 Journal**