Emotion Recognition in Multi-Speaker Conversations through Speaker Identification, Knowledge Distillation, and Hierarchical Fusion

Xiao LI, Kotaro FUNAKOSHI, and Manabu OKUMURA
Institute of Science Tokyo
{lixiao, funakoshi, oku}@lr.first.iir.isct.ac.jp

Abstract

Emotion recognition in multi-speaker conversations faces significant challenges due to speaker ambiguity and severe class imbalance. We propose a novel architecture that addresses these issues through three key innovations: (1) a speaker identification module that leverages audio-visual synchronization to accurately identify the active speaker, (2) a knowledge distillation strategy that transfers superior textual emotion understanding to audio and visual modalities, and (3) hierarchical attention fusion with composite loss functions to handle class imbalance. Comprehensive evaluations on MELD and IEMOCAP datasets demonstrate superior performance, achieving 67.75% and 72.44% weighted F1 scores respectively, with particularly notable improvements on minority emotion classes.

1 Introduction

Human emotion recognition has emerged as one of the most fundamental and challenging problems in artificial intelligence (Dzedzickis et al., 2020; Saxena et al., 2020; Deng and Ren, 2021), with important implications for human-computer interaction, social robotics, mental health monitoring, and conversational AI systems (Zhao et al., 2025; Pereira et al., 2025). The development of emotionally intelligent systems has become increasingly critical as AI applications expand into domains requiring nuanced understanding of human emotional states (Spezialetti et al., 2020; Younis et al., 2024).

Unlike traditional pattern recognition tasks that focus on static objects or isolated signals, emotion recognition requires understanding the complex interplay of multiple communication channels through which humans naturally express their emotional states (Guo et al., 2024). Research in psychology and cognitive science has demonstrated that emotional communication is inherently multimodal, with different modalities providing complementary and sometimes redundant information about a person's emotional state (Manalu and Rifai, 2024). For instance, while spoken words may convey neutral content, facial expressions might reveal underlying frustration or sarcasm (Zupan and Eskritt, 2024).

The computational modeling of this multimodal emotion recognition process has gained significant attention in recent years, driven by advances in deep learning and the availability of large-scale multimodal datasets, such as IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018). These resources have enabled the development of increasingly complex neural architectures that integrate textual, auditory, and visual signals through sophisticated fusion mechanisms (Baltrušaitis et al., 2018).

However, early approaches to emotion recognition focused primarily on single modalities, especially on context modeling and information extraction from the text modality, while neglecting the wealth of visual and auditory information available in video and speech. Such information includes facial keypoint trajectories, expression dynamics (Zhang and Chai, 2021), intonation patterns, and prosodic variations (El Ayadi et al., 2011), all of which are fundamental to human emotion perception in real-world interactions.

The transition toward multimodal emotion recognition has been motivated by the observation that combining multiple information sources typically leads to more robust and accurate emotion classification (Poria et al., 2017). Studies (Mao et al., 2020; Wu and Goodman, 2019) have shown that multimodal approaches can achieve significantly better performance compared to their unimodal counterparts, particularly in challenging conditions such as low-quality audio, poor lighting, or partial occlusion. Furthermore, multimodal systems can better handle the inherent ambiguity in emotional expressions, where the same facial

Findings of the Association for Computational Linguistics: EACL 2026, pages 4091–4106

expression might convey different emotions depending on the context and accompanying vocal cues (Hinton et al., 2015).

Despite these advances, conversational emotion recognition in multi-speaker scenarios presents unique challenges that remain inadequately addressed: **(1) Speaker Disambiguation** - accurately identifying the active speaker in multi-party conversations where multiple faces may be visible; **(2) Modality Performance Gaps** - significant performance disparities between text-based and audio-visual emotion recognition methods; and **(3) Severe Class Imbalance** - real-world conversational datasets often exhibit heavy bias toward certain emotions while underrepresenting others.

We propose a comprehensive framework that addresses these challenges through three main contributions:[1] **(1) Speaker-Centric Processing:** We introduce LipSyncNet, an approach that identifies active speakers through audio-visual synchronization patterns, enabling precise speaker-specific emotion analysis. **(2) Cross-Modal Knowledge Distillation:** We systematically transfer knowledge from high-performing text models to audio and visual counterparts, bridging modality performance gaps through graph-based architectures. **(3) Hierarchical Fusion with Composite Loss:** We develop sophisticated attention mechanisms combined with composite loss functions that effectively handle severe class imbalance in conversational datasets.

2 Related Work

Conversational Emotion Recognition. Conversational emotion recognition extends beyond isolated utterance analysis to consider the contextual and interactive nature of human dialogue (Poria et al., 2019). Unlike traditional emotion recognition tasks that process individual samples independently, conversational scenarios require modeling temporal dependencies, speaker interactions, and emotional dynamics throughout the conversation.

Key characteristics of conversational emotion recognition include **context dependency**, where the emotional interpretation of an utterance depends on preceding context (Hazarika et al., 2018); **speaker modeling**, which requires maintaining separate emotional states and characteristics for different participants (Majumder et al., 2019); and **emotion dynamics**, involving the modeling of emotional transitions and contagion effects between speakers (Ghosal et al., 2019).

Recent approaches have employed various neural architectures to capture conversational dynamics. Recurrent Neural Networks with attention mechanisms model temporal dependencies in conversations (Hazarika et al., 2018). Hierarchical approaches use separate encoders for utterance-level and conversation-level representations (Majumder et al., 2019). Graph-based methods represent conversations as graphs where nodes represent utterances and edges capture relationships between them (Ghosal et al., 2019).

The integration of multiple modalities in conversational settings presents additional challenges, as different modalities may have varying temporal resolutions and alignment issues (Tsai et al., 2019). State-of-the-art approaches include Multimodal Transformer architectures that use cross-modal attention to align and fuse information across modalities (Rahate et al., 2022).

Multimodal Fusion Strategies. Multimodal emotion recognition combines information from multiple input channels to achieve more robust and accurate emotion classification than any single modality alone (Baltrušaitis et al., 2018). The fusion process can occur at different levels of the processing pipeline, each with distinct advantages and limitations.

Early fusion concatenates features from different modalities before classification (Wöllmer et al., 2013). This approach allows the classifier to learn joint representations and cross-modal correlations but may suffer from the curse of dimensionality and differences in feature scales across modalities (Poria et al., 2017).

Late fusion trains separate classifiers for each modality and combines their outputs through techniques such as weighted voting, product rule, or learned combination functions. While this approach is more robust to modality-specific noise, it cannot capture low-level cross-modal interactions (Ramakrishna et al., 2023).

