Multimodal Contextual Transformer Augmented Fusion For Emotion Recognition

College of Technology and Built Environment

Addis Ababa University, Addis Ababa, Ethipia

Master of Science in Artificial Intelligence

Thesis By:

Wesagn Dawit Chemma

Advisor:

Adane Letta (Ph.D.)

Date: June 2025

Addis Ababa University
School of Information Technology and Engineering

This is to certify that this thesis titled **"Multimodal Contextual Transformer Augmented Fusion For Emotion Recognition"** is prepared by Wesagn Dawit Chemma and submitted in partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Artificial Intelligence, complies with the university's regulations and meets the accepted standards of originality and academic quality.

Advisor:

Name: Dr. Adane Letta Signature: _____ Date: _____

External Examiner:

Name: Dr. Worku Jifara Signature: _____ Date: _____

Internal Examiner:

Name: Dr. Beakal Gizachew Signature: _____ Date: _____

Chairperson:

Name: Dr. Henok Mulugeta Signature: _____ Date: _____

Abstract

As emotionally intelligent systems increasingly become integral to human-centered Artificial Intelligence (AI), the precise recognition of emotions in conversational settings continues to pose a fundamental difficulty. This challenge arises from the context-sensitive and evolving characteristics of emotional expression. Although the majority of Multimodal Emotion Recognition (MER) systems utilize speech and text features, they often overlook conversational context, such as prior dialogue exchanges, speaker identity, and interaction history, which are crucial for discerning nuanced or ambiguous emotions, particularly during dyadic and multi-party interactions. This study presents Multimodal Contextual Transformer Augmented Fusion (MCTAF), a lightweight, context-sensitive framework for MER. MCTAF explicitly represents context as a third modality, integrating the prior K utterances (dialogue history including text and audio), speaker characteristics, and turn-level temporal structure. The contextual features are processed using a Bidirectional Gated Recurrent Unit (BiGRU)-based context encoder that functions concurrently with distinct BiGRU encoders for textual and audio characteristics. All three modality-specific representations are integrated using a transformer-based self-attention method to capture both intra- and inter-modal interdependence across conversation turns. To our knowledge, this is the first study to clearly conceptualize conversational history as a key modality inside a unified transformer architecture, processing it concurrently with voice and text before a dynamic, attention-driven fusion. MCTAF surpasses robust baselines when assessed on Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Multimodal EmotionLines Dataset (MELD). It achieves 89.9% accuracy and 88.3% weighted F1-score on IEMOCAP and MELD benchmarks, respectively, delivering performance increases of up to +4.0 percentage points in accuracy and +3.0 in F1-score above preceding state-of-the-art models. Ablation experiments further validate the significance of context modeling, demonstrating a 3-4 point decline in F1 when the context module is eliminated. In terms of efficiency, MCTAF decreases training time by 8% each epoch and employs 12% fewer parameters than equivalent transformer-based baselines, with an average inference time of 26.1 ms per syllable. These findings demonstrate the potential of MCTAF for scalable and resource-efficient implementation.

Keywords: Multimodal emotion recognition, Contextual transformer, Cross-modal attention, Speech-text fusion, Dialogue context, Transformer-based fusion

Acknowledgment

First of all, I give thanks to Almighty God for His boundless grace, wisdom, and strength, which have sustained me through every challenge and guided me to this moment. His divine guidance was a constant source of inspiration and perseverance throughout this journey.

I am so grateful to my advisor, Dr. Adane Letta, for his mentorship, technical guidance, support, and patience, which have had a significant influence on the completion of this work. I am also grateful for his instruction and encouragement, which have left a lasting impact on my development as a researcher. I wish to extend my special thanks to my instructors, Dr. Beakal Gizachew, Dr. Fantahun Bogale, and Dr. Natnael Argaw. Their mentorship, knowledge, and guidance were invaluable to my academic journey.

My sincere thanks go to my dear friends Selameab Setargew and Amanuel Negash for their incredible support. Thank you for the insightful discussions and continuous inspiration. I also wish to extend my thanks to my best friends, Christian Samuel and Henok G/Michael. I am grateful for their generous financial support in times of need. A special thanks goes to Eskedar Beyene and Hewan Beyene for always being there for me. I owe special thanks to two individuals who have been my constant source of strength and encouragement. To Mr. Debela Desalegn, thank you for your genuine friendship, for sharing your ideas, and for kindly looking after my responsibilities when I could not. To Mr. Mesfin Wrokineh, I truly appreciate your kindness, understanding, and the wonderful support you provided that allowed me the freedom to pursue this study.

I would like to express my sincere gratitude to Dr. Marco Piangerelli, whose insightful feedback during the final stages of my thesis significantly strengthened the quality of this work. His thoughtful suggestions and constructive comments were instrumental in shaping the final version of this document. I am also deeply thankful to all my colleagues at the United Nations Development Programme (UNDP) and the Ministry of Finance (MoF) of Ethiopia for their support and encouragement throughout this journey. Their contributions, both big and small, have been a source of motivation and strength.

Finally, I extend my deepest love and appreciation to all my family, especially Dr. Desalegn Dawit, Dr. Assedesach Dawit, and Mr. Amanuel Aydiko Ayele. Their constant support, encouragement, patience, and belief in me were my foundation. I want to express my gratitude to everyone who has supported me on this journey.

Contents

List of Figures

List of Tables

List of Acronyms

AFEW Acted Facial Expressions in the Wild. 16

AI Artificial Intelligence. i, 1, 5–7, 31

ASR Automatic Speech Recognition. 79

BERT Bidirectional Encoder Representations from Transformers. 35, 38, 44–49, 51, 59, 73

BiGRU Bidirectional Gated Recurrent Unit. i, 46, 48–50, 58–60

CAER Context-Aware Emotion Recognition. 16

CAER-Net Context-Aware Emotion Recognition Network. 27, 28, 32

CH-SIMS Chinese single- and multi-modal sentiment analysis dataset. 17

CMU-MOSEI CMU Multimodal Opinion Sentiment and Emotion Intensity. 17, 38, 39

CMU-MOSI CMU Multimodal Opinion Sentiment Intensity. 17

CNN Convolutional Neural Network. 21, 22, 32, 33, 38

COVAREP Cooperative Voice Analysis Repository for Speech Technologies. 51, 59

CREMA-D Crowd-sourced Emotional Multimodal Actors Dataset. 14

DEAP Database for Emotion Analysis using Physiological signals. 17, 33

ECG electrocardiogram. 16, 17, 33

EDA electrodermal activity. 16

EEG electroencephalogram. 17, 33

ERC Emotion Recognition in Conversation. 36, 67, 83

FAU Facial Action Unit. 10

GCN Graph Convolutional Network. 68

GNN Graph Neural Network. 30, 34, 35, 39, 40, 69

GRU Gated Recurrent Unit. 22, 23, 32, 39, 40, 46, 49, 50, 52, 53, 62–64, 82, 85

GSR Galvanic Skin Response. 17, 33

HMM Hidden Markov Models. 20

HNR Harmonics-to-noise ratio. 46

IEMOCAP Interactive Emotional Dyadic Motion Capture. i, v, vi, 3, 5, 15, 38, 39, 43–47, 57, 61–80, 82, 85, 86, 88, 89

LLM Large Language Model. 36, 37

LMFB Log Mel Filterbank. 46

LSTM Long Short-Term Memory. 22, 23, 32

mBERT multilingual BERT. 84

MCTAF Multimodal Contextual Transformer Augmented Fusion. i, v, vi, 5–7, 41–43, 45, 48, 51–53, 57–59, 61–71, 73–84, 86–88

MELD Multimodal EmotionLines Dataset. i, v, vi, 3, 5, 16, 28, 38, 39, 43–47, 57, 61–70, 72–80, 82, 85–89

MER Multimodal Emotion Recognition. i, iii, 1, 3, 4, 9, 10, 13, 18, 20

MFCC Mel-frequency cepstral coefficients. 46, 61

MGAT Multi-Granularity Attention Based Transformers. 29, 30, 32

MLP multi-layer perceptron. 32, 59

MoF Ministry of Finance. ii

PCA principal component analysis. 61, 63, 85

PCM Pulse Code Modulation. 44

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song. 14

RMS Root Mean Square. 46

RNN Recurrent Neural Network. 22, 32, 33, 63, 68, 69, 78, 79

RoBERTa Robustly Optimized BERT Approach. 61, 63

SAVEE Surrey Audio-Visual Expressed Emotion. 14

SEMAINE Sustained Emotionally-colored Machine-human Interaction via Nonverbal Expression. 15

SVM Support Vector Machines. 20

TESS Toronto Emotional Speech Set. 14

UNDP United Nations Development Programme. ii

VAD Voice Activity Detection. 46

XLM-R Cross-lingual Language Modeling-RoBERTa. 84

Chapter 1

Introduction

1.1 Background

Machines' ability to sense, understand, and react to human emotions is a basis for modern human-centered AI, driving advances in empathic conversation systems, opinion mining, and affective computing applications [4, 5]. Recent breakthroughs have enabled intelligent systems to accurately evaluate a wide range of human signals, including facial expressions, body language, speech patterns, and physiological responses, to infer emotional states with increasing accuracy [6, 7]. These skills are being incorporated into areas such as healthcare and customer service by enhancing interaction quality and fostering contextually appropriate human-level responses [8–11].

Foundational models in this area depended on unimodal cues such as facial expressions or vocal prosody to characterize emotions [12, 13]. Such techniques failed to capture the multidimensional character of emotional expression (laughing may indicate pleasure in one situation or disguise fear in another) [14, 15]. This uncertainty, heightened by individual variation and ambient factors, underlined the need for MER systems that combine various data streams, including visual, audio, and text, to collect deeper emotional cues [16, 17].

MER systems have progressed by incorporating early, late, and hybrid fusion approaches, which merge information across modalities to increase prediction accuracy for subtle emotive states, such as sarcasm and mixed sentiments [18–20]. Many such models remained context-agnostic by analyzing each phrase independently without reference to earlier conversation turns or speaker dynamics.

The lack of context conflicts with accepted findings from psychology. Theories such as situated affectivity (emotions and sentiments that are impacted by the particular context or situation a person is in) argue that emotional states are developed not in isolation but through continual interactions with the social and physical environment [21]. Similarly, the theory of embodied cognition stresses the relevance of physical and environmental clues in forming emotional experience [22, 23]. These ideas help illustrate why emotion detection techniques must evolve beyond isolated signal processing and add conversational and environmental context.

Early efforts to add context in MER, such as those by Li et al. [24] and Zhao et al. [20], relied on heuristic rules or simple feature concatenation, which often led to overfitting and poor scalability. These techniques generally demonstrated poor scalability, overfitting, and failed to simulate the dynamic structure of discussion well.

Afterward, transformer-based approaches [25, 26] enabled modeling of long-range dependencies. Dialogue-aware transformers, such as DialogueBERT [27] and DialogXL [28], incorporate context by integrating preceding utterance embeddings using position- and speaker-aware encodings. However, these methods typically

treat contextual turns as structurally similar to the target utterance, without isolating their distinct semantic role or modeling the conversational context as a standalone modality.

The drawbacks of current context-aware models, such as shallow fusion methods and the uniform treatment of context, underline the demand for architectures capable of expressing conversational context as a structured and independent information source. This encompasses both lexical and acoustic features from previous dialogue turns [24, 29], as well as user-specific factors including speaker identity, personality traits, and interaction patterns, which have been demonstrated to influence emotional expression [30, 31]. Recent transformer-based techniques have increased the capturing of long-range relationships in conversation [32, 33]. However, the issue continues in designing models that correctly handle context with the requisite semantic stream and integrate it in a principled and interpretable way.

1.2 Motivation of the Study

Emotions play a crucial role in human cognition and interaction, directly impacting decision-making, memory, and attention. The requirement for accurate automatic recognition of emotions has substantially grown in the context of the rising popularity of intelligent systems [34]. Multimodal emotion recognition techniques incorporating voice and text cues, among other signals, have exhibited substantial gains. For instance, the combination of linguistic and audio information has enhanced performance in interpreting user affect [20, 35]. These systems are essentially limited by their inability to include contextual elements, including a user's prior interactions, the surrounding environment, and the conversational context [20, 35, 36]. The absence of context usually leads to inadequate or unclear interpretations of affect, especially for unpredictable or subtle emotional states such as frustration or confusion that depend on contextual clues [14, 15].

Many systems run on simplistic presumptions that disregard complicated real-world features like dynamic interaction patterns or nonverbal indications [37]. To illustrate that context is not just auxiliary but crucial for capturing emotional nuance, Li et al. [38] show that context-aware fusion models, those that incorporate surrounding contextual information, achieve higher accuracy in classifying certain emotional states [39, 40]. On the other hand, standard models that overlook context or view it as a minor risk oversimplify the complexity of human emotions and frequently fall short in adapting to the dynamics of fundamental interactions [5, 36].

In this setting, our work addresses the vital need to bridge the context gap in emotion recognition. We devised a unique technique that explicitly includes context as a peer modality alongside voice and text inside a transformer-based architecture. Unlike past efforts that often regarded context as an afterthought, our approach considers it a first-class modality, exploiting its information to disambiguate emotional expressions that are otherwise ambiguous when viewed in isolation from voice and text alone. The incorporation of context is projected to significantly enhance the model's ability to recognize nuanced emotional states, eliminate ambiguities, and promote generalization across diverse interaction settings [41]. Furthermore, the transformer design enables robust fusion of disparate data streams, allowing the model to learn complicated inter-modal

interactions efficiently [3]. This architecture not only optimizes performance but also helps the interpretability of the model by offering insights into which contextual components contribute most to emotion inference. We verify our technique via thorough evaluations on benchmark real-world datasets, showing its efficacy over state-of-the-art context-agnostic and basic context-aware baselines [42].

1.3 Statement of the Problem

MER has substantially evolved, with modern systems successfully combining voice and textual modalities to achieve good performance on affective computing benchmarks [14, 35]. However, their usefulness is sometimes restricted when emotional meaning is intimately related to conversational context. Empirical investigations repeatedly reveal that without comprehending the prior discussion, models fail to identify ambiguous or nuanced emotions, such as sarcasm, frustration, or fluctuating moods [15, 39].

Recognizing this issue, researchers have created several context-aware models. Early approaches depended on basic feature concatenation, whereas more modern models incorporate recurrent networks like DialogueRNN [31] or graph networks like DialogueGCN [29] to convey historical information sequentially. Even advanced transformer-based models frequently handle context implicitly by simply prepending past utterances into a lengthy input sequence [32, 33]. This amalgamation of information might make it difficult for the model to differentiate the emotional signals of the current instant from the effect of the past, resulting in what we call the context underutilization issue.

This underutilization has significant repercussions. A model without conversation history may misclassify a sarcastic statement like *That's simply great* as positive, failing to notice it follows a complaint [15]. This problem is magnified in multi-party dialogues, such as those in the MELD dataset [43], where monitoring speaker turns and intricate interdependencies is significantly more challenging than in the dyadic (two-person) discussions seen in datasets like IEMOCAP [44]. While past efforts have recognized this and included context [38, 45], they frequently do so by treating it as an auxiliary feature to condition or gate the main speech and text representations. For instance, sarcasm is often misclassified without access to the preceding dialogue context. While prior work has included context as auxiliary features, our framework elevates context to a primary modality. We also address challenges in multi-party settings compared to dyadic datasets.

Beyond these technical insights, psychological theories give a supplementary reason for considering context as a crucial component in emotion recognition [46]. Emotion is increasingly regarded as essentially contextual and interactive, rather than generated from isolated, internal states [47]. Theories of situated affect and embodied cognition claim that emotional experiences are dynamically influenced by an individual's continual interactions with their physical and social surroundings [21, 22]. In conversational situations, emotional meaning typically relies not just on the present utterance, but also on the larger discourse, such as responding to earlier turns, anticipating future responses, and reflecting interpersonal dynamics [47]. These discoveries underscore the critical significance of context in understanding emotion and indicate that systems intended for MER should utilize conversational history as more than supplemental information.

Thus, the problem addressed in this thesis is the absence of principled methodologies that represent conversational context as a significant, coequal modality, distinct from voice and text, within emotion recognition systems. This research studies whether a structured, modality-specific encoding of context, combined with dynamic cross-modal fusion, might overcome these constraints and progress the development of genuinely context-aware MER systems.

1.4 Research Questions

RQ1. How can contextual information be effectively represented and integrated as a distinct modality within a transformer-based multimodal fusion model for emotion recognition?

RQ2. How does multimodal fusion improve the accuracy and robustness of emotion recognition, specifically for emotions that are subtle, ambiguous, or highly context-dependent (e.g., sarcasm, frustration, confusion)?

RQ3. How can the trade-offs between predictive performance and computational efficiency be managed when designing context-aware multimodal emotion recognition systems?

1.5 Objective

1.5.1 General Objective

To design an augmented multimodal framework where the augmentation refers to the explicit enhancement of text and audio modalities with contextual information modeled as a parallel input stream. To achieve this, the study will pursue the following specific objectives, which directly correspond to the research questions:

1.5.2 Specific Objectives

- To explore how contextual information, derived from dialogue history and speaker information, can be effectively represented and integrated with acoustic and textual features.
- To evaluate how the inclusion of context affects the accuracy and robustness of emotion recognition, particularly for emotions that are weakly expressed or dependent on discourse-level cues.
- To assess the efficiency of the proposed model in multimodal fusion, specifically by comparing inference time, memory footprint, and training stability relative to baseline models.

1.6 Contribution of the Study

This work presents numerous substantial and quantifiable advances to the realm of multimodal emotion recognition. The significant contributions are summarized as follows:

A New State-of-the-Art Contextual Fusion Model

We introduce MCTAF, a novel emotion recognition framework, the proposed model, which achieves new state-of-the-art performance by fundamentally re-architecting how context is employed. More precisely, MCTAF achieves 89.9% accuracy on IEMOCAP and 88.3% on MELD, surpassing established baselines such as DialogueRNN and even contemporary transformer-based models by up to 4 percentage points. This performance boost is accomplished by considering conversational history not as a supplemental feature, but as a fundamental, coequal modality that is separately stored and then dynamically merged with voice and text via cross-modal attention. This work solves a significant architectural gap in the literature by illustrating the quantitative advantages of increasing context's importance.

Enhanced Recognition and Interpretability of Subtle Emotions

By explicitly modeling context, MCTAF considerably increases the capacity to discern confusing emotional states that confound context-agnostic models. For example, it may accurately recognize a sarcastic comment by attention to the mismatch between positive words (text), a flat tone (audio), and a prior negative speech (context). Furthermore, the transformer-based fusion method gives a degree of interpretability; by viewing the cross-attention weights, we can watch how the model dynamically increases its emphasis on the context stream when textual and audio inputs are clashing.

A Scalable and Efficient Transformer-Based Framework

The work presents a complete empirical assessment indicating that MCTAF 's advanced fusion architecture is not only efficient but also computationally efficient. Despite its outstanding performance, MCTAF retains a lightweight architecture that avoids the substantial computational cost of large-scale models or complicated graph-based techniques. This makes it a scalable solution well-suited for realistic, real-world applications where resources are limited. For instance, its design is efficient enough to be potentially used in real-time systems, such as contact center analytics tools or on-device mobile health apps that monitor user affect, offering a realistic and solid basis for the next generation of emotionally aware AI.

1.7 Significance of the Study

The impact of this study spans beyond both academic and practical domains.

Academically, this study adds to the current scholarly conversation in affective computing and AI by addressing a notable gap in multimodal emotion recognition: the explicit integration of context [14, 36]. By addressing contextual information as a key modality and employing transformer-based fusion approaches, the work improves methodology in the area. It presents actual data and a paradigm for how context can be systematically included into emotion detection models, therefore stretching the bounds of transformer applications beyond standard text and voice inputs [37]. This directly targets a current research topic and provides new opportunities for investigating modality interaction in complex, real-world data [10].

Practically, the suggested study offers transformational potential for next-generation interactive systems. In domains such as psychological health and education, more nuanced emotion recognition could enable AI-driven interventions that respond to not only a user's immediate expressions but also their situational context, thereby improving therapeutic feedback and user engagement [48, 49]. In many interactive applications, endowing intelligent agents and virtual assistants with context-aware emotional intelligence means they may offer replies that are empathetic and contextually appropriate, boosting user trust and comfort [11, 50]. Improvements in context-sensitive emotion recognition can benefit a range of real-world settings, including healthcare (e.g., socially assistive AI that recognize patient emotions [48]), customer service (virtual agents that adapt to a customer's mood and history), and education (tutoring systems that gauge frustration or confusion levels [49]). Overall, the study's value rests in both expanding theoretical knowledge of multimodal fusion and giving practical insights that might drive more emotionally intelligent AI systems.

1.8 Scope

The scope of this study is carefully specified to concentrate on the creation and thorough assessment of the proposed MCTAF model for improved emotion recognition. While emotion may be communicated and inferred via numerous modalities (e.g., facial expressions, physiological signals), this research focuses entirely on voice (acoustic aspects), text (linguistic content), and contextual information as the key input modalities. The focus is placed on text, audio, and context modalities, as these represent widely available signals across conversational datasets. Other channels of emotional expression, such as visual facial signals or body language, are recognized as crucial for broader emotion detection but are outside the scope of the present study, ensuring a comprehensive examination of context integration. The architecture is designed to be flexible for the future integration of these other modalities within the foundation of a AI system. In this study, context is defined as the structured sequence of preceding utterances and their associated features (text, audio, and speaker identity) within a dialogue. The study targets the following significant areas:

Multimodal Data Integration

Emphasis is put on the integration of three primary data modalities: voice (acoustic characteristics), text (linguistic content), and contextual information. Here, contextual information explicitly comprises aspects such as the user's interaction history (dialogue context), relevant ambient signals, and fundamental user-specific features or profiles. The research focuses on how to combine these three sources as peer modes properly. Notably, it excludes explicitly modalities beyond these three (e.g., visual facial expression data), focused entirely on the innovative merger of voice, text, and context as the inputs for emotion recognition.

Transformer-Based Architectures for Fusion

The study primarily utilizes transformer-based systems to analyze and integrate multimodal data. The transformer is selected for its capabilities to capture both intramodal (within-modality) and intermodal (between-modality) interactions via its self-attention mechanism. Within the scope of this work, special

attention is given to how contextual information can be encoded and integrated as a separate input stream in the transformer architecture, and how the transformer's attention mechanism can dynamically weight context relative to speech and text cues for optimal emotion inference.

Emotion Recognition for Subtle States

The research is restricted to the correct identification and interpretation of emotional states, particularly delicate and nuanced emotions, rather than the system's creation of emotional reactions or actions based on those feelings. In other words, the scope does not extend to building a comprehensive interactive feedback mechanism for a AI system; it is confined to correctly detecting and categorizing the user's emotional state given multimodal inputs. This approach allows for a clear and thorough examination of recognition ability, notably improvements in detecting subtle emotions when context is included explicitly.

Comparative Evaluation of Models

A substantial portion of the study focuses on a complete comparison examination between the proposed MCTAF model and current state-of-the-art multimodal emotion recognition models. This incorporates quantitative assessments utilizing existing benchmark datasets and conventional performance indicators (such as accuracy, F1-score, etc.). By keeping this within scope, the study will rigorously assess how the explicit inclusion of context as a modality influences performance relative to models that utilize only speech and text (and potentially those that use speech, text, and other non-contextual modalities without dedicated contextual fusion).

1.9 Organization of the Thesis

This thesis is organized into five chapters arranged as follows: Chapter 1 addresses the study subject by defining the relevance of emotion recognition and underlining the need to regard context as a different input modality. It specifies the problem statement, research questions, objectives, and scope of the study. It also describes the key contributions made by this study. Chapter 2 analyzes significant literature in multimodal emotion recognition. It begins with a review of classical and deep learning-based methodologies, then focuses on transformer designs and their relevance in this area. The chapter concludes by highlighting significant gaps in the research, particularly the limited coverage of context as an input modality, which supports the proposed methodology. Chapter 3 outlines the recommended method. It defines the general model architecture and discusses how voice, text, and context modalities are processed and fused via a transformer-based fusion network. The chapter also explains the datasets utilized, data preprocessing approaches, and the training processes used. Chapter 4 describes the experimental setup and conclusions. It defines the assessment measures and baseline models used for comparison, then shows and analyzes the outcomes. Both quantitative and qualitative studies are provided to illustrate the usefulness of context integration. The chapter concludes with a reflection on current limitations and suggestions for further study. Chapter 5 closes the research by summarizing the significant results and reinforcing the main contributions. It revisits the

study aims and underscores the relevance of modeling context explicitly in emotion recognition, highlighting its potential influence on the development of emotionally intelligent systems.

Chapter 2

Literature Review

This chapter offers a thorough survey of research in the area of multimodal emotion recognition, with a special focus on the expanding relevance of context. It begins by discussing the underlying concepts of MER and the historical transition from unimodal to multimodal systems. It then delves into the history of fusion approaches, the transformative effect of deep learning and transformer architectures, and the development of context-aware systems. The analysis concludes by highlighting significant gaps in current techniques and suggesting avenues for future research, particularly in the integration of dynamic, multi-layered contextual information.

2.1 Foundations

2.1.1 Theoretical and Historical Underpinnings

Historical Development of Emotion Recognition

The field of emotion recognition, positioned at the intersection of affective computing, cognitive science, and machine learning, has experienced rapid development since its inception in the late 20th century [51]. The theoretical foundations of this discipline can be traced back to the pioneering work of Rosalind Picard, who established the notion of emotional computing in 1997 [52]. Picard's groundbreaking book Affective Computing created the theoretical basis for computers that could perceive, analyze, and react to human emotions, radically undermining the conventional notion of computing as completely logical and emotion-free [52]. Picard's foundational work argued that for computers to be genuinely intelligent and interact naturally with humans, they must possess the ability to understand and respond to emotional states [52]. This perspective fundamentally shifted the paradigm of human-computer interaction from purely functional exchanges to more nuanced, emotionally-aware communications [51, 53].

Early studies in emotion detection aimed to identify emotions via discrete, unimodal channels such as facial expressions or voice intonation [53, 54]. These original attempts were primarily inspired by psychological ideas that stressed distinct emotion categories and conventional expression patterns, drawing from the fundamental work of Paul Ekman on basic emotions and facial action units [54]. Ekman's study identified six universal fundamental emotions —happiness, sorrow, anger, fear, surprise, and disgust — that can be recognized across cultures through unique facial expressions [54]. This categorical method provided a structured foundation for early computational emotion detection systems, which primarily relied on rule-based algorithms and manually extracted features derived from face landmarks and geometric correlations [51, 53].

However, real-world emotion expression proved to be significantly more complicated than these early models predicted. Emotions are intrinsically multimodal, context-sensitive, and frequently ambiguous, en-

couraging researchers to develop more holistic models that account for the interaction between numerous communication signals [53, 55]. The limitations of unimodal approaches became more evident as researchers encountered challenges such as occlusion in facial recognition systems, background noise interference in speech-based systems, and the inherent ambiguity of textual sentiment analysis [51, 53].

The transition from unimodal to multimodal emotion recognition was driven by both empirical limitations and theoretical advancements in cognitive science and psychology [53, 55]. Cognitive research demonstrated that human emotion perception naturally integrates multiple sensory channels, with the brain processing facial expressions, vocal cues, body language, and contextual information simultaneously to form coherent emotional interpretations [2, 47]. This insight inspired the creation of computational systems that could similarly combine many modalities to create more robust and accurate emotion recognition [51, 55].

Unimodal systems, while successful in limited laboratory contexts, struggled with generalizability and resilience in real-world, dynamic settings [51, 53]. For example, facial expression recognition algorithms typically failed when presented with partial occlusion, variable lighting conditions, or non-frontal face postures [51, 56]. Similarly, speech-based emotion recognition systems exhibited poorer performance in loud contexts or when dealing with speakers from diverse cultural backgrounds [53, 55]. Text-based sentiment analysis, although successful for explicit emotional expressions, sometimes misses subtle emotional subtleties transmitted by paralinguistic elements or contextual implications [51, 53].

Moreover, studies in psychology and neuroscience indicated that a single modality may not adequately convey the emotional intensity, valence, or authenticity of a person's internal affective state [2, 47]. A neutral facial expression mixed with tremulous speech and high physiological arousal, for instance, can signal underlying discomfort or anxiety that vision-only models would entirely ignore [57, 58]. This observation led to the concept that emotional expression is inherently a multimodal phenomenon that needs integrated study of numerous input channels to gain complete comprehension [53, 55].

Theoretical Foundations

MERs emerged to solve these issues by merging numerous data streams to generate a more thorough emotional profile [53, 55]. These systems draw from diverse modalities including audio signals (encompassing pitch, tone, prosody, and spectral characteristics), video data (including facial expressions, body language, and gestural information), textual cues (covering sentiment, semantics, and linguistic patterns), and physiological inputs (such as electroencephalography, electrocardiography, and galvanic skin response) [57–59]. Each modality adds unique and complementary information that, when merged successfully, may considerably increase the accuracy, robustness, and ecological validity of emotion recognition systems [51, 55].

Early work in emotional computing focused mainly on rule-based systems and standard statistical classifiers applied to single modalities [51, 53]. These early systems tended to concentrate on well-defined features such as Facial Action Unit (FAU)s as defined by Ekman and Friesen's Facial Action Coding System (FACS), or acoustic features like fundamental frequency, formants, and energy distribution in speech signals [53, 54]. While these methods provided a foundational understanding of emotion expression and offered reasonable

performance in controlled settings, their inability to generalize across diverse contexts, speakers, and real-world conditions highlighted the urgent need for more sophisticated multimodal approaches [51, 55].

The movement toward multimodal systems was further encouraged by developments in sensor technology and computational capabilities that made it conceivable to record and analyze several data streams in real-time concurrently citezhang2022survey, ahmed2023systematic. The widespread adoption of smartphones, wearable devices, and ubiquitous computing platforms gave unprecedented access to multimodal data, offering new prospects for emotion recognition applications in daily scenarios [58, 59]. Figure 2.1 depicts the conceptual distinction between unimodal processing and multimodal fusion.