Hybrid fusion strategies attempt to combine the benefits of early and late fusion by performing fusion at intermediate levels of the processing pipeline (Zadeh et al., 2017). Recent approaches have explored attention-based fusion mechanisms that dynamically weight the contribution of different modalities based on their relevance to the current sample (Liang et al., 2018).

Advanced fusion architectures include Tensor

[1] https://github.com/llllxx1628/multimodalERC

Fusion Networks (Zadeh et al., 2017), which model multi-modal interactions through tensor operations, and Memory Fusion Networks (Zadeh et al., 2018), which use attention mechanisms to selectively retrieve relevant cross-modal information. Graph-based fusion approaches (Zhang et al., 2019) represent multimodal data as graphs and use Graph Neural Networks (GNNs) to capture complex relationships between modalities.

Speaker Diarization. Speaker diarization is the process of determining "who spoke when" in an audio recording containing multiple speakers (Malik et al., 2020). This task is fundamental to multi-speaker emotion recognition as it provides the necessary speaker attribution required for accurate emotion analysis.

Traditional speaker diarization systems follow a clustering approach: speech segments are first extracted through Voice Activity Detection (VAD), then speaker embeddings are computed for each segment, and finally clustering algorithms group segments belonging to the same speaker (Garcia-Romero et al., 2017). Common clustering techniques include k-means, hierarchical clustering, and spectral clustering (Sell et al., 2018).

Modern approaches leverage deep learning for improved speaker embeddings. i-vectors (Dehak et al., 2010) and x-vectors (Snyder et al., 2018) have become standard speaker representations, with x-vectors showing superior performance through deep neural network training. End-to-end neural diarization systems (Fujita et al., 2019) jointly optimize speaker embedding and clustering components.

SyncNet (Chung and Zisserman, 2017) pioneered the use of lip-sync information for speaker identification by learning to associate lip movements with corresponding audio signals. Subsequent work has explored various audio-visual fusion strategies and improved synchronization detection methods (Afouras et al., 2018).

However, existing audio-visual diarization methods often focus on face detection and tracking without consideration of multimodal emotion recognition. The precise temporal alignment needed for emotion analysis remains a challenge in current diarization systems (Bredin et al., 2020).

Limitations of Existing Work. Current approaches face three key limitations: (1) assumption of perfect speaker identification or treating it as preprocessing, leading to error propagation in multi-speaker scenarios; (2) insufficient handling of modality performance imbalances, particularly the superior performance of text over audio/visual modalities in conversational contexts; and (3) inadequate solutions for severe class imbalance in real-world conversational datasets, where minority emotions are crucial for comprehensive emotion understanding. Our work addresses these limitations through integrated speaker-centric processing, systematic cross-modal knowledge distillation, and composite loss functions specifically designed for conversational class imbalance.

3 Method

We propose a comprehensive multimodal conversational emotion recognition framework that addresses three key challenges: speaker identification in multi-party scenarios, cross-modal knowledge transfer, and class imbalance. Our approach consists of four main components: (1) speaker-centric processing via LipSyncNet, (2) graph-based knowledge distillation, (3) hierarchical attention fusion, and (4) composite loss functions.

3.1 System Overview

Given a dataset $\mathcal{D}=\{(u_i, s_i, e_i)\}_{i=1}^{N}$ where u_i represents the i-th utterance, s_i is the speaker identity, and e_i is the emotion label, our framework extracts multimodal features $\mathbf{F}_t^{(i)} \in \mathbb{R}^{d_t}$, $\mathbf{F}_a^{(i)} \in \mathbb{R}^{d_a}$, and $\mathbf{F}_v^{(i)} \in \mathbb{R}^{d_v}$ for text, audio, and visual modalities respectively. As illustrated in Figure 1, the system processes these features through modality-specific graph networks with knowledge distillation, hierarchical fusion. The system is optimized via the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{fusion} + \lambda_{dis}\mathcal{L}_{dis} + \lambda_{sync}\mathcal{L}_{sync}. \quad (1)$$

The three components $\mathcal{L}_{sync}$, $\mathcal{L}_{dis}$, and $\mathcal{L}_{fusion}$ are to be explained below in this order.

3.2 Speaker-Centric Processing

Multi-party conversations require accurate speaker identification to extract relevant visual cues. However, most conventional speaker diarization or localization methods, e.g., (OShaughnessy, 2025; Singh et al., 2023; Chung and Zisserman, 2016; Fujita et al., 2020) are designed for long continuous recordings and rely on multi-stage pipelines, including voice activity detection, speaker embedding extraction, and global clustering. These procedures are not well suited for utterance-level multimodal emotion recognition, as they introduce additional complexity and may propagate errors

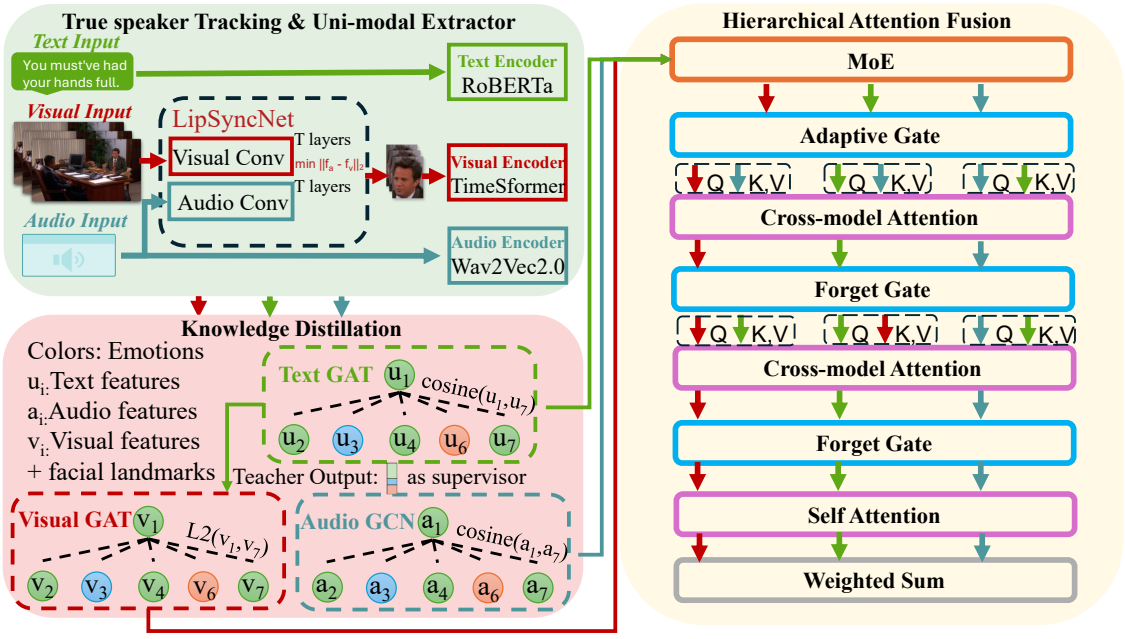

Figure 1: Overall architecture of the proposed multimodal conversational emotion recognition system.

across segments, especially in short and noisy conversational turns.

To address these limitations, we design LipSyncNet, a speaker-centric module for utterance-level processing. LipSyncNet takes short video clips from each detected face track together with the corresponding audio segment, and selects the face whose visual features are most synchronized with the audio track. The model is implemented as a lightweight dual-stream convolutional network and is applied only once per utterance to identify the active speaker. This design enables efficient and reliable speaker filtering while remaining closely aligned with the requirements of multimodal emotion recognition in multi-party conversations.

3.2.1 Architecture and Training

LipSyncNet employs a 3D CNN video encoder and a 2D CNN audio encoder that produce 256-dimensional L2-normalized features $\mathbf{f}_v$ and $\mathbf{f}_a$ respectively. The model is trained using a combined ranking and alignment loss:

$$\mathcal{L}_{sync} = \alpha_{sync}\mathcal{L}_{rank} + (1 - \alpha_{sync})\mathcal{L}_{align}, \quad (2)$$

$$\mathcal{L}_{rank} = \max(0, m + d_{pos} - d_{neg}), \quad (3)$$

$$\mathcal{L}_{align} = \mathbb{E}[d_{pos}], \quad (4)$$

where $d_{pos} = \|\mathbf{f}_v^+ - \mathbf{f}_a\|_2$ denotes the distance between synchronized audio-visual pairs and $d_{neg} = \|\mathbf{f}_v^- - \mathbf{f}_a\|_2$ denotes the distance between non-synchronized pairs. m is the margin. $\mathcal{L}_{rank}$ encourages separation between positive and negative audio-visual pairs. $\mathcal{L}_{align}$ minimizes synchronized pair distances.

3.2.2 Speaker Identification

The synchronization score is defined as the negative embedding distance between a candidate face video F_j and the audio A:

$$\text{Score}(F_j, A) = -\|\mathbf{f}_v^{(j)} - \mathbf{f}_a\|_2. \quad (5)$$

The true active speaker is identified as the one with the highest synchronization score:

$$j^* = \arg\max_j \text{Score}(F_j, A). \quad (6)$$

This ensures that subsequent visual feature extraction only relies on the true speaker rather than background participants.

3.3 Multimodal Feature Extraction

3.3.1 Contextual Text Features

Conversational emotion recognition requires understanding not only the current utterance but also the surrounding conversational context that significantly influences emotional interpretation. Unlike isolated text classification, emotions in conversations are heavily dependent on prior dialogue history, speaker relationships, and conversational flow (Hazarika et al., 2018).

Inspired by Shi and Huang (2023), we construct conversational context through a symmetric expan-

sion strategy that captures both historical and future context around the target utterance u_t. This approach is motivated by the observation that emotional expressions often build upon previous exchanges while also being influenced by anticipated responses in interactive scenarios.

The context construction follows an iterative expansion process:

$$\mathbf{C}_t = \mathbf{L}_k \oplus [\text{SEP}] \oplus u_t \oplus [\text{SEP}] \oplus \mathbf{R}_k, \quad (7)$$

where $\mathbf{L}_k$ and $\mathbf{R}_k$ represent left and right context expansions within a 512-token limit imposed by RoBERTa (Liu et al., 2019)'s input constraints. The special separator tokens [SEP] help the model distinguish the target utterance from its context.

Following (Yun et al., 2024), we incorporate an emotion-aware prompt that guides the model's attention to emotional cues. The prompt takes the form: "[Speaker] feels [MASK]" where [Speaker] is the current speaker identity and [MASK] is the prediction target. This design encourages the model to explicitly consider speaker-specific emotional patterns within the conversational context.

The fine-tuned RoBERTa model produces contextualized hidden states $\mathbf{H}_{text} \in \mathbb{R}^{T_t \times 1024}$ for each token in the input sequence of length T_t. The hidden state of the final token, $\mathbf{H}_{text}[T_t] \in \mathbb{R}^{1024}$, is projected through a linear transformation $\mathbf{W} \in \mathbb{R}^{1024 \times 768}$ to obtain the utterance-level representation $\mathbf{F}_t^{(i)} \in \mathbb{R}^{768}$.

3.3.2 Audio Features

We employ Wav2Vec2.0 (Baevski et al., 2020) for audio feature extraction due to its superior performance in capturing both phonetic and prosodic information essential for emotion recognition.

The model processes raw audio waveforms and generates contextualized hidden states $\mathbf{H}_{audio} \in \mathbb{R}^{T_a \times 1024}$, where T_a represents the temporal length. To obtain utterance-level representations suitable for our graph-based processing, we apply temporal average pooling:

$$\mathbf{F}_a^{(i)} = \frac{1}{T_a} \sum_{t=1}^{T_a} \mathbf{H}_{audio}[t, :] \in \mathbb{R}^{1024}. \quad (8)$$

This pooling strategy preserves global acoustic patterns while maintaining computational efficiency for subsequent graph network processing.

3.3.3 Visual Features

For visual modality, we utilize TimeSformer (Bertasius et al., 2021) specifically designed for spatiotemporal video understanding. TimeSformer's divided attention mechanism separately models spatial and temporal relationships, making it particularly effective for capturing facial expression dynamics and micro-expressions that evolve over time in conversational contexts. This approach is superior to conventional 3D CNNs as it can model long-range temporal dependencies without the burden of dense 3D convolutions.

The model processes speaker-filtered video sequences (obtained from our LipSyncNet module) to ensure that visual features correspond to the actual speaker rather than background individuals. TimeSformer generates frame-level representations which are aggregated through temporal pooling:

$$\mathbf{F}_v^{(i)} = \frac{1}{T_v} \sum_{t=1}^{T_v} \mathbf{H}_{visual}[t, :] \in \mathbb{R}^{768}. \quad (9)$$

The temporal pooling captures holistic facial expression patterns across the entire utterance duration, providing robust representations that are invariant to minor head movements and temporary occlusions while preserving emotional expression dynamics essential for accurate classification.