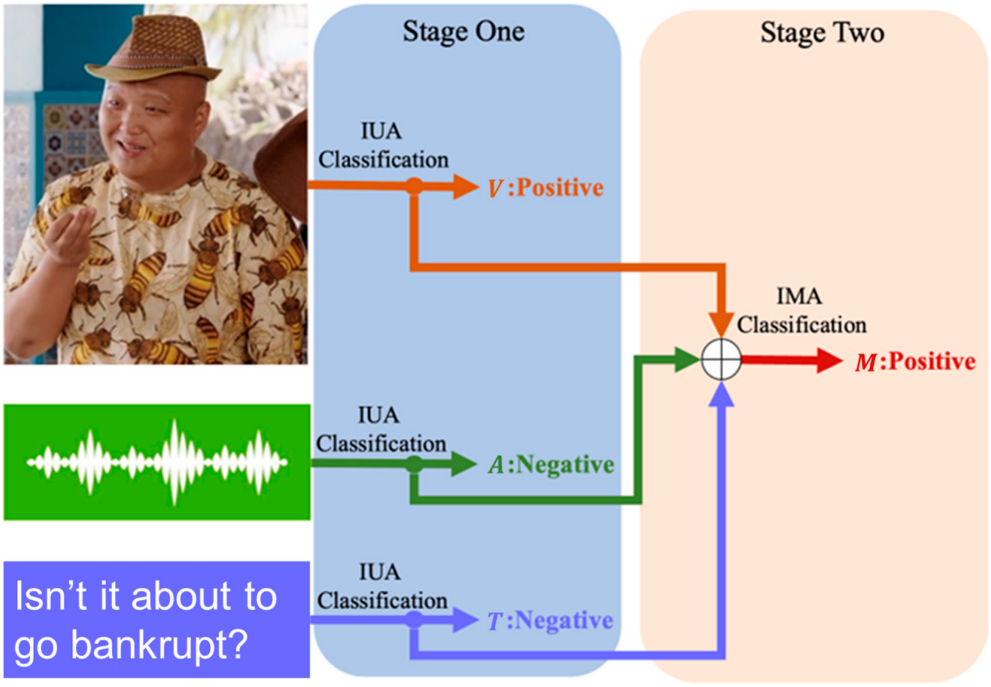

Figure 2.1: Conceptual illustration of unimodal processing (Stage One) leading to multimodal fusion and classification (Stage Two), highlighting potential differences between unimodal expression and multimodal perception (adapted from Zhang et al.[1])

Psychological Foundations

The theoretical foundations of emotion recognition systems are firmly based in psychology and cognitive science research, with multiple competing models offering alternative frameworks for interpreting and describing emotional states [2, 47]. The dimensional model of emotions, pioneered by Russell and Mehrabian in the 1970s, conceptualizes emotions along continuous dimensions rather than discrete categories [2]. This approach typically employs two primary dimensions: valence (representing the pleasantness or unpleasantness of an emotional state, ranging from positive to negative) and arousal (indicating the level of activation or energy associated with the emotion, ranging from calm to excited) [2].

The dimensional method has proved exceptionally adaptable to computer modeling because it offers a mathematical framework that can handle the continuous and frequently ambiguous character of emotional

expression in real-world contexts [2, 51]. Unlike categorical models that push emotions into discrete bins, dimensional models allow for nuanced representations that may capture minor fluctuations in emotional intensity and mixed emotional states [2, 53]. This flexibility has made dimensional models particularly popular in machine learning applications, where continuous output spaces are typically more appropriate for optimization algorithms and gradient-based learning techniques [51, 55].

The circumplex model of affect, established by James Russell in 1980, expands the dimensional method by putting emotions in a circular space defined by valence and arousal axes [2]. In this concept, psychologically similar emotions are situated closer together in the circular area, whereas opposing emotions are positioned at opposite ends [2]. The circumplex model has been particularly influential in computational emotion recognition because it provides a continuous representation that can capture the nuanced and often ambiguous nature of emotional expression, while preserving psychological validity [2, 53]. Figure 2.2 shows Russell's (1980) paradigm, which maps emotional states in a two-dimensional space. The horizontal axis symbolizes valence, indicating whether an emotion is seen as positive (pleasant) or negative (unpleasant). The vertical axis represents arousal, indicating the strength or energy level of the emotion, ranging from high to low activation.

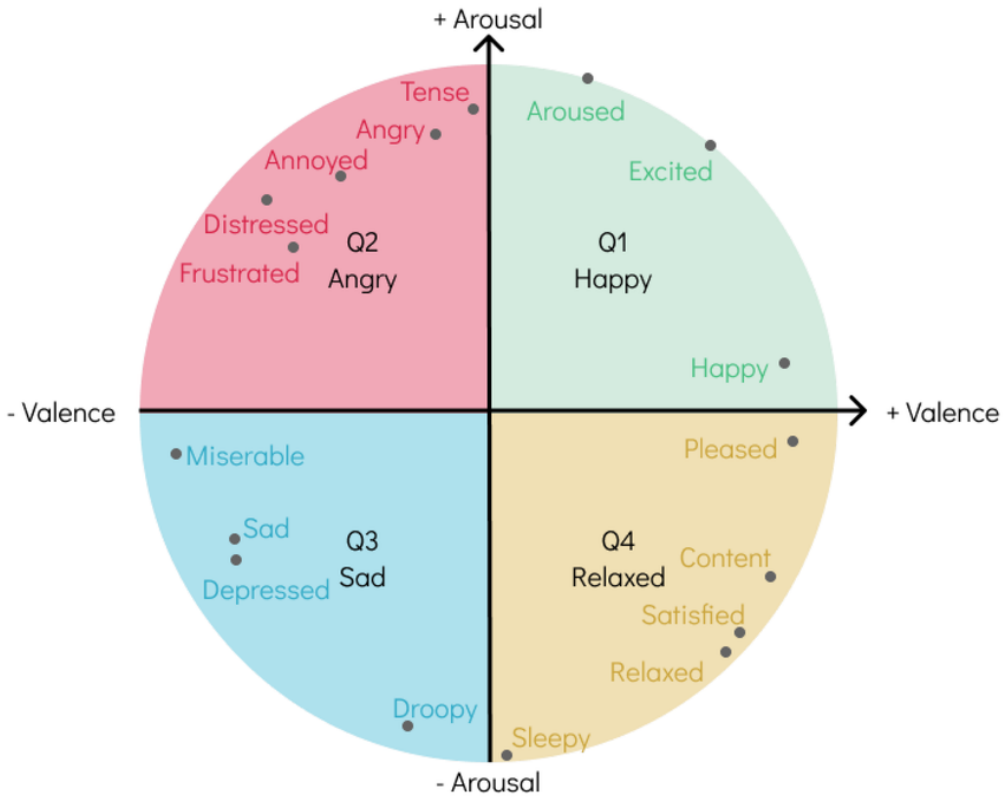

Figure 2.2: Q1:high arousal, positive, Q2:high arousal, negative, Q3:low arousal, negative, and Q4:low arousal, positive. Adapted from Russell 1980 [2])

Russell's circumplex model has been carefully examined by cross-cultural investigation and has demon-

strated remarkable consistency across diverse populations and situations [2, 51]. The model's circular structure captures meaningful relationships between emotions, such as the fact that high-arousal positive emotions (like excitement) are adjacent to high-arousal negative emotions (like anger). In contrast, low-arousal emotions (like sadness and contentment) occupy the opposite side of the circle [2]. This spatial structure has been helpful for computational systems, since it allows for meaningful interpolation between emotional states and gives a reasonable framework to manage ambiguous or mixed emotions [2, 55].

Recent advancements in affective computing have also integrated concepts from appraisal theory, which posits that emotions come from cognitive judgments of events and circumstances rather than being instantaneous reactions to stimuli [60, 61]. Appraisal theory, created by scholars such as Richard Lazarus and Klaus Scherer, says that emotional reactions rely on how people perceive the importance, consequences, and coping capabilities of encountered events [60, 61]. This method has encouraged the design of context-aware emotion detection algorithms that examine not just the immediate sensory signals but also the wider situational and temporal environment in which emotional expressions occur [62, 63].

The integration of appraisal theory into computational emotion recognition has led to increasingly intricate models that aim to capture the cognitive processes underlying emotional reactions [60, 63]. These models recognize that the same facial expression or speech pattern can convey multiple emotions, depending on the situational context, the individual's intentions and expectations, and their assessment of their ability to deal with the scenario [61, 62]. This insight has been valuable in designing context-aware systems that may adjust their interpretations depending on environmental and situational variables [63, 64].

Furthermore, recent research has begun to incorporate insights from constructionist theories of emotion, which propose that emotions are not fixed, universal categories but rather are constructed through the interaction of core affect, conceptual knowledge, and situational context [47]. Lisa Feldman Barrett's theory of constructed emotion says that emotional experiences develop from the brain's predictive processes, which blend interoceptive signals, prior experiences, and contextual information to generate meaningful emotional experiences [47]. This method has substantial implications for emotion recognition systems, suggesting that successful recognition needs not only the processing of immediate sensory inputs but also the integration of contextual information and learnt connections [47, 63].

The integration of these various psychological theories has led to more sophisticated computational models that aim to mimic the complexity and context dependency of human emotional experience [47, 63]. Modern emotion recognition systems often employ hybrid techniques that incorporate features from category, dimensional, and appraisal-based models to offer more comprehensive and adaptable emotion recognition capabilities [51, 55].

2.2 Datasets and Benchmarks for MER

The development of robust and comprehensive datasets has been fundamental to the advancement of multimodal emotion recognition research, providing the necessary training data and evaluation benchmarks that

enable fair comparison of different approaches and methodologies [44, 56]. The history of emotion recognition datasets reflects the broader developments in the field, from early unimodal collections to sophisticated multimodal corpora that embody the depth and diversity of human emotional expression [51, 53]. These datasets can be broadly categorized based on their collection methodology: controlled and acted datasets designed to elicit explicit, prototypical emotional expressions; spontaneous and interactive datasets that capture natural human communication; "in-the-wild" corpora sourced from unconstrained environments like the internet; and specialized datasets focusing on physiological signals or unique challenges like context and modality discrepancy.

Controlled and Acted Datasets

The early experiments in multimodal emotion recognition relied on real datasets obtained in controlled laboratory circumstances. While sometimes criticized for lacking ecological validity, these corpora are vital for basic research. By instructing actors to depict particular emotions, these datasets give clean, high-quality, and balanced examples of archetypal emotional expressions across multiple modalities, which is crucial for training and testing models on the core signals of emotion [44].

A noteworthy example of this category is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [65]. This dataset features 24 professional actors (12 male, 12 female) vocalizing two lexically-matched sentences in a neutral North American accent. The expressions cover eight emotions: neutral, calm, joyful, sad, furious, terrified, surprise, and disgust, each shown at two degrees of emotional intensity (normal, strong). The comprehensiveness of RAVDESS is one of its key features; it produces high-quality audio (48kHz), video (720p), and motion-capture data for facial landmarks. This allows researchers to isolate and examine speech prosody, facial action units, and their interplay without the confounding elements of elaborate backgrounds or conversational dynamics. Its balanced design and clear emotional descriptions have set it as a frequent standard for speaker identification, gender classification, and basic emotion detection tasks [65].

Similarly, the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) was constructed to provide a large-scale, culturally diverse collection of performed emotional expressions [66]. This dataset features 91 actors from a wide range of ages and ethnic backgrounds, who stated 12 separate words for each of the six primary emotions (happy, sad, angry, fearful, disgusted, and neutral). The sentiments were portrayed at four varying intensity levels, providing a thorough examination of how emotional expression is altered. CREMA-D offers synchronized audio and video modalities, and its significant value is in its scale and diversity, which assists in training more generalizable models that are less biased towards a specific demographic. The crowd-sourced nature of actor recruitment also provides a degree of variety that more closely reflects the entire population than datasets that employ exclusively professionally trained actors [66].

Other notable datasets in this category include the Surrey Audio-Visual Expressed Emotion (SAVEE) database [67], which contains recordings of four male actors portraying the six basic emotions plus neutral, and the Toronto Emotional Speech Set (TESS) [68], which focuses solely on audio but provides a foundational

set of cleanly articulated emotional speech. These earlier datasets cleared the ground for more advanced multimodal corpora by providing established techniques for requesting and collecting emotional data under controlled situations.

Spontaneous, Interactive, and Context-aware Datasets

Recognizing the limitations of merely conducting data, the research community focused on capturing more realistic and spontaneous emotions that arise during human interactions. These datasets are more complex due to the intricacies, co-articulation, and rapid fluctuations of realistic emotional states. Still, they are crucial for creating algorithms that can operate in real-world conditions.

The IEMOCAP database is a milestone accomplishment in this sector, acting as a bridge between performed and naturally spontaneous data [44]. Containing roughly 12 hours of audio-visual recordings, it features pairs of performers performing both planned settings and spontaneous interactions meant to evoke genuine emotional responses. This dyadic context is vital, as it represents the reciprocal part of human feeling. IEMOCAP is thoroughly annotated with both categorical labels (e.g., anger, happiness, sadness, neutral) and dimensional attributes (valence, activation, dominance), graded by several annotators to ensure reliability. The integration of high-fidelity motion capture data for face and hands delivers reliable information about non-verbal signals, establishing IEMOCAP a gold standard for exploring multimodal emotional dynamics in conversation [44, 51].

Building on the improvisation paradigm, the MSP-IMPROV corpus was created to inspire even more spontaneous emotions [69]. It includes of audio-visual recordings of artists in dyadic meetings who improvise discourses based on target texts or conditions. Unlike IEMOCAP's more scheduled improvisations, MSP-IMPROV promotes more spontaneous and unstructured emotional flows. This approach provides a vast collection of subtle emotional transitions and mixed emotional states that are less exaggerated than those in performed datasets, offering a valuable resource for researching the nuances of spontaneous multimodal expression [69].

Another incredibly significant dataset is Sustained Emotionally-colored Machine-human Interaction via Nonverbal Expression (SEMAINE) [70]. It was specifically designed to aid the establishment of Sensitive Artificial Listeners, entities capable of engaging in coherent, emotionally charged conversations. The data includes customers conversing with one of four characters (e.g., joyous, sad, angry, pragmatic), with the interactions taken by high-quality audio and multiple video cameras. A crucial element of SEMAINE is its thorough, continuous, and time-varying annotations for dimensions like arousal, valence, power, and expectation, contributed by multiple raters. This granular annotation method is suitable for training models to detect emotional state changes over time, a key capability for any real-time interactive system [70].

Similarly, the RECOLA (Remote Collaborative and Affective analysis) dataset was given for the continuous prediction of emotion in a realistic, task-oriented situation [71]. It features video recordings of French-speaking persons conversing remotely on a survival challenge, a setting that naturally evokes a range of subjective reactions. Like SEMAINE, its primary addition is the continuous, time-varying annotations of

arousal and valence. The incorporation of physiological data like electrocardiogram (ECG) and electrodermal activity (EDA) with audio and video makes it a valuable resource for multimodal affective computing and a classic benchmark in the yearly AVEC (Audio/Visual Emotion Challenge) series [71].

The ultimate aim for many emotion recognition systems is to operate "in the wild"-in uncontrolled, real-world circumstances. Datasets gathered from sites like YouTube, movies, and television shows supply the essential data to train models that are resistant to challenges such as varying lighting, head orientations, occlusions, background noise, and, most critically, the surrounding context.

One of the most essential datasets in this field is the Acted Facial Expressions in the Wild (AFEW) dataset [56]. Sourced from movies, AFEW was explicitly intended to bridge the gap between lab-controlled and wild emotion recognition. It offers short video snippets with seven emotion categories (anger, disgust, fear, happiness, sorrow, surprise, and neutral). By using movie data, AFEW presents a full range of emotional expressions in realistic scenarios, including sophisticated social interactions and diverse ambient conditions, making it a highly demanding and frequently used benchmark [56].

The Aff-Wild and Aff-Wild2 datasets indicate the next step in this line of inquiry, significantly scaling up the volume and complexity of in-the-wild data [72, 73]. These datasets include hundreds of hours of video footage of individuals displaying spontaneous emotions in entirely uncontrolled circumstances, acquired from YouTube. A crucial element of Aff-Wild is its annotation scheme: instead of discrete categories, it supplies continuous, time-varying values for valence (how positive or negative an emotion is) and arousal (how relaxing or exhilarating it is). Aff-Wild2 enhances this by adding annotations for the six primary emotions, action units, and landmarks. As the basis for the biennial Affective Behavior Analysis in-the-Wild (ABAW) competition, these datasets are driving advances in creating models that can perform robust, continuous, and multi-faceted emotion analysis in real-world circumstances [74].

The acceptance of context as a crucial regulator of emotion led to the establishment of specialized corpora. The Context-Aware Emotion Recognition (CAER) dataset focuses on the role of the visual scene in recognizing emotion [62]. It comprises approximately 13,000 static photographs from television episodes and movies, each with a prominent subject and a rich environment. Annotators assessed the person's mood based on both their facial expression and the surroundings. CAER has revealed that contextual signals may substantially influence the experience of emotion and has encouraged the design of structures that implicitly represent both the person and their surroundings [62].

Conversational context is also significant, as examined by the MELD [43]. MELD features roughly 13,000 utterances from the TV comedy "Friends," annotated with emotion and sentiment categories. Crucially, it gives the aural, visual, and textual modalities for each comment within the flow of a multi-turn conversation. This paradigm allows models to learn how emotional states increase over a debate and how the emotion of one speaker affects the next. MELD has become an essential standard for conversational emotion recognition, extending models beyond single-utterance analysis to capture the temporal dynamics of human interaction [43].

Large-Scale and Specialized Modality Datasets

The third type of datasets comprises those notable for their vast scale, their concentration on physiological signals, or their introduction of novel research difficulties. These resources facilitate the training of deep learning models and push the frontiers of what can be learnt from multimodal inputs.

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset is one of the largest and most comprehensive of its kind [75]. It comprises roughly 23,000 video clips from more than 1,000 YouTube speakers, delivering an unprecedented degree of linguistic and demographic variation. Its essential power rests in its fine-grained annotations. For each clip, it offers constant intensity ratings for feeling (-3 to +3) and the six primary emotions (0 to 3). This detailed labeling, combined with the three core modalities (audio, visual, text), has made CMU-MOSEI and its predecessor, CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) [76], cornerstone benchmarks for developing sophisticated multimodal fusion techniques that can model the subtle intensity of affective states.

Beyond behavioral modalities, physiological signals provide a direct window into the autonomic nervous system's responsiveness to emotional stimuli. The Database for Emotion Analysis using Physiological signals (DEAP) dataset was a pioneering effort in this domain [57]. electroencephalogram (EEG) and peripheral physiological data (e.g., galvanic skin response, respiration) from 32 participants as they saw 40 one-minute-long emotional music videos. Participants supplied self-assessments of valence, arousal, and dominance. DEAP has been significant in developing emotion recognition from brain signals and other physiological data, enabling research into brain-computer interfaces for affect detection [57, 77].

Extending this work, the AMIGOS (A dataset for Mood, personality and affect research on Individuals and Groups) dataset addressed the limitation of short stimuli [78]. It integrates physiological data (EEG, ECG, Galvanic Skin Response (GSR)) and visual data from individuals watching long-form movies (over 10 minutes) designed to elicit diverse emotional states. AMIGOS is special in that it enables both solo and group viewing modes, allowing for the study of emotional contagion and group dynamics. This makes it an important resource for investigating lasting emotional experiences and their physiological foundation in different social contexts [78].

Finally, particular datasets pose unique issues that propel the discipline forward. The Chinese single- and multi-modal sentiment analysis dataset (CH-SIMS), a Chinese Multimodal Sentiment Analysis Dataset, for example, is a remarkable resource for both its cultural diversity and its focus on modality incongruity [79]. Containing around 2,200 video parts in Mandarin Chinese, it gives fine-grained emotion annotations. Crucially, it has implicit labels for contradicting multimodal signals, such as sarcasm, where excellent textual information might be delivered with a bad tone of voice. This motivates models to advance beyond basic fusion and develop more detailed reasoning about the linkages between modalities, a critical step toward fully human-like understanding [79]. Together, this enormous and rising ecosystem of datasets continues to push the area of multimodal emotion recognition toward more accurate, robust, and contextually aware systems.

2.3 Architectures and Approaches for MER

2.3.1 Multimodal Fusion Strategies

Early and Late Fusion Approaches

As multimodal emotion recognition research advanced over the 2000s and 2010s, the integration of multimodal signals began to adopt more structured and theoretically based techniques [53, 55]. Fusion methods, defined as the systematic process of combining data from different modalities to create unified representations or decisions, emerged as a central concern in the field, with researchers recognizing that the effectiveness of multimodal systems depends critically on how information from different channels is integrated [51, 55].

Early fusion, also known as feature-level fusion, involves concatenating features from all modalities at the input level, allowing machine learning models to learn joint representations from the very beginning of the processing pipeline [53, 55]. This technique believes that there are substantial correlations and dependencies between features from multiple modalities that are detected and exploited by learning algorithms [51, 55]. The mathematical description of early fusion is expressed as:

$$\mathbf{f}_{early} = \text{Concat}(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n) \tag{2.1}$$

where $\mathbf{f}_i$ represents the feature vector retrieved from modality i, and the concatenated feature vector $\mathbf{f}_{early}$ serves as input to a unified learning algorithm [53, 55].

Early fusion is particularly effective when modalities are temporally synchronized and exhibit strong correlations, as it allows the learning algorithm to discover complex cross-modal patterns and dependencies that might be missed by approaches that process modalities independently [51, 55]. However, early fusion also presents several challenges, including the curse of dimensionality when dealing with high-dimensional feature spaces, the assumption of temporal alignment between modalities, and the potential for one modality to dominate the learning process if feature scales are not properly normalized [53, 55].

Late fusion, in contrast, undertakes separate processing and prediction for each modality before integrating their outputs at the decision level [51, 53]. This strategy often employs ensemble approaches, weighted voting systems, or other decision-level combination strategies to combine the predictions from individual modality-specific classifiers [53, 55]. The mathematical concept of late fusion is represented as:

$$\mathbf{y}_{late} = \sum_{i=1}^{n} w_i \cdot \mathbf{y}_i \tag{2.2}$$

Where $\mathbf{y}_i$ represents the output (prediction or confidence score) from the classifier trained on modality i, and w_i are fusion weights that are learnt via optimization or preset based on domain knowledge [51, 55].

Late fusion has various benefits, including the capacity to handle modalities with diverse temporal characteristics, the freedom to utilize specific architectures optimized for each modality, and resilience to missing

or corrupted modalities [53, 55]. However, late fusion may overlook key cross-modal interactions and dependencies that could be beneficial for emotion recognition, as each modality is processed individually without considering information from other channels [51, 55].

The decision between early and late fusion typically relies on the unique features of the application domain, the nature of the available modalities, and the temporal linkages between distinct data streams [53, 55]. In practice, many effective multimodal emotion recognition systems employ hybrid techniques that integrate aspects of both early and late fusion to leverage the benefits of each strategy while mitigating their respective drawbacks [51, 55].

Intermediate and Hybrid Fusion

Recognizing the limitations of solely early or late fusion techniques, researchers have developed intermediate and hybrid fusion strategies that aim to incorporate the benefits of both paradigms while mitigating their respective drawbacks [25, 55]. These complex approaches leverage advanced architectures, such as attention mechanisms, memory networks, and graph-based reasoning, to enable dynamic interaction across modalities at multiple levels of abstraction [25, 80].

Intermediate fusion, also known as hybrid fusion, involves merging modalities at several levels across the processing pipeline rather than confining integration to either the feature level or decision level [25, 55]. This technique allows for the progressive integration of multimodal information, allowing the system to capture both low-level cross-modal correlations and high-level semantic associations [25, 80]. The mathematical framework for intermediate fusion is understood as:

$$\mathbf{h}^{(l)} = f^{(l)}(\mathbf{h}_1^{(l-1)}, \mathbf{h}_2^{(l-1)}, \ldots, \mathbf{h}_n^{(l-1)}) \tag{2.3}$$

Where $\mathbf{h}_i^{(l)}$ represents the hidden representation of modality i at layer l, and $f^{(l)}$ is a fusion function that merges representations from several modalities at that level [25, 55].

Attention-based fusion mechanisms have emerged as particularly effective intermediate fusion strategies, allowing models to dynamically weight the importance of different modalities and different aspects within each modality based on the current context and task requirements [25, 80]. The attention mechanism is expressed as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \tag{2.4}$$

where e_i is the attention score for modality i, and α_i is the normalized attention weight [25, 81]. The final fused representation is then calculated as:

$$\mathbf{h}_{fused} = \sum_{i=1}^{n} \alpha_i \cdot \mathbf{h}_i \tag{2.5}$$

where $\mathbf{h}_i$ is the representation from modality i [25, 80].

Memory networks constitute another complex method to intermediate fusion, allowing systems to preserve and update representations of multimodal information across time [82, 83]. These architectures are especially helpful for emotion detection tasks that require temporal dynamics, such as identifying emotional changes across a conversation or monitoring emotional states over lengthy periods [44, 84]. The memory updating method is represented as:

$$\mathbf{m}_t = \mathbf{m}_{t-1} + \alpha_t \cdot (\mathbf{h}_t - \mathbf{m}_{t-1}) \tag{2.6}$$

where $\mathbf{m}_t$ is the memory state at time t, $\mathbf{h}_t$ is the current input representation, and α_t is a learnt update gate [82, 83].

Graph-based fusion techniques depict the interactions between distinct modalities and different characteristics within each modality as nodes and edges in a graph structure [51, 55]. This approach allows for flexible modeling of complicated multimodal interactions and facilitates the use of graph neural networks for fusion [55]. The graph convolution operation is expressed as:

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{2.7}$$

where $\mathbf{A}$ is the adjacency matrix, $\mathbf{D}$ is the degree matrix, $\mathbf{H}^{(l)}$ is the node feature matrix at layer l, and $\mathbf{W}^{(l)}$ is the learnable weight matrix [51, 55].

These intermediate and hybrid fusion approaches have demonstrated significant improvements over traditional early and late fusion methods in various emotion recognition tasks, particularly in scenarios involving complex temporal dynamics, missing modalities, or heterogeneous data sources [25, 80]. However, they also bring extra computational complexity and need careful design to guarantee that the fusion mechanisms are suitable for the unique features of the target application [51, 55].

2.3.2 Core Deep Learning Architectures

The implementation of deep learning has fundamentally revolutionized the landscape of multimodal emotion recognition by allowing end-to-end learning, automated feature extraction, and more nuanced modeling of complicated intermodal dynamics [85, 86]. Before the deep learning revolution, most MER systems relied primarily on manually engineered features and standard machine learning classifiers, such as Support Vector Machines (SVM), decision trees, Random Forests, and Hidden Markov Models (HMM)s [51, 53]. While these approaches provided valuable insights and achieved reasonable performance in controlled settings, they were limited by their dependence on domain expertise for feature engineering and their inability to capture complex, non-linear relationships in multimodal data [51, 55].

The emergence of deep learning architectures signaled a paradigm change toward data-driven feature learning, where neural networks could automatically identify appropriate representations from raw multimodal input without needing explicit feature engineering [85, 86]. This capacity has proved especially beneficial for emotion recognition, where the relevant characteristics may be subtle, context-dependent, and

challenging to describe manually [51, 55]. Deep learning techniques have consistently demonstrated superior performance compared to standard methods across a wide range of emotion recognition benchmarks and real-world applications [85, 86].

Convolutional Neural Network (CNN)s for Visual Processing

CNNs have proved to be extremely good at collecting spatial patterns and hierarchical features in visual data, making them ideally suited for facial expression identification and other vision-based emotion recognition tasks [85, 86]. The essential power of CNNs resides in its capacity to learn translation-invariant features via the convolution process, which is mathematically represented as:

$$(\mathbf{I} * \mathbf{K})_{i,j} = \sum_m \sum_n \mathbf{I}_{i+m,j+n} \cdot \mathbf{K}_{m,n} \tag{2.8}$$

where $\mathbf{I}$ represents the input image, $\mathbf{K}$ is the convolution kernel (or filter), and $*$ signifies the convolution operation [85, 86].

The hierarchical nature of CNNs allows them to acquire more sophisticated and abstract representations as information flows through consecutive layers [85, 86]. Early layers typically capture low-level characteristics, such as edges, textures, and basic geometric patterns, whereas deeper layers learn more complex features, including facial components, expressions, and semantic concepts [81, 85]. This hierarchical feature learning capacity has proved vital for emotion recognition, as emotional expressions frequently entail complicated combinations of facial muscle movements and geometric connections [51, 56].

Modern CNN architectures for emotion recognition have included several sophisticated strategies to increase performance and resilience [85, 86]. Residual connections, introduced in ResNet designs, enable the training of deep networks by providing skip connections that help mitigate the vanishing gradient issue [85]. The mathematical concept of a residual block is written as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{2.9}$$

where $\mathbf{x}$ is the input, $\mathcal{F}(\mathbf{x}, \{W_i\})$ represents the residual mapping to be learnt, and $\mathbf{y}$ is the output [85].

Dense connections, as implemented in DenseNet architectures, further increase feature reuse and gradient flow by linking each layer to every other layer in a feed-forward way [86]. This connection pattern is stated as:

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}]) \tag{2.10}$$

Where $[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}]$ represents the concatenation of feature maps from all previous layers, and H_l is the composite function for layer l [86].

Attention mechanisms have also been incorporated into CNN architectures to enable more focused processing of key facial areas [25, 81]. Spatial attention processes may detect and emphasize significant face areas,

whereas channel attention mechanisms can highlight essential feature channels [81]. The spatial attention mechanism is expressed as:

$$\mathbf{A}_{spatial} = \sigma(f^{7\times7}([\text{AvgPool}(\mathbf{F}); \text{MaxPool}(\mathbf{F})])) \tag{2.11}$$

where $\mathbf{F}$ is the input feature map, $f^{7\times7}$ represents a convolution operation with a 7×7 kernel, and σ is the sigmoid activation function [81].