3.4 Knowledge Distillation

To address the performance gap between text and audio-visual modalities, we employ a teacher-student distillation framework using modality-specific graph networks. The fundamental motivation behind this approach comes from the observation that text-based emotion recognition consistently outperforms audio and visual modalities due to the explicit semantic content and rich contextual information available in linguistic expressions.

Our graph-based approach is motivated by the inherent relational nature of conversational data, where utterances are interconnected through temporal dependencies, speaker relationships, and semantic similarities. Traditional sequence-based models often fail to capture these complex interdependencies, whereas graph neural networks can explicitly model the conversational structure and propagate information across related utterances, leading to more robust emotion recognition.

The teacher-student paradigm enables systematic knowledge transfer from the high-performing text modality to weaker audio/visual modalities. This approach not only improves individual modality performance but also enhances the overall multimodal fusion by providing more balanced and informative representations from each modality.

3.4.1 Graph Architecture Design

The design of modality-specific graph architectures reflects the fundamental differences in how information is structured and processed across different input channels.These architectural choices are crucial for effective knowledge transfer and optimal representation learning.

The **teacher text model** uses Graph Attention Networks (GAT) (Velickovic et al., 2017) with 4 layers and multi-head attention to capture semantic dependencies. GATs are particularly suitable for textual data due to their ability to dynamically weight the importance of different neighboring utterances based on semantic similarity and contextual relevance. The attention mechanism allows the model to focus on emotionally significant contextual information while filtering out irrelevant conversational noise.

Student models employ modality-appropriate architectures that respect the unique characteristics of their input domains. Graph Convolutional Networks (GCN) (Kipf and Welling, 2016) are used for audio processing due to homogeneous temporal patterns where adjacent audio segments typically exhibit smooth transitions and consistent acoustic properties. The spectral convolution in GCNs is well-suited for capturing these gradual variations in acoustic features across time.

For visual processing, we employ GAT to handle heterogeneous facial region relationships where different facial components (eyes, mouth, eyebrows) may have varying importance for emotion expression. The attention mechanism enables dynamic weighting of different facial regions based on their relevance to the current emotional state, allowing the model to adapt to person-specific expression patterns and cultural differences in emotional display.

3.4.2 Distillation Loss

The knowledge distillation process requires careful balance between preserving the student model's ability to learn from ground truth labels and incorporating the rich knowledge embedded in teacher predictions. Our composite distillation loss addresses this challenge through a weighted combination of classification and knowledge transfer objectives.

The composite distillation loss combines classification and knowledge transfer:

$$\mathcal{L}_{dis} = \alpha_{dis}\mathcal{L}_{ce}(f_s(\mathbf{x}_s), y) + (1-\alpha_{dis})\cdot \mathcal{L}_{KL}\left(f_s(\mathbf{x}_s)/\tau_{dis}, f_t(\mathbf{x}_t)/\tau_{dis}\right)\tau_{dis}^2, \quad (10)$$

where f_s and f_t are student and teacher models.

$\mathcal{L}_{ce}$ and $\mathcal{L}_{KL}$ are the standard cross-entropy loss and Kullback-Leibler divergence (see Appendix A for the definitions).

3.5 Hierarchical Attention Fusion

Our fusion framework addresses the fundamental challenge of integrating heterogeneous multimodal information with varying reliability, temporal alignment, and semantic granularity. The hierarchical design enables progressive information refinement through multiple processing stages, each addressing specific aspects of multimodal integration. The framework integrates multimodal features through five stages: projection, quality assessment, cross-modal attention, transformer encoding, and ensemble prediction. The five-stage design reflects a principled approach to multimodal fusion: first standardizing feature representations, then assessing their quality, enabling cross-modal information exchange, modeling complex dependencies, and finally generating robust predictions with uncertainty estimation. This progression ensures that each stage builds upon previous refinements while adding specific capabilities essential for robust emotion recognition.

3.5.1 Adaptive Fusion Gates

Quality assessment and adaptive gating are critical for handling the varying reliability of different modalities under different conditions. For instance, visual features may be unreliable under poor lighting conditions, while audio features may be degraded by background noise. Our adaptive gating mechanism dynamically adjusts the contribution of each modality based on multiple quality indicators:

$$Q^{(m)} = \text{clamp}(\sigma(\mathbf{W}_q^{(m)}[q_{stats}^{(m)}; q_{entropy}^{(m)}; q_{neural}^{(m)}]), 0.1, 1.0). \quad (11)$$

We define three complementary indicators:

$$q_{stats}^{(m)} = \frac{\sigma_{\text{inter}}^{(m)}}{\sigma_{\text{intra}}^{(m)} + \epsilon_{\text{stats}}}, \quad (12)$$

where $\sigma_{\text{inter}}^{(m)}$ and $\sigma_{\text{intra}}^{(m)}$ are inter-class and intra-class variances for modality m, and ϵ_{stats} is a small con-

stant added for numerical stability.

$$q_{entropy}^{(m)} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p_c^{(i,m)} \log p_c^{(i,m)}, \quad (13)$$

where $p_c^{(i,m)}$ is the probability of class c from modality m for sample i.

$$q_{neural}^{(m)} = \sigma(\mathbf{w}_q^{(m)} \cdot \bar{\mathbf{H}}^{(m)} + b_q^{(m)}). \quad (14)$$

where $\bar{\mathbf{H}}^{(m)}$ is the mean feature representation of modality m.

Dynamic gates control information flow using global context as:

$$\mathbf{H}_{gated}^{(m)} = \mathbf{G}^{(m)} \odot \mathbf{H}^{(m)}, \quad (15)$$

$$\mathbf{G}^{(m)} = \sigma(\mathbf{W}_g^{(m)}[\mathbf{H}^{(m)}; \mathbf{C}_{global}]), \quad (16)$$

$$\mathbf{C}_{global} = \sum_m Q^{(m)}\mathbf{H}^{(m)} / \sum_m Q^{(m)}. \quad (17)$$

3.5.2 Cross-Modal Attention and Integration

Multi-head cross-modal attention enables information exchange between modalities:

$$\mathbf{H}_{cross}^{(m)} = \alpha_{cross}^{(m)}\text{CrossAttn}(\mathbf{H}_{gated}^{(m)}) + \beta_{cross}^{(m)}\mathbf{H}_{gated}^{(m)}. \quad (18)$$

Before cross-modal integration, we further employ a Mixture-of-Experts (MoE) (Cai et al., 2025) layer in the fusion stage. These layers dynamically route modality-specific representations across multiple experts, allowing the model to capture diverse feature subspaces and provide richer inputs to the subsequent transformer encoder. Features are then processed through transformer encoders and hierarchical attention pooling with learnable queries to generate final predictions. The transformer encoders capture complex dependencies within the fused representations, while hierarchical attention pooling extracts complementary aspects of emotional information that contribute to robust classification.

3.6 Composite Classification Loss

Class imbalance and hard sample mining represent significant challenges in real-world conversational emotion datasets. Standard cross-entropy loss tends to be dominated by frequent classes, leading to poor performance on minority emotions that are often crucial for comprehensive emotional understanding. To address these, we propose a composite classification loss function combining polynomial focusing, label smoothing, and supervised contrastive learning.

The polynomial loss formulation (Leng et al., 2022) provides an approach to hard sample mining compared to traditional focal loss. While focal loss uses exponential decay that can be too aggressive for extremely imbalanced datasets, polynomial loss offers more controlled focusing that adapts to the specific characteristics of conversational emotion data. Our main classification loss extends cross-entropy with polynomial focusing:

$$\mathcal{L}_{comp} = \mathcal{L}_{ce} + \alpha_{poly}(1 - p_{y_i})^{1+\gamma_{poly}}, \quad (19)$$

where p_{y_i} is the predicted probability for true class y_i, and the polynomial term provides adaptive focusing on difficult examples.

The integration of multiple loss components requires balancing to ensure that each component contributes effectively to the overall training objective without creating conflicting gradients or unstable training dynamics. The final training objective for modality fusion combines classification, label smoothing, and supervised contrastive learning:

$$\mathcal{L}_{fusion} = \alpha_{comp}\mathcal{L}_{comp} + (1 - \alpha_{comp})\mathcal{L}_{smooth} + \lambda_{cont}\mathcal{L}_{cont}. \quad (20)$$

The label smoothing loss (Müller et al., 2019) is defined as cross-entropy between the predicted distribution and a smoothed target distribution. Supervised contrastive learning (Khosla et al., 2020) uses normalized feature projections to enhance discrimination between similar emotions. We follow the conventional use of the two objectives. Implementation details of $\mathcal{L}_{smooth}$ and $\mathcal{L}_{cont}$ are described in Appendix B to be self-contained.

4 Experiments

4.1 Datasets and Experimental Setup

We evaluate our framework on two widely-used conversational emotion recognition benchmarks:

MELD (Poria et al., 2018): A large-scale dataset extracted from TV show *Friends*. The dataset exhibits severe class imbalance with neutral (47.1%) dominating, while disgust (2.7%) and fear (2.7%) are severely under-represented.

IEMOCAP (Busso et al., 2008): A carefully constructed dataset from laboratory-recorded dyadic conversations. This dataset provides higher annotation quality and more balanced class distribution compared to MELD.

Following standard protocols, we use the official train/validation/test splits for MELD and 5-fold cross-validation for IEMOCAP. We report weighted F1 (WF1) scores as the primary metric due to class imbalance considerations, along with class-specific F1 scores for detailed analysis.

Methods	Anger	Disgust	Fear	Joy	Neut.	Sad.	Surp.	Average	
	(1109)	(271)	(268)	(1743)	(4710)	(683)	(1205)	Acc.	WF1
DialogueGCN	36.6	6.7	2.9	44.3	54.7	21.9	41.8	41.8	42.5
DialogueRNN	41.5	1.7	1.3	**73.6**	55.9	24.0	50.7	41.1	49.4
QMNN	43.2	0.0	0.0	51.4	77.0	11.7	53.2	59.7	59.0
IterativeERC	48.9	19.4	3.3	56.6	77.5	23.6	53.7	61.7	60.1
MultiEMO	54.0	28.0	24.0	64.8	80.0	43.5	58.3	67.4	66.6
TelME	55.9	22.2	22.2	65.2	**80.9**	43.4	60.2	68.4	67.3
Our Method	**57.8**	**29.0**	**30.5**	66.7	79.7	**44.6**	**61.2**	**68.8**	**67.8**

Table 1: Performance comparison on MELD. Numbers in parentheses indicate sample counts.

Methods	Happy	Sad	Neut.	Angry	Excited	Frus.	Average	
	(392)	(739)	(1167)	(711)	(620)	(1149)	Acc.	WF1
DialogueGCN	42.7	**84.5**	63.6	64.1	63.2	67.0	65.3	64.2
CTNet	51.3	79.9	65.8	67.3	**78.7**	58.9	68.0	67.5
COGMEN	52.0	81.8	68.7	66.0	75.4	68.2	68.3	67.7
IterativeERC	50.2	77.2	61.3	61.5	69.2	60.9	66.2	64.4
QMNN	39.7	68.3	55.3	62.6	66.7	62.2	61.5	59.9
TelME	50.8	80.4	66.5	66.2	73.5	67.1	66.9	68.6
JOYFUL	56.5	84.3	68.4	66.9	73.4	67.6	70.6	71.0
MultiEMO	52.7	83.2	70.0	65.7	72.9	70.0	71.5	71.2
Our Method	**57.8**	83.0	**71.6**	**68.3**	72.8	**71.9**	**71.9**	**72.4**

Table 2: Performance comparison on IEMOCAP. Numbers in parentheses indicate sample counts.

Configuration	MELD	IEMOCAP
w/o Speaker Identification	-2.6	-1.4
w/o Fusion Loss	-2.0	-1.2
w/o Knowledge Distillation	-1.9	-1.2
w/o Cross-Modal Attention	-1.1	-0.4
w/o Adaptive Gating	-1.0	-0.7
w/o Polynomial Loss	-0.9	-0.8
w/o Contrastive Learning	-0.9	-0.6

Table 3: Ablation results showing the performance drop in Δ WF1 (points) when removing each component from the full model, relative to the results reported in Table 1 and Table 2.

Depending on each dataset, we compare our method against representative past works (Ghosal et al., 2019; Majumder et al., 2019; Lu et al., 2020; Li et al., 2021a,b; Joshi et al., 2022; Li et al., 2023; Shi and Huang, 2023; Yun et al., 2024).

4.2 Implementation Details

Our method is implemented using PyTorch 1.12+ and trained on a NVIDIA RTX 3090 GPU. We use the AdamW optimizer. Model training except for data pre-processing (i.e., KD & Fusion parts with 3M parameters) requires 30 to 50 minutes. Appendix C shows all hyperparameter values. We report results averaged over 5 independent runs.