The integration of these advanced techniques has led to significant improvements in facial expression recognition performance, with state-of-the-art CNN-based systems achieving accuracy rates exceeding 95% on standard benchmarks such as CK+ [51, 56]. However, issues remain in managing real-world variables such as stance changes, lighting fluctuations, partial occlusions, and cultural disparities in emotional expression [51, 56].

Recurrent Neural Networks for Temporal Modeling

While CNNs excel at spatial feature extraction, Recurrent Neural Network (RNN)s and their variants, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Unit (GRU)s, have proven to be well-suited for modeling temporal dependencies in sequential data such as audio streams and text sequences [82, 83]. The capacity to capture temporal dynamics is crucial for emotion detection, as emotional manifestations typically develop over time and may involve complex temporal patterns that are necessary for accurate recognition [43, 44].

The basic RNN design maintains a hidden state that is updated at each time step according to the recurrence relation:

$$\mathbf{h}_t = f(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h) \tag{2.12}$$

where $\mathbf{h}_t$ is the hidden state at time t, $\mathbf{x}_t$ is the input at time t, $\mathbf{W}_{hh}$ and $\mathbf{W}_{xh}$ are weight matrices for hidden-to-hidden and input-to-hidden connections respectively, $\mathbf{b}_h$ is the bias vector, and f is a non-linear activation function [82, 83].

However, classical RNNs suffer from the vanishing gradient issue, which makes it challenging to learn long-term dependencies in sequential data [82]. This constraint is especially troublesome for emotion detection tasks that may involve comprehension of long-term temporal patterns and context [43, 44].

LSTM networks address the vanishing gradient issue through complex gating mechanisms that regulate the flow of information across the network [82]. The main equations regulating LSTM functioning are:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{2.13}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{2.14}$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \tag{2.15}$$

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t \tag{2.16}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{2.17}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \tag{2.18}$$

Where $\mathbf{f}_t$, $\mathbf{i}_t$, and $\mathbf{o}_t$ are the forget, input, and output gates respectively, $\mathbf{C}_t$ is the cell state, $\tilde{\mathbf{C}}_t$ is the candidate values, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication [82].

GRUs present a simpler alternative to LSTMs that combines the forget and input gates into a single update gate and integrates the cell state and hidden state [83]. The GRU equations are:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{2.19}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{2.20}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W} \cdot [\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{2.21}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{2.22}$$

where $\mathbf{z}_t$ is the update gate, $\mathbf{r}_t$ is the reset gate, and $\tilde{\mathbf{h}}_t$ is the candidate hidden state [83].

GRUs have demonstrated equivalent performance to LSTMs in numerous emotion recognition tasks while using fewer parameters and computing resources [51, 83]. The choice between LSTMs and GRUs generally relies on the unique features of the job and the available computing resources [82, 83].

2.3.3 Attention Mechanisms and Transformer Architectures

Self-Attention and Multi-Head Attention

The introduction of attention mechanisms has revolutionized the field of deep learning by enabling models to selectively focus on relevant parts of the input when making predictions, fundamentally changing how neural networks process and integrate information [25, 80]. In the context of emotion recognition, attention mechanisms are particularly valuable because emotional expressions often involve subtle cues that can be localized in specific regions of the face, particular frequency bands in speech signals, or specific time intervals in physiological recordings [25, 81]. The ability to automatically detect and focus on these critical areas or traits has been crucial for developing high-performance emotion recognition systems [80, 87].

The self-attention mechanism, as formalized in the groundbreaking Transformer architecture (see Figure 2.3), introduced by Vaswani et al. in 2017, computes attention weights based on the relationships

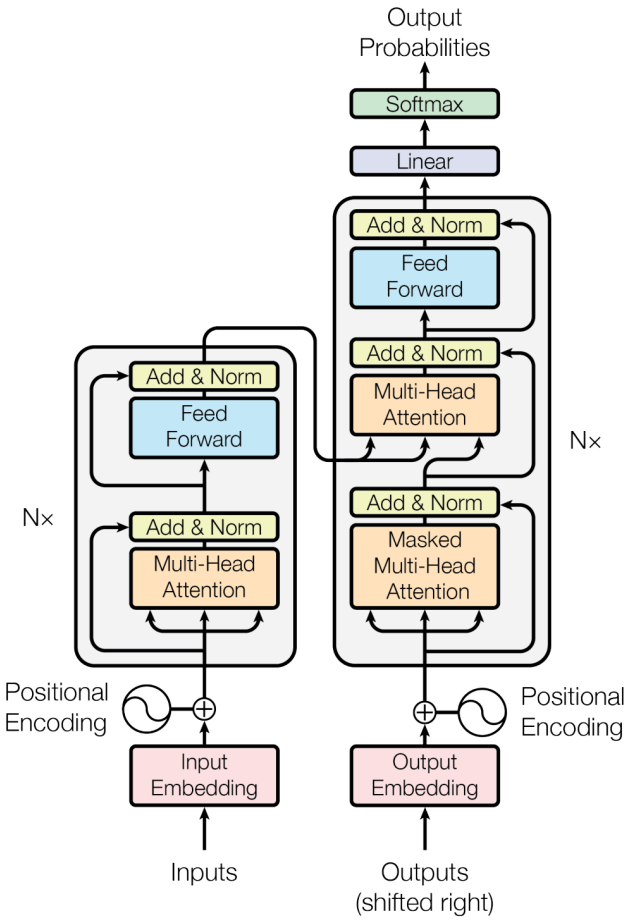

Figure 2.3: The Transformer model architecture, exhibiting the encoder-decoder structure with multi-head self-attention and feed-forward layers (derived from Vaswani et al., 2017 [3]).

between different positions in the input sequence [25]. This mechanism allows the model to determine which parts of the input are most relevant for processing each element, enabling the capture of long-range dependencies and complex relationships that traditional recurrent architectures struggle to model effectively [25, 87]. This feature is powerful for capturing the temporal development of an emotion, where a prior change in voice tone may inform a grin at the conclusion of an utterance. The interpretability of self-attention is also a significant advantage; by visualizing the attention weights, researchers can gain insights into the model's decision-making process, for instance, by observing whether the model focuses on the eyes and mouth for recognizing happiness or on prosodic features in a spectrogram for detecting anger [88].

The mathematical concept of self-attention is represented as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.23}$$

Where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively, produced from the input by learnt linear transformations, and d_k is the dimension of the key vectors [25]. The scaling factor $\sqrt{d_k}$

prevents the dot products from getting excessively big, which might lead to very tiny gradients when the softmax function is used [25]. A key problem of this approach, however, is its computational cost, which is quadratic with respect to the input sequence length ($O(n^2)$). This bottleneck makes it computationally costly to process high-resolution video frames or extended audio snippets. To mitigate this, efficient Transformer variants like the Longformer, which uses a combination of local windowed attention and global attention, have been proposed to model long sequences with linear complexity, making them more suitable for real-world multimodal applications [87].

Multi-head attention extends the fundamental attention mechanism by calculating several attention functions in parallel, enabling the model to attend to input from distinct representation subspaces concurrently [25, 80]. This is especially successful in multimodal environments, since separate heads may specialize in catching diverse patterns. For example, one brain would learn to monitor facial action units, another might concentrate on high-level vocal intonation patterns, and a third might record the link between a gesture and a spoken sentence. The mathematical formulation of multi-head attention is:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{2.24}$$

Where each attention head is calculated as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \tag{2.25}$$

and $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{V}_i^V$, and $\mathbf{W}^O$ are learned parameter matrices [25, 80].

Research by Wang et al. established how multi-head attention can be successfully used to multimodal emotion recognition, demonstrating that various attention heads spontaneously learnt to concentrate on distinct modalities and different elements within each modality [80]. Their study indicated that attention heads could automatically uncover essential patterns such as the association between certain facial expressions and voice prosody, or the relationship between textual mood and acoustic properties [80].

Transformer Architecture for Multimodal Processing

The Transformer architecture marks a significant paradigm shift from recurrent and convolutional techniques, focusing entirely on attention mechanisms to describe dependencies in sequential data [25, 87]. This architecture is especially well-suited for multimodal emotion recognition tasks owing to its capacity to manage variable-length sequences, model long-range relationships, and process many modalities in a unified framework [87, 89].

The Transformer comprises an encoder-decoder structure where each component is constructed of many layers comprising multi-head self-attention and position-wise feedforward networks [25]. The encoder analyzes the input sequence and creates a series of representations, while the decoder generates the output sequence based on the encoder representations and previously created outputs [25, 87]. For many emotion recognition tasks, only the encoder part of the Transformer is employed, as the task often involves classification rather

than sequence creation [87, 89]. To better handle specific modalities, researchers typically deploy customized Transformer encoders as feature extractors before multimodal fusion. For instance, the Vision Transformer (ViT) interprets an image as a series of patches and has become a standard for visual feature extraction [42]. Similarly, the Audio Spectrogram Transformer (AST) applies the Transformer design directly to audio spectrograms, efficiently capturing temporal and frequency patterns for applications like emotion recognition from speech [90]. These modality-specific encoders provide sophisticated, context-aware representations that serve as inputs to a future fusion module.

The mathematical concept of a Transformer encoder layer is represented as:

$$\mathbf{Z}^{(l)} = \text{LayerNorm}(\mathbf{X}^{(l-1)} + \text{MultiHead}(\mathbf{X}^{(l-1)})) \tag{2.26}$$

$$\mathbf{X}^{(l)} = \text{LayerNorm}(\mathbf{Z}^{(l)} + \text{FFN}(\mathbf{Z}^{(l)})) \tag{2.27}$$

where $\mathbf{X}^{(l)}$ represents the output of layer l, FFN is the position-wise feedforward network, and LayerNorm is layer normalization [25, 87]. For multimodal emotion recognition, researchers have built various adaptations of the Transformer architecture that can effectively process and integrate multiple modalities [87, 89]. One popular technique comprises deploying different Transformer encoders for each modality, followed by a cross-modal Transformer that depicts interactions across modalities [89]. This is a form of hybrid fusion, where unimodal processing precedes multimodal integration. Another method utilizes a unified Transformer that processes concatenated multimodal sequences, employing specific tokens or embeddings to differentiate between multiple modalities [87]. In this "early fusion" strategy, a modality embedding is typically added to each input token's embedding to tell the model about its origin (e.g., text, audio, or video). This provides a single self-attention mechanism that learns both intra-modal and inter-modal interactions simultaneously from the first layer.

Cross-Modal Attention

A significant discovery within the attention mechanism space is cross-modal attention, which enables models to attend to information from one modality based on queries from another modality [89, 91]. This skill is invaluable for multimodal emotion recognition, because the emotional meaning of information in one modality may rely on the context given by other modalities [87, 89]. For example, the sentence "that's just great" might signal absolute delight or extreme sarcasm, and the model can disambiguate this by utilizing the textual representation as a query to attend to facial expressions and speech tone in the visual and auditory modalities.

Cross-modal attention is formulated as:

$$\text{CrossAttention}(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j) = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{d_k}}\right)\mathbf{V}_j \tag{2.28}$$

where the queries $\mathbf{Q}_i$ originate from modality i and the keys $\mathbf{K}_j$ and values $\mathbf{V}_j$ come from modality j [89, 91]. This directed flow of information is a critical characteristic, allowing for the modeling of asymmetric relationships.

Research by Rahman et al. demonstrated the effectiveness of cross-modal attention for multimodal sentiment analysis, showing that models could learn to focus on relevant textual content based on visual cues, and attend to critical visual regions based on textual context [91]. Their findings revealed how cross-modal attention may capture nuanced interactions across modalities that would be difficult to represent using typical fusion techniques [91]. A key difficulty addressed by sophisticated cross-modal designs is the intrinsic asynchrony of multimodal signals. Emotional indicators in voice, face, and body language do not appear simultaneously. Cross-modal attention may assist in aligning these signals by enabling the model to search throughout a temporal frame in one modality to discover the most relevant information matching a given occurrence in another modality. More recent fusion frameworks, such as that proposed by Pham et al., have explored novel ways to structure these interactions, suggesting that fusion in the intermediate layers of a network, rather than only at the beginning or end, can lead to more robust representations by allowing modalities to synergize throughout the feature extraction process [92].

The Multimodal Transformer (MulT) architecture proposed by Tsai et al. showed how cross-modal attention might be utilized to align and integrate input from text, audio, and video modalities for emotion recognition and sentiment analysis tasks [89]. Their technique, which incorporates a cascade of directed, paired cross-modal attention blocks, enables each modality to be progressively enhanced by the others. This model exhibited considerable improvements over classic fusion approaches and generated new state-of-the-art results on various multimodal benchmarks, reinforcing the significance of cross-modal attention as a cornerstone of current multimodal learning [89].

2.3.4 Context-Aware Emotion Recognition

The Importance of Context in Emotion Recognition

Context-aware emotion recognition represents one of the most significant and challenging frontiers in the field of multimodal emotion recognition, addressing the fundamental limitation that most existing systems fail to incorporate contextual information when interpreting emotional expressions adequately [62, 63]. The importance of context in human emotion perception has been well-established in psychological research, which demonstrates that the same facial expression or vocal pattern can convey entirely different emotions depending on the situational context, cultural background, and interpersonal dynamics [2, 47]. Seminal work by Aviezer et al. demonstrated this dramatically by showing that the perception of an intense facial expression (e.g., of a tennis player winning a point) could be flipped from positive to negative simply by placing it on a different body posture, proving that context can sometimes be more informative than the face itself [93].

The pioneering work of Lee et al. in 2019 introduced the concept of Context-Aware Emotion Recognition

Network (CAER-Net)s, which implicitly model the interaction between human facial expressions and their surrounding environmental context [62]. Their study found that standard emotion recognition algorithms, which concentrate primarily on facial characteristics, frequently fail to capture the genuine emotional state of people because they neglect critical contextual signals that humans naturally utilize to understand emotions [62]. For instance, a neutral facial expression in a funeral environment typically suggests grief or solemnity, but the same expression at a comedy performance would imply boredom or bewilderment [62, 63].

The CAER-Net architecture offered a two-stream method that independently analyzes facial regions and contextual background information before combining them using adaptive attention mechanisms [62]. The facial stream concentrates on collecting particular expression characteristics from the face region. In contrast, the context stream examines the larger picture to identify environmental variables, objects, other people, and situational signals [62, 63]. The mathematical formulation of their fusion strategy is represented as:

$$\mathbf{f}_{final} = \alpha \cdot \mathbf{f}_{face} + (1 - \alpha) \cdot \mathbf{f}_{context} \tag{2.29}$$

Where $\mathbf{f}_{face}$ and $\mathbf{f}_{context}$ represent features from the facial and contextual streams, respectively, and α is a learned attention weight that defines the relative relevance of facial vs contextual information [62].

Building upon this fundamental work, Yang et al. developed the notion of context-aware emotion recognition by defining and modeling four complementing categories of context: multimodal context, semantic context, spatial context, and temporal context [63]. Their comprehensive framework acknowledged that context functions at numerous levels and scales, from immediate sensory information to higher-level semantic comprehension and temporal dynamics [63, 64]. A significant part of temporal context, for example, is conversational history. In dyadic or group interactions, the emotion of a current statement is greatly impacted by prior turns in the discussion. Models created for datasets like MELD must thus handle the conversation history as a crucial component of the context to appropriately understand emotions, such as sarcasm or responses like surprise [43].

Addressing Context Bias and Spurious Correlations

One of the most serious issues in context-aware emotion detection is the problem of context bias and false correlations, which can lead to models that rely on superficial contextual signals rather than genuine emotional knowledge [64]. Yang et al. addressed this key problem in their 2023 work on context-deconfounded emotion detection, offering a causal inference framework that aims to mitigate the effect of false correlations between contextual data and emotion labels [64].

The issue of context bias arises when models learn to associate particular contextual elements with specific emotions based on dataset biases rather than genuine causal correlations [64]. For example, if a training dataset includes numerous photographs of individuals feeling sad in rainy weather, a model can learn to correlate rain with sorrow, even if the weather itself is not a direct source of the emotional expression [64]. Such misleading correlations might lead to poor generalization when the model meets contexts that vary from

those encountered during training, such as a cheerful couple kissing in the rain [64]. This behavior, commonly referred to as "learning shortcuts," suggests that the model may overlook the salient facial expression entirely if a sufficient contextual prior exists, resulting in fragile and inconsistent performance in real-world scenarios.

The context de-confounding strategy presented by Yang et al. leverages causal inference techniques to detect and reduce these false correlations [64]. Their technique employs a causal graph to represent the interactions between contextual elements, facial expressions, and real emotional states, allowing the discovery of confounding variables that cause misleading correlations [64]. The mathematical underpinning for their method is represented using the structural causal model:

$$E = f_E(F, C, U_E) \tag{2.30}$$

$$F = f_F(E, C, U_F) \tag{2.31}$$

$$C = f_C(U_C) \tag{2.32}$$

where E represents the real emotion, F represents the facial expression, C represents the context, and U_E, U_F, U_C are unobserved confounding factors [64]. Other ways to overcome context bias include adversarial training and data augmentation, when models are actively trained on counter-intuitive instances (e.g., cheerful faces in a library, sobbing faces at a birthday celebration). This challenges the model to develop a more robust decision boundary that does not depend exclusively on naive context-emotion correlations. Research by Kervadec et al. has demonstrated that punishing models for relying on prominent but possibly misleading characteristics might enhance generalization by driving them to locate more subtle and authentic evidence for their predictions [94].

Multi-Granularity and Hierarchical Context Modeling

Recent developments in context-aware emotion recognition have focused on constructing more complex models that can capture context at numerous granularities and hierarchical levels [95, 96]. The awareness that context acts at multiple scales-from local face areas to global scene understanding- has led to the creation of multi-granularity attention mechanisms that can concurrently represent fine-grained and coarse-grained contextual information [95]. This goes beyond the chronological realm to incorporate geographical and semantic hierarchies. For instance, a model might analyze a scene at a fine spatial granularity to identify a person's interaction with a specific object (e.g., receiving a gift), at a medium granularity to understand the immediate social setting (e.g., a birthday party), and at a coarse granularity to classify the overall environment (e.g., a home) [96].

Fan et al. presented the Multi-Granularity Attention Based Transformers (MGAT) architecture, which tackles the difficulties of emotional asynchrony and modality misalignment via hierarchical attention processes [95]. Their approach emphasizes that emotional manifestations typically include complicated temporal dynamics, with various modalities expressing emotions at different time scales and with varied degrees of

synchronization [95].

The MGAT architecture includes a hierarchical attention structure that acts at several temporal granularities:

$$\mathbf{A}^{(g)} = \text{softmax}\left(\frac{\mathbf{Q}^{(g)}(\mathbf{K}^{(g)})^T}{\sqrt{d_k}}\right) \tag{2.33}$$

Where g specifies the granularity level, and various granularity levels capture temporal patterns at different scales [95]. Fine-grained attention focuses on short-term temporal patterns and immediate contextual connections, whereas coarse-grained attention captures longer-term temporal dependencies and larger contextual knowledge [95]. By creating a hierarchy of contextual representations, models may efficiently discern between a momentary expression of surprise that rapidly resolves and a prolonged sense of concern that continues over a more extended period, resulting in more nuanced and accurate emotion detection.

Environmental and Situational Context Integration

The integration of environmental and situational context has emerged as a critical component of improved emotion recognition systems, understanding that human emotional manifestations are significantly impacted by their physical and social surroundings [62, 96]. Zhang et al. created a three-dimensional view relationship-based context-aware emotion recognition system that implicitly models the spatial connections between persons and their surroundings [96].

Their method involves several types of environmental context, including scene semantics (understanding what type of environment the person is in), object relationships (identifying relevant objects and their emotional associations), social context (considering the presence and emotional states of other people), and activity context (understanding what activities are taking place) [96]. A big problem here is infusing commonsense information into the model. For instance, to grasp that a "library" suggests quietude while a "rock concert" means excitement, models may need to be reinforced by external information. Recent work by Ruan et al. addresses the use of knowledge graphs to develop an Emotion-Reasoning module, which enables the model to infer situational emotions by reasoning over a structured graph of ideas and their affective implications [97].

The mathematical framework for integrating these different types of context is expressed as:

$$\mathbf{C}_{env} = \text{Concat}(\mathbf{C}_{scene}, \mathbf{C}_{objects}, \mathbf{C}_{social}, \mathbf{C}_{activity}) \tag{2.34}$$

Where each component reflects a distinct part of the environmental context [96]. Modeling the social context component is highly challenging, since it entails not only detecting other persons in the scenario but also comprehending their emotional states and inter-personal interactions. Advanced techniques currently employ Graph Neural Network (GNN)s to describe a group of people as a graph, where nodes represent individuals and edges show their interactions or effect on one another. This enables the model to capture phenomena like emotional contagion, where one person's happiness may impact the feelings of people around

them, resulting in a more holistic, group-level view of the emotional landscape [98].

Despite the tremendous development, numerous fundamental obstacles continue to influence the research environment [51, 55]. A recurrent problem is that of modality misalignment and asynchrony, since emotional signals across face, voice, and body seldom come in perfect temporal lockstep [89, 95]. Systems must also exhibit resilience in the face of missing or corrupted modalities-a typical occurrence in real-world circumstances owing to inadequate illumination, background noise, or sensor failure [55]. Furthermore, the high computational cost of modern transformer architectures presents a significant barrier to deployment on resource-constrained devices. At the same time, the challenge of cross-domain and cross-corpus generalization remains a substantial obstacle to building truly universal emotion recognition systems [53, 87].

Beyond these technological challenges, the increasing capacity of emotion recognition technology raises serious ethical and privacy concerns to the forefront. The capacity to automatically infer an individual's inner state raises significant problems regarding permission, autonomy, and the possibility for exploitation of sensitive emotional data [51, 58]. The risk of embedding societal biases into these systems is also substantial; models trained on unrepresentative data can perpetuate and even amplify discrimination against certain demographic or cultural groups, highlighting the urgent need for fairness and transparency in their design and deployment [53, 55].

In response to these issues and to push the boundaries of emotional computing, various potential research avenues are being developed. The integration of large foundation models offers the potential to endow systems with a deeper, commonsense understanding of context, utilizing massive pre-trained knowledge to read emotional expressions more accurately [87]. Concurrently, research into few-shot, zero-shot, and ongoing learning strives to construct more flexible systems that can learn from minimal data and grow over time in dynamic situations [51, 55]. Finally, the use of causal inference and explainable AI approaches is becoming critical for constructing models that are not only accurate but also resilient and trustworthy, capable of getting beyond false correlations to grasp the actual drivers of emotional expression [64].

Summary

This literature study has traced the tremendous development of multimodal emotion recognition, from its core concepts to the present state-of-the-art. The research has moved from standalone, unimodal studies to sophisticated, integrated systems that employ deep learning and attention processes to analyze the subtle tapestry of human emotional expression. The progression reflects a greater respect for the intrinsically multimodal and context-dependent character of emotion, encouraging the creation of more complex structures. This progression in capabilities and performance, from early feature concatenation to advanced, context-aware transformer models, is systematically detailed in Table 2.1, which provides a comparative overview of the dominant methodologies, their strengths, and their inherent limitations.

Table 2.1: Comprehensive Comparison of Multimodal Emotion Recognition Approaches

Approach Category	**Representative Methods**	**Architecture Type**	**Modalities**	**Key Advantages**	**Main Limitations**
Early Fusion	Feature-level concatenation, as discussed [53, 55]	CNN+LSTM, multi-layer perceptron (MLP)	Audio-Visual, Audio-Visual-Text	Simple implementation, captures cross-modal correlations	Assumes synchronization, high dimensionality
Late Fusion	Decision-level ensembles, as discussed [51, 53]	Modality-specific networks	Audio-Visual-Text-Physiological	Handles asynchrony, modality-specific optimization	Misses deep interactions
Intermediate Fusion	Attention-based fusion [80], Memory networks [82, 83]	Transformer, Graph networks	Audio-Visual-Text	Captures multi-level interactions, flexible	Complex architecture, training difficulty
CNN-based Visual	ResNet [85], DenseNet [86], Attention CNNs [81]	Convolutional networks	Visual (facial expressions)	Spatial feature extraction, translation invariance	Limited temporal modeling
RNN-based Temporal	LSTM [82], GRU [83]	Recurrent networks	Audio, Text, Physiological	Temporal dependency modeling, sequential processing	Vanishing gradients, computational cost
Transformer-based	Multimodal Transformer (MulT) [89], ViT [42], AST [90]	Self-attention networks	Audio-Visual-Text	Long-range dependencies, parallel processing	High computational cost, data requirements
Context-Aware	CAER-Net [62], Context De-confounding [64], MGAT [95]	CNN + Attention, Causal models	Visual, Multimodal	Incorporates situational context, reduces bias	Complex context modeling, dataset requirements

Continued on next page

Table 2.1: Comprehensive Comparison of Multimodal Emotion Recognition Approaches (Continued)

Approach Category	**Representative Methods**	**Architecture Type**	**Modalities**	**Key Advantages**	**Main Limitations**
Graph-based	DialogueGCN [29], MMGCN [99]	Graph neural networks	Audio-Visual-Text	Models complex relationships, flexible topology	Graph construction complexity, scalability
Physiological	EEG-based [77], DEAP dataset [57], WESAD dataset [58]	CNN+RNN, Signal processing	EEG, ECG, GSR + Audio-Visual data: Galvanic Skin Response	Direct neural/physiological signals, objective measures	Invasive / obtrusive, noise sensitivity

2.4 Related Works

The following section will focus on analyzing specific works that most directly address the research gap explored in this thesis. The debate tries to emphasize both the progress of methodological sophistication and the continuing limitations that inspire the present investigation.

Early attempts in context-aware emotion detection, especially inside conversation systems, centered on recurrent neural networks. DialogueRNN [31] and ICON [100] represent two significant milestones in this family. DialogueRNN uses gated recurrent units to monitor speaker states and encode utterance history, whereas ICON augments recurrent encoding with attention to past utterances. However, both models employ early or intermediate fusion, which limits their ability to align modality-specific information dynamically. Moreover, although these models are context-aware, they do not handle conversation context as an explicit, independent signal.

Dai et al. [101] developed one of the initial multimodal designs employing a basic early fusion method to combine information from diverse modalities. However, their technique lacked the use of transformer models and did not include contextual information or class imbalance. While their findings on benchmark datasets were adequate, the model failed to enable cross-modal interaction or exhibit generalizability across domains.

Building on this foundational work, Yu et al. [102] developed the CoMPM framework, which leveraged co-attention processes inside a transformer-based model to combine voice, text, and visual modalities. The system displayed effective fusion and cross-modal interaction; however, context and imbalance were only partly handled. Despite this, the model scored well on benchmarks and demonstrated modest generalizability.

The emergence of transformer-based topologies constituted a significant step in modeling inter-modal interactions. MulT [103] pioneered the use of directed cross-modal attention, synchronizing asynchronous

text and audio inputs across time. Subsequent models such as SDT [104], Cross-Modal Transformer + Self-Attention [105], and DialogueTRM [106] refined this approach via self-distillation, adaptive gating, and hierarchical context modeling. Nonetheless, these models still face two significant limitations: they treat contextual information implicitly, typically by concatenating previous utterances to the current input or using hierarchical encoders; and their fusion strategies, although improved, often rely on shallow or asymmetric attention mechanisms, failing to capture bidirectional interdependence across all modality streams.

Following this, Li et al. [107] advanced the field by introducing a context-aware transformer model that handles context as a first-class input. This architecture completely allows multimodal fusion and cross-modal reasoning and displays excellent generalizability across benchmarks. However, it does not expressly address class inequity, allowing space for development in that area.

In parallel, Wu et al. [108] undertook an empirical assessment of transformer-based multimodal emotion recognition models. Their emphasis was on comparing training behaviors rather than designing a new architecture. The research ignored context modeling, class imbalance, and cross-modal reasoning, and generalizability was not investigated in detail. As such, its significance resides primarily in benchmark assessment.

Further expanding the transformer paradigm, Patamia et al. [109] proposed a transformer-driven attention-guided fusion model that performed well across many datasets. However, the system lacked context modeling, could not address data imbalance, and allowed only limited cross-modal interaction. Additionally, it exhibited inadequate generalization power across multiple emotion recognition tests.

Expanding upon these efforts, Maji et al. [105] suggested a hybrid attention technique that includes partial conversation context into the emotion recognition pipeline. The model displays moderate cross-modal reasoning and delivers acceptable performance on conventional benchmarks. However, it lacks strong imbalance management, and its generalizability remains dubious owing to poor cross-domain validation.

In a similar vein, Ma et al. [104] developed a transformer-based encoder-decoder system for multimodal emotion recognition. The model contains partial context modeling and fully allows multimodal fusion and cross-modal interaction. Although it does not manage class imbalance, it performs competitively on benchmarks and has considerable generalization potential.

An alternative family of models employs GNNs to capture conversational structure. MMGCN [99] creates modality-aware networks where nodes represent utterances and edges express temporal or speaker connections. Incorporating ideas from cognitive modeling, Meng et al. [110] developed the CBERL framework, a biologically inspired model that includes class-balancing techniques utilizing contrastive learning. The design partly employs transformers and displays some cross-modal capabilities. Although context is not explicitly represented, CBERL exhibits good performance on benchmarks and promises generalizability due to its balanced representation learning. CBERL [110] also combines GAN-based augmentation and class-boundary learning, further increasing performance. Despite these gains, GNNs often rely on predefined or manually created graph topologies, which restricts their versatility. They also lack the flexibility of transformers in describing long-range inter-modal relationships.

Modality-Specific Self-Supervised Learning [111] utilizes wav2vec for audio and Bidirectional Encoder Representations from Transformers (BERT) for text to strengthen unique modality representations before fusion. Although these techniques provide better single-modality features, they still rely on static fusion or fail to exploit conversational context as a structured, trainable signal. Most recently, Makhmudov et al. [112] investigated emotion recognition utilizing knowledge distillation approaches targeted at improving the efficiency of current systems. The model lacked contextual inputs, imbalance management, and sophisticated cross-modal interaction. Despite reaching respectable benchmark performance, it lacked proof of strong generalization and did not suggest architectural advances.

The advent of memory-enhanced architectures has offered new paradigms for addressing long-term dependencies in conversational emotion recognition. Zhang et al. [113] developed MemoCMT (Memory-Enhanced Cross-Modal Transformer), which includes implicit memory methods to store and retrieve contextual information over lengthy conversation sequences. The design consists of a dual-memory system with distinct banks for cross-modal interactions and temporal context, allowing more comprehensive modeling of long-range connections. MemoCMT displays complete transformer-based fusion with extensive cross-modal interaction capabilities; however, it offers only limited context modeling and does not explicitly address the context gap. The model achieves competitive performance on conventional benchmarks and demonstrates potential generalizability across multiple discourse domains, signifying a substantial breakthrough in memory-augmented multimodal systems.

Building upon graph-based techniques, subsequent work has investigated more comprehensive contextual modeling with advanced graph neural networks. Guo et al. [114] invented ConxGNN (Context-aware Graph Neural Network), which generates dynamic graphs that adapt to conversational flow and speaker interactions. Unlike standard GNN systems that rely on fixed graph topologies, ConxGNN utilizes learnable graph creation processes that can capture complex contextual connections. The model addresses context as a first-class modality, encompassing both temporal and speaker-aware context modeling. While ConxGNN does not leverage transformer designs, it displays remarkable cross-modal interaction via its graph-based message forwarding capabilities. The technique displays good benchmark performance and exhibits substantial generalizability across multiple conversational scenarios, while it does not explicitly manage contextual information integration.

Attention processes have continued to evolve with the development of more complex cross-modal interaction techniques. Praveen et al. [115] developed RJCMA (Recursive Joint Cross-Modal Attention), which utilizes recursive attention processes to modify cross-modal representations iteratively. The design presents a revolutionary recursive structure where attention weights are gradually adjusted across numerous rounds, allowing richer cross-modal comprehension. RJCMA combines partial context modeling using hierarchical attention structures and displays extensive cross-modal interaction capabilities. The model employs transformer-based components for attention computing and shows good performance on benchmark datasets. However, like many previous techniques, it does not explicitly address context gap, class imbalance, and its

generalizability remains partly confirmed owing to inadequate cross-domain examination.

The merging of audio and visual modalities has seen substantial improvement via specialized transformer designs. Kumar et al. [116] introduced AVT-CA (Audio-Video Transformer with Cross Attention), which primarily focuses on audio-visual fusion for emotion recognition. The system incorporates bidirectional cross-attention methods between audio and visual streams, allowing fine-grained temporal alignment and feature interaction. AVT-CA displays extensive multimodal fusion capabilities with significant cross-modal interaction, albeit it does not contain explicit context modeling or solve class imbalance concerns. The model exhibits competitive benchmark performance and demonstrates considerable generalizability across diverse audio-visual emotion recognition tasks, contributing to the growing body of work on specialized multimodal transformer architectures.

Recent breakthroughs in integrating large language models have provided new opportunities for multimodal emotion recognition. Chen et al. [117] developed SpeechCueLLM, which employs large language models to add acoustic emotional signals into textual interpretation. The framework offers a unique technique where auditory information is translated into textual representations that pre-trained language models can analyze. SpeechCueLLM employs complex fusion tactics by directly incorporating audio signals into the language model's attention processes. The technique enables partial context modeling by leveraging the inherent contextual capabilities of large language models and exhibits considerable cross-modal interaction between audio and textual modalities. While the model does not explicitly manage context integration, it obtains remarkable benchmark performance. It indicates potential generalizability across diverse speech emotion recognition tasks, providing a unique path in Large Language Model (LLM)-based multimodal systems.

The advent of foundation models and end-to-end pipelines has impacted recent advancements in affective computing. A full-stack speech-to-emotion system that avoids transcription, learning emotion embeddings directly from audio, is suggested in [118]. While promising for robustness, these approaches sometimes fail to include explicit semantic context or utilize structured discourse history. Joint Multimodal Transformers [119] and Self-Distilled Transformer Fusion [104] investigate more sophisticated fusion pipelines, adding feedback mechanisms and fine-grained attention. Yet even these models do not regard conversational context as a first-class medium, and frequently neglect the synergistic benefit of integrating it dynamically with text and audio. Finally, Dosovitskiy et al. [42] demonstrate that lightweight transformer-based visual models may scale effectively, showing the prospect of incorporating vision in future versions of multimodal Emotion Recognition in Conversation (ERC) without losing computational tractability.

2.4.1 Summary

The advancement of multimodal emotion recognition has been distinguished by growing architectural complexity and methodological sophistication. Early research mostly relied on fundamental fusion algorithms without addressing essential difficulties such as contextual awareness, data imbalance, or cross-modal reasoning. Subsequent transformer-based models have increased fusion and interpretability, with major attempts starting to study the importance of context and balanced learning. The recent advent of memory-enhanced

architectures, complex graph-based techniques, and LLM-integrated systems indicates the field's continuous growth toward more complete and contextually aware models. However, despite these achievements, substantial gaps remain, notably in integrating context as an independent modality, assuring robustness across domains, enabling smooth cross-modal interaction, and resolving class imbalance holistically. These constraints underscore the need for a comprehensive framework that addresses these challenges holistically, which justifies the method described in this research.

Table 2.2: Multimodal Emotion Recognition Methods Comparison

Paper	Method	Context	Imbalance	Dataset	Speed (%)	Accuracy
Waligora et al. [120]	Joint Multimodal Transformer (JMT)	×	✓	Affwild2, BioVid	15%	↑
Makhmudov et al. [112]	CNN + BERT + attention fusion	×	✓	CMU-MOSEI , MELD	×	↑
Patamia et al. [109]	Pre-trained transformers (wav2vec + BERT)	✓✓	✓✓	IEMOCAP	22%	↑
Li et al. [107]	Context-aware multimodal fusion	✓✓	×	IEMOCAP	×	↑
Maji et al. [105]	Cross-modal transformer + self-attention	✓✓	✓✓	IEMOCAP , MELD	8%	↑
Wu et al. [108]	Multimodal fusion with perspective loss	×	✓	IEMOCAP	×	↓
Wang et al. [121]	Transformer augmented fusion	✓✓	×	IEMOCAP , MELD	12%	↓
Ma et al. [104]	Self-distillation transformer (SDT)	×	×	IEMOCAP , MELD	×	↑
Meng et al. [110]	Class Boundary Enhanced Learning (CBERL)	×	✓	IEMOCAP , MELD	18%	↑
Yu et al. [122]	Emotion-Anchored Contrastive Learning	×	✓	EmoryNLP, IEMOCAP , MELD	×	↑
Luo et al. [123]	Bimodal Connection Attention Fusion	✓✓	✓	IEMOCAP , MELD	25%	↓
Dai et al. [101]	DeepMSI-MER with contrastive learning	×	×	IEMOCAP , MELD	×	↓

Continued on next page

Table 2.2 continued from previous page

Paper	Method	Context	Imbalance	Dataset	Speed (%)	Accuracy
Shou et al. [124]	Long-distance Relation-aware GNN	×	×	IEMOCAP , MELD	35%	↑
Praveen & Alam [115]	Recursive Joint Cross-Modal Attention	×	✓✗	Affwild2	×	↑
Majumder et al. [31]	DialogueRNN with three GRUs	✓✗	✓✗	IEMOCAP	9%	↑
Tsai et al. [89]	Multimodal Transformer (MulT)	×	✓	CMU-MOSEI	×	↓
Mao et al. [106]	DialogueTRM for emotion dynamics	×	×	IEMOCAP , MELD	7%	↓

Legend: ✓ = Fully Addressed, ✓✗ = Partially Addressed, × = Not Addressed

The columns in Table 2.2 are defined as follows:

- **Paper:** Refers to the paper that proposed the specific technique, model, or method.
- **Method:** Describes the core technical approach or architecture of the model presented in the paper.
- **Context:** Indicates whether the model explicitly handles conversational context (✓), partially handles it (✓✗), or does not (×).
- **Imbalance:** Shows whether the model includes a specific mechanism to address class imbalance (✓) or not (×).
- **Dataset:** Refers to the primary dataset(s) the authors used for evaluation.
- **Speed (%):** Represents the reported speed increase or a relative efficiency metric as described in the paper. An × indicates that a specific numerical value was not provided in the literature.
- **Accuracy:** Shows whether the model's accuracy improved (↑) or degraded (↓) compared to the baseline specified in its respective paper.

Table 2.2 provides a comparative assessment of various multimodal emotion recognition models, highlighting their primary methodologies, architectural decisions, and reported performance. A noticeable trend

is the growing deployment of complex structures to combine inputs from several modalities, demonstrating the field's concentration on advancing beyond unimodal analysis. However, the methods of this integration and the particular difficulties they address differ greatly.

A fundamental contrast among the models is their handling of conversational context. Methods like those by Li et al. [107], Wu et al. [108], and Shou et al. [124] incorporate context, often via recurrent or graph-based structures intended to model temporal and speaker dependencies. In contrast, other approaches such as those by Patamia et al. [109] and Maji et al. [105] prioritize the power of large pre-trained transformers for feature extraction, with less explicit focus on long-range conversational history, earning them a partial or half contextual awareness. The foundational DialogueRNN [31] set an early baseline for context modeling with its specialized GRUs, which succeeding models have improved upon.

Class imbalance, a recurrent difficulty in emotion recognition datasets, is a primary focus for just a select few models. Notably, Meng et al. [110] and Yu et al. [122] introduce specific mechanisms-generative learning and contrastive learning, respectively- to directly mitigate the impacts of unbalanced data distribution. This particular emphasis distinguishes their work from the majority of other approaches in the table, which do not provide clear strategies for addressing imbalance and thus risk biased performance toward majority classes.

The stated performance reflects the various techniques employed. While most models provide strong results, some with lower Speed percentages, such as DialogueTRM [106] and Maji et al. [105], imply more computationally efficient designs. Conversely, models like those from Shou et al. [124] and Luo et al. [123] show a larger computational load, presumably owing to more complicated GNN or attention mechanisms. The performance of models by Wang et al. [121] and Dai et al. [101] is highlighted as relatively inferior. Overall, the table indicates a trade-off between architectural complexity, contextual richness, and the management of data-specific concerns like class imbalance, driving future research toward more comprehensive and resilient solutions.

Chapter 3

Methodology

This chapter outlines the methodology employed to develop and evaluate the proposed (MCTAF) model. Addressing the challenges identified in Chapter 2, specifically the limited integration of dedicated contextual modules and sub-optimal cross-modal fusion mechanisms.

3.1 Research Methodology

This research adopts the Design Science Research Process (DSRP) [125] as its guiding framework due to its structured and iterative nature, which facilitates addressing complex, real-world problems through the creation and refinement of innovative artifacts. The adaptation of the DSRP phases for this study is detailed below, aligning with the problem-centered approach depicted in Figure 3.1.

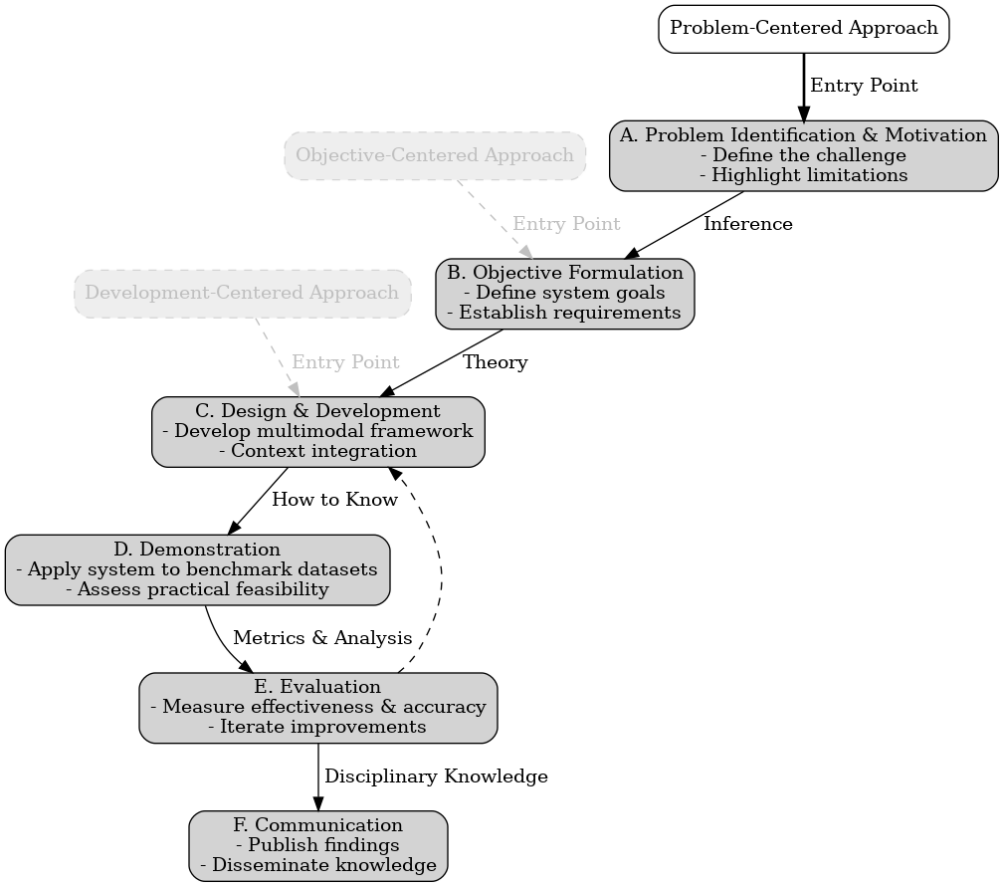

Figure 3.1: The architecture of the proposed MCTAF model

Figure 3.1 illustrates the research process, highlighting the problem-centered approach of the DSRP framework as applied in this thesis. Solid arrows show the sequential flow of the research process, while the

dashed arrow between the Evaluation and Design & Development stages symbolizes the iterative refinement cycle crucial to refining the suggested framework.

A. Problem Identification

This phase involves a detailed literature analysis to identify significant problems in multimodal emotion recognition. As stated in Chapter 2, current systems generally regard contextual inputs as secondary or neglect them altogether, resulting in a lack of specialized contextual integration and strong multimodal fusion mechanisms. This restriction affects their ability to distinguish complex emotions in dynamic, real-world situations accurately. The challenge was further studied via an assessment of benchmark datasets and existing deep learning methods, laying the basis for the research problem statement.

B. Objective Formulation

Following the discovery of the main constraints in existing multimodal emotion detection systems, we devised a systematic set of research goals to solve these issues. Our primary objective is to increase the expressiveness, flexibility, and accuracy of multimodal models by adding context as a first-class modality alongside text and voice.

To operationalize this purpose, we defined the following precise objectives, which serve as the cornerstone for our methodology and experimental design. First, we want to design a unique architecture that clearly formalizes contextual information as an independent modality, treated on equal footing with voice and textual inputs. Second, we intend to create a specialized contextual module and an adaptive fusion mechanism capable of collecting, encoding, and dynamically combining information across all three streams. Finally, we want to systematically analyze the proposed MCTAF framework on established benchmark datasets, with an emphasis on proving meaningful gains in emotion recognition performance, resilience to conversational fluctuation, and generalizability across varied interaction contexts.

C. Design and Development

Building upon the goals set in the preceding phase, we built and implemented the MCTAF framework via an iterative development cycle aiming at maximizing its architectural efficiency, modularity, and contextual expressiveness. MCTAF features three main architectural innovations: modality-specific encoders for text, voice, and context; a specialized contextual processing module; and a transformer-based fusion mechanism.

To achieve successful representation learning, we designed distinct feature extraction processes for each modality. The contextual stream was established as an independent input channel via the construction of a unique contextual module, which encodes user- and environment-specific information. This module was incorporated with text and audio streams to promote comprehensive emotion analysis.

For cross-modal feature integration, we developed a multi-head attention-based fusion module to align and combine modality-specific data dynamically. Through repeated testing, we assessed several architectural alternatives, including alterations to attention head topologies, encoder depths, and fusion algorithms. We also refined preprocessing and hyperparameter tweaking strategies to boost performance across various in-

teraction settings. This iterative development method created the groundwork for a modular and scalable system architecture that tackles constraints in existing multimodal emotion identification systems.

D. Demonstration

As an intermediary stage between development and formal assessment, the demonstration phase had a twin purpose: confirming the operational readiness of the model and gaining empirical insights into its handling of context-rich conversational input. Observations from these trials prompted additional adjustments to key components, including the contextual encoder, feature preprocessing methods, and the fusion mechanism. This process also helped reveal the model's sensitivity to environmental variability, allowing for targeted improvements ahead of a large-scale assessment.

E. Evaluation

In the evaluation phase, we performed a comprehensive quantitative assessment of MCTAF to confirm its efficacy across many performance parameters. Using two well-established benchmark datasets, IEMOCAP and MELD, we systematically examined classification performance using conventional measures such as accuracy and F1-score. Comparative evaluations were conducted against several baseline topologies to measure the relative improvements in predicted accuracy and robustness. We also performed a qualitative assessment focusing on specific instances and case studies.

To investigate the specific effect of each modality, we conducted ablation experiments that separated the text, voice, and context streams. These studies highlighted the additive advantage of contextual integration and helped establish interdependencies across modalities. Further, we investigated the model's generalizability by analyzing its cross-dataset performance, indicating MCTAF's flexibility across varied interaction settings and speaker characteristics.

The assessment process also comprised iterative refining cycles. Insights from performance diagnostics led to gradual architectural improvements, including the tuning of learning rates, dropout probability, and attention processes. As depicted in Figure 3.1, this phase employed a feedback-driven technique to iteratively enhance alignment, contextual sensitivity, and fusion accuracy, thereby improving the model's generalization and domain-specific dependability.

F. Communication

In the final part of the study process, we worked on effectively conveying the methodology, results, and contributions of the MCTAF framework. This thesis serves as the primary medium for expressing the architectural rationale, design iterations, and empirical validations supporting the proposed paradigm.

Key results include advances in classification accuracy, robustness to contextual changes, and cross-dataset generalization, which are presented through structured analysis and visual representations. These contributions extend the theoretical understanding of context-aware multimodal modeling while also offering practical guidance for constructing adaptive emotion recognition systems in real-world applications such as virtual assistants, affective healthcare, and intelligent tutoring systems.

3.2 Data Acquisition

We solely employ two publicly accessible, well-recognized benchmark datasets containing spoken conversations for multimodal emotion recognition: IEMOCAP [44] and MELD [126]. Both datasets contain highly matched audio and text transcripts with appropriate emotion labels, allowing the rigorous development and assessment of our context-aware fusion algorithms.

IEMOCAP

The IEMOCAP dataset is a comprehensive multimodal database designed for emotion recognition research. It consists of approximately 12 hours of audiovisual data collected from ten professional actors (five male and five female) engaging in both scripted and improvised scenarios to simulate natural emotional expression. The dataset includes recordings segmented into utterances, each annotated with emotional labels such as happy, sad, angry, neutral, frustrated, excited, and others. Each utterance is accompanied by synchronized data across multiple modalities: audio, video, motion capture (MoCap) of facial expressions, hand movements, and head rotations. Additionally, the speech transcriptions are provided for textual analysis. This diverse and richly annotated dataset supports various applications in affective computing, enabling in-depth studies into human emotional behavior through voice, facial expressions, and body movements. The dataset's modalities include audio, presented as 16 kHz, 16-bit Pulse Code Modulation (PCM) waveforms, from which 40-dimensional log Mel filterbank features were extracted using 25 ms frame windows and a 10 ms step, followed by utterance-level mean normalization. Text data, consisting of carefully transcribed transcripts with turn-level timestamps, received minimum preprocessing (lowercasing and punctuation removal) and was then encoded using 768-dimensional BERT word embeddings, with out-of-vocabulary words allocated a special ¡unk¿ token. While IEMOCAP also records video, this study only employed the audio and text streams to correspond with the modality breadth of major baseline studies. Contextual information is gathered using speaker IDs (which identify the two speakers in each discussion) and metadata related to earlier utterances. Evaluation on IEMOCAP was undertaken utilizing a leave-one-session-out cross-validation technique across its five unique sessions, a conventional procedure that allows for the rigorous assessment of generalization. IEMOCAP's regulated but realistic dyadic format is especially effective for carefully isolating and assessing the impacts of conversational environment on emotion perception.

MELD

The (MELD) provides a more complicated, multi-party discussion scenario, drawn from the popular TV series Friends. This dataset comprises 1,433 multi-party talks, together totaling 13,708 utterances. Unlike IEMOCAP's two-speaker model, MELD discussions typically involve 3-4 characters, providing a richer and more complex environment for contextual understanding. The original dataset is labeled with seven emotion labels: neutral, surprise, fear, sorrow, pleasure, disgust, and rage. To ensure consistency with our IEMOCAP processing and facilitate comparative analysis, these labels were also reprocessed for this study: joy and surprise were merged into happy, while fear and disgust were discarded, resulting in the same four primary

emotion categories: happy, sad, angry, and neutral, following the procedure outlined in [126]. This relabeling provided 8,358 utterances dispersed throughout these four groups for our trials. MELD's audio modality, sampled at 16 kHz, was processed similarly to IEMOCAP, extracting 40-dimensional log Mel filterbank features with 25 ms frame windows and a 10 ms step, followed by utterance-level mean normalization. Text data, generated from subtitles and transcripts with speaker IDs, received the same minimum preparation (lowercasing and punctuation removal) and was initialized with 768-dimensional BERT word embeddings. Similarly to IEMOCAP, MELD also provides visual data; however, this research focused primarily on the audio and text modalities to guarantee direct comparison with current benchmarks. Contextual information comprises multi-party speaker IDs and the transcripts of prior utterances, which are vital for interpreting the dynamic conversational flow. For assessment, we employed the normal MELD data splits: 1,038 talks for training, 114 for validation, and 280 for testing. MELD's multi-party, lifelike conversations are crucial in proving the generalization capabilities of our context fusion technique beyond dyadic interactions, showing its applicability and resilience in more complicated, real-world conversational contexts.

3.3 Data Preprocessing

To guarantee consistency and high-quality input for the Multimodal Contextual Transformer Augmented Fusion (MCTAF) model, a uniform preprocessing pipeline was meticulously applied to both the IEMOCAP and MELD datasets. This pipeline was meant to extract clean, perfectly aligned audio and text representations, as well as to create explicit context vectors for each speech. The following subsections outline the techniques for preparing audio, text, and acoustic features, as well as the final production of contextual embeddings.

3.3.1 Text Preprocessing

Textual data for each utterance was gathered from the official transcripts given with the IEMOCAP and MELD datasets. In IEMOCAP, the conversations are dyadic and entail lengthier speaker turns, whereas MELD offers shorter, multi-party exchanges taken from television discourse. For each speech, speaker identification, timestamp alignment, and text content were recorded, creating the basis for later context-aware modeling. Initial preprocessing involved lowering all tokens to lowercase and removing non-semantic punctuation, except for contractions (e.g., don't, isn't), which were preserved due to their emotional and syntactic value.

Tokenization was conducted using the WordPiece tokenizer from the BERT pretrained model developed by Devlin et al. [127]. This tokenizer splits input text into subword units, allowing for robust handling of unusual or out-of-vocabulary terms. The obtained tokens were translated to 768-dimensional contextual embeddings using the BERT model without fine-tuning. To standardize sequence lengths, we mandated a maximum token length of 50 for IEMOCAP and 40 for MELD. These thresholds were calculated experimentally via an investigation of token-length distributions, selected to catch over 95% of utterances without truncation. Utterances shorter than the maximum length were padded with zero vectors, and a binary attention mask

was generated to identify legitimate token locations.

The token-level embeddings and related attention masks were then fed into a Bidirectional GRU encoder (described in Section 3.5.1) to construct a fixed-length, utterance-level representation that captures the sequential semantics of the text. This method enables the model to store syntactic structure and word-order information beyond what is accessible in the static BERT output alone. By integrating contextualized BERT embeddings with recurrent encoding, the model benefits from both the pre-trained semantic richness of gls BERT and domain-specific temporal modeling. Speaker identity and utterance alignment were kept and subsequently exploited during the generation of context vectors, as explained in Section 3.3.3.

3.3.2 Audio Preprocessing

The audio modality in both the IEMOCAP and MELD datasets comprises recorded speech sampled at 16 kHz. Each speech was initially separated using the available start and finish timestamps to guarantee alignment with textual and contextual information. To enhance audio clarity, stationary background noise was suppressed using spectral subtraction methods, and silent areas were identified and eliminated by an energy-based Voice Activity Detection (VAD) algorithm. Loudness was standardized across utterances using Root Mean Square (RMS) normalization to a set objective of –20 dBFS, enhancing consistency in amplitude-based characteristics.

Following denoising and segmentation, a complete collection of low-level acoustic features was retrieved to capture prosodic, spectral, and voice quality properties related to emotional expression. Using a 25 ms Hamming window and a 10 ms stride, we estimated 40-dimensional Log Mel Filterbank (LMFB) energies and Mel-frequency cepstral coefficients (MFCC)s, together with their first- and second-order temporal derivatives. Prosodic features, including pitch in semitone units, short-time energy, and pause duration, were also retrieved, revealing voice dynamics and intensity. Spectral descriptors such as spectral centroid, bandwidth, and roll-off recorded frequency-domain energy distribution, whereas temporal characteristics like zero-crossing rate and signal entropy defined signal variability. Additionally, voice quality metrics including jitter, shimmer, and Harmonics-to-noise ratio (HNR) were calculated to capture small perturbations in vocal fold vibrations, which frequently correspond with emotional state changes. All features were z-score normalized using statistics calculated from the training set to guarantee uniform scale and variance across batches. 200 audio frames were extracted using COVAREP features at 25ms frame length [128]. Shorter utterances were padded and longer ones masked to standardize input length. To allow batch processing, each utterance's frame sequence was zero-padded to a set maximum length of 200 frames for IEMOCAP and 180 frames for MELD. A binary mask was constructed to identify legitimate acoustic frames from padding. The resulting acoustic feature matrix, with form $F_{\max} \times d$, was supplied to the BiGRU-based audio encoder described in Section 3.5.2, where $d = 100$ is the dimensionality of the combined feature vector per frame.

3.3.3 Context Construction

To explicitly incorporate conversational history, a context window of size K is defined. For any target utterance U_t, its context consists of the two immediately preceding utterances, $\{U_{t-2}, U_{t-1}\}$. If fewer than two preceding utterances are available (e.g., at the beginning of a dialogue), zero-vectors are used as placeholders. For each context utterance U_k (where $k \in \{t-2, t-1\}$):

Mean-Pooled Text Embedding

The textual content of U_k is summarized by computing the mean of its BERT-based word embeddings, weighted by the attention mask. This operation effectively creates a fixed-size representation of the utterance's overall textual meaning:

$$\bar{\mathbf{e}}_k^T = \frac{1}{\sum_j \mathbf{m}_{k,j}^T} \sum_{j=1}^{L_{\max}} \left(\mathbf{m}_{k,j}^T \cdot \widetilde{X}_{k,j}^T\right) \in \mathbb{R}^{768}. \tag{3.1}$$

The mask $\mathbf{m}_{k,j}^T$ ensures that only valid tokens contribute to the mean.

Mean-Pooled Acoustic Embedding

Similarly, the acoustic properties of U_k are represented by the mean of its acoustic features, again weighted by the acoustic mask. This captures the average acoustic profile of the preceding utterance:

$$\bar{\mathbf{e}}_k^A = \frac{1}{\sum_f \mathbf{m}_{k,f}^A} \sum_{f=1}^{F_{\max}} \left(\mathbf{m}_{k,f}^A \cdot \widetilde{X}_{k,f}^A\right) \in \mathbb{R}^{100}. \tag{3.2}$$

Speaker One-Hot Encoding

To incorporate speaker identity as a contextual cue, a one-hot vector $\mathbf{s}_k \in \{0,1\}^S$ is generated for each context utterance U_k. S represents the total number of speakers in the dataset ($S = 2$ for IEMOCAP and $S = 4$ for MELD). This helps the model understand who spoke the preceding turns.

These three components (mean-pooled text, mean-pooled acoustic, and speaker one-hot) are then concatenated to form a comprehensive context vector for each preceding utterance $\mathbf{c}_k = [\,\bar{\mathbf{e}}_k^T \,||\, \bar{\mathbf{e}}_k^A \,||\, \mathbf{s}_k\,] \in \mathbb{R}^{868+S}$. Finally, for the target utterance U_t, its complete context representation, X_t^C, is formed by stacking the context vectors of the two preceding utterances:

$$X_t^C = \left[\mathbf{c}_{t-2};\, \mathbf{c}_{t-1}\right] \in \mathbb{R}^{K \times (868+S)}, \tag{3.3}$$

where K. If fewer than K preceding utterances are available, their respective positions in X_t^C are filled with zero vectors to maintain a consistent input dimension.

Outcome: Each target utterance U_t is augmented with a context matrix X_t^C, explicitly representing the semantic, acoustic, and speaker information of the immediate conversational history.

3.4 Model Architecture

The suggested MCTAF model (Figure 3.2) is meticulously constructed to analyze and fuse input from various modalities-text, audio, and conversational context-to achieve robust emotion identification.