4.3 Performance Analysis

MELD Dataset Performance: Table 1 presents comprehensive per-class results on MELD, revealing the effectiveness of our approach on severely imbalanced classes. On MELD, our framework achieves 67.8% weighted F1, representing significant improvements over the strongest baseline TelME (Yun et al., 2024). Our results demonstrate statistically significant improvements with $p < 0.01$ (Appendix F), validating the robustness of our approach. More importantly, we demonstrate substantial improvements on severely underrepresented emotions: disgust (29.0% vs. 28.0% best baseline) and fear (30.5% vs. 24.0%), addressing critical limitations in minority class recognition.

Our framework demonstrates particularly strong performance on challenging minority classes, with notable improvements in disgust (+1.0% point) and fear (+6.5% points) recognition compared to the best baselines. These improvements are crucial for practical applications where minority emotion detection is often most critical.

IEMOCAP Dataset Performance: Table 2 shows our balanced performance across IEMOCAP's emotion categories, maintaining strong recognition across all emotion types. On IEMOCAP, we achieve 72.4% weighted F1, a significant 1.2% improvement over MultiEMO (Shi and Huang, 2023), with consistent performance gains across emotion categories while maintaining balanced recognition across all classes.

Additional Analysis: The comparison result between unimodal and multimodal models is provided in Table 5 in Appendix D. Confusion matrices for the main results are shown in Appendix E

4.4 Ablation Study and Error Analysis

Table 3 presents an ablation study in which each component is removed from the complete system. This setting allows a direct interpretation of how each module contributes to reducing specific types of errors.

Removing speaker identification results in the largest performance degradation on both datasets. Without explicit speaker identification, the model always selects the first detected face, which often corresponds to a non-speaking participant. Error inspection shows that this leads to frequent attribution of emotions to the wrong speaker, especially in multi-party conversations. This issue is more severe on MELD, where speaker turns are shorter and the number of participants is larger.

The fusion loss and knowledge distillation are the next most influential components. When the fusion loss is removed and the objective degenerates to standard cross-entropy, confusion among minority emotion categories increases noticeably, indicating reduced robustness under class imbalance. Similarly, removing knowledge distillation weakens cross-modal alignment, resulting in more errors when one modality is unreliable or contradictory. These errors are often observed in samples where acoustic and textual cues convey different emotional tendencies.

Cross-modal attention and adaptive gating mainly affect cases with modality ambiguity. Without cross-modal attention, the model struggles to resolve conflicts between modalities, leading to misclassification in samples where emotions are subtly expressed. Removing adaptive gating further increases modality dominance errors, with the model relying excessively on textual features, particularly on MELD, where audio and visual signals carry strong emotional information.

Finally, removing the polynomial loss causes a moderate but consistent drop in performance. This suggests that the polynomial formulation helps control overconfident predictions on hard samples, reducing misclassification of less frequent emotions. In comparison, contrastive learning shows a smaller impact, indicating that it plays a complementary but less central role in the overall error reduction.

Overall, this ablation and error analysis clarifies that speaker identification and fusion-related objectives primarily address structural errors in multi-party settings, while cross-modal attention, adaptive gating, and polynomial loss reduce modality bias and class-specific misclassification. Representative qualitative visualizations that concretely illustrate these error patterns are presented in Appendix G.

5 Conclusion

We present a comprehensive framework for multimodal conversational emotion recognition that systematically addresses speaker identification and class imbalance challenges through three key innovations. Our LipSyncNet-based speaker identification integrates audio-visual synchronization learning directly into the emotion recognition pipeline, eliminating error propagation from preprocessing steps. Cross-modal knowledge distillation successfully transfers superior textual understanding to audio and visual modalities through graph-based architectures. Hierarchical attention fusion with composite loss functions effectively handles severe class imbalance while maintaining strong overall performance.

Experimental results demonstrate state-of-the-art performance on both MELD and IEMOCAP, with particularly notable improvements on challenging minority emotions that are crucial for comprehensive emotional understanding. The framework shows practical viability with reasonable computational requirements and robust performance across different experimental conditions.

Key contributions include: (1) integrating speaker identification as a learnable component rather than preprocessing step, (2) systematic knowledge transfer across modalities using graph-based teacher-student frameworks, and (3) composite loss functions that effectively address severe class imbalance in conversational scenarios.

Limitations

Our evaluation focuses primarily on English conversational data from specific domains (TV shows, laboratory recordings). Cross-lingual and cross-cultural generalization requires further investigation, particularly for languages with different prosodic patterns or cultural expression norms.

The composite classification loss function introduces additional hyperparameters requiring careful tuning for different datasets and domains.

In addition, both evaluation benchmarks present limited challenges in terms of overlapping speech. IEMOCAP contains no overlapping speech, and overlaps are relatively rare in MELD. As a result, our speaker-centric processing is designed mainly for turn-level speaker identification and does not explicitly address severe speaker overlap, visual occlusion, or partial face visibility. Handling such complex interaction patterns remains an open direction and will be addressed in future work.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP24K02974.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The conversation: Deep

audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 7124–7128. IEEE.

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

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, page 120.

J. S. Chung and A. Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Joon Son Chung and Andrew Zisserman. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019. End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*.

Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification. *arXiv preprint arXiv:2003.02966*.

Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. 2017. Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934. IEEE.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Runfang Guo, Hongfei Guo, Liwen Wang, Mengmeng Chen, Dong Yang, and Bin Li. 2024. Development and application of emotion recognition technologya systematic literature review. *BMC psychology*, 12(1):95.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. 2022. Polyloss: A polynomial expansion perspective of classification loss functions. *Preprint*, arXiv:2204.12511.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021a. Quantum-inspired neural network for conversational emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13270–13278.

Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. 2021b. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9904–9917.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who speaks next? turn change and next speaker prediction in multimodal multiparty interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 349–354. IEEE.

Haposan Vincentius Manalu and Achmad Pratama Rifai. 2024. Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intelligent Systems with Applications*, 21:200339.

Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Douglas OShaughnessy. 2025. Speaker diarization: A review of objectives and methods. *Applied Sciences*, 15(4):2002.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2025. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):1–37.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239.

Raksha Ramakrishna, Anna Scaglione, Tong Wu, Nikhil Ravi, and Sean Peisert. 2023. Differential privacy for class-based data: A practical gaussian mechanism. *IEEE Transactions on Information Forensics and Security*, 18:5096–5108.