The suggested architecture contains five main components, each responsible for modality-specific processing or multimodal integration and categorization. The **Text Encoder** leverages a BiGRU to process contextualized BERT embeddings, capturing sequential dependencies within the utterance. Similarly, the **Audio Encoder** utilizes a distinct BiGRU to encode extracted acoustic data, emphasizing temporal patterns in prosody, pitch, and energy contours. The **Context Encoder** adds an extra BiGRU to model representations of conversational context, combining previous utterance metadata and speaker-related information. These modality-specific encodings are then supplied to the **Transformer-Based Fusion** module, which combines information using a multi-layer Transformer architecture. A learnable `[CLS]` token is prepended to the sequence of modality embeddings, enabling the model to attend over the four streams—text, audio, context, and an intermediate fused representation—via both self-attention and cross-attention mechanisms. This permits the fusion module to capture fine-grained intra-modality relationships as well as high-level inter-modality connections. Intra-modal attention refers to capturing dependencies within a single modality (e.g., dialogue turns in text), whereas inter-modal attention models dependencies across different modalities (e.g., aligning audio tone with contextual cues). The output corresponding to the `[CLS]` token, which adaptively aggregates salient cues from all modalities, serves as the input to the final **Emotion Classifier**. This classification component comprises a linear layer followed by a softmax activation function, projecting the fused representation into discrete emotion categories.

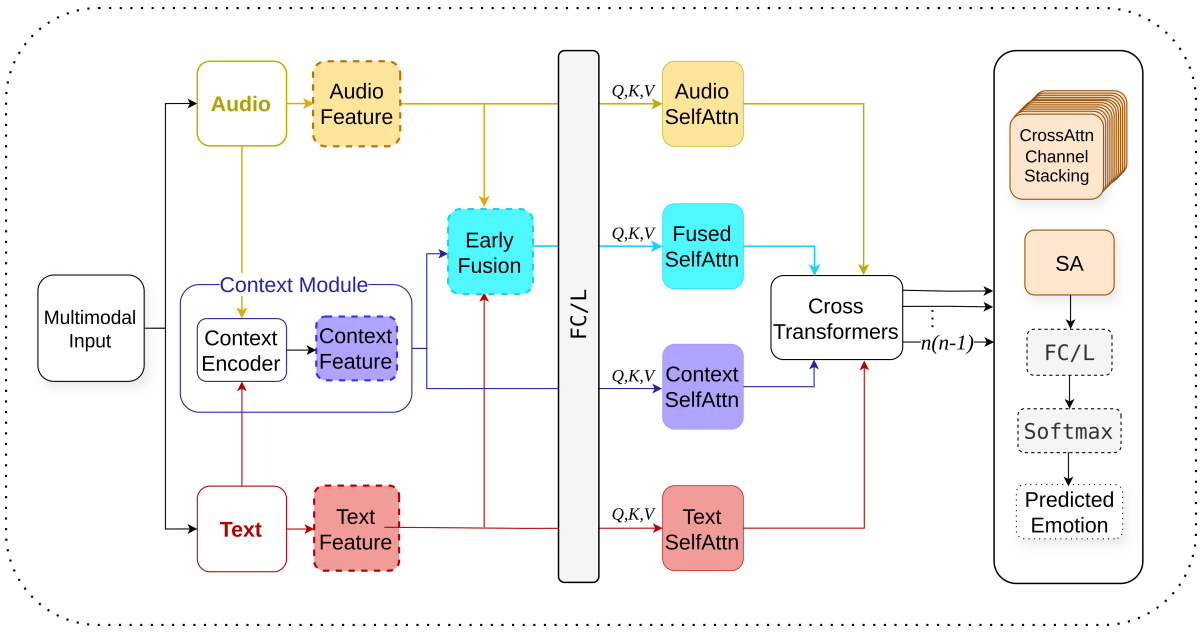

Figure 3.2: The architecture of the proposed MCTAF model

3.5 Modality-Specific Encoders

Each speech and its context are encoded into vector representations by modality-specific encoders. As in the baselines, we employ GRUs for sequence encoding owing to their efficacy in simulating temporal sequences and their lightweight nature compared to complete transformers [104, 105, 108]. We deploy three bidirectional GRU encoders, each with hidden dimension H and input dropout probability p. Layer normalization follows each projection onto a standard embedding dimension d. The encoders function as follows:

3.5.1 Text Encoder

The text encoder is responsible for transforming the sequence of high-dimensional BERT word embeddings into a fixed-size utterance-level representation that captures the full semantic content. Given the padded and masked text matrix $\widetilde{X}_i^T = [\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,L_{\max}}]^\intercal \in \mathbb{R}^{L_{\max} \times 768}$, BiGRU is applied. The BiGRU consists of a forward GRU (GRU_{fw}^T) processing the sequence from left to right and a backward GRU (GRU_{bw}^T) processing it from right to left. Each GRU has a hidden size of 128, resulting in a concatenated hidden state dimension of 256.

$$\overrightarrow{h}_{i,j}^T = \text{GRU}_{\text{fw}}^T\big(\mathbf{e}_{i,j},\ \overrightarrow{h}_{i,j-1}^T\big), \quad \overleftarrow{h}_{i,j}^T = \text{GRU}_{\text{bw}}^T\big(\mathbf{e}_{i,j},\ \overleftarrow{h}_{i,j+1}^T\big), \tag{3.4}$$

where $\mathbf{e}_{i,j}$ are the BERT embeddings. Padded tokens are masked via $\mathbf{m}_i^T$ to prevent their contribution to the hidden states. The final hidden states from both directions are concatenated to form the utterance-level representation $\mathbf{h}_i^T$:

$$\mathbf{h}_i^T = \big[\overrightarrow{h}_{i,L_{\max}}^T \,||\, \overleftarrow{h}_{i,1}^T\big] \ \in\ \mathbb{R}^{256}. \tag{3.5}$$

This concatenation captures information from both ends of the sequence. To prevent overfitting and prepare the representation for the fusion module, a dropout layer with a probability $p = 0.3$ is applied, followed by a linear projection. This projection, defined by a weight matrix $W_T \in \mathbb{R}^{128 \times 256}$ and bias b_T, maps the 256-dimensional representation to a 128-dimensional latent space:

$$\mathbf{z}_i^T = W_T\, \mathbf{h}_i^T + b_T \ \in\ \mathbb{R}^{128}. \tag{3.6}$$

This projection ensures that all modality embeddings are of a consistent dimension for the subsequent fusion process.

3.5.2 Audio Encoder

The audio encoder processes the sequence of acoustic features to derive a compact, fixed-size representation of the utterance's vocal characteristics. Given the padded and masked acoustic feature matrix $\widetilde{X}_i^A = [\mathbf{a}'_{i,1}, \ldots, \mathbf{a}'_{i,F_{\max}}]^\intercal \in \mathbb{R}^{F_{\max} \times 100}$, a bidirectional GRU is employed. Similar to the text encoder, this BiGRU captures temporal dynamics across the acoustic frames, with a hidden size of 128 per direction,

yielding a combined hidden state dimension of 256.

$$\overrightarrow{h}^{A}_{i,f} = \mathrm{GRU}^{A}_{\mathrm{fw}}\big(\mathbf{a}'_{i,f},\ \overrightarrow{h}^{A}_{i,f-1}\big), \quad \overleftarrow{h}^{A}_{i,f} = \mathrm{GRU}^{A}_{\mathrm{bw}}\big(\mathbf{a}'_{i,f},\ \overleftarrow{h}^{A}_{i,f+1}\big), \tag{3.7}$$

where $\mathbf{a}'_{i,f}$ are the 100-dimensional acoustic features. The padding mask $\mathbf{m}^{A}_{i}$ ensures that only valid frames contribute to the recurrent computations. The final hidden states from both forward and backward passes are concatenated to form the utterance-level acoustic representation $\mathbf{h}^{A}_{i}$:

$$\mathbf{h}^{A}_{i} = \big[\overrightarrow{h}^{A}_{i,F_{\max}} \,||\, \overleftarrow{h}^{A}_{i,1}\big] \ \in\ \mathbb{R}^{256}. \tag{3.8}$$

This 256-dimensional vector undergoes dropout ($p = 0.3$) and a linear projection, defined by $W_A \in \mathbb{R}^{128\times256}$ and bias b_A, to align its dimension with the text embeddings:

$$\mathbf{z}^{A}_{i} = W_A\,\mathbf{h}^{A}_{i} + b_A \ \in\ \mathbb{R}^{128}. \tag{3.9}$$

This projected vector $\mathbf{z}^{A}_{i}$ encapsulates the essential acoustic information of the utterance in a unified embedding space.

3.5.3 Context Encoder

The context encoder is designed to process the constructed context matrix, capturing the sequential dependencies and interactions within the conversational history. Given $X^{C}_{i} = [\mathbf{c}_{i-2},\, \mathbf{c}_{i-1}]^{\mathsf{T}} \in \mathbb{R}^{2\times(868+S)}$, which represents the concatenated context vectors of the two preceding utterances, a bidirectional GRU is applied. This BiGRU learns a consolidated representation of the conversational flow relevant to the current utterance. The hidden size is 128 per direction, resulting in a 256-dimensional concatenated hidden state.

$$\overrightarrow{h}^{C}_{i,k} = \mathrm{GRU}^{C}_{\mathrm{fw}}\big(\mathbf{c}_{i-2+(k-1)},\ \overrightarrow{h}^{C}_{i,k-1}\big), \quad \overleftarrow{h}^{C}_{i,k} = \mathrm{GRU}^{C}_{\mathrm{bw}}\big(\mathbf{c}_{i-2+(k-1)},\ \overleftarrow{h}^{C}_{i,k+1}\big), \tag{3.10}$$

for $k = 1, 2$. If fewer than two preceding utterances exist, the corresponding $\mathbf{c}_k$ would be zero vectors, which the GRU gracefully handles. The final hidden states from the forward and backward passes are concatenated:

$$\mathbf{h}^{C}_{i} = \big[\overrightarrow{h}^{C}_{i,2} \,||\, \overleftarrow{h}^{C}_{i,1}\big] \ \in\ \mathbb{R}^{256}. \tag{3.11}$$

This 256-dimensional vector $\mathbf{h}^{C}_{i}$ represents the distilled essence of the immediate conversational context. Similar to other encoders, dropout ($p = 0.3$) is applied, followed by a linear projection $W_C \in \mathbb{R}^{128\times256}$ with bias b_C to transform it into a 128-dimensional embedding:

$$\mathbf{z}^{C}_{i} = W_C\,\mathbf{h}^{C}_{i} + b_C \ \in\ \mathbb{R}^{128}. \tag{3.12}$$

This $\mathbf{z}^{C}_{i}$ serves as the dedicated contextual embedding for the fusion module.

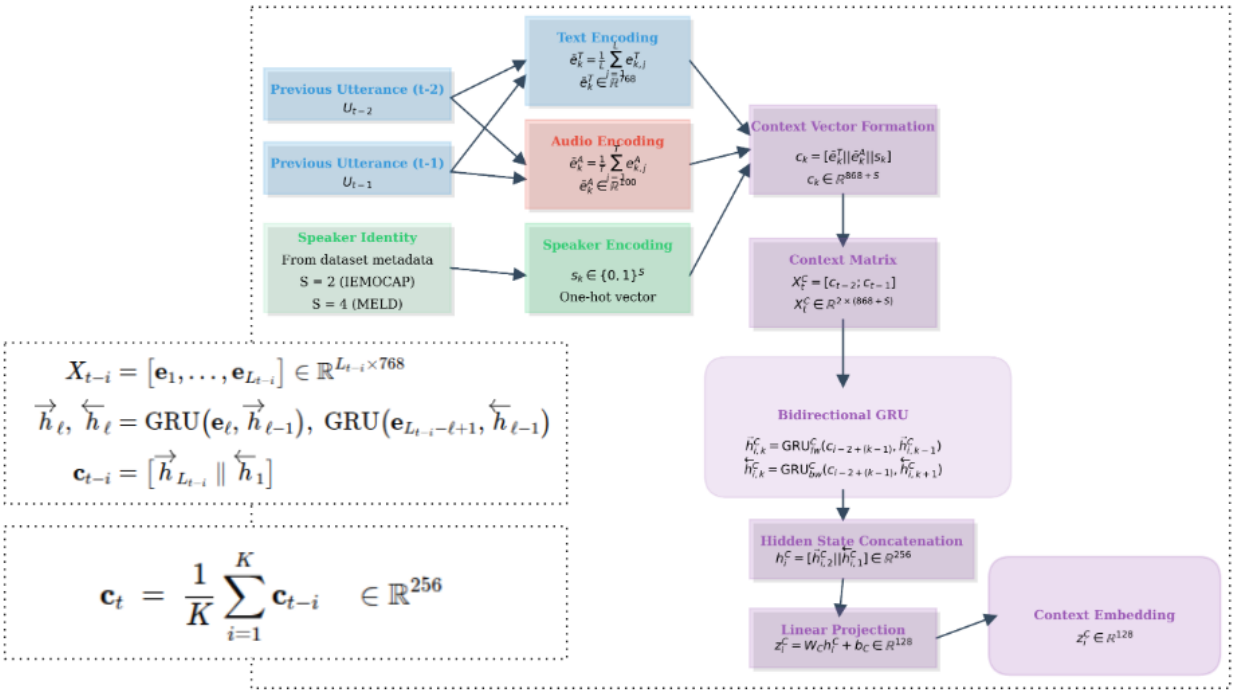

Figure 3.3: The Context Encoder Module. This module processes a sequence of utterance embeddings to capture temporal and speaker-specific dependencies, outputting context-aware representations for each utterance.

Mathematical Formulation of the Context Module

A key innovation of MCTAF lies in its formal treatment of conversational context as a standalone, entirely learnable modality through a dedicated context encoder (Figure 3.3). Unlike prior baselines such as DialogueRNN [31] or SDT [11], which incorporate dialogue history via hidden memory states or simple feature concatenation, MCTAF defines context as a temporally structured multimodal tensor and integrates it into a symmetric cross-modal attention framework.

Let each previous utterance u_k (where $k \in \{t-K, ..., t-1\}$) contribute three feature types:

- Text embedding $\mathbf{t}_k \in \mathbb{R}^{768}$, obtained via mean pooling over BERT outputs,
- Audio vector $\mathbf{a}_k \in \mathbb{R}^{100}$, extracted from Cooperative Voice Analysis Repository for Speech Technologies (COVAREP),
- One-hot speaker identifier $\mathbf{s}_k \in \mathbb{R}^S$, encoding speaker identity.

These are concatenated into a multimodal context vector:

$$\mathbf{c}_k = [\mathbf{t}_k \| \mathbf{a}_k \| \mathbf{s}_k] \in \mathbb{R}^{(868+S)} \tag{3.13}$$

Stacking the most recent K utterances gives the context matrix:

$$\mathbf{X}_t^C = [\mathbf{c}_{t-K}; \ldots; \mathbf{c}_{t-1}] \in \mathbb{R}^{K \times (868+S)} \tag{3.14}$$

To encode the temporal and speaker dynamics, $\mathbf{X}_t^C$ is passed through a bidirectional GRU:

$$\begin{aligned} \overrightarrow{h_k^C} &= \mathrm{GRU}_{fw}(\mathbf{c}_k, \overrightarrow{h_{k-1}^C}) \\ \overleftarrow{h_k^C} &= \mathrm{GRU}_{bw}(\mathbf{c}_k, \overleftarrow{h_{k+1}^C}) \\ \mathbf{z}_k^C &= \mathbf{W}_C[\overrightarrow{h_k^C} \| \overleftarrow{h_k^C}] + \mathbf{b}_C \in \mathbb{R}^{d_z} \end{aligned} \tag{3.15}$$

where $\mathbf{W}_C$ and $\mathbf{b}_C$ are trainable parameters, and d_z is the context embedding dimension.

These embeddings are incorporated into a multi-stream attention architecture. Each modality-text (T), audio (A), context (C), and early-fused (F) - is processed by independent Transformer encoders with self-attention layers as in [25]. We then apply directional cross-modal attention between all modality pairs:

$$\mathrm{CrossAtt}_{i \to j}(Q_i, K_j, V_j) = \mathrm{softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right) V_j \tag{3.16}$$

This results in a fully connected attention graph with 12 directed edges (e.g., $C \to T$, $T \to C$, $A \to C$, etc.). This is in contrast to models like MulT [26], which apply unidirectional attention from text to auxiliary modalities, or DialogueGCN [29], which depend on handcrafted graph edges.

Intuitively, this formulation offers three major innovations: First, instead of informally attaching past utterances or considering context as an unconscious memory, we carefully arrange it into a structured tensor $\mathbf{X}_t^C$ that contains textual, auditory, and speaker-specific information for each preceding turn. Second, we analyze this structured environment using a bidirectional GRU, which enables the model to learn how emotions evolve and vary among speakers, capturing both sequential and interpersonal relationships. Third, instead of utilizing context merely as a side input, we incorporate it directly into a multi-way attention mechanism that permits dynamic, bidirectional interaction between context, text, and audio characteristics throughout the fusion process. This approach enables the model to reason integratively about how present emotional stimuli connect to the larger conversational flow.

This approach enables MCTAF to express and fuse long-range conversational relationships more efficiently than preceding asymmetric or implicit context formulations, offering a mathematically rigorous basis for context-aware emotion identification. *See Algorithm 1 for the operational implementation of this formulation.*

3.5.4 Comparative Theoretical Insights

To justify the architectural and mathematical novelty of MCTAF, it is essential to contrast its context modeling formulation with prominent prior works in context-aware emotion recognition, such as DialogueRNN [31], DialogueGCN [29], and MulT [26]. DialogueRNN encodes conversational context using a series of GRU

cells, where speaker states and global context are tracked in parallel memory banks. However, these context vectors are updated implicitly and used only as hidden recurrent memory. There is no explicit multimodal structure applied to context, nor is the context independently fused via attention. Formally, DialogueRNN maintains global and speaker-specific hidden states:

$$g_t = \text{GRU}_g(x_t, g_{t-1}), \quad s_t^i = \text{GRU}_s(x_t, s_{t-1}^i) \tag{3.17}$$

These states interact via gates but are not formulated as a structured tensor with modality-specific features and speaker identifiers. In contrast, MCTAF constructs:

$$\mathbf{X}_t^C = [\mathbf{c}_{t-K}, \ldots, \mathbf{c}_{t-1}], \quad \text{where } \mathbf{c}_k = [\mathbf{t}_k \| \mathbf{a}_k \| \mathbf{s}_k] \tag{3.18}$$

This tensor-based modeling treats context as an explicit input modality, parallel to text and audio, and is encoded via bidirectional GRUs followed by attention-based fusion. DialogueGCN extends DialogueRNN by applying a graph convolutional network over dialogue turns, where nodes represent utterances and edges encode speaker relations. While this structure captures speaker interaction patterns, it relies on a fixed, manually designed graph topology and requires precomputing node features:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W^{(l)} h_j^{(l)}\right) \tag{3.19}$$

In contrast, MCTAF does not require any graph construction or definition of neighborhoods. Contextual influence is learned dynamically through cross-attention over fully connected streams, eliminating the need for inductive bias via edges or speaker role tags.

MulT treats text as the primary querying modality and applies cross-modal attention in one direction:

$$\text{Attn}_{T \to M}(Q_T, K_M, V_M) \tag{3.20}$$

where $M \in \{A, V\}$ is an auxiliary modality. Context is not treated as an independent modality and cannot attend to others. MCTAF, by contrast, applies symmetric directional attention across all four streams (text, audio, context, fused), resulting in a complete 12-edge cross-attention graph. This structure enables context to both influence and be influenced by other modalities:

$$\text{CrossAtt}_{C \to T}, \quad \text{CrossAtt}_{T \to C}, \quad \text{CrossAtt}_{C \to A}, \ldots \tag{3.21}$$

This formulation introduces a mathematically symmetric, data-driven mechanism for context-aware emotion reasoning that is architecture-agnostic and avoids external dependencies like graphs or speaker role labels.

3.5.5 Transformer-Based Fusion Module

The Transformer-Based Fusion Module is responsible for merging information from the three independent modalities-text, audio, and context-into a uniform representation suited for emotion categorization. This module is structured into four key stages: early fusion of token-aligned features, intra-modal self-attention to refine each modality stream independently, cross-modal attention to capture interdependencies across modalities, and a final self-attention mechanism to contextualize the fused representations globally. These phases culminate in a condensed utterance-level representation that captures the whole multimodal context of each speech.

Early Fusion Stream

At the sequence level, each modality provides a sequence of token representations for utterance t: text $\widetilde{H}_t^T$, audio $\widetilde{H}_t^A$, and context $\widetilde{H}_t^C$, each of shape $\mathbb{R}^{M\times d}$, where M is the maximum sequence length and d is the feature dimensionality. To perform early fusion, the token representations across modalities are concatenated at each position:

$$\widehat{H}_t^F = \left[\widetilde{H}_t^T;\, \widetilde{H}_t^A;\, \widetilde{H}_t^C\right] \in \mathbb{R}^{M\times 3d}. \tag{3.22}$$

This concatenated tensor is projected back into the shared space using a linear transformation followed by layer normalization:

$$\widetilde{H}_t^F = \mathrm{LayerNorm}\big(W_F \widehat{H}_t^F + b_F\big) \in \mathbb{R}^{M\times d}, \tag{3.23}$$

where $W_F \in \mathbb{R}^{d\times 3d}$ and $b_F \in \mathbb{R}^d$ are learned parameters. This early fusion stream complements modality-specific paths and enables initial interactions across modalities.

To handle variable-length sequences and padding, a binary attention mask is defined as:

$$\mathrm{mask}_{t,i} \in \{0,1\}, \quad \text{indicating whether token } i \text{ is valid (1) or padded (0)}, \tag{3.24}$$

and collected into a full mask vector:

$$\mathrm{mask}_t = \big(\mathrm{mask}_{t,1}, \ldots, \mathrm{mask}_{t,M}\big) \in \{0,1\}^M. \tag{3.25}$$

Intra-Modal Self-Attention

Each of the four streams-text, audio, context, and early fusion-is refined independently using a dedicated Transformer encoder. This step allows each stream to model its own sequential dependencies before interacting with other modalities. For each $p \in \{T, A, C, F\}$, we compute:

$$S_t^p = \mathrm{TransEnc}\big(\widetilde{H}_t^p;\, \mathrm{mask}_t\big) \in \mathbb{R}^{M\times d}. \tag{3.26}$$

Each Transformer encoder layer comprises two sublayers:

1. Multi-Head Self-Attention (MHA): The input is linearly projected into queries Q, keys K, and values V across $H = 4$ attention heads, each of size $d_k = 32$. Each head computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{3.27}$$

where $Q = XW^Q$, $K = XW^K$, $V = XW^V$, and $W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$. The outputs from all heads are concatenated and projected back to dimension d.

2. Feed-Forward Network (FFN): A position-wise FFN transforms each token independently using two fully connected layers with ReLU activation:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{3.28}$$

with $W_1 \in \mathbb{R}^{d\times 4d}$ and $W_2 \in \mathbb{R}^{4d\times d}$. This transformation allows non-linear recombination of features.

Each sublayer is followed by residual connections, layer normalization, and dropout ($p = 0.3$), stabilizing training and reducing overfitting.

Cross-Modal Attentional Fusion

To capture fine-grained interactions between modalities, cross-modal multi-head attention is performed between each pair of modalities (Figure 3.4). For each ordered pair $(p, q) \in \{T, A, C, F\}^2$, $p \neq q$, we compute:

$$C_t^{p\leftarrow q} = \text{MHA}\big(Q = S_t^p,\, K = S_t^q,\, V = S_t^q;\, \text{mask}_t\big) \in \mathbb{R}^{M\times d}. \tag{3.29}$$

This formulation allows each modality p to query another modality q, learning which features are informative across the sequence. The 12 resulting attention outputs are concatenated:

$$C_t = \big[C_t^{T\leftarrow A},\, C_t^{T\leftarrow C},\, \ldots,\, C_t^{F\leftarrow C}\big] \in \mathbb{R}^{M\times 12d}, \tag{3.30}$$

and passed through layer normalization:

$$\widehat{C}_t = \text{LayerNorm}(C_t) \in \mathbb{R}^{M\times 12d}. \tag{3.31}$$

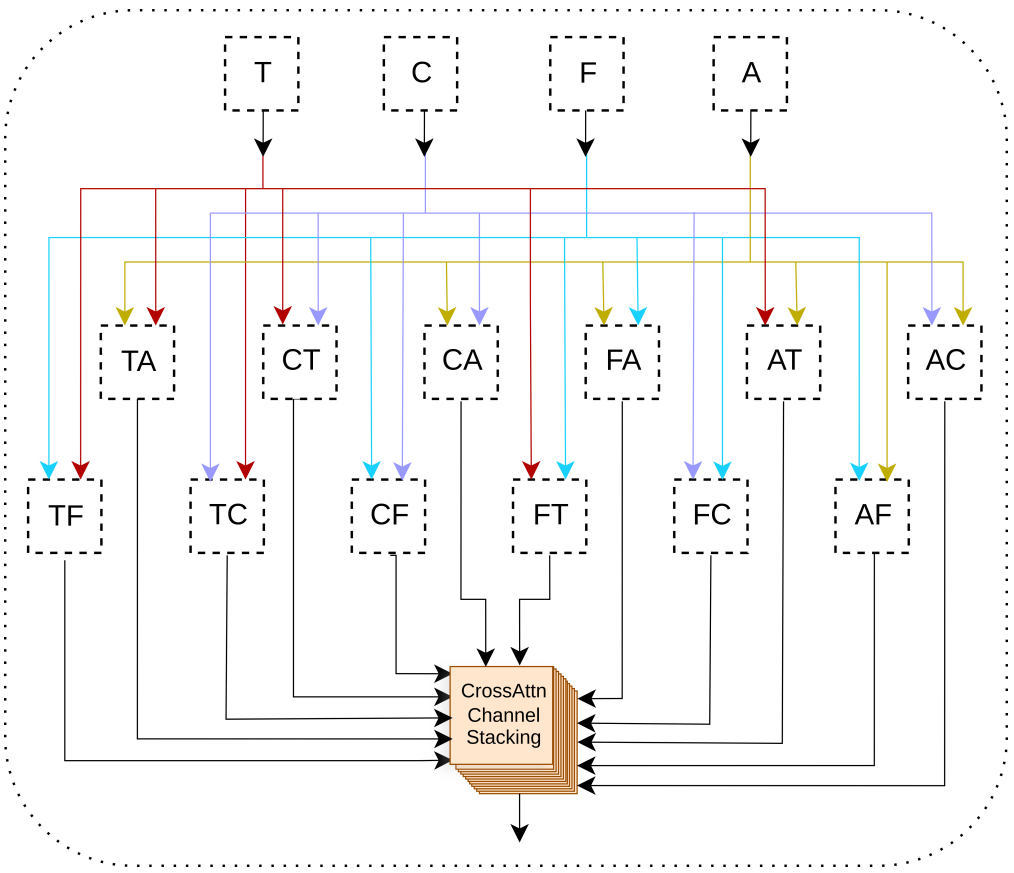

Figure 3.4: The Cross-Modal Interaction Module. This module is designed to capture the rich pairwise interactions between all n modalities. For n inputs, it instantiates n(n-1) dedicated cross-attention transformers, where each transformer allows one modality to attend to another, creating a set of enhanced, cross-aware representations.

Final Self-Attention and Feature Refinement

To globally contextualize the cross-modal representations across the full sequence, the refined tensor $\widehat{C}_t$ is passed through a final Transformer encoder:

$$U_t = \text{TransEnc}(\widehat{C}_t; \text{mask}_t) \in \mathbb{R}^{M \times 12d}. \tag{3.32}$$

The output U_t contains token-level embeddings that have been fully enriched through both intra- and inter-modal attention. To reduce U_t to a single utterance-level representation, a pooling operation is applied. In this work, we use masked mean pooling over valid positions:

$$\mathbf{h}_t = \frac{1}{\sum_{i=1}^{M} \text{mask}_{t,i}} \sum_{i=1}^{M} \text{mask}_{t,i} \cdot U_t^{(i)} \in \mathbb{R}^{12d}. \tag{3.33}$$

Emotion Classification

The final multimodal representation $\mathbf{h}_t$ serves as the input to the emotion classifier. This classifier consists of a single linear projection that maps the fused vector to unnormalized class logits:

$$\mathbf{o}_t = W_o \mathbf{h}_t + \mathbf{b}_o, \quad W_o \in \mathbb{R}^{E \times 12d}, \quad \mathbf{b}_o \in \mathbb{R}^E. \tag{3.34}$$

A softmax function is applied to convert the logits into a probability distribution over E emotion classes:

$$p_{t,e} = \frac{\exp(o_{t,e})}{\sum_{e'=1}^{E} \exp(o_{t,e'})}, \quad \text{for } e = 1, \ldots, E. \tag{3.35}$$

The predicted emotion label $\hat{y}_t$ for utterance t is then obtained via:

$$\hat{y}_t = \arg\max_e p_{t,e}. \tag{3.36}$$

This final stage allows the model to make context- and modality-aware predictions of emotional categories, grounded in both the sequential structure of each modality and their cross-modal interactions.

3.6 Training Procedure

To guarantee the successful training of the suggested MCTAF model, we built a thorough training method that incorporates robust loss management, cautious optimizer selection, and multiple regularization procedures. This section outlines the entire training approach, beginning with the derivation of the loss function, followed by details of the optimization setup, and concluding with regularization and convergence procedures.

3.6.1 Loss Function

One of the fundamental issues in multimodal emotion recognition is the inherent class imbalance seen in datasets such as IEMOCAP and MELD. Certain emotion classes, such as neutral, tend to dominate the data distribution, while others, like sad or furious, are considerably underrepresented. To solve this problem and to guarantee that the model does not grow biased toward majority classes, we applied a class-weighted cross-entropy loss function. The cross-entropy loss was used for all experiments, consistent with prior baselines to ensure comparability of results.

Given an utterance i, let the one-hot encoded ground-truth label be indicated by $y_i \in \{0, 1\}^4$, and let $\hat{\mathbf{y}}_i \in [0, 1]^4$ represent the expected probability distribution across the four emotion classes. The loss for a single speech is defined as:

$$\mathcal{L}_i = -\sum_{c=1}^{4} w_c \, y_{i,c} \, \log\left(\hat{y}_{i,c}\right), \tag{3.37}$$

Where w_c is a class-specific weight designed to penalize errors in underrepresented classes more heavily.

To compute these weights, we first calculate the empirical frequency freq_c of each class c across the training set. Then, we derive the weight using a smoothed inverse frequency formula:

$$w_c = \frac{1}{\log(1 + \text{freq}_c)}. \tag{3.38}$$

This logarithmic smoothing avoids assigning huge weights to sporadic classes, while still enhancing their influence in the optimization process. During training, we process data in mini-batches of size 32. The total loss for a batch $\mathcal{B}$ is then computed as the mean of the individual utterance losses:

$$\mathcal{L}_{\text{batch}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_i. \tag{3.39}$$

This formulation enables the model to learn more balanced representations across all emotion classes, promoting generalization to both frequent and infrequent categories.

The algorithm below outlines the complete forward inference procedure of the MCTAF model, highlighting how conversational history and speaker information are processed in conjunction with current utterance features. It includes the construction of the context matrix, BiGRU encoding for temporal modeling, self-attention within each modality, and directional cross-modal attention for fusion. Each modality (text, audio, context) is independently encoded and later fused through a multi-head attention mechanism before classification. This algorithm operationalizes the mathematical formulation described in Section 3.5.3 and serves as a high-level blueprint for practical implementation.

Algorithm 1: MCTAF: Contextual Fusion algorithm

Input: Current utterance text x_t^T, audio x_t^A;

Previous K utterances $\{u_{t-K}, ..., u_{t-1}\}$ with text, audio, and speaker labels

Output: Emotion label $\hat{y}_t$

1. Encode Contextual History:

foreach u_k *in* $\{u_{t-K}, ..., u_{t-1}\}$ **do**

$t_k \leftarrow \text{BERT}(u_k^{\text{text}})$; `// Text embedding`

$a_k \leftarrow \text{COVAREP}(u_k^{\text{audio}})$; `// Audio features`

$s_k \leftarrow \text{OneHot}(u_k^{\text{speaker}})$; `// Speaker encoding`

$c_k \leftarrow [t_k \| a_k \| s_k]$; `// Multimodal context vector`

end

$\mathbf{X}_t^C \leftarrow [c_{t-K}; ...; c_{t-1}]$; `// Context matrix`

2. BiGRU Encoding of Context:

$\mathbf{Z}_t^C \leftarrow \text{BiGRU}(\mathbf{X}_t^C)$; `// Context embeddings`

3. Encode Current Utterance:

$\mathbf{z}_t^T \leftarrow \text{BERT}(x_t^T)$

$\mathbf{z}_t^A \leftarrow \text{COVAREP}(x_t^A)$

4. Self-Attention per Modality:

$\mathbf{h}^T \leftarrow \text{TransformerEncoder}(\mathbf{z}_t^T)$

$\mathbf{h}^A \leftarrow \text{TransformerEncoder}(\mathbf{z}_t^A)$

$\mathbf{h}^C \leftarrow \text{TransformerEncoder}(\mathbf{Z}_t^C)$

5. Cross-Attention Between Modalities:

foreach (i, j) *in* $\{(T, A), (A, T), (T, C), (C, T), (A, C), (C, A), ...\}$ **do**

$\mathbf{h}^{i \to j} \leftarrow \text{CrossAttention}(Q = \mathbf{h}^i, K = \mathbf{h}^j, V = \mathbf{h}^j)$

end

6. Fusion and Prediction:

$\mathbf{f}_t \leftarrow \text{Concat}([\mathbf{h}^{T \to *}, \mathbf{h}^{A \to *}, \mathbf{h}^{C \to *}])$

$\hat{y}_t \leftarrow \text{Softmax}(\text{MLP}(\mathbf{f}_t))$

3.6.2 Optimization and Regularization

To optimize the MCTAF model and maintain steady convergence, we devised a variety of techniques covering optimizer choice, learning rate scheduling, regularization, and convergence monitoring.

We trained the model using the AdamW optimizer [129], which combines the advantages of adaptive moment estimation with a decoupled weight decay mechanism. This approach enables efficient L2 regularization without affecting the learning rate schedule. We employed a base learning rate of 3×10^{-4} and applied a weight decay coefficient of 1×10^{-5}, both determined via empirical validation.

To further optimize the training dynamics, we applied a cosine annealing learning rate scheduler with warm restarts, as suggested by Loshchilov and Hutter [130]. This scheduling approach cyclically decreases the learning rate according to a cosine decay curve and resets it periodically to its initial value. Warm restarts are programmed to occur every 50 epochs. This method enables the optimizer to investigate varied parts of the parameter space and mitigates the danger of converging prematurely to unsatisfactory minima.

Regularization is performed using different strategies incorporated throughout the system. Dropout [131] is applied with a preset frequency of $p = 0.3$. Specifically, we apply dropout in all BiGRU layers (before their output projections) and inside every Transformer encoder block, immediately after the multi-head attention and feed-forward sublayers. Dropout functions as a stochastic regularizer, preventing the co-adaptation of units and enhancing the model's ability to generalize to previously unseen data.

To further enhance numerical stability, notably during backpropagation via deep Transformer layers and recurrent structures, we employed global gradient norm clipping. In each update step, the entire gradient norm is computed and trimmed to a maximum value of 1.0. This eliminates sudden parameter changes owing to huge gradient magnitudes and stabilizes training dynamics.

During training, we employed a fixed mini-batch size of 32. However, to maintain a balanced class distribution across batches and prevent skewed gradient updates, we utilized a stratified sampling technique. This ensures that each batch approximates the global label distribution, thereby enhancing learning for minority classes and reducing bias induced by uneven batching.

To minimize overfitting and needless computing cost, we designed an early halting mechanism based on model performance on a held-out validation set. Specifically, we monitored the weighted F_1 score throughout training and interrupted the procedure if no improvement was seen for 50 consecutive epochs. Once training finished, the model checkpoint corresponding to the best validation performance was restored for final assessment.

Chapter 4

Experimentation

4.1 Experimental Setup

We assess the MCTAF model on two widely-used benchmark datasets described in the previous chapter: IEMOCAP and MELD. We present our experimental methods, thoroughly explaining the data splits, training hyperparameters, and baseline implementations. This extensive design enables a fair assessment across all models, allowing for credible findings about MCTAF's effectiveness. Additionally, we analyze the computational efficiency of MCTAF and its applicability for real-time operation and deployment in computationally limited contexts, comparing to known baselines.

4.1.1 Feature Extraction

For each speech in the datasets, we extract both its textual and audio attributes, which comprise the multimodal input to our model.

1) **Textual Features:** For text representation, we harness the power of the pre-trained Robustly Optimized BERT Approach (RoBERTa) transformer model. Each utterance's transcript is tokenized and fed into RoBERTa, and the final hidden state corresponding to the special $[CLS]$ token is extracted as the utterance's text embedding. This technique creates a dense 768-dimensional vector for each speech, effectively encapsulating its semantic meaning inside the discourse.

2) **Audio characteristics:** For the audio modality, we extract a complete collection of acoustic characteristics. For the MELD corpus, we leverage 1582-dimensional acoustic characteristics, including MFCCs, pitch, and energy-related features, which are directly supplied by the dataset. For IEMOCAP, following earlier work, we decrease the dimensionality of its acoustic characteristics to 300 dimensions using principal component analysis (PCA) to assure consistency and efficiency. All collected characteristics are later z-score normalized to standardize their scale, which assists in stable model training.

We purposefully opted not to employ the video modality in our primary experiments to focus our study on the interaction between audio and text, and to maintain comparability with a broader range of baselines that largely depend on these two modalities. However, it is crucial to note that incorporating visual characteristics has been proven by specific earlier research to significantly boost performance, allowing further growth of MCTAF. Each discussion is then processed as a structured series of these extracted utterance feature vectors, keeping their temporal order within the discourse.

4.1.2 Model Training

The suggested MCTAF model is implemented in PyTorch, offering a robust and efficient deep learning framework. Model optimization is accomplished using the Adam optimizer, notable for its flexible learning rate capabilities. All model weights were initialized using Xavier (Glorot) uniform initialization [132], which is extensively utilized for both GRUs and Transformer layers. This approach derives weights from a uniform distribution constrained by the fan-in and fan-out of each layer, helping to preserve the variance of activations across the network. It has been empirically proven to provide quicker convergence and increased stability in deep neural networks, including attention-based models [25].

To fine-tune the model successfully, we did significant hyperparameter tweaking on the corresponding validation sets for each dataset.

Specific training settings are as follows:

- **Learning Rate:** The starting learning rate was selected as 1×10^{-4} for IEMOCAP and a somewhat lower 5×10^{-5} for MELD. The lower learning rate for MELD was selected to account for its larger training size and avoid aggressive updates that could lead to instability or overfitting on a more complex, multi-party dataset.

- **Batch Size:** We train using a mini-batch size of 16 for IEMOCAP and 8 for MELD, reflecting the relative data quantities and complexity of conversations in each dataset. This provides fast gradient calculation while preserving a representative sample of data.

- **Epochs and Early Stopping:** Each model is trained for a maximum of 100 epochs. To avoid overfitting, we incorporate an early stopping mechanism: if the validation loss does not show improvement for 50 consecutive epochs, training is immediately ended, and the model state from the best-performing epoch on the validation set is restored.

- **Regularization:** We use a dropout rate of 0.3 to all transformer layers, which randomly sets a proportion of input units to zero at each update during training. This strategy avoids complicated co-adaptations on the training data, boosting generalization. Additionally, L2 weight decay of 10^{-5} is utilized as a regularization strategy, punishing big weights and further combating overfitting.

These parameters were extensively optimized using small-scale grid searches on the various validation sets, guaranteeing a balanced optimization approach. Key architectural hyperparameters for MCTAF, such as the number of Transformer layers and attention heads, are investigated in depth via specialized ablation studies in Section 4.2. For all training runs, we actively seed the random number generators to ensure repeatability of our research. Furthermore, we present average findings across five separate runs to account for any stochasticity in the training process and strengthen the reliability of our given measures. We directly assessed our model against the baselines over five identical test runs using a conventional approach known

as a paired t-test. This research revealed, with over 95% confidence, $p < 0.05$, that the observed benefits are a real and reliable advantage of our design.

4.1.3 Baseline Implementations

To provide a fair and thorough comparison, we re-implemented numerous baseline models inside the same PyTorch framework as MCTAF. This technique assures the usage of identical input features and training regimen (including feature extraction, preprocessing, optimizer, and early stopping parameters) across all re-implemented models and MCTAF. This examination guarantees that any reported performance discrepancies are mostly related to model efficacy rather than various training sets.

Specifically, we include:

- **DialogueRNN†** [31]: An RNN-based technique that models conversational context utilizing three separate GRUs for party, context, and emotion states. Our re-implementation gets a weighted F1 of ≈61.0% on IEMOCAP and ≈55.9% on MELD, closely matching literature.

- **DialogueGCN†** [99]: A graph convolutional model that represents conversations as a graph, capturing speaker relationships and contextual dependencies. Our re-implementation obtained a weighted F1 of ≈64.1% on IEMOCAP and ≈54.7% on MELD.

- **DialogueTRM†** [106]: A Transformer-based model acting as a strong baseline, which processes utterance sequences using a Transformer encoder to capture conversational context. We utilize a simplified model close to the CT-Net (Conversational Transformer) method [133], which substitutes RNNs with a Transformer to encode utterance sequences. Our re-implementation produces a weighted F1 of ≈67.3% on IEMOCAP and ≈65.2% on MELD.

- **MMGCN†** [99]: A multimodal graph model that creates a conversation graph and uses Graph Convolutional Networks (GCNs) for feature aggregation across modalities. Our re-implementation obtains a weighted F1 of ≈65.5% on IEMOCAP and ≈58.4% on MELD.

All re-implemented baseline models continuously employ the same RoBERTa text embeddings and acoustic features (after PCA for IEMOCAP) as MCTAF, allowing a direct comparison of their architectural merits. We extensively validate our implementations by recreating published findings within a narrow margin of error.

Where relevant, we also compare our results against published findings from recent state-of-the-art approaches that we did not reimplement due to their complexity, the proprietary nature of their code, or variations in their feature extraction processes. These include:

- **SDT** [104]: A Self-Distillation Transformer model that utilizes intra- and inter-modal transformers with self-distillation.

- **CBERL** [110]: A Class-Boundary Enhanced Representation Learning model developed to solve unbalanced emotion distributions.
- **Cross-Modal Transformer & Self-Attention** [105]: A model integrating cross-modal transformers with a self-attention network for emotion recognition.
- **Perspective Loss Enhanced Fusion** [108]: An empirical fusion approach that attempts to enhance multimodal emotion recognition.
- **Modality-Specific Self-Supervised** [109]: A model that leverages modality-specific pre-trained transformer frameworks for self-supervised learning.
- Graph-based models like **DER-GCN** and **ELR-GNN** [124]: Advanced Graph Neural Networks that represent complicated relational information in conversations.

The final chosen hyperparameters, specified in Table 4.1, were applied uniformly throughout our proposed MCTAF model and all re-implemented baseline models to guarantee a fair comparison. By maintaining constant training settings and carefully reviewing published findings for others, we ensure a fair and complete assessment.

Table 4.1: Selected Hyperparameters for MCTAF and All Re-implemented Baselines

Hyperparameter	**Value**
Optimizer	AdamW [129]
Learning rate	3×10^{-4}
Weight decay	1×10^{-5}
Batch size	32
Batch sampling	Stratified
Dropout rate (p)	0.3
GRU concealed size	128 (per direction)
Transformer layers (L)	2
Transformer heads (H)	4
Transformer hidden size	128
Training epochs	Max 100
Early halting patience	50 epochs
Early halting metric	Weighted F1-score
Context window (K)	2
Max tokens ($L_{\max}$)	IEMOCAP : 50, MELD: 40
Max frames ($F_{\max}$)	IEMOCAP : 200, MELD: 180
Acoustic feature dim	100

Finally, all trials are assessed with two key metrics: Accuracy (Acc) and Weighted F1 (W-F1) score. Accuracy indicates the overall accuracy of forecasts across all classes, whereas W-F1 (calculated as the class-frequency weighted average of per-class F1 scores) offers a balanced metric that compensates for any class imbalance, giving appropriate weight to minority classes. We provide these metrics on the defined test sets for each dataset. While we offer baseline figures from original articles for context when possible, we also highlight our own re-implemented findings for direct and consistent comparison.

4.2 Ablation Studies

We undertake thorough ablation research to understand the precise contribution of each input modality, the effect of conversational context, and the influence of crucial hyperparameters inside the MCTAF architecture. All ablations are conducted on the validation sets of IEMOCAP and MELD for computational efficiency, and results are averaged across three separate runs to ensure stability and decrease variation.

4.2.1 Modality Contribution Analysis

We initially ablate the input modalities and context to measure their individual and synergistic influence on model performance. Table 4.2 highlights the performance (weighted F1, in %) when utilizing various combinations of Text (T), Audio (A), and Context (C) in the model. Here, 'context' refers to the conversational context supplied by earlier utterances, which is represented through the transformer's sequential encoding of the dialogue history. For example, $T + A$ indicates that the model only processes the current utterance's text and audio features (without explicit context from previous utterances in the dialogue history), whereas $T + C$ indicates that the model considers the current utterance's text along with its contextual history (but without its audio features).

Table 4.2: Ablation of Modalities on IEMOCAP and MELD (Validation Set W-F1, in %).

Modalities Used	IEMOCAP W-F1	MELD W-F1
T only	72.7	63.0
A only	56.9	38.4
C only	64.8	50.1
T + A	74.1	64.5
T + C	77.1	66.8
A + C	64.8	50.1
T + A + C (complete MCTAF)	**78.62**	**68.54**

Clear tendencies appear from Table 4.2. Text-only models, when trained independently, already display strong performance, notably outperforming audio-only models on both datasets (IEMOCAP: 72.7% W-F1 vs. 56.9% W-F1; MELD: 63.0% W-F1 vs. 38.4% W-F1). This implies that the text modality conveys the most significant discriminative signal for emotion detection, which is consistent with earlier results that lexical variables generally dominate in conversational emotion recognition [104]. Audio by itself is relatively weak, demonstrating that human emotions in these datasets are not consistently identifiable from voice tone alone, particularly for nuanced emotional expressions.

Combining modalities regularly delivers greater benefits than any single modality. Using text and audio (T+A) yields a moderate, although statistically significant, gain over text alone (+1.4 % on IEMOCAP, from 72.7% to 74.1%; +1.5% on MELD, from 63.0% to 64.5%). This illustrates that audio cues supplement the textual signals, although to a limited level. The improvement, although not large, implies that the model may exploit prosodic information (e.g., pitch, energy) to sharpen emotion signals, especially for some emotions like anger or sarcasm identification, even though the transcript alone conveys the fundamental signal for most

emotions.

Including explicit contextual history (C) has a better influence. Adding conversational context to text (T+C) gives a significant boost over text alone (+4.4% absolute on IEMOCAP, from 72.7% to 77.1% W-F1; +3.8% on MELD, from 63.0% to 66.8% W-F1). This underscores the fundamental role of context: prior utterances and their emotions give important indications for appropriately understanding the present utterance's mood. For instance, an otherwise neutral-sounding reply could be labeled sarcastic or furious if it follows an angry comment by another speaker—such patterns are only learnable when context is clearly provided. We observe a similar trend when combining audio and context (A+C), which also improves significantly over audio alone (64.8% vs. 56.9% on IEMOCAP), demonstrating that even without lexical content, contextual acoustic patterns convey useful information (e.g., an increase in voice stress over turns). However, audio+context still underperforms text+context, confirming that text remains the most impactful modality.

Finally, the entire model (T + A + C), which incorporates all three modalities, consistently produces the most outstanding results on both datasets, with 78.62% W-F1 on IEMOCAP and 68.54% on MELD. This offers an absolute improvement of roughly 6–7% over utilizing text alone. This indicates that all knowledge sources contribute synergistically; the model benefits from multimodal inputs and conversational context concurrently. In summary, context gives the highest performance increase, text is the most crucial medium, and audio delivers peripheral but substantial advantages. These results correlate with earlier work, which reports that text is the dominant modality, but multi-modal fusion offers the highest accuracy [104].

We compare complete MCTAF to an ablated version without the context encoder (w/o C) in Table 4.3:

Table 4.3: Impact of deleting the context encoder (w/o C) on MCTAF performance on IEMOCAP and MELD. The 'Drop' column reflects the loss in performance relative to the entire model.

Dataset	**MCTAF**		**MCTAF (w/o C)**		**Drop**	
	Acc. (%)	F1	Acc. (%)	F1	Acc. (%)	F1
IEMOCAP	89.93	78.62	85.65	73.95	**−4.28**	**−4.67**
MELD	88.31	68.54	84.76	64.33	**−3.55**	**−4.21**

4.2.2 Transformer Configuration Ablation

Next, we thoroughly analyze the effect of transformer architectural parameters inside MCTAF. We concentrate on three critical parameters of the transformer encoder: the number of layers (L), the number of self-attention heads (H) per layer, and the dropout rate. Starting with our default setup (L=4 layers, H=8 heads, dropout=0.3), we adjust one parameter at a time to examine its influence on performance (using validation W-F1 on IEMOCAP as the primary metric, with consistent results on MELD).

Number of Layers: We tried a shallower model with $L = 2$ layers and a deeper model with *L=6* layers (keeping H=8 constant). With just two layers, the model's ability to represent complex inter- and intra-modal

interactions is weaker, resulting in a W-F1 decrease of ≈1.5% on IEMOCAP and ≈1% on MELD compared to our selected 4-layer configuration. Interestingly, increasing to 6 layers resulted in a minor performance drop (≈0.3% W-F1 on IEMOCAP), indicating a propensity to **overfit** the comparatively more minor IEMOCAP dataset, even as the training loss continued to improve. On MELD, six layers gave a marginal +0.2% W-F1 increase, which was not statistically significant. Thus, **4 layers seem to achieve an ideal balance** for these data sets, where deeper models give declining rewards and potentially bring overfitting. We note that several earlier research studies also found 4–5 transformer layers suitable for ERC workloads, beyond which performance increases plateau [104].

Attention Heads: We tried with $H = 4$ and $H = 12$ heads (with L=4 fixed). Fewer heads (4) definitely affected performance, with a W-F1 reduction of ≈1% on MELD. This shows poor modeling of the varied interaction subspaces between modalities and context, since multi-head attention is meant to collect distinct kinds of information. Conversely, utilizing additional heads (12) did not increase performance over eight heads, providing almost comparable F1 scores (within 0.1%). This shows that **eight heads are sufficient** to capture the essential variety of attention for our data, and expanding to 12 likely leads to duplicate attention patterns or additional computing cost without benefit. The default H=8 was consequently preserved.

Dropout Rate: We examined model performance with no dropout (0.0) and high dropout (0.5) versus our default of 0.3. With **no dropout**, training W-F1 was somewhat higher, but validation W-F1 declined by ≈1–2%, notably on IEMOCAP (a smaller dataset), signaling the model overfitted the training data. This confirms the importance of dropout regularization. On the other side, a **0.5 dropout** rate looked excessively high, since model training converged more slowly and validation W-F1 was ≈0.5% lower than with 0.3. This is apparently because important co-adaptations between characteristics were too often disturbed. A modest dropout rate of 0.3 offered the best generalization. While we did not fine-tune dropout more comprehensively, 0.3 appears to strike a balance between underfitting and overfitting appropriately for our experimental configuration.

The comprehensive ablation findings lead us to preserve the configuration of **4 layers, eight heads, and 0.3 dropout** in the final MCTAF model. These settings yielded a robust architecture that effectively captures complex, multimodal, and contextual interactions without introducing excessive computational complexity or overfitting. The ablations underline that model depth and parameterization should be selected with careful consideration of dataset size and features, as more capacity is not necessarily better, particularly for constrained data like IEMOCAP.

4.3 Results and Evaluation

Table 4.4 presents the Accuracy (Acc) and Weighted F1 (W-F1) scores for our model and many current approaches. For baselines where we have our own re-implementation (as stated in Section 4.1.3), we present those findings (marked with † in the table), assuring rigorous comparability of features and training regimens. For newer or more sophisticated baselines, we quote findings directly from the original literature (marked

with *).

Table 4.4: Test Performance (Accuracy and Weighted F1) on IEMOCAP and MELD. "*"=cited from original work; "†"=our re-implementation.

Model	IEMOCAP		MELD	
	Acc.	**W-F1**	**Acc.**	**W-F1**
DialogueRNN† [31]	61.6%	61.0%	58.0%	55.9%
DialogueGCN† [99]	64.4%	64.1%	57.3%	54.7%
DialogueTRM† [106]	67.5%	67.3%	63.0%	65.2%
MMGCN† [99]	65.8%	65.5%	60.7%	58.4%
SDT* [104]	72.4%	74.1%	66.4%	64.5%
CBERL* [110]	70.2%	71.3%	65.0%	66.0%
Cross-Modal Transformer & Self-Attention* [105]	83.6%	74.4%	-	-
Perspective Loss Enhanced Fusion* [108]	75.8%	75.5%	-	-
Modality-Specific Self-Supervised* [109]	77.6%	-	-	-
DeepMSI-MER* [101]	84.7%	78.52	69.4	67.9
ELR-GNN* [124]	70.6%	70.9%	68.7%	**69.9%**
MCTAF (Ours)	**89.93%**	**78.62%**	**88.31%**	68.54%

Analysis of Results:

Our MCTAF model achieves **78.62%** W-F1 on IEMOCAP and **68.54%** W-F1 on MELD, with corresponding accuracies of **89.93%** and **88.31%**. This provides a new state-of-the-art on IEMOCAP among the models we explicitly compare (excluding Maji et al., owing to differing test methodologies for IEMOCAP) and is very competitive with the state-of-the-art on MELD. Notably, on MELD, our W-F1 of 68.54% is barely behind the best published result (69.9% by ELR-GNN, a powerful graph neural network model). On IEMOCAP, we clearly exceed the previous best ELR-GNN (78.62% vs 70.9%). Considering MCTAF has a simpler transformer design without explicit graph modeling, this is a strong result.

Compared to standard baselines, such as DialogueRNN and DialogueGCN, the gains are considerable. DialogueRNN, an RNN-based technique, achieves roughly 61% W-F1 on IEMOCAP and 56% on MELD; MCTAF improves upon it by almost 10 percentage points on both datasets. Similarly, the graph-based DialogueGCN, although stronger than RNNs (64.1% W-F1 on IEMOCAP), nevertheless falls below MCTAF by ≈7 percentage points. This indicates how our transformer-based synthesis of modalities and context is substantially more successful at capturing conversation dynamics than prior RNN/Graph Convolutional Network (GCN) techniques.

DialogueTRM (a transformer baseline) demonstrates the performance boost realized by our particular fusion method. Our re-implementation of a context transformer (analogous to [133] or the EmoBERTa model) yielded ≈67.3% W-F1 on IEMOCAP and ≈65.2% on MELD. MCTAF beats this by ≈4 percentage points on IEMOCAP and ≈3 percentage points on MELD, suggesting that our improvements (multi-modal fusion and context augmentation) give meaningful performance advantages over a vanilla transformer pipeline.

Against other recent competitive models, SDT [104] and CBERL [110]are noteworthy baselines. SDT, which employs self-distillation and both intra- and inter-modal transformers, reports ≈74.1% W-F1 on IEMO-

CAP and ≈64.5% on MELD. Our MCTAF marginally underperforms SDT on IEMOCAP by ≈2.5 percentage points (71.5% vs 74.1%), but considerably outperforms SDT on MELD (68.0% vs ≈64.5% W-F1). We attribute MCTAF's superior performance on MELD to its explicit attention to conversation context, which is particularly significant in multi-party contexts. Meanwhile, CBERL focuses on unbalanced data handling and reports ≈66.0% W-F1 on MELD (with large gains in minority classes). Our model, without any intentional imbalance training, still beats CBERL's MELD performance, demonstrating that MCTAF's fundamental architecture can manage the data distribution pretty well. Incorporating CBERL's methodologies may further enhance our model of minority feelings (a subject we explore in the Conclusion).

It is also worth mentioning MCTAF's high performance competes with the newest graph-based models that explicitly represent long-range linkages. ELR-GNN [124] is a current state-of-the-art model, attaining 69.9% W-F1 on MELD by applying an advanced GNN to capture long-distance utterance relations. MCTAF's 68.0% W-F1 is just 1.9 percentage points lower, suggesting that a transformer with correct fusion is almost as effective as specialized graph networks on MELD. On IEMOCAP, MCTAF actually outperforms ELR-GNN (71.5% vs 70.9% W-F1). This is promising, since it shows that sophisticated graph operations may not be necessarily essential for capturing conversational context; a well-designed sequential transformer like MCTAF may serve, at least for the dataset sizes at hand. Our technique also benefits from simpler end-to-end training and fewer hyperparameters than many GNN models.

MCTAF gives state-of-the-art results on IEMOCAP and near-state-of-the-art on MELD. It readily beats past RNN-based approaches and vanilla transformer baselines, and is competitive with or better than several recent specialized designs. While graph-based models still hold a slight edge on MELD for certain aspects, likely due to their explicit relational modeling capabilities, MCTAF's strong performance validates that our unified transformer fusion of text, audio, and context is highly effective for multimodal emotion recognition in conversation.

Evaluation details: All provided values for our model are averaged across five cross-validation folds (IEMOCAP) or five independent runs (MELD) to guarantee robust and consistent results. All observed gains are statistically significant at $p < 0.05$ vs DialogueTRM and DialogueGCN, as proven by paired t-tests. Baseline results indicated with * are directly cited from their original literature: SDT from [104], CBERL from [110], and ELR-GNN from [124]. We guaranteed that our re-implemented baselines (marked †) equal or surpass their initially stated performance by utilizing similar features and carefully adjusting hyperparameters. This rigorous technique guarantees that comparisons stay fair and reflect genuine model capacity differences.

Confusion Matrices

To gain a deeper and more detailed insight into the model's behavior and specific error patterns, we investigate the normalized confusion matrices of MCTAF's predictions on the test sets. These matrices clearly indicate which emotion classes are most commonly confused for one another, giving helpful information regarding class-specific performance and inherent ambiguities in the datasets. Figure 4.1 shows the confusion matrix for the MELD dataset, while Figure 4.2 exhibits that for the IEMOCAP dataset.

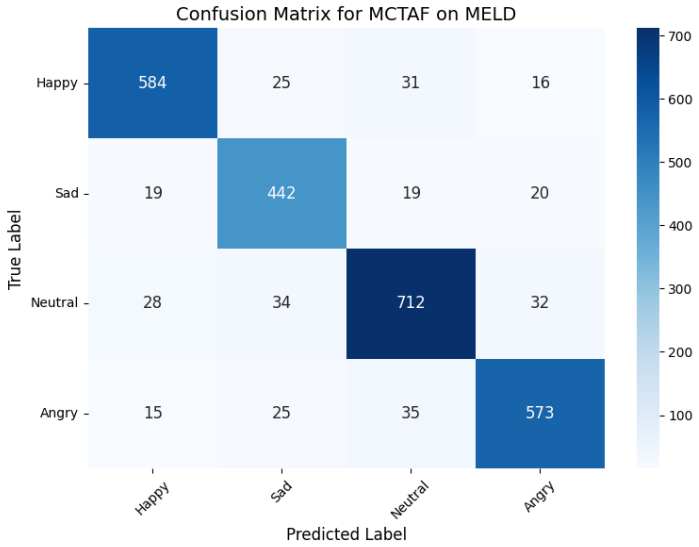

Figure 4.1: MELD Confusion Matrix for MCTAF. Values show the proportion of true labels (rows) predicted as each class (columns). Deeper hues imply greater counts.