Anvita Saxena, Ashish Khanna, and Deepak Gupta. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and 1 others. 2018. Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge. In *Interspeech*, pages 2808–2812.

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

Prachi Singh, Amrit Kaul, and Sriram Ganapathy. 2023. Supervised hierarchical clustering using graph neural networks for speaker diarization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, and 1 others. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Mike Wu and Noah Goodman. 2019. Multimodal generative models for compositional representation learning. *arXiv preprint arXiv:1912.05075*.

Eman MG Younis, Someya Mohsen, Essam H Houssein, and Osman Ali Sadek Ibrahim. 2024. Machine learning for human emotion recognition: a comprehensive review. *Neural Computing and Applications*, 36(16):8901–8947.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. *arXiv preprint arXiv:2401.12987*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

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

Haidong Zhang and Yekun Chai. 2021. Coin: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 12–18.

Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. 2019. Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6. IEEE.

Mingyi Zhao, Linrui Gong, and Abdul Sattar Din. 2025. A review of the emotion recognition model of robots. *Applied Intelligence*, 55(6):1–33.

Barbra Zupan and Michelle Eskritt. 2024. Facial and vocal emotion recognition in adolescence: a systematic review. *Adolescent Research Review*, 9(2):253–277.

A Cross-entropy loss and KL divergence

.

The cross-entropy loss is defined as:

$$\mathcal{L}_{ce} = -\sum_{c=1}^{C} y_c \log p_c, \tag{21}$$

where C is the number of classes, $y_c \in \{0, 1\}$ is the ground-truth one-hot label, and p_c is the predicted probability.

The KullbackLeibler divergence is defined as:

$$\mathcal{L}_{KL}(p_s, p_t) = \sum_{c=1}^{C} p_t(c) \log \frac{p_t(c)}{p_s(c)}. \tag{22}$$

where $p_t(c)$ and $p_s(c)$ are teacher and student probabilities after temperature scaling.

B Label Smoothing and Supervised Contrastive Learning

The label smoothing loss (Müller et al., 2019) is defined as cross-entropy between the predicted distribution and a smoothed target distribution:

$$\mathcal{L}_{smooth} = -\sum_{c=1}^{C} q_c \log p_c. \tag{23}$$

where p_c is the predicted probability for class c, C is the number of classes, and q_c is the smoothed target distribution.

$$q_c = \begin{cases} 1 - \epsilon_{\text{smooth}} + \frac{\epsilon_{\text{smooth}}}{C}, & \text{if } c = y, \\ \frac{\epsilon_{\text{smooth}}}{C}, & \text{otherwise.} \end{cases} \tag{24}$$

where y is the ground-truth class, and ϵ_{smooth} is the smoothing ratio. It prevents overconfidence and improves generalization by encouraging the model to be less certain about its predictions, which is particularly beneficial for minority classes where limited training data may lead to overfitting (Müller et al., 2019).

Supervised Contrastive learning (Khosla et al., 2020) uses normalized feature projections to enhance discrimination between similar emotions. Let $\mathbf{h}_i \in \mathbb{R}^{d_f}$ denote the fused representation of the i-th sample before the classifier (where d_f is the fusion dimension), and let $g(\cdot) : \mathbb{R}^{d_f} \to \mathbb{R}^{d_z}$ be a two-layer MLP projection head. We obtain L2-normalized projections as

$$\mathbf{z}_i = \frac{g(\mathbf{h}_i)}{|g(\mathbf{h}_i)|_2}, \quad \mathbf{z}_i^+ = \frac{g(\tilde{\mathbf{h}}_i)}{|g(\tilde{\mathbf{h}}_i)|_2}, \tag{25}$$

where $\tilde{\mathbf{h}}_i$ is a positive sample for $\mathbf{h}i$ obtained either from an augmentation of the same instance or from another instance with the same ground-truth label within the mini-batch. All negatives $\mathbf{z}j$ are the projections of the remaining instances in the mini-batch with $j \neq i$. The contrastive objective is

$$\mathcal{L}_{cont} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{z}_i^T \mathbf{z}_i^+ / \tau_{cont})}{\sum_{j \neq i} \exp(\mathbf{z}_i^T \mathbf{z}_j / \tau_{cont})}, \tag{26}$$

with temperature parameter $\tau cont$.

C Hyperparameter Settings

Parameter	Value
Batch size	16
Number of epochs	30
Dropout rate	0.35
Fusion dimension	256
Gradient clipping (maximum norm)	1.0
Learning rate (Text)	8×10^{-5}
Learning rate (Audio)	6×10^{-5}
Learning rate (Visual)	6×10^{-5}
Learning rate (Fusion)	4×10^{-5}
Weight decay (all modules)	0.001
λ_{dis}	0.3
λ_{sync}	0.15
margin m	1.5
α_{sync}	0.7
α_{poly}	1.2
α_{cross}	0.7
β_{cross}	0.3
γ_{poly}	1.2
α_{comp}	0.8
λ_{cont}	0.1
τ_{cont}	0.07
ϵ_{stats}	1×10^{-6}
ϵ_{smooth}	0.1
Distillation temperature τ_{dis}	2.0
Distillation balance α_{dis}	0.65

Table 4: Hyperparameter configurations used in our experiments.

The balance parameter $\alpha_{dis} = 0.65$ gives slightly more weight to ground truth supervision, ensuring that student models maintain their discriminative ability while benefiting from teacher guidance. The temperature parameter $\tau_{dis} = 2.0$ softens the probability distributions, allowing the student to learn from the teacher's uncertainty patterns and confidence levels, which often contain valuable information about decision boundaries and class relationships.

D Modality Performance Analysis

D.1 Modality-Specific Performance Analysis

Table 5 presents a comprehensive analysis of individual modality contributions and their fusion effectiveness. This analysis provides crucial insights into the relative importance of different modalities in conversational emotion recognition tasks.

Models	MELD	IEMOCAP
Only Visual	37.8	32.5
Only Visual$_{\text{w/o LipSyncNet}}$	33.4	30.1
Only Audio	47.3	48.4
Only Text	66.1	68.7
Fusion Model	**67.8**	**72.4**

Table 5: Unimodal vs. Multimodal Performance Comparison (WF1 scores).