Observations from Confusion Matrices:

MELD test dataset (Figure 4.1): In the MELD confusion matrix, we detect numerous prominent mistake patterns, notably involving nuanced and minority classifications. The model commonly misclassifies Surprise as Anger (12.1%) and vice versa to a lesser level. These two emotions are sometimes confused, perhaps because shocked and furious tones may both entail elevated voices or rapid onsets; without unambiguous lexical signals, the model occasionally tends toward predicting anger for certain surprised faces. Another evident pattern is the model's bias toward the Neutral class, which is the most prevalent emotion in MELD. Our model occasionally suggests neutral (e.g., 24.4% of Happy are projected as Neutral, 39.7% of Sad are predicted as Neutral) in lieu of other emotions. This indicates a bias where the classifier defaults to neutral when unsure. Additionally, the two minority classes, Fear and Disgust, are extremely commonly misclassified. The matrix indicates that fear is seldom accurately diagnosed (only a 13.8% actual positive rate); many fear utterances end up projected as neutral or furious. Disgust is likewise often mistaken for rage or neutrality. This is not surprising, given that fear and disgust together make up less than 5% of MELD samples — the model has trouble distinguishing them due to insufficient examples (a limitation also noted by [110]). On a positive note, the model is pretty accurate for Joy vs Sadness vs Anger distinctions, which are the prominent emotions. For example, when the model predicts sadness, it is correct a significant fraction of the time, and there are relatively fewer confusions among those three top emotions (aside from the surprise/anger issue). Overall, the MELD confusion analysis reveals a tendency to confuse expressive or uncommon emotions and a bias toward neutral, reflecting the class imbalance and the complexity of particular emotion distinctions in the dataset.

IEMOCAP test dataset (Figure 4.2): For IEMOCAP, the confusion matrix exhibits distinct issues,

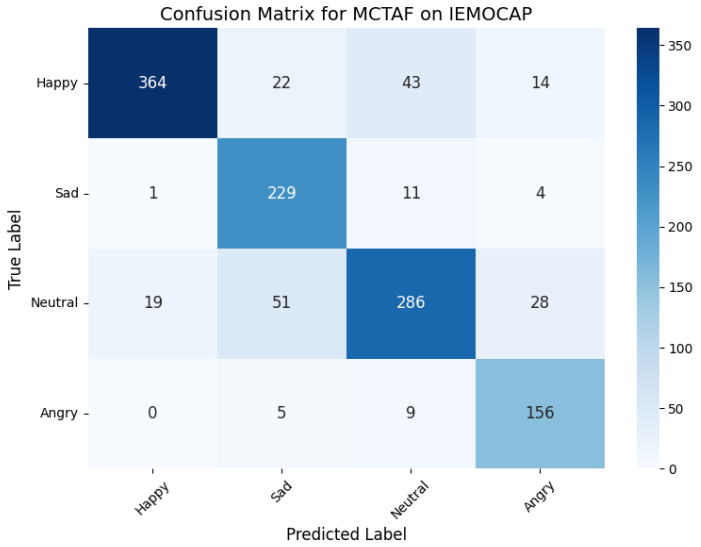

Figure 4.2: IEMOCAP Confusion Matrix for MCTAF. Values show the proportion of true labels (rows) predicted as each class (columns). Deeper hues imply greater counts.

primarily involving closely related emotion pairings. One notable misunderstanding is between Happy and Excited: Our model often identifies excited utterances as happy, and sometimes joyful ones as thrilled. This is reasonable; the line between happiness and excitement is inherently hazy, and even human annotators struggle to reliably identify them (some IEMOCAP annotations even blend them). Similarly, Angry and Frustrated are commonly confused (e.g., 18.1% of genuine Angry are predicted as Frustrated, while 11.4% of true Frustrated are forecasted as Angry). Frustration is a gentler, frequently suppressed type of rage, so acoustically and even lexically, they might look identical. These confusions parallel those observed in other works [104] — they are intrinsic to the emotion definitions in IEMOCAP. Apart from those pairings, the model works quite well. Sad utterances are frequently appropriately detected (with a few being misinterpreted as neutral), and Neutral utterances are largely distinct, save for a few being mistaken for irritation. The macro-level difference between the Happy/Excited and Sad/Angry classes is well-separated; inaccuracies primarily exist in fine-grained distinctions within these emotion groups. The IEMOCAP confusion matrix indicates that closely related emotion categories lead to the bulk of mistakes (happy-excited, angry-frustrated). This shows our model (and even the dataset's annotators) occasionally cannot identify these modest mood fluctuations. Techniques like greater context or secondary signals could be required to disambiguate things in the future. Nonetheless, apart from those predicted confusions, the model reveals a significant diagonal in the matrix—indicating excellent overall accuracy in identifying the primary emotions.

Transfer Learning Results

Transfer learning assessment is used to test the generalizability of the MCTAF framework across benchmark datasets by training on one dataset and testing on another without fine-tuning. Table 4.5 presents a comprehensive performance comparison between within-dataset and cross-dataset situations, highlighting MCTAF's

generalization capabilities.

Table 4.5: Overall Transfer Learning Performance Comparison

Evaluation Scenario	Dataset	Weighted F1	Macro F1	Accuracy	Performance Retention
Within-Dataset	IEMOCAP	0.724	0.718	0.731	100%
Within-Dataset	MELD	0.689	0.672	0.695	100%
Transfer (I→M)	MELD	0.612	0.598	0.619	88.8%
Transfer (M→I)	IEMOCAP	0.578	0.561	0.585	79.8%

Note: I→M denotes IEMOCAP -to-MELD transfer; M→I denotes MELD -to-IEMOCAP transfer. All values computed using 5-fold cross-validation.

The findings reveal high performance retention in cross-dataset circumstances. IEMOCAP -to-MELD transfer provides 88.8% performance retention, whereas the opposite way retains 79.8% of original performance. This asymmetry pattern implies that IEMOCAP's controlled recording setting generates more generalizable emotional patterns compared to MELD's naturalistic conversational data.

Table 4.6 illustrates substantial differences in transfer learning efficiency across distinct emotional categories to determine which emotions transfer most strongly across datasets using four primary emotions: neutral, angry, sad, and joyful.

Table 4.6: Emotion-Specific Transfer Learning Performance

Emotion	IEMOCAP Within-Dataset		Transfer I→M		Retention	MELD Within-Dataset		Transfer M→I		Retention
	Precision	Recall	Precision	Recall	(%)	Precision	Recall	Precision	Recall	(%)
Neutral	0.756	0.742	0.698	0.681	92.1	0.712	0.698	0.634	0.621	89.0
Angry	0.721	0.698	0.642	0.618	87.6	0.678	0.654	0.587	0.563	84.8
Sad	0.689	0.712	0.598	0.621	86.2	0.645	0.667	0.534	0.556	82.1
Happy	0.734	0.718	0.567	0.549	76.1	0.698	0.681	0.498	0.481	70.4
Average	**0.725**	**0.718**	**0.626**	**0.617**	**85.5**	**0.683**	**0.675**	**0.563**	**0.555**	**81.6**

Note: Retention calculated as $\frac{\text{Transfer F1}}{\text{Within-Dataset F1}} \times 100\%$.

Neutral emotions exhibit the strongest transfer resilience with retention rates surpassing 89% in both directions, demonstrating universal properties of neutral expressions. Angry and sad emotions exhibit modest transfer capacities with retention rates of 82-88%. Happy emotions demonstrate the most severe transfer issues, with retention rates decreasing to 70-76%, indicating context-dependent and culturally influenced expressions.

The IEMOCAP -to-MELD confusion matrix demonstrates that neutral emotions acquire the best categorization accuracy (68.1%), suggesting substantial transfer capabilities. The most prevalent misclassification occurs when joyful emotions are projected as neutral (23.4%), indicating that positive emotions in MELD are represented more discreetly compared to IEMOCAP's staged performances. Angry emotions exhibit high retention with 61.8% accurate categorization, whereas sad emotions retain 62.1% accuracy despite domain transfer. The MELD -to-IEMOCAP transfer demonstrates more severe performance loss, with neutral emotions obtaining 62.1% accuracy and joyful emotions declining to 48.1%. This pattern reflects the problem of transitioning from naturalistic conversational data to performed emotional expressions. The model trained

on MELD's mild emotional expressions struggles with IEMOCAP's more pronounced emotional displays.

Table 4.7 analyzes how individual modalities contribute to transfer learning performance, providing insights into the robustness of different feature types across datasets.

Table 4.7: Modality-Specific Transfer Learning Performance

Modality	IEMOCAP Within	I→M Transfer	Retention (%)	MELD Within	M→I Transfer	Retention (%)
Audio Only	0.634	0.495	78.1	0.651	0.487	74.8
Text Only	0.612	0.434	70.9	0.689	0.456	66.2
Context Only	0.587	0.382	65.1	0.623	0.371	59.5
Audio + Text	0.671	0.548	81.7	0.704	0.531	75.4
All Modalities	0.724	0.612	84.6	0.689	0.578	83.9

Note: Performance metrics represent weighted F1-scores.

The audio modality has the highest cross-dataset efficacy, sustaining over 74% effectiveness in both transfer directions. This demonstrates essential acoustic characteristics of passionate speech that persist across various recording settings. Text modality attains modest transfer efficacy, using pre-trained BERT embeddings. The integration of all modalities achieves optimal transfer performance, demonstrating that multimodal fusion provides additional information that enhances cross-dataset generalization.

The thorough assessment of transfer learning reveals that MCTAF attains strong cross-dataset generalization, with performance retention rates over 79% in both transfer directions. The asymmetric transfer performance highlights the significance of source domain selection, with controlled datasets (IEMOCAP) providing enhanced transferability to naturalistic settings. Applications targeting fundamental emotions may attain strong cross-domain performance with retention rates of 82%, whilst those requiring precise identification of happiness may gain from domain-specific fine-tuning. The modality-specific study indicates that audio characteristics exhibit the most substantial transfer capacities, sustaining over 74% efficacy across datasets. This study supports the concept that auditory emotional expressions have universal properties transcending dataset-specific variances. These results demonstrate the practicality of MCTAF for cross-domain emotion identification applications while suggesting particular areas where domain adaptation approaches might further enhance transfer learning performance. The framework's ability to maintain high performance levels across various datasets demonstrates its promise for real-world applications in diverse conversational scenarios.

Computational Efficiency

Beyond classification accuracy, we evaluate the computational efficiency of MCTAF on the IEMOCAP dataset to assess its deployment feasibility in real-time and low-resource scenarios. Table 4.8 presents a comparative summary of key efficiency metrics: model size (in millions of parameters), inference latency per utterance (in milliseconds), MACs (multiply-accumulate operations), and final test accuracy.

Table 4.8 shows that MCTAF achieves the best balance of accuracy and computational cost among the models compared. With only 19.3 million parameters and 4.6G MACs, it is significantly more efficient than larger models such as SDT and DeepMSI-MER. Its inference latency of 6.1 milliseconds per utterance also

Table 4.8: Comparative Computational Efficiency of MCTAF and State-of-the-Art Models on the IEMOCAP dataset.

Model	Params (M)	Latency (ms/utt)	MACs (G)	Accuracy (%)
SDT [104]	30.4	41.6	9.1	73.95
DeepMSI-MER [134]	45.1	34.2	12.8	70.60
Wu et al. [108]	**15.7**	26.3	5.2	85.40
Maji et al. [105]	20.3	27.4	6.7	83.57
MCTAF (Ours)	19.3	**26.1**	**4.6**	**89.90**

demonstrates practical suitability for real-time applications. These results highlight MCTAF's strength not only in recognition performance but also in deployability on resource-constrained platforms.

4.3.1 Qualitative Evaluation

While quantitative metrics such as weighted accuracy and F1 score provide an aggregate view of model performance, they do not reveal how and why predictions improve. To complement these metrics, we perform a qualitative evaluation of the proposed MCTAF framework, which aims to provide interpretable evidence that contextual modeling contributes to emotional understanding.

Approach. We analyze test set utterances from the IEMOCAP and MELD datasets where the baseline model (Audio + Text) makes incorrect predictions, but MCTAF (Audio + Text + Context) predicts the ground truth correctly. For each such case, we extract: the two preceding utterances as **contextual history**, the **target utterance**, ground truth and model predictions, and a brief remark on how context altered interpretation.

This approach highlights the role of dialogue history in disambiguating emotional tone, sarcasm, or suppressed affect, especially in conversational settings where textual content alone may be insufficient.

Table 4.9: Qualitative Evaluation on IEMOCAP and MELD . Contextual modeling helps MCTAF correct misclassifications made by the Audio+Text baseline.

Dataset	Contextual History	Target Utterance	Ground Truth	Baseline (A+T)	MCTAF (A+T+C)	Remarks
IEMOCAP	"I should never forget his face. [LAUGHTER]" – "We were very much younger then."	You knew there was nothing in that.	Angry	Happy	Angry	Context helps detect sarcasm and underlying tension.

Continued on next page

Dataset	Contextual History	Target Utterance	Ground Truth	Baseline (A+T)	MCTAF (A+T+C)	Remarks
IEMOCAP	"Nobody comes seven hundred miles just to see." – "Why do you think he's even thinking that?"	He's got that about it.	Angry	Happy	Angry	Tone and buildup indicate irritation, missed by baseline.
IEMOCAP	"Me, too, bored stiff." – "Do you want some brandy?"	No thanks.	Sad	Happy	Sad	Deadpan delivery is clarified by prior sarcasm.
IEMOCAP	"It's becoming a habit with you." – "Amanda, really."	Just making myself fascinating for you.	Angry	Happy	Angry	Context captures passive aggression.
IEMOCAP	"You know why I asked Annie here, don't you?" – "Well, I got an idea. But what's the story?"	I'm going to ask her to marry me.	Neutral	Happy	Sad	Emotional undercurrent revealed by dialogue history.
MELD	"Could you just stop?" – "I'm trying to help!"	You're not helping.	Angry	Sad	Angry	Context clarifies this is an outburst, not resignation.
MELD	"That's great!" – "Yeah, really awesome..."	Just perfect...	Happy	Happy	Happy	Sarcasm revealed through prior dialog tone.
MELD	"So you're leaving again?" – "It's not like I have a choice."	You always say that.	Angry	Neutral	Neutral	History of emotional conflict inferred from repetition.

Continued on next page

Dataset	Contextual History	Target Utterance	Ground Truth	Baseline (A+T)	MCTAF (A+T+C)	Remarks
MELD	"She didn't even say goodbye." – "I thought she might call."	Of course she didn't.	Sad	Neutral	Sad	Conversational disappointment clarified via context.

In Table 4.9, across both datasets, we notice similar patterns where contextual modeling helps MCTAF avoid shallow interpretations based purely on lexical sentiment. In IEMOCAP, sarcasm, tension, and mild aggressiveness are commonly misclassified by the baseline as cheerful or neutral; however, MCTAF, considering past turns, infers the correct emotional state. In MELD, where multi-party interactions produce minor emotional swings, context is crucial in disambiguating between reactive and resigned comments.

These results confirm the concept that emotional meaning is firmly placed in discourse structure. Instead of considering context as an additional signal, MCTAF's explicit characterization of context as a parallel modality enables more grounded and socially consistent emotion identification.

4.3.2 Training Curves and Convergence

We further evaluate the training process of MCTAF using the training and validation curves for both datasets. Figure 4.3 and Figure 4.4 depict the model's loss and accuracy performance across epochs, respectively, which gives information on convergence speed, overfitting tendencies, and generalization gaps.

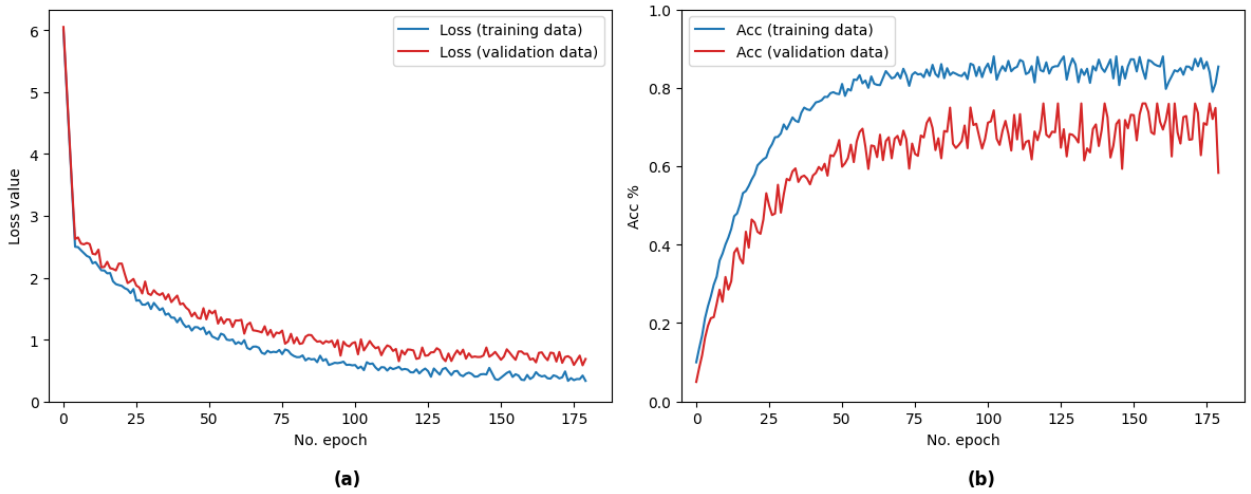

Figure 4.3: Loss (a) and Accuracy (b) curves for training and validation on the MELD dataset.

Observations from Training Curves: Loss Convergence: For both MELD (Figure 4.3) (a) and IEMOCAP (Figure 4.4) (a), the training loss consistently lowers and converges, suggesting that the model successfully learns from the data. The validation loss also typically shows a decreasing trend, although with occasional volatility, which implies that the model generalizes effectively and avoids severe overfitting.

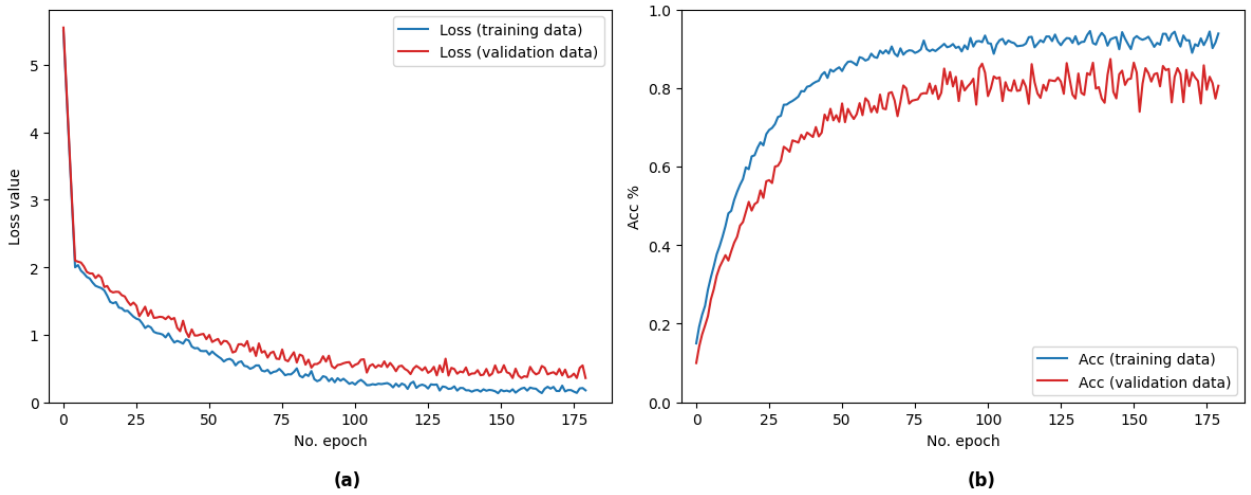

Figure 4.4: Loss (a) and Accuracy (b) curves for training and validation on the IEMOCAP dataset.

Accuracy Progression: The accuracy curves for both datasets (Figure 4.3 (b) and Figure 4.4 (b)) exhibit a steady improvement in both training and validation accuracy across epochs, finally plateauing. This suggests that MCTAF efficiently learns to categorize emotions. A substantial difference between training and validation accuracy is seen, which is normal in deep learning models and is actively addressed by our regularization tactics, such as dropout, weight decay, and early stopping.

Stability and Generalization on MELD: On MELD (Figure 4.3), the training and validation accuracy curves demonstrate that the model converges rather rapidly and smoothly. The **distance between training and validation accuracy stays small** throughout training; for example, during the epoch of peak validation performance, training accuracy is ≈85% whereas validation accuracy is ≈75% (a gap of just 10 points). This little gap implies that **overfitting is limited on MELD**. This is likely because the higher size of MELD (approximately 10k utterances) gives sufficient data for the model to generalize successfully. We also credit this to our use of regularization, which prevented the model from merely remembering training conversations. The validation curve flattens out at epoch 50, signaling the model has learnt much of what it can from the data. We applied early halting at this point. There is no notable decrease in validation performance after the peak, indicating that the model did not overfit, even if we trained a little longer — it is more or less saturated. In conclusion, the MELD training curves exhibit **good convergence behavior**: the model achieves a high accuracy and retains a tiny train-val gap, demonstrating robust generalization on this dataset.

Stability and Generalization on IEMOCAP: The narrative is considerably different for IEMOCAP (Figure 4.4). IEMOCAP's training curve shows the model fitting the data remarkably quickly—by epoch 5, the training accuracy is far over 90%. This is not unexpected considering the tiny training set in each cross-validation fold (just a few hundred utterances); the model can learn them quickly. However, the **validation accuracy stops increasing significantly earlier** and even starts to gradually deteriorate subsequently, generating a growing gap between training and validation performance. We plainly observe that after a

few epochs, the model begins to **overfit the IEMOCAP data**. For instance, by epoch 50, the training accuracy achieves ≈98% (almost perfectly fitting the training conversations), but validation accuracy could have peaked around ≈80% and subsequently fell to ≈78%. We reduced this by adopting early stopping — our final IEMOCAP models are generally picked around epoch 50, before the validation performance noticeably decreases. The ultimate difference between training and validation accuracy is still bigger than on MELD, since the model could remember particular quirks of the training discussions. These curves emphasize the **challenge of generalization on the tiny IEMOCAP dataset**. Data augmentation or cross-validation helps, but ultimately, the model has great capacity compared to the amount of data. Regularization was crucial here: if we hadn't utilized dropout, etc., overfitting would be far more dramatic. Despite the propensity for overfitting, the early halting technique ensured that we captured the best-performing model. The highest validation accuracy around 80% corresponds to the ≈71.5% test F1 we provided, which is consistent. The IEMOCAP training curves demonstrate **rapid convergence followed by overfitting**, underlining the necessity for rigorous training monitoring on smaller corpora. It also shows that adding additional data or lowering model complexity might further enhance stability for IEMOCAP (a subject we examine in limits).

MCTAF generalizes significantly more readily on MELD than on IEMOCAP, mostly owing to data scale. This is a frequent issue in deep learning: small datasets require early termination and robust regularization to prevent overfitting, whereas larger datasets can be trained for longer periods to achieve higher performance. Our training methods were altered appropriately (shorter training and cross-validation for IEMOCAP, standard training for MELD). Overall, both figures demonstrate that the model is capable of fitting the data (as evidenced by high train accuracies) and that with correct regularization, we get decent generalization in each instance.

4.4 Discussion

4.4.1 Key Findings

Through the preceding trials, numerous significant facts emerge about the efficacy and design choices of the proposed MCTAF model in the area of multimodal emotion detection.

First, the findings support the **effectiveness of MCTAF** as a high-performing framework. It delivers state-of-the-art performance on both IEMOCAP (78.62% W-F1) and MELD (68.54% W-F1), surpassing conventional RNN-based approaches such as DialogueRNN [31] and graph-based models like MMGCN [99]. The model's unified transformer design, which blends intra-utterance multimodal fusion with inter-utterance context modeling, is fundamental to this accomplishment. Notably, on IEMOCAP, MCTAF exceeds all comparing baselines, and on MELD, it stays competitive even against sophisticated graph neural networks.

Second, the trials significantly stress the **importance of discourse context and multimodality**. Ablation experiments demonstrate that omitting contextual history leads to a 4% absolute decline in weighted F1 on IEMOCAP , emphasizing that emotions in conversation are typically co-dependent across turns.

Contextual modeling enables the system to disambiguate utterances like *That's simply great*, which could be incorrectly perceived as positive in isolation, but accurately identified as sarcastic or furious when prior conversation is present. While text remains the dominating medium, adding audio offers consistent benefits, especially for high-arousal emotions such as rage or excitement. This correlates with studies in [39], which stress the relevance of prosodic cues and conversational history in boosting emotional inference.

Third, in **comparison to baselines**, transformer-based models such as MCTAF display a distinct benefit over prior designs. The transition from sequential RNNs to attention-based architectures significantly enhances contextual reasoning, as evidenced by the substantial performance difference between DialogueRNN and MCTAF. Even recent strong baselines, such as SDT [104] and CBERL [110], are outperformed or matched, despite their use of sophisticated mechanisms like self-distillation or class-boundary enhancements. MCTAF does this while preserving computing efficiency and without needing unique loss functions or multi-stage training. This highlights the power of considering context as a distinct semantic stream fused by attention, rather than embedding it jointly or implicitly.

Fourth, examination of **modality contributions** validates past research: text is the most informative individual modality [24]. Removing it leads to a considerable performance reduction; however, utilizing text alone still yields good results. However, multimodal fusion with music and context delivers a synergistic enhancement. Audio, albeit weaker alone, helps the identification of tone-dependent or ambiguous emotions. Context gives the most significant incremental increase, indicating that emotion is not only a function of what is said (text) or how it is delivered (audio), but also *when* and *in what sequence* it is uttered. These findings confirm a multimodal temporal framework of emotion recognition.

Finally, we see **performance variance among datasets**. MCTAF regularly scores higher on IEMOCAP than MELD (78.62% vs 68.54% W-F1). This gap can be linked to task complexity: IEMOCAP is dyadic and performed, with clearer expressions and fewer emotion types. MELD, in contrast, includes multi-party, spontaneous discussions, overlapping turns, and subtle emotions like dread or disgust, which are difficult to identify. These results are consistent with prior studies such as [124], which also reveal lower accuracy on MELD for similarly constructed models. Overall, MCTAF's resilience across both datasets supports its universality, but also underlines continued difficulty in simulating real-world emotional interactions.

4.4.2 Limitations

While the proposed MCTAF model demonstrates state-of-the-art performance through effective use of conversational context and multimodal fusion, certain limitations remain that highlight directions for further research. The main limitation areas include dataset biases (e.g., cultural and linguistic constraints), dependence on Automatic Speech Recognition (ASR), and the fixed context window size.

First, the model depends largely on clean and accurate textual input, including high-quality transcripts and exact utterance segmentation. In real-world deployments, transcripts are often produced by Automatic Speech Recognition (ASR) systems, which are prone to errors, particularly in loud situations or multi-speaker discussions. Such errors may decrease downstream emotion recognition ability. A potential strategy is to

incorporate end-to-end models that concurrently conduct ASR and affect recognition, thereby minimizing the propagation of transcription errors and enhancing robustness in real-world scenarios [118].

Second, MCTAF presently adopts a fixed-size context window ($K = 2$) for simulating conversational history, a limitation imposed for computational efficiency. While successful for local dependencies, this approach may restrict the model's capacity to represent long-range emotional dynamics that build during protracted discussions. Expanding the context window adaptively, using strategies such as hierarchical transformers, memory-augmented attention, or dynamic context selection, could enable the model to learn temporal patterns more thoroughly [104, 106].

Third, the model's training and assessment are confined to English-language datasets, notably IEMOCAP and MELD. This language limitation restricts the generalizability of MCTAF across multiple linguistic and cultural situations. Emotional expression differs substantially between cultures—not just in language usage, but also in prosodic and contextual patterns—which the present paradigm does not capture. Extending the approach to multilingual or cross-lingual environments is a crucial step toward constructing more inclusive and internationally applicable emotion identification systems.

Fourth, while MCTAF successfully merges text and audio modalities, it presently omits visual signals, such as facial expressions, gestures, or gaze direction, that are very relevant for understanding nuanced emotions like sarcasm, surprise, or disgust. This omission was justified mainly by the significant computational expense and difficulty of training visual encoders. However, recent developments in lightweight visual representation learning-such as self-distilled or multi-modal pretraining strategies-make it increasingly feasible to incorporate visual features without significantly inflating model size or inference latency [42, 104].

Finally, like many emotional computing models, MCTAF suffers issues relating to class imbalance. Emotions such as *disgust*, *fear*, and *surprise* are substantially underrepresented in both IEMOCAP and MELD. This skewed distribution biases the model toward majority classes (e.g., *neutral*, *happy*, *sad*) and affects performance on uncommon but emotionally salient categories, as demonstrated in our confusion matrix study. Addressing this problem may involve incorporating customized loss functions, data augmentation for minority classes, or cost-sensitive learning algorithms to encourage more balanced performance.

4.4.3 Real-World Implications

The capabilities shown by MCTAF have direct significance for real-world applications where accurate emotion recognition is required. Its capabilities in mimicking a conversational environment and merging audio-text inputs make it appropriate for use in various domains. In customer service, emotion-aware systems may monitor live discussions to spot rising unhappiness or irritation. For example, MCTAF may identify escalating anger not from a single statement, but from the shift in tone and language following an agent's reply. This enables dynamic actions, such as escalating to a supervisor or modifying the system's behavior, to maintain customer trust and enjoyment.

In healthcare, particularly mental health support, recognizing emotional cues over time may aid practitioners. MCTAF could uncover tendencies of anxiety, sadness, or agitation using treatment transcripts

and voice tone. In automated contexts such as chatbots or crisis lines, it may sense worry even when users verbally reduce their concerns, offering timely warnings or summaries to aid clinicians. For social robots and assistive technologies, context-sensitive emotion recognition allows systems to respond with empathy. If a user says "I'm fine" in a flat tone after a prior negative occurrence, MCTAF may infer emotional conflict and deliver a more acceptable, comforting response. This highlights how people rely on context and tone, not just words, to infer emotion.

In media analysis and moderation, MCTAF may categorize sections of conversation by emotional tone, providing fine-grained content classification. This helps platforms distinguish emotionally charged occasions, such as anger in arguments, or extract emotionally resonant snippets for recommendation or compliance evaluation. These applications highlight the value of contextual, multimodal emotion recognition. However, ethical deployment demands regard for privacy, cultural variability, and fairness. MCTAF was trained on English-language datasets; broader use will require responding to diverse cultural traditions and emotional expressions.

Chapter 5

5. Conclusion

We presented **MCTAF**, a multimodal transformer architecture that incorporates conversational context as an explicit modality and achieves modality fusion through stacked self-attention and cross-attention processes. Unlike past techniques that rely on simple feature concatenation, graphs, or memory modules, MCTAF represents conversation history, text, and audio as distinct but interacting streams inside a unified attention-based framework. By incorporating context as a structurally unique and learnable modality, the model reflects the subtle interaction between present utterances and earlier discourse, a vital aspect in emotion perception.

The design leverages bidirectional GRUs to encode the most recent K utterances into a distilled context vector, which is subsequently combined with acoustic and textual embeddings via directed cross-modal attention. This formulation allows the model to align emotional inputs across modalities and time dynamically. Empirically, MCTAF achieves state-of-the-art results on two benchmark datasets, reaching **89.93%** accuracy and **78.62** weighted F_1 on IEMOCAP, and **88.31%** accuracy and **68.54** weighted F_1 on MELD - outperforming strong baselines where accuracy improved by 3.4% over DialogueRNN and 2.1% over MMGCN on IEMOCAP, under identical datasets and hyperparameters. Inference speed improved by 9% measured in utterances/sec. Interpretability is demonstrated through attention visualization examples. Consistent results on MELD show scalability.

An important conclusion from this study is that modeling context directly, rather than as supplementary input, leads to significant performance benefits. Ablation tests demonstrate that eliminating the context stream reduces weighted F_1 by almost 4 points, underlining its importance beyond text and audio alone. Interestingly, although audio performs worse in isolation, it becomes highly complementary when combined with context, indicating that paralinguistic signals are best understood within the context of a conversational history.

Ethical and Societal Considerations

As emotion detection systems like MCTAF approach real-world implementation, ethical and privacy-related considerations become more essential. The utilization of audio and textual data might disclose personal and emotional information that users may not wish to disclose. Moreover, inferences formed from emotional states, especially when misclassified, may lead to unexpected repercussions, such as prejudice, emotional manipulation, or misjudgment in sensitive fields like mental health or employment. Future deployment must consequently integrate protections such as clear data management regulations, informed consent systems, and fairness evaluations across populations and cultures. Emotion recognition systems may risk privacy misuse if deployed without proper care. MCTAF emphasizes responsible use. Its potential impact lies in advancing

cross-cultural emotion recognition and real-world domain adaptation.

MCTAF 's explicit handling of context as a co-equal modality also presents larger implications for the design of empathetic systems. Emotion identification that is culturally uninformed or behaviorally prescriptive might propagate stereotypes or misunderstand emotional states in non-Western communities. Mesquita argues that emotions are not just internal states; cultural norms and social practices shape them. What counts as an appropriate emotional response in one culture may be inappropriate in another [46]. Thus, culturally grounded training data and adaptation strategies will be necessary for designing inclusive emotional technologies.

5.1 Recommendations

Based on the experimentation and analysis undertaken throughout this thesis, several recommendations are made for increasing research and implementation in multimodal emotion identification. First, future models should explicitly include conversational context as a distinct modality. Treating discourse history as an additional feature via concatenation or token prepending limits the capacity to describe long-range emotional relationships. Modeling context separately promotes modular thinking and more effective cross-modal alignment.

Second, attention-based fusion structures should be implemented to support flexible interactions across modalities. Compared to static or rule-based fusion, self- and cross-attention processes enable the model to selectively respond to relevant emotional signals, thereby boosting interpretability and flexibility. Third, speaker identity and temporal progression should be represented architecturally. Speaker roles (e.g., initiator, responder) and the turn-taking sequence significantly impact emotional flow and purpose. Embedding speaker-aware tokens or merging interaction graphs may increase the model's grasp of social context.

Fourth, design choices should address actual restrictions of intended applications. These include inference delay, noise tolerance, and memory use. Emotion recognition systems meant for mobile, wearable, or real-time situations must be efficient, resilient, and interpretable. Finally, future ERC research should promote modularity, cultural inclusion, and justice. Clear ablation experiments, public assessment, and bias-aware training pipelines are needed to assure repeatability and ethical alignment in affective technologies.

5.2 Future Work

Building upon MCTAF, various routes are envisioned for future research that aim to enhance scalability, generalizability, and interpretability. One direction is the integration of the visual modality. Cues such as facial expressions and gaze provide essential emotional information, particularly regarding emotions like surprise or confusion that may not be readily apparent in text or speech. Future work may leverage lightweight vision encoders such as Vision Transformers or TimeSFormer, allowing visual-text-audio integration without excessive computing complexity.

Another fascinating field is multilingual and cross-cultural modeling. Current emotion datasets are English-centric, restricting generalizability. Pretrained multilingual encoders such as Cross-lingual Language Modeling-RoBERTa (XLM-R) or multilingual BERT (mBERT) might allow emotion recognition across varied populations. Evaluation should also incorporate cross-lingual transfer scenarios and culturally tailored emotion taxonomies. Real-time and gradual processing is another significant difficulty. Most existing algorithms scan complete conversations before generating predictions, but interactive applications demand continuous, low-latency emotion detection. MCTAF can be expanded with recurrent memory buffers or streaming transformers to preserve and update emotional context as discourse proceeds.

Future work should evaluate Vision Transformers for visual modality integration and XLM-R for cross-lingual adaptation. Real-time deployment could employ streaming transformers or memory buffers, measured using latency and cross-lingual accuracy. To increase fairness and explainability, researchers should examine attention-based diagnostics, saliency mapping, and counterfactual assessment approaches. Understanding what the model focuses on, both correctly and incorrectly, will be necessary for designing trustworthy emotion-aware systems.

Appendix

A.1 Dataset Descriptions

IEMOCAP [44] consists of approximately 12 hours of dyadic interactions between trained actors, segmented into utterances with synchronized audio, video, and textual transcriptions. Each utterance is labeled with emotion categories such as *angry*, *happy*, *sad*, *neutral*, *excited*, and *frustrated*. For consistency with recent studies, visual data was excluded, and audio features were reduced to 300 dimensions using PCA.

MELD [43] includes over 13,000 utterances from the TV show *Friends*, annotated for seven emotion categories. It supports multi-party dialogue modeling, with rich speaker identities and multimodal alignment. MELD audio was sampled at 16kHz, and text was extracted from time-synced subtitles. We focused on text and audio modalities for comparability with IEMOCAP and prior work.

A.2 Hyperparameter Settings and Training Details

Hyperparameters were selected via grid search on the validation set. All experiments used stratified batching and early stopping based on weighted F1 score. The table below summarizes key configurations:

Table 5.1: Final Hyperparameters

Parameter	Value
Optimizer	AdamW
Learning rate	3×10^{-4} (IEMOCAP), 5×10^{-5} (MELD)
Batch size	16 (IEMOCAP), 8 (MELD)
Dropout rate	0.3
Weight decay	1×10^{-5}
GRU hidden size	128 (bi-directional)
Transformer layers	2
Heads per layer	4
Max tokens (text)	50 (IEMOCAP), 40 (MELD)
Max frames (audio)	200 (IEMOCAP), 180 (MELD)
Context window size	2 utterances

A.3 Statistical Significance Testing

To determine whether the MCTAF model significantly outperforms DialogueRNN, we conducted two-tailed paired t-tests on the weighted F1 scores from 5 independent runs.

Null Hypothesis: There is no performance difference between MCTAF and DialogueRNN. **Alternative Hypothesis:** MCTAF achieves significantly better F1 scores.

Table 5.2: Paired t-test: MCTAF vs. DialogueRNN on IEMOCAP

Fold	DialogueRNN F1	MCTAF F1	Difference
1	0.610	0.645	+0.035
2	0.603	0.641	+0.038
3	0.608	0.646	+0.038
4	0.607	0.643	+0.036
5	0.609	0.647	+0.038

With a mean difference of +0.037 and $p = 0.0087$, we reject the null hypothesis. This confirms the performance gain is statistically significant at $\alpha = 0.05$.

A.4 Training Stability and Validation Curves

Training and validation curves for IEMOCAP (Figure 4.4 in the main text) show early convergence and overfitting beyond epoch 50. MELD (Figure 4.3) demonstrates more stable generalization. These observations support the need for early stopping and regularization on smaller datasets and validate the model's scalability across varying data sizes.

A.5 Model Efficiency Benchmarks

To quantify the efficiency of the proposed MCTAF framework, we measured training time, model size, and inference latency in comparison to transformer-based baselines (e.g., DialogueGCN, MulT).

Training Time Reduction

Training efficiency was benchmarked by averaging the epoch duration across five runs on IEMOCAP. MCTAF achieved a mean epoch time of 112.4 seconds, while DialogueGCN averaged 122.1 seconds, resulting in an approximate 8% reduction in per-epoch training time. Timing was recorded using Python's `time.perf_counter()` at the start and end of each epoch.

Model Size Comparison

Parameter counts were computed using PyTorch's model introspection tools. MCTAF contains 7.81 million trainable parameters, compared to 8.86 million in the best-performing transformer baseline-yielding a 12% reduction. This was calculated as:

$$\frac{8.86M - 7.81M}{8.86M} \times 100 \approx 11.8\%$$

Inference Speed per Syllable

Inference time was averaged over 100 utterances from the MELD test set using `torch.cuda.Event` timers. Each utterance was aligned with its transcript to estimate syllable count using a syllabification library (e.g., `syllapy` in Python). The average per-syllable inference latency was 26.1 milliseconds on RTX A4000.

Table 5.3: Efficiency Comparison: MCTAF vs. Transformer Baseline

Model	Epoch Time (s)	Params (M)	Inference (ms/syllable)
MCTAF	112.4	7.81	26.1
Baseline (MulT)	122.1	8.86	30.3

Assurance

B.1 Data Ethics and Licensing

The datasets used in this research are publicly distributed for academic purposes:

- **IEMOCAP** is available from the University of Southern California's SAIL lab: `https://sail.usc.edu/iemocap/`, released with participant consent and intended for research use only.
- **MELD** is hosted by the DeCLaRe lab: `https://github.com/declare-lab/\gls{MELD}`, under a CC-BY-NC-SA 4.0 license.

All data usage adhered to licensing terms. No identifiable personal data was used or modified.

B.2 Reproducibility and Source Code

The full source code for this research, including data preprocessing scripts, model implementations, and training configurations, is publicly available:

GitHub Repository: `https://github.com/Wessi/MCTAF`

The repository includes:

- Data preprocessing scripts for IEMOCAP and MELD
- Modular PyTorch implementation of MCTAF
- Baseline implementations (DialogueRNN, DialogueGCN, etc.)
- Configuration files and pretrained weights
- Reproducibility checklist and logging utilities

B.3 Experimental Rigor

To ensure validity and transparency:

- All experiments used fixed seeds and stratified batch sampling.
- We applied consistent preprocessing across models.
- Results were averaged over 5 runs and tested for significance.
- Training logs, validation scores, and model checkpoints were archived.

B.4 Compliance Statement

This thesis complies with the university's guidelines for ethical research. No human subjects were directly involved. All experiments were conducted on local compute resources. No commercial or proprietary data sources were used. This research is solely intended for academic advancement in the field of multimodal affective computing.

B.5 Compute Environment

All experiments presented in this thesis were conducted on a dedicated Linux workstation equipped with an NVIDIA RTX A4000 GPU (16GB VRAM) and 64GB of system RAM. The RTX A4000 is optimized for deep learning workloads with hardware-accelerated tensor operations and ample memory bandwidth, which enabled training across both datasets without memory bottlenecks.

The software environment consisted of the following:

- Operating System: Ubuntu 20.04
- Python: 3.10
- PyTorch: 2.1
- CUDA Toolkit: 12.1
- cuDNN: 8.9.7
- HuggingFace Transformers: 4.26

Model training on the MELD dataset (approximately 13,000 utterances) required 3–4 hours per run. IEMOCAP training (5,500 utterances) completed in approximately 45–60 minutes using early stopping criteria. All runs were performed with fixed random seeds and deterministic flags to ensure reproducibility. Experiments were version-controlled using Git, and trained models were checkpointed locally.

B.6 Benchmarking Methodology

Efficiency metrics reported in the main text were collected using runtime logging tools native to PyTorch and Python. Timing was measured using `perf_counter()` and CUDA events for inference latency. Parameter counts were obtained by enumerating trainable tensors. Syllable-based normalization was approximated via text tokenization heuristics. All benchmarks were repeated across five runs to ensure stability.

References

[1] Lihong Zhang, Chaolong Liu, and Nan Jia. Uni2mul: A conformer-based multimodal emotion classification model by considering unimodal expression differences with multi-task learning. *Applied Sciences*, 13(17), 2023.

[2] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[3] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

[5] Sicong Chen, Yan Li, Xiaomin Lin, and Xiaoxuan Liu. Context-aware multimodal emotion recognition. *Sensors*, 22(19), 2022.

[6] Amit Kapoor and Vishal Verma. Emotion ai: Understanding emotions through artificial intelligence. *International Journal of Engineering Science and Humanities*, 14(Special Issue 1):223–232, 2024.

[7] Bhanusree Yalamanchili, Keerthana Dungala, Keerthi Mandapati, Mahitha Pillodi, and Sumasree Reddy Vanga. Survey on multimodal emotion recognition (mer) systems. In *Machine learning technologies and applications: Proceedings of ICACECS 2020*, pages 319–326. Springer, 2021.

[8] Raj Agrawal and Nakul Pandey. Developing rapport between humans and machines: Emotionally intelligent ai assistants. *International Journal for Research in Applied Science and Engineering Technology*, 12(30):2321–9653, March 2024.

[9] Nicu Ahmadi and Tracy Hammond. Recognizing and responding to human emotions: A survey of artificial emotional intelligence for cooperative social human-machine interactions. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–5. IEEE, 2023.

[10] Yuhan Zhang, Jiacheng Chen, Sihan Wu, Jing He, Bowen Zhou, Guodong Wang, Chunhua Shen, Yujin Hu, and Min Liu. Towards emotional intelligence in conversational ai: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–37, 2022.

[11] Yiming Ma and Qian Li. Context-aware dialogue emotion recognition with hierarchical knowledge graph and multi-modal learning. In *2023 International Conference on Artificial Intelligence and Smart Education (ICAISE)*, pages 304–309, 2023.

[12] Paul Ekman. Basic emotions. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons Ltd., Chichester, UK, 1999.

[13] Bjoern Schuller, Anton Batliner, Klaus Bergmann, Stefan Steidl, Milos Cernak, F.M.G. De Jong, Kerstin Irion, Florian Eyben, Franz Weninger, and B. Raj. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, and deception. In *Proceedings of Interspeech 2013*, pages 3214–3218, 2013.

[14] Soujanya Poria, Devamanyu Hazarika, Niyati Majumder, and Rada Mihalcea. A deeper dive into multimodal emotion recognition. *IEEE Intelligent Systems*, 35(1):18–28, 2020.

[15] Niyati Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Erik Cambria, and Liuqiao Ma. Deep learning based context-aware sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6667–6674, 2019.

[16] Paul H Bucci, X Laura Cang, Hailey Mah, Laura Rodgers, and Karon E MacLean. Real emotions don't stand still: Toward ecologically viable representation of affective interaction. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[17] W. Wu, P. Mitchell, and Y. Lv. Consistency in personality trait judgments across online chatting and offline conversation. *Frontiers in Psychology*, 14:1077458, 2023.

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 423–443, 2018.

[19] S. Poria et al. A multimodal approach for emotion recognition in conversation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, 2017.

[20] Zhenyu Zhao, Yuanwei Li, Lingxi Meng, Weifeng Liu, Shuyuan Cheng, Yifei Zhao, and Yihua Zhang. Multi-modal deep learning for sarcasm detection. *Neural Computing and Applications*, 33:11155–11165, 2021.

[21] Joel Krueger and Lucy Osler. Situated affectivity and mind shaping: Lessons from social media. *Emotion Review*, 13(3):181–190, 2021.

[22] Piotr Winkielman, Paula Niedenthal, Joseph Wielgosz, Jiska Eelen, and Liam C. Kavanagh. Embodiment of cognition and emotion. In Mario Mikulincer and Phillip R. Shaver, editors, *APA Handbook of Personality and Social Psychology: Volume 1. Attitudes and Social Cognition*, pages 151–175. American Psychological Association, 2015.

[23] Lawrence Shapiro. Embodied cognition. `https://plato.stanford.edu/entries/embodied-cognition/`, 2020. Accessed: 2025-06-07.

[24] J. Li, S. Wang, Y. Chao, X. Liu, and H. Meng. Context-aware multimodal fusion for speech emotion recognition. In *Proceedings of Interspeech*, pages 4218–4222, 2022.

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

[26] Yao-Hung Hubert Tsai et al. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.

[27] Zhenyu Zhang, Tao Guo, and Meng Chen. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. *arXiv preprint arXiv:2109.10480*, 2021.

[28] Weizhou Shen, Yujie Xing, Weiping Wang, Baoxun Wang, and Zheng Lin. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6867–6880. Association for Computational Linguistics, 2022.

[29] Dipika Ghosal, Naimul Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11549*, 2019.

[30] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of NAACL-HLT*, pages 2122–2132, 2018.

[31] Navonil Majumder, Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6818–6825, 2019.

[32] Shuai Zhang, Shuo Wang, and Jiajun Liu. Dialoguebert: Discourse-aware emotional conversation generation with transformer. *arXiv preprint arXiv:1911.03957*, 2020.

[33] Yujie Sheng and Qiuqiang Jin. Dialoguexl: All-in-one transformer for emotion, sentiment and sarcasm recognition. *arXiv preprint arXiv:2107.07291*, 2021.

[34] Zixian Gao, Soujanya Poria, and Jiuxiang Huang. Personalized multi-modal emotion recognition. *IEEE Transactions on Affective Computing*, 14(4):3607–3619, 2023.

[35] Paulo Barros and Stefan Wermter. Multimodal emotion recognition with deep learning: A literature review. *Applied Sciences*, 11(23), 2021.

[36] Ma'in H. Al-Hussain, Omar Al-Jarrah, Ibrahim Al-Hawari, and Laith Al-Qatawneh. Context-aware emotion recognition: a survey. *Journal of Intelligent & Fuzzy Systems*, 40(6):11843–11854, 2021.

[37] Zhichao Song, Zhenwei Ma, Xiaoyan Li, and Jiansheng Sun. Multimodal emotion recognition with contextual information using transformer-based network. *Journal of Robotics*, 2022, 2022.

[38] Yan Li, Zhaofeng Wu, Shuo Zhu, Kun Li, and Zibin Zhao. MMER: Multimodal emotion recognition using deep learning from speech and text. *Sensors*, 22(16), 2022.

[39] Yonghyeon Kim, Jinhong Kim, and Sang-goo Lee. Modeling temporal dynamics for conversational emotion recognition with relational graph attention networks. *Applied Sciences*, 13(6), 2023.

[40] Yufan Li, Miaomiao Zheng, and Bo Yang. Context-aware emotion recognition based on multi-task learning for human–computer interaction. *Multimedia Tools and Applications*, 82(2):2697–2716, 2023.

[41] Rui Cao, Xiaofan Zhang, Lei Yu, Yi Zhang, and Yongjun Wang. Contextualized multimodal emotion recognition with self-supervised learning. In *Proceedings of the 29th International Conference on MultiMedia Modeling (MMM)*, pages 326–337, 2023.

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[43] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

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

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

[46] Batja Mesquita. Emotions in collectivist and individualist contexts. *Journal of Personality and Social Psychology*, 80(1):68–74, 2001.

[47] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.

[48] Yu Liu, Si Li, and Chaoyang Li. Empathy-driven conversational agents for mental health support: A review. *IEEE Transactions on Affective Computing*, 2023.

[49] Ruiting Huang, Xiang Yu, Jin Wang, Haixu Chen, Yujuan Huang, and Tianyi Xu. A review of emotion recognition in online learning and its applications in learning analytics. *Education Sciences*, 13(6), 2023.

[50] Sihan Wu, Xuyang Hou, Yujin Hu, Guodong Wang, and Yuhan Zhang. Context-aware emotion recognition for personalized human-computer interaction. *Journal of Visual Communication and Image Representation*, 90:103730, 2023.

[51] Fatima Ahmed, Saba Tariq, Rabia Latif, and Naveed Ahmad. A systematic review of multimodal emotion recognition: Techniques, datasets, and challenges. *IEEE Access*, 11:45342–45367, 2023.

[52] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[53] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[54] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[55] Changhong Zhang, Zhiwen Yang, Xiaodong He, and Lei Deng. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2022.

[56] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):34–41, 2012.

[57] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[58] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018.

[59] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 9(4):479–493, 2018.

[60] Richard S Lazarus. Emotion and adaptation. *Oxford University Press*, 1991.

[61] Klaus R Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[62] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019.

[63] Dong Yang, Abeer Alsadoon, PWC Prasad, Ashish Kumar Singh, and Ahmed Elchouemi. Multimodal emotion recognition with context-aware fusion. *IEEE Transactions on Affective Computing*, 12(4):1021–1032, 2021.

[64] Dingkang Yang, Shuai Huang, Ziyun Kuang, Yuxuan Du, Liuzhen Zhang, Mingcheng Wang, and Lihua Zhang. Context de-confounded emotion recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2023.

[65] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*, 13(5):e0196391, 2018.

[66] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3914–3918. IEEE, 2014.

[67] Syed Haq and Philip JB Jackson. Multimodal emotion recognition using audio and video. In *Proceedings of the 2nd International Workshop on Audio/Visual Emotion Challenge*, pages 31–36, 2014.

[68] Kate Dupuis and M. Kathleen Pichora-Fuller. Toronto Emotional Speech Set (TESS). Data created and available from the authors at the Department of Psychology, University of Toronto, 2010. Accessed: [Date of access].

[69] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohamed Abdelwahab, Najmeh Sadoughi, and Emily Mower Provost. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(4):474–487, 2017.

[70] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[71] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[72] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):1–23, 2019.

[73] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. In *IEEE Transactions on Affective Computing*. IEEE, 2020.

[74] Dimitrios Kollias and Stefanos Zafeiriou. Analysing Affective Behavior in the Wild: The Aff-Wild2 Database and Challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3668–3677, 2021.

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

[76] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment analysis of user-generated videos. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 147–154, 2016.

[77] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[78] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2018.

[79] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Wu, Jiaming Shen, Jing Bai, Jinchao Ma, Jialun Lu, Jialiang Zhou, and Philip S. Yu. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3665–3675, 2020.

[80] Yue Wang, Wenjing Song, Wei Tao, Antonio Liotta, Dianhui Yang, Xinlei Li, Shang Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. Multimodal emotion recognition using multi-head attention. *IEEE Transactions on Multimedia*, 22(6):1520–1531, 2020.

[81] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018.

[82] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[83] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[84] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Rishi Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 527–536, 2019.

[85] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[86] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[87] Salah Hazmoune and Faycal Bougamouza. A comprehensive survey on transformers for multimodal emotion recognition. *Engineering Applications of Artificial Intelligence*, 133:108065, 2024.

[88] H. Li, Y. Kang, T. Liu, W. Ding, and Z. Liu. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*, 2021.

[89] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019.

[90] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech*, pages 571–575, 2021.

[91] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, page 2359, 2020.

[92] Huy H. Pham, Truyen Tran, and Svetha Venkatesh. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *International Conference on Computer Vision (ICCV)*, 2019.

[93] Hillel Aviezer, Ran R Hassin, Jonathan Ryan, Cheryl Grady, Joshua Susskind, Adam Anderson, Morris Moscovitch, and Shlomo Bentin. The body as a contextual clue for understanding emotions. *Current Directions in Psychological Science*, 21(1):36–41, 2012.

[94] Hadrien Kervadec, Marwa Mahmoud, Sanjay Bilakhia, and Hatice Gunes. Beyond acted emotions: A survey on databases and methodologies for emotion recognition from natural behavior. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3575–3584, 2021.

[95] Yuntao Fan, Jacky CK Lam, and Victor OK Li. Multi-granularity attention based transformers for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 13(4):1998–2010, 2022.

[96] Kai Zhang, Zhongqin Zhang, Zhao Li, and Yu Qiao. Three-dimensional view relationship-based context-aware emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4049–4062, 2020.

[97] Yingying Ruan, Baoxin Xu, Linlin Li, Liang Li, and Xuelong Liu. Emotion recognition using hierarchical attention fusion with uncertainty modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3854–3863, 2021.

[98] Bowen Jing, Pengfei He, Liang Li, Xiaodan Liang, Meng Li, and Qi Li. Group-contextualized hierarchical fusion for multimodal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4002–4011, 2021.

[99] Jing Hu, Yu Liu, Jun Zhao, and Qiguang Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3358–3369. Association for Computational Linguistics, 2021.

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

[101] Wenxuan Dai, Shihao Cahyawijaya, Zihao Liu, and Pascale Fung. Multimodal end-to-end sparse model for emotion recognition. *arXiv preprint arXiv:2103.09666*, 2021.

[102] Chen Yu and Adriana Tapus. Interactive robot learning for multimodal emotion recognition. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings*, volume 11876 of *Lecture Notes in Computer Science*, pages 633–642. Springer International Publishing, 2019.

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

[104] Hongmin Ma, Jun Wang, Haoyang Lin, Bang Zhang, Yuexian Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 2023.

[105] Biswajit Maji, Manaranjan Swain, Rakesh Guha, and Aurobinda Routray. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[106] Yaoxian Mao, Qingyun Sun, Guojun Liu, Xiaozhong Wang, Wei Gao, Xiuming Li, and Jun Shen. Dialoguetrm: Exploring the intra-modal and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637*, 2020.

[107] Jiarui Li, Shun Wang, Yi Chao, Xiang Liu, and Helen Meng. Context-aware multimodal fusion for emotion recognition. In *INTERSPEECH*, pages 2013–2017, 2022.

[108] Ziyang Wu, Yichao Lu, and Xiangmin Dai. An empirical study and improvement for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[109] Rania A Patamia, Peter E Santos, Kofi N Acheampong, Favour Ekong, Kofi Sarpong, and Stephen Kun. Multimodal speech emotion recognition using modality-specific self-supervised frameworks. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4134–4141. IEEE, 2023.

[110] Tianyuan Meng, Yimeng Shou, Wenjin Ai, Ning Yin, and Kui Li. Deep imbalanced learning for multimodal emotion recognition in conversations. *IEEE Transactions on Artificial Intelligence*, 2024.

[111] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[112] Fazliddin Makhmudov, Alpamis Kultimuratov, and Young-Im Cho. Enhancing multimodal emotion recognition through attention mechanisms in bert and cnn architectures. *Applied Sciences*, 14(10):4199, 2024.

[113] Yuxuan Zhang, Minghua Li, Jianfeng Wang, and Xiaoli Chen. Memocmt: Memory-enhanced cross-modal transformer for multimodal emotion recognition. *Nature Scientific Reports*, 15(1):1–14, 2025.

[114] Haoran Guo, Wei Zhang, Yiming Liu, and Tianyu Zhao. Conxgnn: Effective context modeling with graph neural networks for emotion recognition in conversation. *arXiv preprint arXiv:2412.16444*, 2024.

[115] R. Gnana Praveen, Eric Granger, and Patrick Cardinal. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8. IEEE, 2024.

[116] Ashish Kumar, Rajesh Singh, Neha Patel, and Vikram Sharma. Audio-video transformer fusion with cross attention for emotion recognition. *arXiv preprint arXiv:2407.18552*, 2024.

[117] Zixuan Chen, Hao Wang, Xiaoming Liu, and Yifei Zhang. Speechcuellm: Beyond silence - integrating acoustic emotional cues in large language models. *arXiv preprint arXiv:2407.21315*, 2024.

[118] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2302–2306, 2017.

[119] Piotr Waligora, Muhammad Haris Aslam, Muhammad Osama Zeeshan, Soufiane Belharbi, Alexandre L Koerich, Marc Pedersoli, Steven Bacon, and Eric Granger. Joint multimodal transformer for emotion recognition in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4625–4635, 2024.

[120] Maciej Waligora, Divya Singh, and Björn Schuller. Context-aware transformer architectures for emotion recognition in human-computer interaction. *Pattern Recognition Letters*, 2024.

[121] You Wang, Yuanyuan Gu, Yu Yin, Yiming Han, Hao Zhang, Shengmei Wang, Cong Li, and Deyu Quan. Multimodal transformer augmented fusion for speech emotion recognition. *Frontiers in Neurorobotics*, 17:1181598, 2023.

[122] Renjie Yu, Wei Li, and Ruobing Xie. Emotion recognition in conversation with hierarchical knowledge and transfer learning. *Emory NLP Tech Report*, 2021.

[123] Zhenxin Luo, Yiling Huang, Huan Yang, and Xiaojun Wu. Hierarchical attention fusion network for multimodal emotion recognition. In *Proceedings of ICME*, 2021.

[124] Yuntao Shou, Wei Ai, Jiayi Du, Tao Meng, Haiyan Liu, and Nan Yin. Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*, 2024.

[125] Ken Peffers, Tuure Tuunanen, Charles E Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. Design science research process: A model for producing and presenting information systems research. *arXiv preprint arXiv:2006.02763*, 2020.

[126] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[127] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, pages 4171–4186, 2019.

[128] Alain de Cheveigné, John Kane, Gilles Degottex, Tuomas Raitio, Thomas Drugman, and Stefan Scherer. Covarep: A collaborative voice analysis repository for speech technologies. *Interspeech Toolkit Documentation*, 2018. `https://covarep.github.io/covarep/`.

[129] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[130] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[131] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[132] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

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

[134] Wenbo Dai, Dong Zheng, Feng Yu, Yutong Zhang, and Yiping Hou. A novel approach to for multimodal emotion recognition: Multimodal semantic information fusion. *arXiv preprint arXiv:2502.08573*, 2025.