The results reveal several important findings about modality contributions in conversational emotion recognition: **Text Modality Dominance:** Text emerges as the most informative modality, achieving 66.1% and 68.7% WF1 on MELD and IEMOCAP respectively. This demonstrates that linguistic content carries the primary emotional signals in conversational contexts, consistent with the rich semantic information available in spoken language. **Audio Modality Contribution:** Audio modality shows moderate performance (47.3% on MELD, 48.4% on IEMOCAP), capturing prosodic and paralinguistic cues that complement textual information. The stronger performance on IEMOCAP suggests that laboratory-recorded conversations may preserve more nuanced acoustic features compared to TV show audio. **Visual Modality Challenges:** Visual features show the lowest individual performance (37.8% on MELD, 32.5% on IEMOCAP), highlighting the inherent difficulty of emotion recognition from facial expressions alone in conversational settings. The inclusion of LipSyncNet provides a meaningful contribution (+4.4% on MELD, +2.4% on IEMOCAP), demonstrating the value of audio-visual synchronization features for emotion understanding. **Fusion Benefits:** The multimodal fusion model achieves substantial improvements over the best single modality (text), with gains of +1.7% on MELD and +3.7% on IEMOCAP. These improvements validate our fusion strategy's effectiveness in leveraging complementary information across modalities, particularly the integration of prosodic audio cues and synchronized visual features with semantic text information.

E Detailed Performance Tables and Confusion Matrices

E.1 MELD Dataset Results

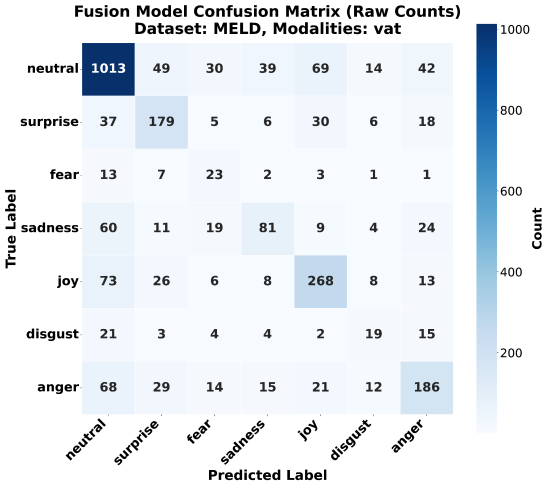

Figure 2: Confusion matrix for emotion classification on MELD dataset showing improved performance on minority emotion classes.

E.2 IEMOCAP Dataset Results

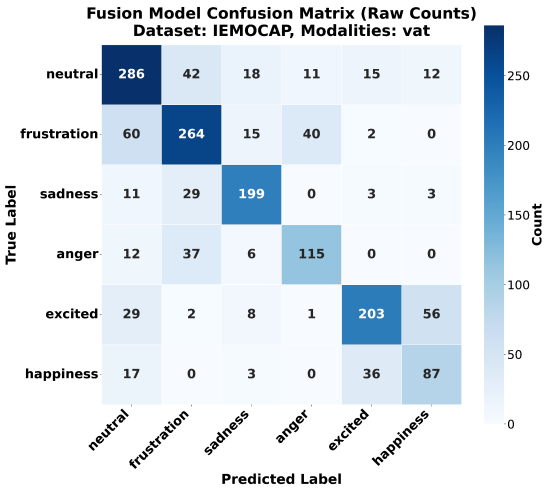

Figure 3: Confusion matrix for emotion classification on IEMOCAP dataset demonstrating balanced performance across emotion categories and clear separation between positive and negative emotional states.

F Statistical Significance Analysis

We validate our improvementon MELD and IEMOCAP through rigorous statistical significance testing against the strongest baselines

Methods	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Average	
	(1109)	(271)	(268)	(1743)	(4710)	(683)	(1205)	Acc.	WF1
DialogueGCN	36.56	6.73	2.85	44.26	54.72	21.87	41.78	41.75	42.46
DialogueRNN	41.53	1.74	1.25	73.56	55.94	23.97	50.73	41.09	49.43
QMNN	43.17	0.00	0.00	51.44	77.00	11.70	53.18	59.66	59.03
IterativeERC	48.88	19.38	3.31	56.63	77.52	23.62	53.65	61.66	60.07
MultiEMO	54.02	28.00	24.00	64.79	80.02	43.45	58.28	67.39	66.55
TelME	55.91	22.22	22.22	65.24	80.88	43.41	60.17	68.35	67.30
Our Method	**57.76**	**28.79**	**30.46**	66.67	79.73	**44.63**	**61.20**	**68.78**	**67.75**

Table 6: Detailed performance comparison on MELD dataset with 2 decimal places. Numbers in parentheses indicate sample counts.

Methods	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average	
	(392)	(739)	(1167)	(711)	(620)	(1149)	Acc.	WF1
DialogueGCN	42.73	84.52	63.56	64.13	63.17	66.95	65.26	64.18
CTNet	51.34	79.93	65.84	67.26	78.72	58.89	68.04	67.53
COGMEN	51.96	81.75	68.68	66.04	75.38	68.24	68.27	67.65
IterativeERC	50.17	77.19	61.31	61.45	69.23	60.92	66.21	64.37
QMNN	39.71	68.30	55.29	62.58	66.71	62.19	61.52	59.88
TelME	50.84	80.36	66.47	66.18	73.54	67.09	66.85	68.56
JOYFUL	56.53	84.33	68.42	66.90	73.40	67.61	70.63	70.90
MultiEMO	52.65	83.18	70.02	65.74	72.88	69.98	71.46	71.18
Our Method	**57.81**	83.04	**71.59**	**68.25**	72.76	**71.93**	**71.94**	**72.44**

Table 7: Detailed performance comparison on IEMOCAP dataset with 2 decimal places. Numbers in parentheses indicate sample counts.

(TelME and MlutiEMO, respectively) using paired t-tests over 5 independent runs with different random seeds.

Dataset	Method	Mean WF1	p-value
MELD	TelME	67.30 ± 0.05	
	Our Method	$\mathbf{67.75 \pm 0.07}$	**0.008**
IEMOCAP	MultiEMO	71.18 ± 0.09	
	Our Method	$\mathbf{72.44 \pm 0.12}$	**0.004**

Table 8: Statistical significance test results.

G Qualitative Error Analysis and Visualization

Figure 4 provides representative qualitative visualizations to complement the quantitative ablation and error analysis discussed in the main paper. For each emotion category, we present one representative sample, including the original video frame, the ground-truth emotion distribution, the prediction of the full model, and the predictions obtained after removing a specific module. The emotion distributions are visualized as horizontal color bars, where each color corresponds to one emotion class and the bar length indicates the predicted probability.

Figure 4: Error Analysis